Volume 34 Number 2 June 2010 **ISSN 0350-5596**

Informatica

An International Journal of Computing and Informatics

Special Issue:

Quantitative Risk Analysis Techniques for Security Applications

Guest Editors:

Christopher Kiekintveld
Janusz Marecki
Praveen Paruchuri
Katia Sycara

1977

EDITORIAL BOARDS, PUBLISHING COUNCIL

Informatica is a journal primarily covering the European computer science and informatics community; scientific and educational as well as technical, commercial and industrial. Its basic aim is to enhance communications between different European structures on the basis of equal rights and international refereeing. It publishes scientific papers accepted by at least two referees outside the author's country. In addition, it contains information about conferences, opinions, critical examinations of existing publications and news. Finally, major practical achievements and innovations in the computer and information industry are presented through commercial publications as well as through independent evaluations.

Editing and refereeing are distributed. Each editor from the Editorial Board can conduct the refereeing process by appointing two new referees or referees from the Board of Referees or Editorial Board. Referees should not be from the author's country. If new referees are appointed, their names will appear in the list of referees. Each paper bears the name of the editor who appointed the referees. Each editor can propose new members for the Editorial Board or referees. Editors and referees inactive for a longer period can be automatically replaced. Changes in the Editorial Board are confirmed by the Executive Editors.

The coordination necessary is made through the Executive Editors who examine the reviews, sort the accepted articles and maintain appropriate international distribution. The Executive Board is appointed by the Society Informatika. Informatica is partially supported by the Slovenian Ministry of Higher Education, Science and Technology.

Each author is guaranteed to receive the reviews of his article. When accepted, publication in Informatica is guaranteed in less than one year after the Executive Editors receive the corrected version of the article.

Editorial

Risk Analysis for Security Applications

Securing critical infrastructure and computer networks is one of the most important challenges of our modern, interconnected society. In real-world security domains it is necessary to predict, mitigate, and react to intentional threats from adversarial agents, often under significant uncertainty.

Decisions must be made quickly, by processing large amounts of information, and taking into account the goals and capabilities of the adversary.

We believe that meeting these challenges will require the development and integration of new methods in multi-agent systems, risk analysis, computational game theory, machine learning, and other related fields. Developing and applying the tools needed to analyze and manage complex security problems will be a high-impact area for decades to come.

Several recent examples of deployed security applications point to progress and potential in this area.

For example, a deployed software system called ARMOR applies game-theoretic reasoning to help police officers at the Los Angeles International Airport make critical decisions about how to schedule both vehicle checkpoints and canine patrols.

This system has been in active use at the airport since 2007, and has received numerous accolades from the police, the popular press, and the research community.

We believe that the success of this application is just the beginning, and that new research will allow exciting new applications to be built and deployed.

For example, there are potential applications in robotic patrolling, automated security cameras, border security screening policies, and intrusion detection tools for computer networks.

Despite recent progress, there is a need for more research to address fundamental challenges in real-world security domains to provide comprehensive risk analysis and management tools.

First, most real domains are extremely large with complex, interacting decisions.

New algorithms that exploit domain structure are needed to apply sophisticated reasoning such as decision-theoretic or game-theoretic analysis to these domains.

These difficulties are further magnified by the need to model and reason about uncertainty in the domain, and to account for the fact that human decision-makers may behave in ways that are not easily captured in mathematical models of perfect rationality.

In many domains, even building a reliable, validated model of the domain is a significant challenge, whether by elicitation of the model from domain experts or through the use of simulation tools or empirical evidence.

Finally, the problem of evaluating deployed security systems poses numerous challenges, including the lack of controlled studies and limited access to data.

We are pleased to bring out this special issue of the Informatica journal which comprises of four papers that were presented at the inaugural workshop on Quantitative Risk Analysis for Security Applications (QRASA). These papers touch on many of the themes outlined above, and each paper makes a significant contribution to this exciting, emerging area of research.

We hope that the reader will be inspired by these papers to explore and participate in this dynamic and growing field of study.

The first paper titled "A Framework for Evaluating Deployed Security Systems: Is There a Chink in your ARMOR?" by Matthew E. Taylor, Christopher Kiekintveld, Craig Western and Milind Tambe, addresses the challenges of evaluating deployed security systems, using the ARMOR system as a case study to raise many of the issues involved in evaluating security systems in general. In addition to discussing these issues, the paper lays out a framework for guiding the evaluation process, giving insights into the different types of analyses that are possible and the kinds of evaluations that could further improve knowledge of a given systems' utility.

The second paper titled "Application of Microsimulation to the Modeling of Epidemics and Terrorist Attacks" by Ian Piper, Daniel Keep, Tony Green and Ivy Zhang describes a simulation tool for agent-based simulation. The authors describe the benefits of this tool and evaluate the tool on a case study, modeling the spread of an infectious disease in a community, based on real historical incident. In addition to the relevance of this study for biological attacks, the authors describe some other possible uses for this tool in modeling terrorist attack scenarios.

The third paper titled "Strategic Modeling of Information Sharing Among Data Privacy Attackers" by Quang Duong, Kristen LeFevre and Michael P. Wellman propose a framework for modeling multiple attackers with heterogeneous background knowledge, supporting analysis of their strategic incentives for sharing information prior to attack. The framework posits a decentralized mechanism by which agents decide whether and how much information to share, and defines a normal-form game representing their strategic choice setting. This paper represents one of the first applications of game theory to study possible attacks against databases.

The fourth paper titled "Planning to Discover and Counteract Attacks" by Tatiana Kichkaylo, Tatyana Ryutov, Michael D. Orosz and Robert Neches develops a set of tools that can provide decision support for recognizing plans in an adversarial setting. The approach is demonstrated in a network security setting, showing how an attacker's plan can be decomposed into separate actions and how recognizing the overall intent of the plan requires complex analysis of these independent plan fragments. The tool described in this paper assists security experts with analyzing various possible attack scenarios.

The special editors would like to thank Professor Matjaz Gams (Managing editor of Informatica) for providing the opportunity to edit this special issue on Risk Analysis for Security Applications. Finally, we would like to thank the authors of the papers for their individual contributions and all of the referees for their helpful comments. Their efforts helped to ensure the high quality of the material presented here.

Christopher Kiekintveld
University of Southern California
Los Angeles, CA 90089
kiekintv@usc.edu

Janusz Marecki
Mathematical Sciences Department
IBM T.J. Watson Research Center
marecki@us.ibm.com

Praveen Paruchuri
Carnegie Mellon University
Pittsburgh, PA 15232
paruchur@gmail.com

Katia Sycara
Carnegie Mellon University
Pittsburgh, PA 15232
katia@cs.cmu.edu

A Framework for Evaluating Deployed Security Systems: Is There a Chink in your ARMOR?

Matthew E. Taylor, Christopher Kiekintveld, Craig Western and Milind Tambe
http://teamcore.usc.edu
Computer Science Department
The University of Southern California, USA

Keywords: physical security, game theory, Stackelberg games, evaluation

Received: September 7, 2009

A growing number of security applications are being developed and deployed to explicitly reduce risk from adversaries' actions. However, there are many challenges when attempting to evaluate *such systems, both in the lab and in the real world. Traditional evaluations used by computer scientists, such as runtime analysis and optimality proofs, may be largely irrelevant. The primary contribution of this paper is to provide a preliminary framework which can guide the evaluation of such systems and to apply the framework to the evaluation of ARMOR (a system deployed at LAX since August 2007). This framework helps to determine what evaluations could, and should, be run in order to measure a system's overall utility. A secondary contribution of this paper is to help familiarize our community with some of the difficulties inherent in evaluating deployed applications, focusing on those in security domains.*

Povzetek: Kako ovrednotiti varnostne aplikacije, kot recimo sistem ARMOR, ki je od 2007 dalje v uporabi na LAX?

1 Introduction

Computer scientists possess many tools that are particularly applicable to security-related problems, including game-theoretic reasoning, efficient algorithmic design, and machine learning. However, there are many challenges when attempting to *evaluate* a security system in a lab setting or after it has been deployed. Traditional evaluations used by computer scientists — such as runtime analysis and optimality proofs — often do not consider the relevance of modeling assumptions or account for how a system is actually used by humans. If there is an error "between the keyboard and the chair," it still needs to be addressed, even if such problems are beyond the scope of some computer programs. An additional complication is that no security system is able to provide 100% protection. Instead, systems must be evaluated on basis of risk reduction, often through indirect measures such as increasing adversary cost and uncertainty, or reducing the effectiveness of an adversaries' attack. Despite these challenges, evaluation remains a critical element of the development and deployment of any security system.

An important challenge in security evaluation is that performance necessarily depends on an adversarial human's behavior and decisions. Controlled laboratory studies can be a valuable component of an evaluation, but the population of test subjects is necessarily different that that of actual attackers. Evaluating a system once it is deployed only increases the experimenter's burden. First, while a system could be alternatively enabled and disabled on different days to measure its efficacy, this is at best impractical and at worst unethical. Second, data related to the configuration and performance of the system may be classified or sensitive, and not available to researches. Third, a key component of many security systems is *deterrence*: an effective system will not only identify and prevent successful attacks, but will also dissuade potential attackers. Unfortunately, it is generally impossible to directly measure the deterrence effect.[1]

This paper introduces a general framework for evaluating deployed systems and then presents a case study of one such security system. While computer scientists traditionally prioritize precise, repeatable studies, this is not always possible in the security community; computer scientists are used to quantitative evaluations in controlled studies, whereas security specialists are more accepting of qualitative metrics on deployed systems. For instance, Lazaric (14) summarized a multi-year airport security initiative by the FAA where the highest ranked evaluation methodology (of seven) relied on averaging *qualitative* expert evaluations.

The primary advantage of quantitative evaluations, rather than qualitative, is that they can be integrated into a cost-benefit analysis. A particular security measure may be effective but prohibitively expensive — consider two extremes in the domain of airport security. Hand searching every passenger who enters an airport and disallowing all luggage would likely increase the security of plane flights,

[1] To measure deterrence, one needs to know how many attacks *did not occur* due to security, a generally unmeasurable counterfactual.

Figure 1: A LAX checkpoint scheduled by ARMOR

but the costs from extra security personnel, increased time in the airport, and lost revenue makes such a draconian policy infeasible. On the other hand, removing all airport screenings and restrictions would reduce costs and delays, but also significantly increase security risks. By carefully weighing costs and benefits, including non-monetary effects like privacy loss, security experts and policy makers can better decide which measures are appropriate in a particular context.

Our ultimate goal is to provide a framework for comprehensive evaluation of deployed systems along multiple attributes, in absolute or relative terms, to facilitate cost-benefit analysis. We examine existing evaluations of the ARMOR system (16) as a case study. Several different kinds of evaluation of this system support the claim that it significantly improves over the previous best practices of uniform randomization or hand-constructed schedules and is cost effective.

The primary contribution of this paper is to provide a framework to evaluate such deployed systems and apply it to ARMOR. This framework helps to determine what to measure, how to measure it, and how such metrics can determine the system's overall utility. A secondary contribution of this paper is to help familiarize our community with some of the difficulties inherent in evaluating deployed applications, particularly for security domains.

2 Case study: ARMOR

The Los Angeles World Airports (LAWA) police at the Los Angeles International Airport (LAX) operate security for the fifth busiest airport in the United States (and largest destination), serving 70–80 million passengers per year. LAX is considered a primary terrorist target on the West Coast and multiple individuals have been arrested for plotting or attempting to attack LAX (19). Police have designed multiple rings of protection for LAX, including vehicular checkpoints, police patrols of roads and inside terminals (some with bomb-sniffing canine units, also known as K9 units), passenger screening, and baggage screening.

There are not enough resources (police and K9 units) to monitor every event at the airport due to the large physical area and the number of passengers served. ARMOR addresses two specific security problems by increasing the unpredictability of security schedules and weighting defensive strategy based on targets' importance. First, there are many roads that are entry points to LAX. When and where should vehicle checkpoints (Figure 1) be set up on these roads? Pertinent information includes typical traffic patterns on inbound roads, the areas each road accesses within LAX, and areas of LAX which may have more or less importance as terrorist targets. Second, how and when should the K9 units (Figure 2) patrol the eight terminals at LAX? Here it is important to consider the time-dependant passenger volumes per terminal, as well as the attractiveness of different terminals. In both cases a predictable pattern can be exploited by an observant attacker.

Figure 2: A K9 patrol

To address the two security problems above, we use game theory to model and analyze the two domains. The police and attackers play a Bayesian Stackelberg game (6), with the police first committing to a (randomized) security policy. Multiple attacher types are modeled. Each attacker type observes this policy and then selects the optimal attack strategy (depending on the defense strategy). Solving this game for a Strong Stackelberg Equilibrium finds an optimal randomized policy for the police, which is sampled as necessary to give specific schedules. ARMOR (*Assistant for Randomized Monitoring Over Routes*) is the software tool that assists police with randomized scheduling using this game-theoretic analysis (16). The software uses an optimized algorithm for solving Bayesian Stackelberg games called DOBSS (15).

The randomized schedules account for three key factors: (1) attackers are able to observe the security policy using surveillance, (2) attackers change their behavior in response to the security policy, and (3) the risk/consequence of an attack varies depending on the target. The end result is a randomized police schedule that is unpredictable, but weighted towards high-valued targets. ARMOR has been in use at LAX since August 2007, marking an important transition from theoretical to practical application. The system has received very positive feedback and is considered an important element of security at the airport.

3 Current ARMOR evaluations

The ARMOR system has undergone multiple evaluations before and after deployment. We summarize the current evaluations below, both from existing publications and novel to this article, grouped by category. It is not difficult to argue that ARMOR is a significant improvement over previous practices: it saves time for human schedulers, it is inexpensive to implement, and humans are known to have difficulty randomizing effectively (21). However, our goal is to take steps towards a more comprehensive understanding of ARMOR that provides as much insight as possible into the value of the system.

3.1 Mathematical

The first category of analyses are mathematical evaluations that use our game-theoretic model to evaluate ARMOR's security policies against other baseline policies. In particular, if we assume attackers act optimally and have the utilities specified in the model, we can predict how they will react to any schedule and therefore compare the expected utility of these schedules. ARMOR uses a game theoretic optimal schedule. Comparing against benchmark uniform random and hand-crafted schedules show that ARMOR's schedule is substantially better than these benchmarks across a variety of different settings. For example, Figure 3(d) shows the expected reward for the police using ARMOR's schedule (calculated using DOBSS) compared with a uniform random benchmark strategy in the canines domain. ARMOR is able to make such effective use of resources that using three canines scheduled with DOBSS yields higher utility than using six canines with uniform random scheduling!

Sensitivity analysis is another important class of evaluations that can be performed using only the mathematical models. In this type of evaluation, important parameters of the model are varied to test how sensitive the output of the model is to the input. One important input to our models is the distribution of different types of attackers. For example, some attackers may be highly organized and motivated, while others are amateurish and more likely to surrender. Different types of attackers can be modeled as having different payoff matrices. Changing the percentages of each attacker can help show the system's sensitivity to assumptions regarding the composition of likely attackers, and (indirectly) the system's dependence on precise utility elicitation. In Figure 3(a)–3(c), there are two adversary types with different reward matrices. Figure 3(a) demonstrates that DOBSS has a higher expected utility than that of a uniform random strategy on a single checkpoint, regardless of the percentage of "type one" and "type two" adversaries. Figures 3(b) and (c) shows that DOBSS again dominates uniform random for two and three checkpoints, respectively.

Further sensitivity analysis can be applied to measure how the optimal strategy computed by DOBSS changes as payoffs are modified. Since the payoff functions are determined through preference elicitation sessions with experts, these payoffs are estimates of true utilities. Game-theoretic models can be quite sensitive to payoff noise, and arbitrary changes in the payoffs can lead to arbitrary changes in the optimal schedule. However, there is some evidence that ARMOR is robust to certain types of variations. In one experiment, we multiplied all of the defender's negative payoffs for successful attacks by a factor of four, essentially increasing the impact of a successful attack. We found that in the one and three checkpoint case, the strategies were unchanged. In the two checkpoint case the actions were slightly different, but the overall strategy and utility were unchanged.

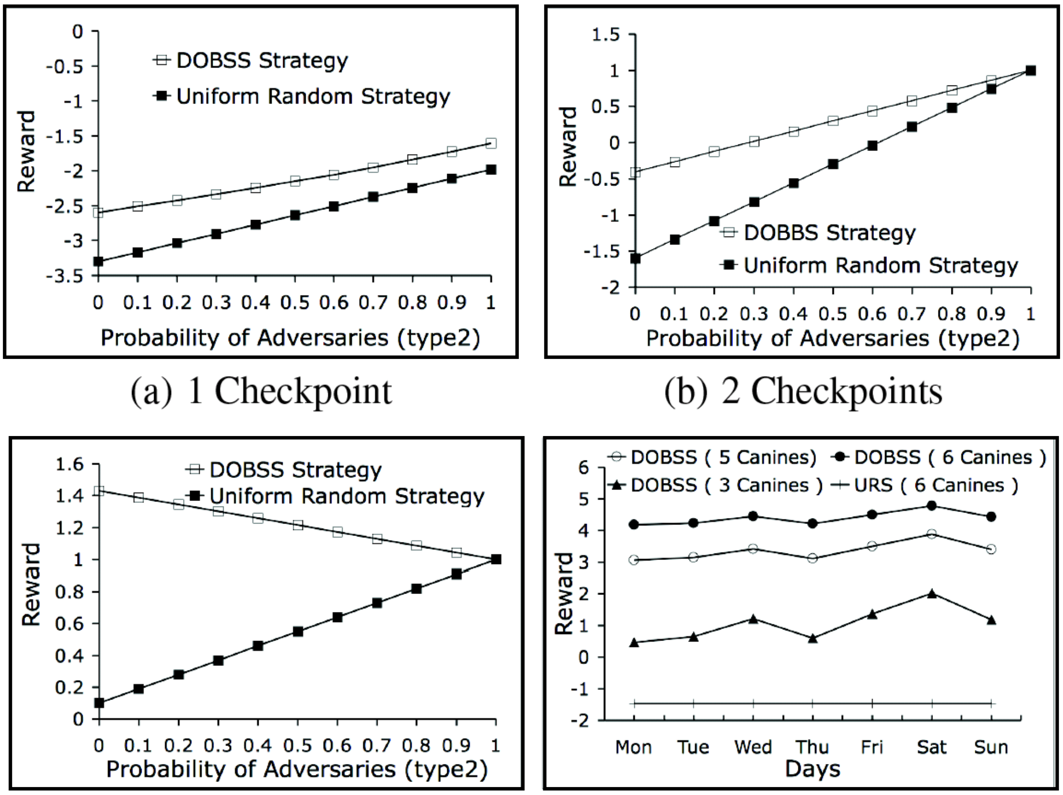
As with any game theoretic analysis, the assumptions regarding the opponent's behavior may dramatically change the outputs and evaluated performance. Figure 4 examines an assumption typically made by Stackelberg solvers. Specifically, such solvers assume that if an adversary is given a set of actions with equivalent payoffs, the attacker will select the action that maximizes the defender's payoff (the Strong Stackelberg Equilibrium, or SSE). We compare this potentially optimistic behavior with two other reasonable choices: the attacker selects randomly from the set of actions with the maximum (equivalent) attacker utility, and the attacker selects the action that minimizes the defender utility from the set of equivalent actions with the maximum attacker utility. The similarity in payoffs of these three ways for attackers to break ties show that this assumption is not critical for ARMOR's success.

Additionally, note that Figure 4 has a roughly linear trend. Resource graphs that have a "knee," or location where the marginal utility improvement sharply decreases, suggest a natural resource allocation. In the case of a linear utility curve, adding an extra resource will return the same marginal expected utility. One benefit of such an analysis is that in budget meetings, security experts can show the expected impact to safety as the budget changes.

Lastly we also note that significant work has gone into speeding up the Bayesian Stackelberg solver. While detailed timing analysis often features prominently in computer science papers, for the purposes of evaluating ARMOR it is sufficient that the system runs "quickly enough" to meet the needs of the LAWA police on the size of problem instances they face on a daily basis. Other speedup techniques may be necessary in much larger domains, such as when scheduling over hundreds of thousands of different targets (11).

3.2 Human behavioral experiments

ARMOR's game-theoretic model uses strong assumptions about the attacker's rationality to predict how they will be-



(a) 1 Checkpoint

(b) 2 Checkpoints

(c) 3 Checkpoints

(d) Canines

Figure 3: Comparisons of ARMOR's schedules with a uniform random baseline schedule. Figures a–c show the utility of schedules for 1–3 vehicle checkpoints varying the relative probability of two different attacker types. The x-axes show the probability of the two attacker types (where 0 corresponds to 0% attack type 2, and 100% attack type 1) and y-axes show the expected utility of ARMOR (using the DOBBS solver) and a uniform random defense strategy. Figure d shows that DOBSS can outperform the baseline, even using many fewer K9 units. The x-axis shows the results from seven different days, and the y-axis shows the expected utility for the different scheduling methods.

have and optimize accordingly. Humans often do not always conform to the predictions of strict equilibrium models (though some other models offer better predictions of behavior (8)). In addition, ARMOR assumes that an attacker can perfectly observe the security policy, which may not be possible in reality.

We have run controlled laboratory experiments with human subjects to address both of these concerns (17). In these experiments, subjects play a "pirates and treasure" game designed to simulate an adversary planning an attack on an LAX terminal, shown in Figure 5. Subjects are given information about the payoffs for different actions and the pirates' strategy for defending their gold (analogous to the security policy for defending airport terminals). Subjects receive payments based on their performance in the game.

These experiments have provided additional support for quality of ARMOR's schedules against human opponents. First, they suggest that the assumptions imposed by the game-theoretic model are reasonable. Second, we have tested many conditions, varying both the payoff structure and the observation ability, ranging from no observation of the defense strategy to perfect observation. The results show that ARMOR's schedules achieve higher payoffs than the uniform random benchmark across all of the experimental conditions tested, often by a large margin.[2] These results demonstrate that ARMOR schedules outperform competing methods when humans are trying to defeat the defender.

3.3 Operational record

A potentially useful test of ARMOR would be to compare the risk level at the airport with and without the system in place. This is problematic for several reasons that we discuss in more depth later on, including the sensitivity of the relevant data and the impossibility of controlling for many important variables. However, there is some public information that can be of use in evaluating the performance of the system, including arrest records. There have been many success stories, prompting significant media coverage. For example, in the month of January this year, the following

[2] New defense strategies developed in this work show even better performance against some (suboptimal) human adversaries by explicitly exploiting the attacker's weaknesses.

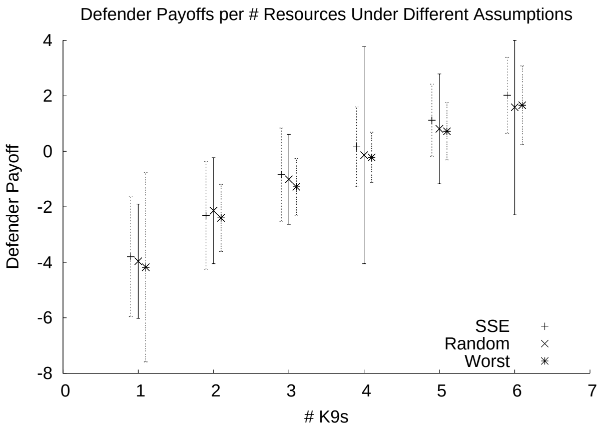


Figure 4: This graph shows game theoretic evaluations of the K9 scheduling program for different numbers of resources, each averaged over 161 trials (error bars show standard errors). First, the SSE assumption is reasonable, as two different defender action selection mechanisms yield little change to the defender payoff. Second, note that the defender utility of additional resources appears approximately linear.

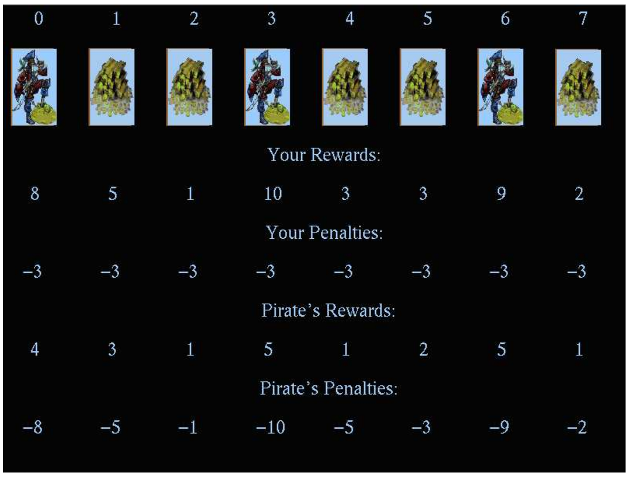
seven stops discovered one or more firearms, resulting in five arrests:

1. January 3, 2009: Loaded 9/mm pistol discovered
2. January 3, 2009: Loaded 9/mm handgun discovered (no arrest)
3. January 9, 2009: 16 handguns, 4 rifles, 1 pistol, and 1 assault rifle discovered — some loaded
4. January 10, 2009: 2 unloaded shotguns discovered (no arrest)
5. January 12, 2009: Loaded 22/cal rifle discovered
6. January 17, 2009: Loaded 9mm pistol discovered
7. January 22, 2009: Unloaded 9/mm pistol discovered (no arrest)

This data, while not conclusive, is encouraging. It appears that potential attackers are being caught at a high rate at ARMOR-scheduled checkpoints.

3.4 Qualitative expert evaluations

Security procedures at LAX are subject to numerous internal and external security reviews (not all of which are public). The available qualitative reviews indicate ARMOR is both effective and highly visible. Director James Butts of the LAWA police reported that ARMOR "makes travelers safer," and Erroll Southers, Assistant Chief of LAWA police, told a Congressional hearing that "LAX is safer today than it was eighteen months ago," due in part to ARMOR. A recent external study by Israeli transportation security experts concluded that ARMOR was a key component of the LAX defensive setup.



	0	1	2	3	4	5	6	7
Your Rewards:	8	5	1	10	3	3	9	2
Your Penalties:	−3	−3	−3	−3	−3	−3	−3	−3
Pirate's Rewards:	4	3	1	5	1	2	5	1
Pirate's Penalties:	−8	−5	−1	−10	−5	−3	−9	−2

Figure 5: Screenshot of the "pirates and treasure" game

ARMOR was designed as a mixed initiative system that allows police to override the recommended policies. In practice, users have not chosen to modify the recommended schedules, suggesting that users are confident in the outputs. While such studies are not very useful for directly quantifying ARMOR's benefit, it would be very hard to deploy the system without the support of such experts. Furthermore, if there were an "obvious" problem with the system, such experts would likely identify it quickly.

We have also compared ARMOR-enabled scheduling with previous LAWA practices (7). First, all checkpoints previously remained in place for an entire day, whereas checkpoints are now are moved throughout the day according to ARMOR's schedule (adding to the adversary's uncertainty). Second, before ARMOR only a single checkpoint was manned on any given day; multiple checkpoints are now used (due to an increased security budget). Third, a fixed sequence of checkpoints was defined (i.e., checkpoints 2, 3, 1, etc.), to create a static mapping from date to checkpoint. This sequence was not optimized according to the importance of different targets and the sequence would repeat (allowing the attacker to anticipate which checkpoint would be manned on any given day).

Expert opinions have said that an important benefit of the system is its transparency and visibility which contribute to deterrence. ARMOR assumes that adversaries are intelligent and have the ability to observe the security policy: knowing about the system does not reduce its effectiveness. The deployment of ARMOR has been quite visible: ARMOR has been covered on local TV stations (including FOX and NBC), in newspapers (including the LA Times and the International Herald Tribune), and in a national magazine (Newsweek).

4 Dimensions of comparison

Evaluating deployed security systems poses many challenges and there is currently no "gold standard" that can be applied in all cases. Our general approach is based on cost-benefit analysis, with the goal of maximizing the utility of the deployed system. A key challenge in applying this methodology to security domains is that many costs and benefits are difficult, or even impossible, to measure directly. For this reason, it is important to carefully consider which metrics of costs and benefits are desirable, and what sources of data are available to estimate these metrics. We thus categorize representative tests in terms of the assumptions they make, relative accuracy, and the cost of running the test. We first discuss three general dimensions of evaluation (Section 4.1) and then a fourth security-specific dimension (Section 4.2). Each type of test has inherent limitations and it is important to draw on as many different categories as possible to provide a compelling validation of a deployed system.

4.1 Test categories

Possible evaluations cover a broad spectrum of evaluation methods. At one end, mathematical analysis is relatively convenient, but requires strong and sometimes questionable assumptions (14; 4). At the other, situated tests using the actual personnel and equipment are very realistic, but also very costly and may not be able to directly measure desired variables. Along this spectrum, we group tests according to their type, their accuracy, and cost:

Test Types:

- Mathematical: formal reasoning using a precise model
- Computational simulation: Computational simulations of varying degrees of abstraction/realism
- Controlled laboratory studies: Testing systems with human subjects can account for human decision making, which may be suboptimal or irrational
- Natural experiments: Observe the behavior the the deployed system by gathering data without intervention
- Situated studies: Testing a deployed system provides the most realistic data, but at high cost
- Qualitative expert studies: Domain experts can examine a system and give a holistic evaluation

Accuracy: Different categories of evaluation offer different tradeoffs in the realism of their assumptions, as well as the precision and repeatability of the results. A mathematical model is typically precise, but dependent on modeling assumptions. On the other hand, real-world tests make fewer assumptions and simplifications, but it may not be possible to draw strong conclusions from a small number of trials and repeatability is often low.

Cost: Test vary dramatically in cost. In addition to monetary costs, situated tests require the time of domain experts and personnel. A special concern for security domains is that simulated attacks where security personnel are not informed before the event may be quite dangerous to participants.

4.2 Quantitative metrics

We now shift our attention to the variety of different metrics that different tests can measure. The fundamental goal of a security system is to maximize utility, which can be decomposed into minimizing deployment cost, attack frequency, and expected damage of attacks. These *primary* metrics are not directly measurable in all types of tests, so we must often fall back on *secondary* metrics that are correlated with one or more primary metrics (and therefore, overall utility). Here we describe a representative set of such secondary metrics, commenting on their benefits and detriments.

- **Attacks Prevented:** How many attacks in progress are interdicted? *Pro:* This metric directly measures the benefit of reduced attack damage/frequency. *Con:* The total number of attempted attacks may be unobservable (for instance, it is not known how many weapons have been smuggled past ARMOR checkpoints) and quite rare.
- **Attacks Deterred:** How many planned attacks are abandoned due to security measures? *Pro:* Attack deterrence may be a primary benefit of security (9; 4). *Con:* Deterrence is generally impossible to measure directly.
- **Planning Cost:** How much time and cost is necessary to plan an attack? *Pro:* Increased planning costs provides deterrence and opportunities to detect terrorist activities before an attack. *Con:* This cost is difficult to measure directly, and motivated attackers may have significant planning resources.[3]
- **Attacking Resources Required:** Can a single attacker with simple equipment cause significant damage? Or is sophisticated equipment and/or multiple attackers required? *Pro:* Like increasing planning cost, increased resources require larger attacker efforts, improving the chance of detection or infiltration. *Con:* Attackers may have sufficient resources, regardless.
- **Attack Damage:** What is the expected consequence of a successful attack? *Pro:* Possible consequences are relatively easy to estimate, as they are less dependent on human decisions. *Con:* Determining which attacks are most likely is still difficult, and there may be high variance. Multiple assumptions must hold

[3] For instance, see http://www.globalsecurity.org/security/profiles/dhiren_barot.htm

about the attackers' behavior and preferences for the reasoning to be correct (8).

- **Implementation Cost:** What are the implementation and maintenance costs for a particular measure, including detrimental effects such as inconvenience to passengers, lower cargo throughput, etc.? *Pro:* Such a measurement can help decide which security measurements to implement. *Con:* All effects, positive and negative, must be quantified.

- **Expert Evaluation:** Are domain experts satisfied with the system? *Pro:* Security experts, who spend their career addressing such issues, have well informed opinions about what works and what does not. *Con:* Expert evaluations may identify security flaws but generally are not quantitative nor consistent across different experts.

5 Evaluation options

The previous sections introduced a classification system for different types of tests and metrics that be useful to measure. We now list and discuss possible evaluations that can be conducted in a security domain, in the context of the above discussion. The evaluation options are situated within the proposed framework and categorized according to the type of test, relative accuracy, cost, and which metric(s) can be measured. The decision of which test(s) to run requires weighing each of these factors.

1. **Game Theoretic Analysis:** Given assumptions about the attacker (e.g., the payoff matrix is known), game theoretic tools can be used to determine the attacker's expected payoff. Additionally, deterrence can be measured by including a "stay home" action, returning neutral reward.[4]

   (a) **Attacker Resources vs. Damage:** A game theoretic analysis can evaluate how attacker observation, equipment, and attack vectors can change the expected attacker payoff. Only defensive measures known by the researcher can be considered, but such an analysis will provide an estimate of attack difficulty, an indirect measure of deterrence.

   (b) **Defense Dollars vs. Successful Attack:** A game theoretic analysis can measure how attacker success varies as security measures are added (e.g., implementing a new baggage screening process), or increasing the strength of an existing measure (e.g., adding checkpoints). Such an analysis may help ensure that resources are not over-committed and provide organizations with quantitative data to assist with budgeting.

2. **Simulated Attacks:** A simulator with more or less detail can be constructed to model a specific security scenario. Such modeling may be more realistic than a game theoretic analysis because structure layout, simulated guard capabilities, and agent-level policies[5] may be incorporated.

3. **Human Studies:** Human psychological studies can help to better simulate attackers in the real world. Evaluations on an abstract version of the game may test base assumptions, or a detailed rendition of the target in a virtual reality setting with physiological stress factors could test situated behavior. Human subjects may allow researchers to better simulate the actions of attackers, who may not be fully rational. Human tests suffer from the fact that participants are not drawn from the same population as the actual attackers.

4. **Foiled Attacks:** The number of attacks disrupted by a security system can provide a sanity check (i.e., it disrupts a non-zero number of attacks). If the metric is correlated with an estimated number of attacks, it may help estimate of the attacker percentage captured. Enabling and disabling the security system and observing how the number of foiled attacks changes would be more accurate, but this methodology is likely unethical in many real world settings.

5. **Red Team:** Tests in which a "Red Team" of qualified security personnel attempt to probe security defenses provides realistic information in life-like situations using the true defenses (including those that are not visible). However, such a test is very difficult to conduct as some security must be alerted (so that the team is not endangered) while remaining realistic, the tests are often not repeatable, and a single test is likely unrepresentative.

6. **Expert Evaluation:** Security experts — internal or external — may holistically evaluate a target's defenses, including both visible and non-visible, and provide a high-level security assessment.

7. **Deterrence Measurement:** Different methods for directly estimating deterrence can be used, such as estimating how likely an attacker is to know about a security precaution and how that knowledge will affect the likelihood of attack. A more quantitative approach would allow attackers to choose between different actions that attack the defended target and actions that attack a different target.[6]

[4] Some attackers may be set on attacking at any cost and may be modeled with a "stay home" action returning a large negative reward.

[5] One exciting direction, as yet unexplored, is to incorporate machine learning into such policies. Such an extension would allow attackers to potentially discover flaws in the system, in addition to modeling known attacker behaviors.

[6] Although this may seem myopic, institution-level security measures are designed to protect a single target; if ARMOR causes attackers to be deterred and attack elsewhere, the security measure has successfully defended LAX. If a measure was designed to cause attackers to *never*

8. **Cost Study:** A cost estimate for an entire location may examine multiple security measures and different levels of staffing, as well as measuring each resource's total cost. Some intangible factors may be very difficult to determine, such as quantifying a decrease in civil liberties.

6 ARMOR evaluation, revisited

This section first re-examines the current evaluations presented in Section 3 to summarize the state of the system's evaluation and then discusses what additional experiments could/should be performed based on the framework presented above.

Existing evaluations of the deployed ARMOR system fall into the Mathematical, Controlled Laboratory, Natural Experiments, and Qualitative categories. These represent a fairly broad range of types of evaluations, showing that ARMOR works well in theory, and that security experts agree it is beneficial. The controlled laboratory experiments, qualitative evaluations, and (sparse) data from natural experiments are particularly interesting in that they go beyond the framework of the game-theoretic model to test it's key assumptions. In many ways, this level of evaluation goes beyond what is typical of applications, even those deployed in real-world settings. Overall, we find strong evidence to support the use of ARMOR over previous methods (notably, hand-crafted or uniform random schedules).

Nevertheless, our framework also suggests new directions that could fill in gaps in the existing evaluation of ARMOR. This is particularly important as we move forward and wish to evaluate ARMOR against more sophisticated alternatives than the hand-crafted and uniform random baselines. In cases where the comparison is less clear-cut, we may need additional metrics to make a compelling argument for one approach or another. Based on our analysis above, we suggest several possible directions for future evaluations of ARMOR and similar systems:

1. None of the current evaluations effectively measure the cost of deploying ARMOR. New analysis should estimate the cost of deploying ARMOR at a new location, both in monetary terms and in side effects. For example, does using ARMOR result in increased congestion or wait times for travelers? It would also be useful to quantify the time required to create hand-crafted schedules instead of using ARMOR.

2. Any additional data we can gather about the effects of ARMOR on risk at LAX would be incredibly valuable for evaluation. Hard numbers are quite difficult to obtain due to security concerns, but efforts to find alternatives should continue. One that is frequently suggested is using security experts to "Red Team" the airport and plan or simulate attacks against it. While this would undoubtedly provide useful information, it is very costly and would require the approval of the airport authorities. Truly live red team operations are generally not conducted due to the risks they create for security personnel.

3. It would be useful to correlate the number of suspected attackers stopped at checkpoints with the number of suspected attackers stopped by other security methods over time. If the number of people detained at checkpoints increases after ARMOR was deployed and the number of people detained by other methods stayed the same (or fell), it is likely that ARMOR is more successful than the previous checkpoint strategy. Currently, such arrest statistics are considered sensitive and are not available to researchers.

4. Another approach for testing the assumptions of our game-theoretic model and the quality of the payoffs elicited from the security experts is to build more detailed computer simulations of airport operations and potential attack scenarios. These simulations themselves also make assumptions, but it would potentially improve reliability to model and understand the domain using two very different modeling frameworks at different levels of abstraction.

5. A weakness of the current evaluation is the lack of an effective measure of deterrence. This is an inherently difficult aspect to capture, as the important variables cannot be observed in practice. One possibility is to explore deterrence more carefully in the game-theoretic model. For example, attackers could be given the option of attacking other targets in addition to LAX. Combined with sensitivity analysis and behavioral experiments, this could be a way to better understand the effects of deterrence.

7 Related work

Security is a complex research area, spanning many disciplines, and policy evaluation is a persistent challenge. Many security applications are evaluated primarily on the basis of theoretical results; situated evaluations and even laboratory experiments with human subjects are relatively rare. In addition, existing general methodologies for risk and security evaluation often rely heavily on expert opinions and qualitative evaluations.

Lazarick (14) is a representative example which relies heavily on expert opinions. In the study, seven tools/approaches used to evaluate airport security were compared as part of a competitive bidding process. At the end of the multi-year security initiative, the highest ranked evaluation methodology relied on averaging qualitative expert evaluations.

A second example of a high-level methodology for per-facility and regional risk assessment, such as described by Baker (1). The methodology relies heavily on expert

attack (or fail at any attack, anywhere), our definition of deterrence would have to be significantly modified.

Evaluation Summary

Test	Type	Accuracy	Cost	Prevented	Deterred	Plan Cost	Resources	Cost	Damage	Qualitative
Game Theory	Mathematic	High	Low	✓	✓				✓	
Attacker Resources/Payoff	Mathematic	High	Low	✓		✓	✓		✓	
Defense Dollars / Damage	Mathematic	High	Low	✓				✓	✓	
Simulated Attacks	Simulation	High	Low	✓					✓	
Human Studies	Human	Med	Med	✓					✓	
Foiled Attacks	Natural	Low	Low	✓						
Red Team	Situated	Low	High	✓					✓	
Expert Evaluation	Qualitative	Low	Med	✓	✓	✓	✓		✓	✓
Deterrence Measurement	Math. / Qual.	Low	Low		✓	✓				✓
Cost Study	Math. / Qual.	Med	Low	✓				✓		✓

Table 1: This table summarizes our proposed evaluation methods by suggesting where each falls along the three general dimensions and which of the seven security-specific metrics are measured.

opinions and evaluations from local technical staff/experts, similar to Lazarick (14). The three key questions in the methodology are: (1) Based on the vulnerabilities identified, what is the likelihood that the system will fail? (2) What are the consequences of such failure (e.g., cost or lives)? (3) Are these consequences acceptable? Such an approach enumerates all vulnerabilities and threats in an attempt to determine what should (or must) be improved. There is no quantitative framework for evaluating risk.

Many in the risk analysis community have recently argued for game theory as a paradigm for security evaluation, with the major advantage that it explicitly models the adaptive behavior of an intelligent adversary. Cox (13) provides a detailed discussion of the common "Risk = Threat × Vulnerability × Consequence" model, including analysis of an example use of the model. There are several arguments raised as weaknesses of the approach, including (1) the values are fundamentally subjective (2) rankings of risk are often used, but are insufficient (3) there are mathematical difficulties with the equation, including dependencies between the multiplied terms, and (4) the model does not account for adaptive, intelligent attackers. One of the main recommendations of the paper is to adopt more intelligent models of attacker behavior, instead of more simple, static, risk estimates.

Bier et al. (3) provide a high-level discussion of game-theoretic analysis in security applications and their limitations. The main argument is that the *adaptive* nature of the terrorist threat leads to many problems with static models — such models may overstate the protective value of a policy by not anticipating an attacker's options to circumvent the policy. They explicitly propose using quantitative risk analysis to provide probability/consequence numbers for game-theoretic analysis.

Beir (4) performs a theoretical analysis of the implications of a Bayesian Stackelberg security game very similar to the one solved by ARMOR, although most of the analysis assumes that the defender does *not* know the attacker's payoffs. The primary goal is to examine intuitive implications of the model, such as the need to leave targets uncovered in some cases so as not to drive attackers towards more valuable targets. There are no "real world" evaluation of the model. Other work (2) considers high-level budget allocation (e.g., to large metropolitan areas). While the study uses real data, its focus is not model evaluation but the implications resulting from the model.

Game theory does have much to offer in our view, but should not be considered a panacea for security evaluation. One difficulty is that human behavior often does not correspond exactly to game-theoretic predictions in controlled studies. Weibull (22) describes many of the complex issues associated with testing game-theoretic predictions in a laboratory setting, including a discussion of the ongoing argument regarding whether people typically play the Nash equilibrium or not (a point discussed at length in the literature, such as in Erev et al. (8)). This is one reason we believe behavioral studies with humans are an important element for security system evaluation.

Many of the issues we describe in acquiring useful real-world data for evaluation purposes are mirrored in other types of domains. Blundell and Costa-Dias (5) describe approaches for experimental design and analysis of policy proposals in microeconomics, where data is limited in many of the same ways: it is often not possible to run controlled experiments and many desired data cannot be observed. They describe several classes of statistical methods for these cases, some of which may be valuable in the security setting (though data sensitivity and sparse observations pose significant additional challenges). In truth, it is often hard to evaluate complex deployed systems in general — in our field a test of the prototype often suffices (c.f., Scerri et al. (18)).

Jacobson et al. (9) describe a deployed model for screening airline passenger baggage. The model includes detailed information regarding estimated costs of many aspects of the screening process, including variables for probability of attack and cost of a failed detection, but these are noted to be difficult to estimate and left to other security experts to determine. One particularly interesting aspect of the approach is that they perform sensitivity analysis on the model in order to assess the effect of different values on the overall decisions. Unfortunately, the authors have little to say about actually setting the input values to their model; in fact, there is no empirical data validating their screening approach.

Kearns and Ortiz (10) introduce algorithms for a class of "interdependent" security games, where the security investment of one player has a positive externality and increases the security of other players. They run the algorithms on data from the airline domain but do not directly evaluate their approach, instead looking at properties of the equilibrium solution and considering the broad insight that this solution yields regarding the benefits of subsidizing security in such games.

Lastly, the field of *fraud detection* (12), encompassing credit card fraud, computer intrusion, and telecommunications fraud, is also related. Similar to the physical security problem, data is difficult to access, researchers often do not share techniques, and deterrence is difficult (or impossible) to measure. Significant differences include:

1. Humans can often classify (in retrospect) false positives and false negatives, allowing researchers to accurately evaluate strategies.

2. Companies have significant amounts of data regarding known attacks, even if they do not typically share the data outside the company. Some datasets do exist for common comparisons (c.f., the 1998 DARPA Intrusion Detection Evaluation data[7]).

3. The frequency of such attacks is much higher than physical terrorist attacks, providing significant training/evaluation data.

4. Defenders can evaluate multiple strategies (e.g., classifiers) on real-time data, whereas physical security may employ only, and evaluate, one strategy at a time.

8 Conclusions

While none of the evaluation tests presented in Section 5 can calculate a measure's utility with absolute accuracy, understanding what each test *can* provide will help evaluators better understand what tests *should* be run on deployed systems. The goal of such tests will always be to provide better understanding to the "customer," be it researchers, users, or policy makers. By running multiple types of tests, utility (the primary quantity) can be approximated with increasing reliability.

At a higher level, thorough cost-benefit analyses can provide information to policy makers at the inter-domain level. For instance, consider the following example from Tengs and Graham (20):

> To regulate the flammability of children's clothing we spend $1.5 million per year of life saved, while some 30% of those children live in homes without smoke alarms, an investment that costs about $200,000 per year of life saved.

While such a comparative cost-benefit analysis is beyond the scope of the current study, these statistics show how such an analysis can be used to compare how effective measures are across very different domains, and could be used to compare different proposed security measures.

In the future we plan to use this framework to help decide which evaluation tests are most important to determine ARMOR's utility, as suggested in Section 6. Additionally, we intend to continue collaborating with security experts to determine if our framework is sufficiently general to cover all existing types of security tests, as well test how the framework can guide evaluation in additional complex domains.

Acknowledgements

The authors would like to thank James Pita and Vicki Bier for helpful suggestions, Prakhar Garag and Rahul Jain for assistance in running experiments, Sergeant Jim Holcomb for supplying ARMOR-related arrest data, and First Sargent Cruz for discussing pre-ARMOR scheduling procedures. This research was supported by the United States Department of Homeland Security through the National Center for Risk and Economic Analysis of Terrorism Events (CREATE) under grant number 2007-ST-061-000001. However, any opinions, findings, and conclusions or recommendations in this document are those of the authors and do not necessarily reflect views of the United States Department of Homeland Security.

References

[1] G. H. Baker. A vulnerability assessment methodology for critical infrastructure sites. In *DHS Symposium: R and D Partnerships in Homeland Security*, 2005.

[2] V. M. Bier, N. Haphuriwat, J. Menoyo, R. Zimmerman, and A. M. Culpen. Optimal resource allocation for defense of targets based on differing measures of attractiveness. *Risk Analysis*, 28(3):763–770, 2008.

[3] V. M. Bier, Jr. L. A. Cox, and M. N. Azaiez. Why both game theory and reliability theory are important in defending infrastructure against intelligent attacks. In *Game Theoretic Risk Analysis and Security Theats*, volume 128. Springer US, 2009.

[7] See `http://www.ll.mit.edu/mission/communications/ist/index.html` for data and program details.

[4] V. M. Bier. Choosing what to protect. *Risk Analysis*, 27(3):607–620, 2007.

[5] R. Blundell and M. Costa-Dias. Alternative approaches to evaluation in empirical microeconomics. *Journal of Human Resources*, 2009.

[6] V. Conitzer and T. Sandholm. Computing the optimal strategy to commit to. In *Proc. of EC*, 2006.

[7] First Sargent Cruz. Personal communication, August 20 2009.

[8] I. Erev, A. E. Roth, R. L. Slonim, and G. Barron. Predictive value and usefulness of game theoretic models. *International Journal of Forecasting*, 18(3):359–368, 2002.

[9] S. H. Jacobson, T. Karnai, and J. E. Kobza. Assessing the impact of deterrence on aviation checked baggage screening strategies. *International J. of Risk Assessment and Management*, 5(1):1–15, 2005.

[10] M. Kearns and L. E. Ortiz. Algorithms for interdependent security games. In *Neural Information Processing Systems (NIPS)*, 2003.

[11] Christopher Kiekintveld, Manish Jain, Jason Tsai, James Pita, Fernando Ordónez, and Milind Tambe. Computing optimal randomized resource allocations for massive security games. In *AAMAS*, 2009.

[12] Y. Kou, C. Lu, S. Sinvongwattana, and Y.P. Huang. Survey of fraud detection techniques. In *Proc. of IEEE Networking*, 2004.

[13] Jr. L. A. Cox. Some limitations of "risk = threat x vulnerability x consequence" for risk analysis of terrorist attacks. *Risk Analysis*, 28(6):1749–1761, 2008.

[14] R. Lazarick. Airport vulnerability assessment – a methodology evaluation. In *Proc. of 33rd IEEE International Carnahan Conference on Security Technology*, 1999.

[15] Praveen Paruchuri, Jonathan P. Pearce, Janusz Marecki, Milind Tambe, Fernando Ordonez, and Sarit Kraus. Playing games with security: An efficient exact algorithm for Bayesian Stackelberg games. In *AAMAS-08*, 2008.

[16] J. Pita, M. Jain, C. Western, C. Portway, M. Tambe, F. Ordonez, S. Kraus, and P. Paruchuri. Deployed ARMOR protection: The application of a game theoretic model for security at the Los Angeles International Airport. In *Proc. of AAMAS*, 2008.

[17] J. Pita, M. Jain, M. Tambe, F. Ordonez, S. Kraus, and R. Magori-Cohen. Effective solutions for real-world stackelberg games: When agents must deal with human uncertainties. In *Proc. of AAMAS*, 2009.

[18] P. Scerri, T. Von Goten, J. Fudge, S. Owens, and K. Sycara. Transitioning multiagent technology to UAV applications. In *Proc. of AAMAS Industry Track*, 2008.

[19] D. Stevens, T. Hamilton, M. Schaffer, D. Dunham-Scott, J. J. Medby, E. W. Chan, J. Gibson, M. Eisman, R. Mesic, C. T. Kelly, J. Kim, T. LaTourrette, and K. J. Riley. Implementing security improvement options at Los Angeles international airport, 2009. `www.rand.org/pubs/documented_briefings/2006/RAND_DB499-1.pdf`.

[20] T. O. Tengs and J. D. Graham. Risks, costs, and lives saved: Getting better results from regulation. In *The opportunity costs of haphazard social investments in lifesaving*. American Enterprise Institute, Washington, 1996.

[21] W. A. Wagenaar. Generation of random sequences by human subjects: A critical survey of literature. *Psychological Bulletin*, 77(2):65–72, 1972.

[22] J. Weibull. Testing game theory. In Steffen Huck, editor, *Advances in Understanding Strategic Behavior: Game Theory, Experiments and Bounded Rationality.*, pages 85–104. Palgrave MacMillan, 2004.

Application of Microsimulation to the Modelling of Epidemics and Terrorist Attacks

Ian Piper and Daniel Keep
School of Computer Science and Software Engineering
University of Wollongong, Australia
E-mail: ian@uow.edu.au

Tony Green and Ivy Zhang
School of Risk and Safety Science
University of New South Wales, Australia
E-mail: a.green@unsw.edu.au

Keywords: modelling, microsimulation, terrorism, epidemiology

Received: August 10, 2009

In this paper, a novel approach to behavioural modelling is presented with reference to biological infection spread in communities. Its potential application to various terrorist-related scenarios is discussed in relation to attack point simulation and interdiction simulation.

Povzetek: Predstavljen je nov način modeliranja bioloških infekcij predvsem v zvezi s terorizmom.

1 Introduction and background

The motivation of this project is to develop a computer model which allows the modelling of human response to a variety of human threat scenarios.

These may range across both natural threats such as disease, fire and flood and unnatural events including accident and deliberate acts such as terrorist attacks.

The tool is being developed in response to a number of requests from third parties, including both govenment and commercial bodies, to address a wide range of roles.

Forensic: can we gain an understanding of events that have already taken place?

Speculative: can we develop a reasonable "what if" model for potential threats?

Decision Support: can we dynamically model and predict the course of an evolving threat?

Training: can we develop realistic models for the training of response personnel?

Evaluative: can we test the validity of assumptions made in other modelling techniques?

Terrorist attacks can take many different forms using a wide range of weapon types and this poses some considerable problems in defining both the risks that arise in modern urban environments as well as optimising interdiction and response strategies. Critical infrastructure is dispersed and provides multiple vulnerabilities and opportunities for attack. As attacks both in London and, more recently, in Mumbai have shown, modern terrorism should be thought of as requiring a 3D spatial vulnerability approach to counter terrorism(7). While both game theory and queuing theory have been used for assessing terrorism events we present an alternative approach to this problem which overcomes their fundamental problem, that of estimating the number that get through to attack.

Microsimulation is a discrete simulation technique which allows for the modelling of the behaviour of single individuals in a complex system (4)(13). It was originally devised for financial and economic modelling (17)(15), but is generally applicable to a wide range of scenarios.

In the current research project, we have created a modular, scalable microsimulation package, called Simulacron, which allows for the rapid creation of microsimulations involving large numbers of people interacting with each other and their environment.

The framework is designed to be scalable and distributable, implementing all interactions in terms of distinct locations and individuals. The state of these locations and individuals is flexible and can be arbitrarily extended. Model behaviour is broken into individual modules which can be combined as needed.

In addition to this simulation framework, we have also developed a number of support tools including a prototype non-linear visualisation package which allows for the creation of complex visualisations by non-programming personnel.

Models are specified using an XML dialect. The creation of very large and complex models is achieved through the use of a preprocessor which allows the instantiation of generalised templates into concrete data sets.

The initial study undertaken was the simulation of an influenza epidemic (8) at the Royal Naval School (RNS) in Greenwich, London in 1920. It was of particular interest due to the relatively complete information available regarding the outbreak, including the progress of the disease over time and its infection mechanism (6) as well as the behaviour of the population in its "normal" state (14). This scenario was also attractive in that it occurred in an essentially closed community for which we have detailed historical documentation (2).

The model of the RNS outbreak involved the creation of nine dormitories, nine reading rooms, around 40 classrooms and roughly 15 other locations including the hospital. Student behaviours were established in seasonal class groups involving a total of 951 students. In addition, the behaviours of 27 staff were also modelled. Each of these individuals is assigned a unique set of infectious parameters based on statistical distributions shared between all the participants. When the simulation starts, each person is sent to their appropriate location based on their schedule. The movements and interactions, including cross infection, of the individuals is then modelled over time by Simulacron.

Behaviours are a mixture of constrained (students must follow their timetables, sleep in assigned dormitories, etc.) and unconstrained (students move about freely at playtime). As stated previously, the modular nature of the framework allows us to use whatever behaviours are most appropriate for the model at hand.

The underlying infection model used in the simulation is conceptually a combination of the "Susceptible, Infective, Recovered, Susceptible" (11) and "Susceptible, Exposed, Infective, Recovered" (5) models (commonly known as SIRS and SEIR respectively). In addition to the standard states, we added two mechanisms to the model. The first, "hero" time, allows the simulation of the "I'm too busy to get sick," or "It's just a little cold," phenomenon. The second, isolation, allows for individuals to be removed from the cross-infection domain once disease is detected. The isolation mechanic can be conditionally applied only during particular hours of the day. The parameters are effectively selected by a convergence methodology against the historical data. While the full procedure is still being developed, the results in the case study presented agree reasonably well with other influenza studies (9).

2 Methodology

The simulation environment currently consists of a number of programs: Simulacron, the simulator, and DSTP, the template preprocessor. Planned future additions include Jazz, a visualisation engine and Refinery, an automated parameter estimation system.

The following sections examine these in more detail.

2.1 Simulacron

Early on it was decided to divorce the development of the simulation models themselves from the surrounding support code. This led to the creation of a general-purpose microsimulation framework called Simulacron. On to this scaffolding, special-purpose simulation modules can be attached to craft purpose-built simulation environments. The current module set is discussed later in this document in sections 2.2, 2.3 and 2.4.

The framework has been written using the D Programming Language. This choice was made because it provided a number of advantages over more traditional choices such as C, C++ or Java. Since it compiles to native machine code, it has the performance advantages of C and C++. It also directly incorporates automatic memory management similar to that of Java. Distinct from the above, D's advanced template support has allowed for the creation of complex, repetitive code to be automated within the language including serialisation and XML parsing libraries.

We have also utilised a customised version of Don Knuth's literate programming environment (12) modified to process D instead of C.

The framework was specifically designed to support very large simulations, leading to the adoption of a master/slave architecture, in which the processing of the simulation can be arbitrarily divided among one or more worker (slave) processes under the supervision of the master process. Each of these processes may operate in a loosely-coupled environment such as that provided by a cluster or network of machines. This architecture allows for the development of models of far greater size and complexity than could reasonably be supported by a monolithic, single-process design.

To date, it has been possible to operate with a single slave running on the local machine. This has permitted models with tens of thousands of individuals.

The only requirement that the framework imposes upon the model is that it must be expressible using a combination of distinct locations (called Cells) and individuals (called Peeps[1]). All interactions and behaviours of these are specified by the simulation modules. Due to its modular nature, the framework itself places no requirements for any particular state information for locations or individuals; their state may be arbitrarily extended by each module, by attaching fields to cells and peeps, to allow for composition of model components into arbitrarily complex simulations.

Time in Simulacron is managed via discrete "ticks", time intervals chosen to be small enough to capture important details in the model but large enough to render large simulations practical. The master is responsible for coordinating the slave processes via issuing tick commands. This mechanism was chosen in preference to the alternative, queueing theory, approach (with free-running slaves) as this was markedly simpler to implement given that there is a poten-

[1] Note that these are unrelated to the popular American marshmallow confection of the same name.

tially unbounded number of future events for every individual object in the simulation at any given time.

It is also worth noting that using queuing theory would make it far more difficult to distribute processing as any event has the potential to affect the state of any object in the simulation. This means that no computational node can proceed *other than* the node which has the next event, thus making distribution of the work pointless.

In addition, the framework allows for the definition of multiple states, each referred to by name. Each cell and peep is in one specific state at any given time, independent of the state of other objects in the simulation. Cells and peeps may have their current state changed by modules at any time.

Input to the simulation is via an XML dialect. Although not our first choice, XML has proven to be a good fit to our needs; the extensible nature of the framework demanding the ability to represent structured data of arbitrary complexity, a task well suited to XML. However, this has proven to be prohibitive in terms of defining large data sets, a problem which is in part resolved by the template preprocessing program DSTP discussed in section 2.5.

Output from the simulator can be encoded in a number of formats. Initially, XML reports were used due to the availability of third party tools which could read and process these reports including web browsers such as Internet Explorer and Firefox, which acted as rudimentary viewers, and Excel, which allowed for more involved analysis.

However, these reports had an undesirable overhead both in terms of disk space and processing time. This has been resolved by moving to a more compact and more easily processed database format based on the SQLite engine.

2.2 Movement modules

In itself, Simulacron does not prescribe any behaviours for either locations of individuals. One basic requirement for modelling people is movement. To this end, there are two modules which provide for different aspects of simulated motion.

The first of these is the scheduling module. This allows for each peep to have one cyclic schedule defined per state for them. These schedules are sequences of time and action pairs. When the trigger time is reached, the defined action is performed. Actions can include moving the peep to another cell and changing their state (both deterministically and at random).

For example, one could use a cyclic schedule to define a full week's worth of movement; going to work on weekdays and engaging in leisure activities on the weekends. Alternately, the same could be done by defining a one-day "workday" schedule, a number of one-day "weekend" schedules and selecting between them via the state mechanism.

There is also a similar scheduling mechanism designed for "one-off" events. Whilst the cyclic schedules define the time for events relative to the start of the cycle, the "one-off" schedule defines the time for events as an absolute date and time.

The second movement module is the dispersal module which, in contrast to the scheduling module, is applied to cells. Simply put, at each simulated tick, the cell will disperse each peep currently located within it to a cell randomly selected from a defined set. The destination cell may, in turn, also be associated with a dispersal set. This is used to simulate, for example, children moving about in a playground or movement of people within a large office building.

2.3 Infection Module

This, the first specialised module to be developed, supports the modelling of a single infectious process. In its simplest form it allows the injection of one or more infected peeps into a population of susceptibles. The infection can then be spread between peeps within the same cell.

The infection model, as stated earlier, is a modified combination of traditional models. The infection progresses through several states: susceptible, latent infected, asymptomatic infective, "heroic" infective and symptomatic infective; following this, the peep will either recover (and possibly become immune) or die.

Each of these is present in various existing models (such as those in (11) and (5)), with the exception of "heroic" infective. This state was originally introduced to model people's tendency to shrug off or hide sickness. It is implemented as a linearly increasing chance of "detection" starting at 0% and ending at 100%, at which point the peep becomes symptomatic infective.

This also plays into another feature of our model: isolation. In its most basic form, this causes symptomatic infective peeps to be forcibly sent to an isolation cell until they either recover or die. The heroic infective state interacts with this to produce an increasing chance of a peep being noticed and sent to isolation as the infection progresses.

Isolation can be configured to only be in effect during particular times of the day. This was used in our RNS simulation to limit isolation to daylight hours; people are unlikely to be isolated at night.

Cross infection is possible at every simulated tick. For each susceptible peep in a cell, a random number in $[0, 1)$ is generated which is then compared to the following:

$$\frac{P(\text{Infection}) \times \text{Time step}}{\text{Average infectious time}} \times N$$

where $P(\text{Infection})$ is the peep's chance of being infected given continuous exposure over "Average infectious time" (a parameter which is specified globally), "Time step" is the amount of simulated time that has passed since the last tick and N is the number of infectious peeps in the cell.

An infection may also be spread via the environment itself; cells can be "infected" at which point they can transmit the infection to peeps passing through them in the same manner. This could be used, as an example, to represent an air conditioning system for modelling the spread of Legionella.

It is worth noting that not only is the progression of the infection recorded via a field attached to each peep, but the various parameters that control the infection are as well. This means that every peep in the simulation can have unique infection parameters, which includes the duration of the various stages of the infection, their chance to contract the infection and their chance to recover.

A recent addition was the implementation of "infection masking," a process whereby each peep can be placed into zero or more groups, with membership managed via a bitmap mask. The groups a peep is capable of infecting can then be restricted to a possibly different set of groups. This allows for the representation of complex multi-vector diseases such as avian influenza.

It is interesting to note that there is nothing in either this module or Simulacron itself that limits the module to modelling the spread of an infectious disease among humans. Peeps can be used to represent anything which is capable of spreading or merely contracting an infection: pets, wild animals, birds and even particulate matter or molecules of an airborne pathogen. Similarly, the module can be used to model any process which shares similar viral propagation. This includes things such as memes or even a terrorist recruiter (i.e. infective) subverting dissatisfied (i.e. susceptible) individuals.

2.4 Terrorism module

A recent development, the "TPC" module was created to model simple terrorist scenarios. It allows one to divide all peeps into one of three categories: terrorists, police and civilians; hence the name. Terrorists are modelled as having three important parameters; the first is the "camouflage factor," which represents how adept that individual is at hiding themselves. The higher the factor, the lower their chance of being "detected."

Secondly, terrorists have an attack time; presently, the model only deals with terrorists engaging in suicide attacks. At the appointed time, the terrorist will explode, killing all peeps within the same cell. Finally, terrorists also have a configurable chance of prematurely detonating themselves if they are detected by police.

Police have only one parameter: their perception factor, which represents their capacity for spotting terrorists[2]. Thus, a terrorist's chance of being detected by any given police officer is $F_p(1 - F_c)$ where F_p is the police officer's perception factor and F_c is the terrorists' camouflage factor.[3]

Citizens have no parameters; their only function in the model at present is to serve as cannon-fodder. This is not as needlessly malicious as it may first appear: given a community derived from real-world census data, they can serve to determine when and where an attack is likely to occur. They also provide meaningful data in the event of a premature detonation.

Note that this model does not cover the motions of any of the peeps involved; this is provided by the movement modules.

The precise nature of the terrorist threat may be adjusted to meet the needs of a specific scenario. These may include the following:

Instantaneous lethality: e.g. the detonation of a bomb with a substantial payload.

Delayed lethality: e.g. the release of some chemical or radiological agent which is not instantaneously lethal.

Probabilistic lethality: e.g. the detonation of a low yield or less reliable explosive device.

Infection: e.g. the release of an infectious agent or the terrorist deliberately infecting himself.

2.5 DSTP

Models are specified using a template language, based on XML, which permits the creation of very large data sets via the instantiation of relatively simple macro templates. For example, this allows the specification of a single template for a statistically average individual which can then be instantiated an arbitrary number of times to create a background population into which unique individuals may be "injected" to simulate specific behaviours.

Note that although specifically designed to aid in the creation of Simulacron data sets, DSTP itself is not limited to generating them. It could conceivably be used to output any XML-based format. DSTP also supports self-recursive behaviour where the output of a template may be another template; an integer suffix can be added to filenames which the processor will then decrement when naming the output file.

Like the data set format, the template language is based on XML. It resembles a relatively simple functional programming language with dynamic scope. Its standard constructs are concerned with the definition and substitution/instantiation of variables and templates. It also contains a few primitives for the random sampling of values from normal and uniform distributions; these can be used anywhere a value is expected, allowing an arbitrary mixing of randomly sampled and fixed values.

The language currently lacks constructs such as conditionals, arbitrary looping and arithmetic. Given the complexity of templates already possible with the language, we view this lack as a blessing.[4]

Another key construct is the ability to instantiate a template multiple times with a single instruction. Templates

[2] There is currently no provision for "false positives."

[3] This is normalised to be the probability of detection over a one-minute time span.

[4] The authors take the position that if the template language ever becomes Turing-complete, it would likely be advisable to simply switch to using LISP instead.

can be expanded a specified number of times, for each value in an integral range or for each value in a list of words. It is using this, in concert with the random sampling constructs, that allows for a statistically "generic" person to be defined and then instantiated into a concrete population.[5]

To aid in the construction of more complex data sets, DSTP supports a module system that allows additional constructs to be added without requiring change to the basic processor. These range from very simple automation (such as generating interlinked webs of dispersion cells) to much more complex processes.

One such complex process is the creation of communities: given a set of parameters (such as number of adults, number of children, number of employed, number of houses, etc.) derived from census data, it will attempt to create a community that matches those requirements. This includes automatic creation of families and homes.

Another example is the "office builder" which, given some basic properties such as number of floors and offices, will generate a complete, interlinked office building complete with lifts and street access.

In addition to plug-in modules, common components or even complete templates can be abstracted into libraries via the import mechanism to aid in reuse.

3 Preliminary results

3.1 Infection

The original 1920 outbreak lasted for roughly 25 days. Each simulation was run over 30 virtual days with a time step of five minutes and output every hour. The output represents a snapshot of the entire population, recording the location and infection state for each individual. For infected individuals it also records the source of the infection, and when they became infected. This data allows comparison with the historical number of new cases every day.

Systematic changes to the model parameters in successive simulations allowed investigation of the sensitivity of parameters against the historical data. Furthermore, by varying only the initial random seed, different instances of the same underlying process can be modelled, allowing the determination of statistical parameters, such as mean and standard deviation for such properties as number of deaths.

The infection model is not a standard compartmental SEIRS type model and is more readily suited to backcasting and forecasting methods required for decision support in live situations. Furthermore, because it simulates actions of the individual it can test alternative policies and social controls that are difficult, if not impossible, to test without making some gross assumptions. Because of these attributes, it can be used to test assumptions, made in other techniques, that are not ordinarily testable.

Our technique also allows the investigation of additional properties of the infection process which would be equally difficult to evaluate with traditional statistical modelling. Chief among these is the ability provided to determine "infection chains", a chronological sequence of who infected who, where and when, essential in tracing the exact progression of a disease and identifying critical peeps and cells.

The results, even allowing for the relative simplicity of the model and the inevitable inexactitude of the parameter estimates, showed remarkably close agreement with the historical data, as seen in Figure 1 below, capturing the development, peak and recovery times with surprising accuracy. Contrast this with the result of the more traditional purely statistical model shown in Figure 2.[6]

An unexpected outcome was the presence, among the simulation series, of runs in which, despite all parameters being the same, no epidemic occurred. This suggests that further investigation may be required into the "known" causes of epidemic spread.

3.2 Terrorism

The studies we are currently conducting are more directly of interest to the present audience as they include models related to terrorist activity. Two of these we briefly describe below.

In support of this, a more comprehensive model has been developed involving 14402 individuals, grouped into families, whose overall properties match the statistical census data for Australia, interacting in an environment of 6935 locations including 6000 homes and 935 workplaces, schools, recreational areas, hospitals, etc. The people move within the community according to schedules that emulate employed, part-time employed, unemployed and home workers, primary and secondary school children and infants that again match statistical data for Australian communities.

3.2.1 Attack point simulation—the "living bomb"

The first proposed model addresses the question "What happens if we vary the point of release chosen by a terrorist conducting a biological attack?"

The required model behaviour for this scenario can be achieved, without change to the previously described infection model, by the injection of a "living bomb" represented as a single individual who remains stationary for the duration of the simulation and becomes highly infectious at a predetermined moment in time.

Four locations in a virtual community were used as a preliminary assessment of the impact of location on release from a "living bomb". The four locations were a cinema complex, a club, a large store, and the community hospital.

[5] The same can be said of generic locations.

[6] It could be argued that the statistical model is, in a sense, more accurate than the simulation results as it provides a precise match for the total number of cases. However, the simulation method clearly provides a much more realistic, if less "exact", result.

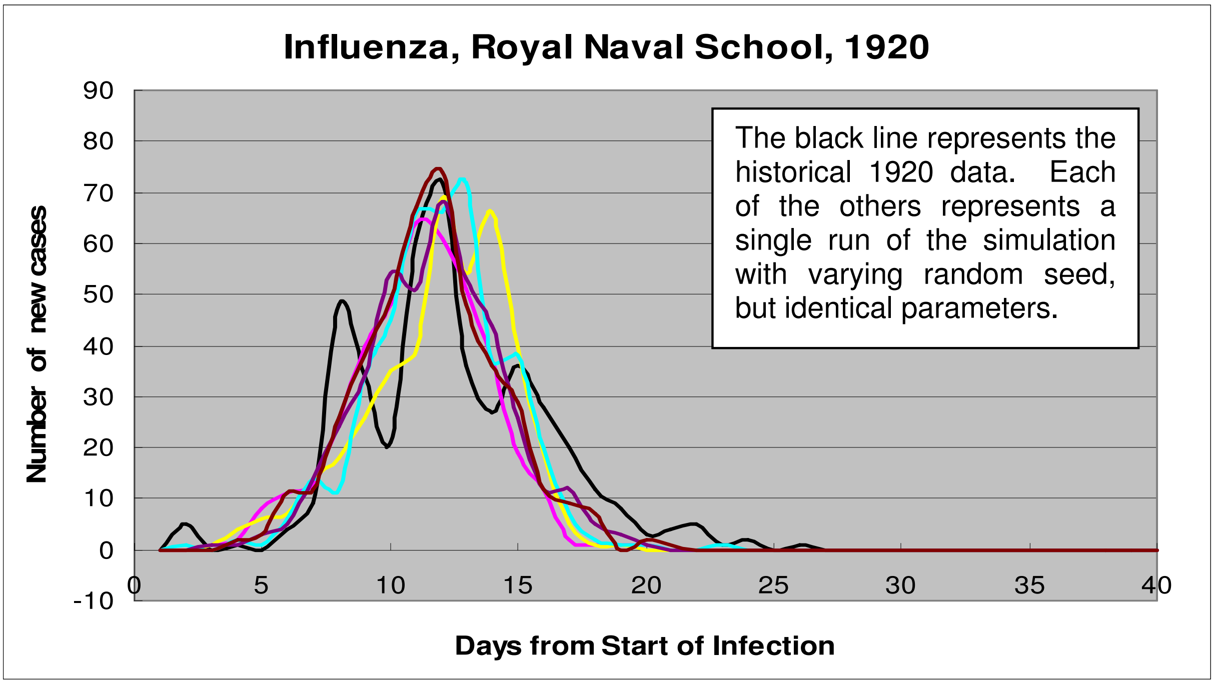


Figure 1: Comparison of simulated runs with historical data

The infection that was used in the simulation was based on smallpox with typical time parameters given by CDC information (3). The probability of infection for each individual was the reproduction rate for an infected person adjusted for the timestep used in the simulation. The initial release emulated a badly constructed device with limited ability to spread infection.

Figure 3 shows the cumulative number of people who visited the four locations. Only two of the simulations resulted in infection spread, the club and the cinema complex with one infection each over the release time. The infection occurred about 10 hours after the release. The pathogen in the absence of the human host was assumed to be viable for approximately 24 hours. The time of first infection is shown in Figure 3. It suggests that in any release there is a threshold of exposure required to spread infection; this can be the number of people or the time exposed. The store and the hospital did not have enough people moving through the building to make it likely that someone contracted the disease.

In the two cases of disease spread, the disease continued to spread through the community. The first appearance of symptoms occurred 14 days after exposure in the cinema and 18 days from the club. By 80 days, 2560 and 2016 people were infected from the cinema and club exposures respectively. With this particular disease, the relatively long incubation period does allow time for intervention, isolation and ring vaccination so long as surveillance systems for the disease identify a case quickly. The second infected person in each simulation became symptomatic 26 and 31 days after the primary exposure by which time there were 4 and 5 additional infecteds.

These early results suggest that the location of release will be extremely important to the number of subsequent cases of infection that occur. While more studies are required to elucidate the sensitivity to population moving through a target and dispersal effectiveness at the point of delivery, the result does have implications for assessment of risk, the provision of resources for dealing with an outbreak and the effectiveness of possible control mechanisms.

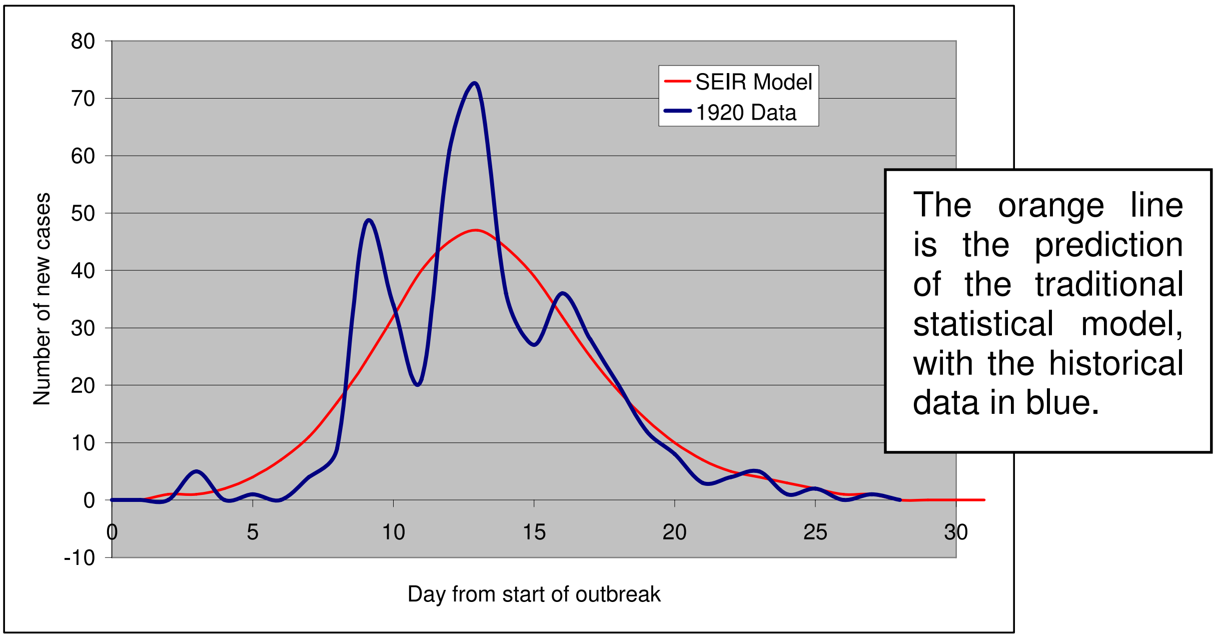
3.2.2 Interdiction simulation

The second proposed model addresses the question “How effective is a specific interdiction regime?”

This basic model may be varied by changing properties in a logical manner. For example, replacing the instantaneous lethality of the terrorist attack with a probabilistic one (a smaller bomb) or replacing it with a conventional infective state simulating the release of a biological agent.

Because of the flexibility of the program, police behaviours may range from completely random to precisely specified, the latter allowing the investigation and validation of predetermined interdiction strategies such as those derived from game-theoretic modelling (16).

The community used in the above “living bomb” scenario was used with the club as a target location for a terrorist attack. The terrorist was embedded in the community undertaking normal activities arriving at the point of explosion just before the time the explosion is due. Police also moving about the community are attempting to stop the attack. In the nine scenarios tested the three factors that are used in the simulation are shown in Table 1. There was one



The orange line is the prediction of the traditional statistical model, with the historical data in blue.

Figure 2: Comparison of SEIR model with historical data

terrorist and ten police in each of these simulations. Figures 4 and 5 show the results of the simulation; the top shows the interception factor as a function of the date and time while the bottom graph shows the location and the number of dead or injured as a function of date and time. The only simulation to get to the target time of 03 Jun 21:23 was run 8. Pre-emptive detonation only occurred in run 0. The other 6 simulations resulted in successful arrest. The decline of the interception factor (top) follows the transition from early to later times in interception eventually resulting in no interdiction and the terrorist reaching the target at the designated time.

These two examples show the potential power in this type of modelling as it can allow the investigation not only of complex attacks but also the requirements for resources, the levels of perception or intelligence assistance and the tactics that are needed to optimise these resources if these types of activities are to be prevented.

Our method differs significantly from other methods for assessing interdiction strategies (1)(10) as it is time rather than event driven and based on the detailed modelling of individual behaviour within population groups rather than more abstract constructs.

Because of this, we can "inject" deterministic behaviour patterns for specific individuals into a background population modelled with randomly varying properties.

The advantage as we see it over existing techniques is that it can produce in simulations both the successes and failures together with full information about the paths to success or failure.

4 Conclusion and future work

A number of improvements to the array of available simulation modules are presently being planned. These include a transportation system including both personal and public transport, the addition of multiple simultaneous infections, chokepoints and other rate-limiting devices and a "baggage" system to assist in accurately modelling airports.

A number of new applications of Simulacron are currently in planning, including the simulation of a large transportation facility and the development of a more comprehensive end-to-end terrorism model including the training of agents, the planning, preparation and conduct of attacks and the subsequent response.

From its inception, the simulation system was intended to be part of an integrated risk modelling and assessment software environment. To this end, it is intended to integrate the Simulacron package with a more fully-realised version of the visualisation environment, allowing a bi-directional real-time flow of data between them.

This package is currently called "Jazz" and is based on a non-linear editor allowing a non-programmer to connect various processing components together to produce a visualisation. This visualisation can then be distributed to other parties and either be used interactively or to "play" back a simulation visually. At the moment, Jazz exists as a prototype with development proper intended to begin before the end of the year.

This will allow dynamic monitoring and modification of the simulation process via an intuitive graphical interface. Key to this process is the planned development of a scriptable supervisor process which will moderate and coordi-

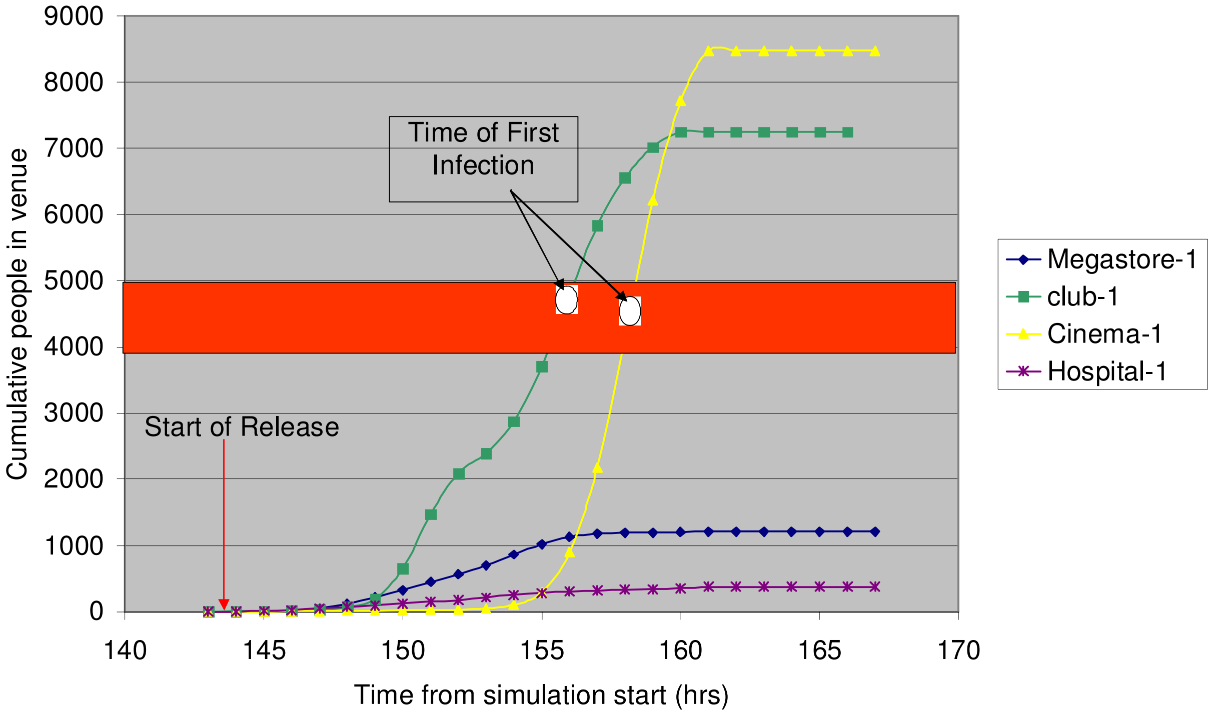


Figure 3: Terrorist smallpox simulation

Run	0	1	2	3	4	5	6	7	8
Camouflage Factor	0.8	0.8	0.8	0.8	0.8	0.8	0.9	0.95	0.99
Perception Factor	0.8	0.6	0.5	0.4	0.2	0.1	0.1	0.1	0.1
Preemptive Factor	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
Interception Factor	0.16	0.12	0.1	0.08	0.04	0.02	0.01	0.005	0.001

Table 1: Interdiction Factors

nate this data exchange.

With this framework, it should be possible to integrate further state of the art simulation packages dealing with such matters as the effects of fire, explosions, etc. This would be achieved by the representation, within Simulacron, of externally simulated events via such mechanisms as a “survivability index” for a set of locations and the dynamic modification of behaviours to represent the response to death, injury and damage to locations. The addition of further Simulacron modules would allow the modelling of emergency response to such events.

One of the key aims of the development of this package was that it be capable of running at better than realtime. This will allow the coupling of non-simulated events to permit the use of the simulation environment to predict reactions and potentially to investigate alternate response strategies in a live system.

Another use is as a forensic tool to analyse past events and to investigate the likely result of alternative intervention strategies.

Work is currently being done on a prototype of the “Refinery” program which will automate the derivation of simulation parameters by iteratively refining them such that the output more closely matches historical or expected results.

Interest in this project has already been expressed by a number of groups within Australia, each of which has seen a different potential use. These range from examination of policy effectiveness in public health, through training scenarios (spot the terrorist) in law enforcement to its use as a decision support environment by emergency response organisations.

Acknowledgement

This work presented in this paper has been supported by the National Security Science and Technology Unit of Prime Minister and Cabinet, Australian Federal Government together with Australian Federal and State Agencies under the NSST CBRN Grant scheme.

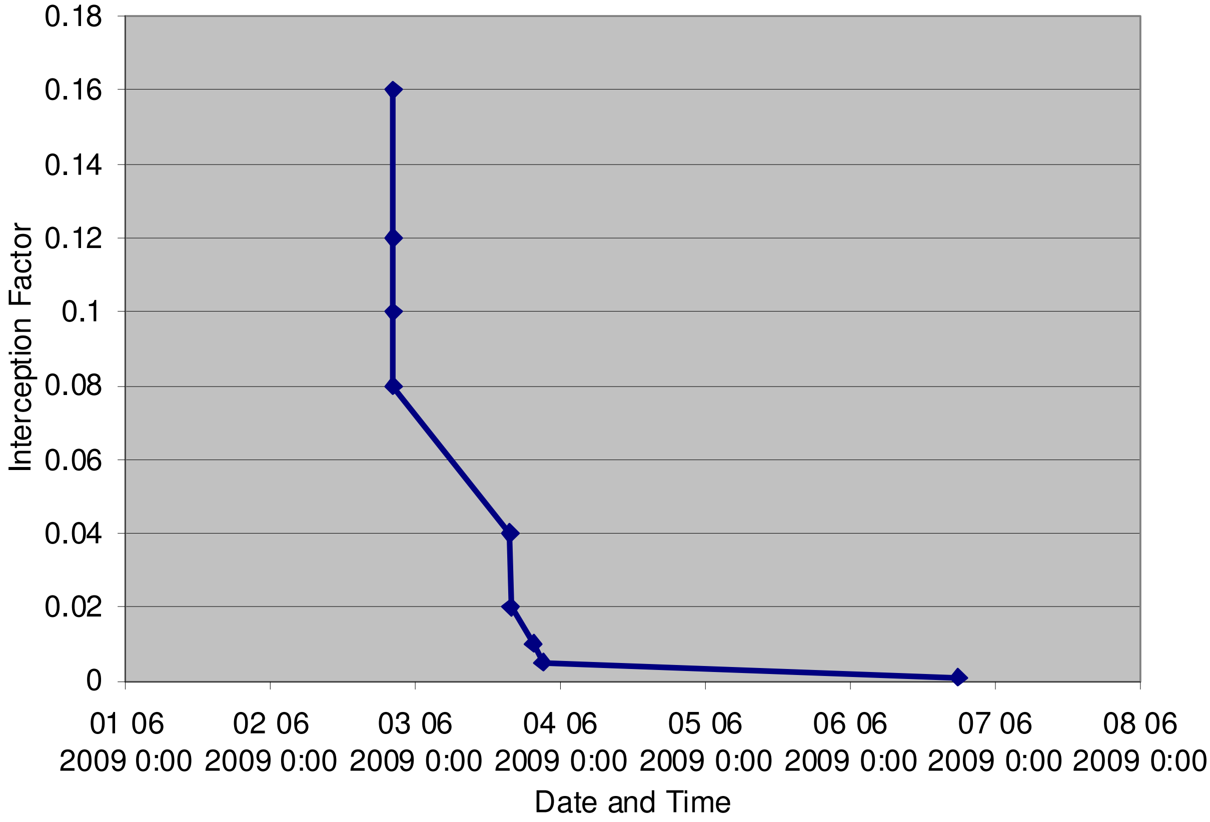


Figure 4: Interception factor

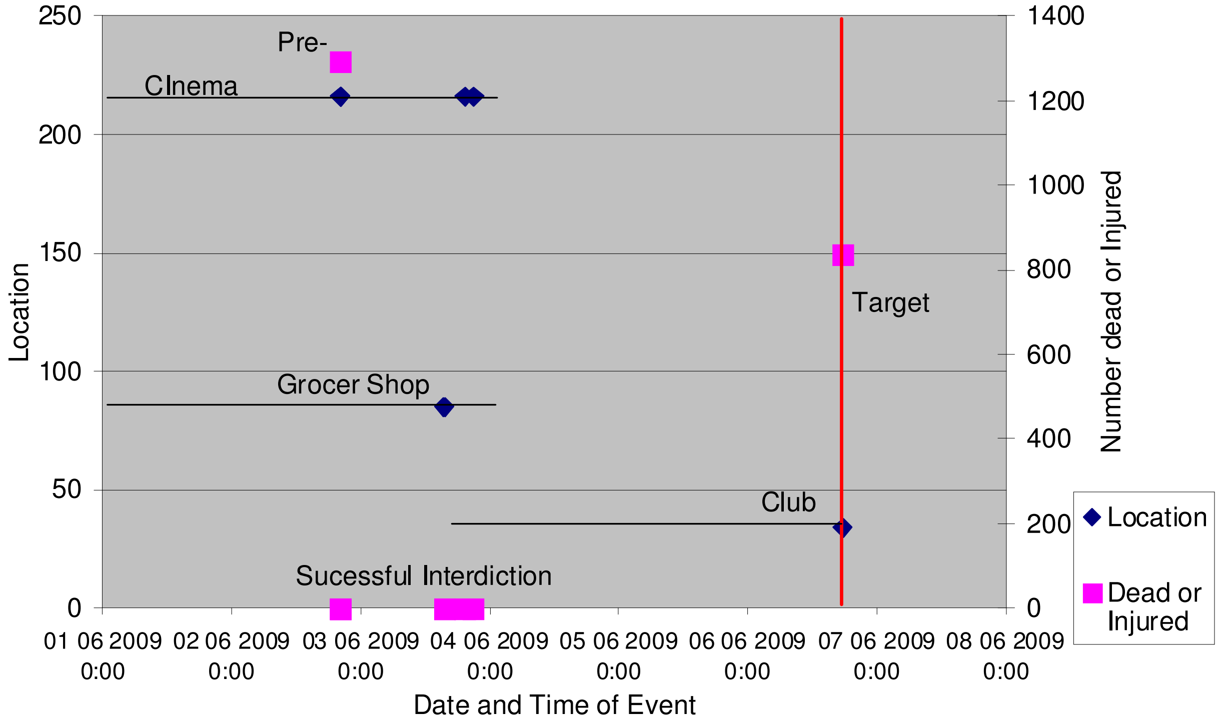


Figure 5: Event location (cell number) and number of dead or injured by interception time

References

[1] Atkinson MP and Wein LM (2008) Spatial Queuing Analysis of an Interdiction System to Protect Cities from a Nuclear Terrorist Attack *Operations Research, 56* pp. 247–254.

[2] Bold J (2000) *Greenwich, An architectural history of the Royal Hospital for seamen and the QueenŠs House* Yale University Press.

[3] CDC (2004) *"Smallpox Overview." Fact Sheet: Smallpox* www.cdc.gov/smallpox.

[4] Connor RJ, Boer R, Prorok PC, and Weed DL (2000) Investigation of Design and Bias Issues in Case-Control Studies of Cancer Screening Using Microsimulation *Am. J. Epidemiol. 151* pp. 991–998.

[5] Daley DJ and Gani J (1999) *Epidemic Modelling: An Introduction.* Cambridge University Press.

[6] Dudley S (1926) The Spread of "Droplet infection" in semi-isolated communities *Medical Research Council Special Report* His Majesty's Stationery Office.

[7] Flaherty C (2008) 3D Tactics and Information Deception *Journal of Information Warfare.* pp. 49–58.

[8] Grist RN (1979) Droplet Infection in Semi-isolated communities, Pandemic Influenza 1918. *British Medical Journal.* pp. 1632–1633.

[9] Halloran EM (2001) *Epidemiologic Methods for the Study of Infectious Diseases* Oxford University Press.

[10] M.G. Hazen, R. Burton, R. Klingbeil, K. Sullivan, M. Fewell, I. Grivell, C. Phips and P. Marland Modelling the Effects of NetCentric Maritime Warfare (NCMW) in Maritime Interdiction Operations (MIO). 8th International Command and Control Research and Technology Symposium, National Defence University, Washington DC, 17-19 June 2003

[11] Kermack WO and McKendrick AG (1927) A Contribution to the Mathematical Theory of Epidemics *Proceedings of the Royal Society of London 115* pp. 700–721.

[12] Knuth DE (1992) *Literate Programming* University of Chicago Press.

[13] Merz J (1991) Microsimulation - a survey of principles, developments and applications *International Journal of Forecasting vol. 7, no. 1* pp. 77–104.

[14] The Cradle of the Navy, National Maritime Museum; 2007. (http://www.nmm.ac.uk/server/show /conWebDoc.6490/viewPage/1).

[15] Orcutt G (1957) A New Type of Socio-Economic Syste. *The Review of Economics and Statistics vol. 39, no. 2* pp. 116–123.

[16] Pita J, Jain M, Janusz Marecki, Fernando Ordóñez, Christopher Portway, Tambe M, Western C, Paruchuri P and Kraus S (2008) Deployed ARMOR Protection: The Application of a Game Theoretic Model for Security at the Los Angeles International Airport *Proc. of 7th Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2008)- Industry and Applications Track* pp. 12–16.

[17] Weinstein MC (2006) Recent Developments in Decision—Analytic Modelling for Economic Evaluation *Pharmacoeconomics* pp. 1043–1053.

Strategic Modeling of Information Sharing Among Data Privacy Attackers

Quang Duong, Kristen LeFevre and Michael P. Wellman
Computer Science & Engineering
University of Michigan
Ann Arbor, MI 48109-2121 USA
E-mail: {qduong,klefevre,wellman}@umich.edu

Keywords: data privacy, game theory, attacker model

Received: September 7, 2009

Research in privacy-preserving data publishing has revealed the necessity of accounting for an adversary's background knowledge when reasoning about the protection afforded by various anonymization schemes. Most existing work models the background knowledge of one individual adversary or privacy attacker, or makes a worst-case assumption that attackers will act as one: colluding through sharing of background information. We propose a framework for modeling multiple attackers with heterogeneous background knowledge, supporting analysis of their strategic incentives for sharing information prior to attack. The framework posits a decentralized mechanism by which agents decide whether and how much information to share, and defines a normal-form game representing their strategic choice setting. Through a simple example, we show that the efficacy of database generalization operations depends on the information-sharing strategies adopted by the attackers. Through analysis of the underlying game model, a database publisher can adopt a generalization level geared to the level of sharing expected among rational attackers.

Povzetek: Predstavljen je model napadov na privatnost s heterogenim ozadjem.

1 Introduction

Many organizations publish non-aggregate personal data, for research purposes including social science, public health, and marketing. At the same time, high-profile incidents have underscored the importance of taking steps to protect individual privacy. In one compelling demonstration, Sweeney (17) showed that by cross-referencing a public voter registration list and a published database of health insurance information, using the combination of birth date, gender, and zip code attributes, an attacker could locate the medical record of the Governor of Massachusetts.

Over the past several years, research in data privacy has sought to provide tools to guard against *identity disclosure* and *attribute disclosure* under this so-called *record linkage* attack model, while preserving the utility of the resulting data. Informally, identity disclosure refers to the ability of an attacker to locate a target individual in the published data, and attribute disclosure refers to the attacker's ability to determine the value of some sensitive attribute associated with a target individual.

One of the principal approaches employed for this purpose is *generalization*, which is best illustrated with a simple example. Consider the input data set shown in Figure 1(a), and consider an attacker who is interested in learning information about Alan. Suppose that the attacker already knows Alan's age, gender, and zip code. Even if the name identifiers are removed from the published data set, the attacker can identify Alan's record (assuming that Alan is included) using this combination of attributes (commonly called *quasi-identifiers*). In the generalized data set of Figure 1(b), however, attribute values are abstracted into coarser-grained equivalence classes. In this instance, the attacker cannot tell which of the first two records is Alan's. Thus, the attacker is also unable to determine whether Alan's disease is AIDS or flu.

In addition to quasi-identifier information, it is also common for an attacker to have access to other instance-level *background knowledge*. In our simple example, suppose that, in addition to Alan's age, gender, and zip code, the attacker also knows that Alan does not have the flu. Using this additional knowledge, in combination with the generalized database, the attacker can determine that Alan has AIDS.

Recent work has proposed incorporating an attacker's background knowledge into the data publication scheme, adopting *worst case* assumptions (3; 15). This is motivated by the practical difficulty for the person deciding what information to publish (the *database publisher*) of modeling the exact information available to an attacker (e.g., "Alan does not have flu"). Further, there may be multiple attackers, each with different background knowledge. Thus, these protocols instead seek to publish generalized data sets that are robust to a certain *amount* of structured background knowledge of a certain form, in the worst case, regardless of the specific content of the knowledge.

This perspective also provides an understandable and objective measure of privacy for the published database—the number of "pieces" of background knowledge that are necessary in order to breach it. However, the database pub-

Name	Age	Gender	Zipcode	Disease
Alan	20	M	12345	AIDS
Bob	24	M	12344	flu
Carol	32	F	12455	flu
Dana	35	F	12411	cancer
Erin	30	F	12455	AIDS

(a) Original data set

	Age	Gender	Zipcode	Disease
(Alan)	2*	M	1234*	AIDS
(Bob)	2*	M	1234*	flu
(Carol)	3*	F	124**	flu
(Dana)	3*	F	124**	cancer
(Erin)	3*	F	124**	AIDS

(b) Generalized data set

Figure 1: Simple attribute disclosure example

lisher often knows very little about potential attackers, so even estimating these background knowledge parameters can be challenging. (This problem is analogous to the task of setting parameter k in k-anonymity (17), a related data privacy requirement.)

When considering the quantities of background knowledge available to an individual attacker, it is helpful for the database publisher to consider two categories of knowledge. First, there are some background facts that are known individually to the attacker. However, the attacker may obtain additional information by colluding (sharing information) with other attackers. At one extreme, the database publisher might optimistically assume that the attackers do not share information, in which case the amount of information available to any particular attacker is relatively low. At the opposite pessimistic extreme, the publisher might assume that attackers share all of their information with one another; thus each has available the collective information originally possessed by individual attackers.

The purpose of this paper is to initiate a study of how information is shared among strategic attackers, which influences how a database publisher should select a data set for publication. In particular, we investigate several possible scenarios of information sharing, and observe that the model of information sharing significantly influences the privacy-preserving data publishing problem.

Paper overview

- We review the idea of generalization-based privacy-preserving data publishing in Section 3. This section illustrates in particular the importance of accounting for attackers' background knowledge when reasoning about privacy.
- We present the motivation and observations supporting our approach to strategic modeling of data privacy attacks in Sections 4 and 5.
- Our empirical study in Section 6 illustrates how to construct and analyze the strategic model as a way of evaluating a database publisher's decision about generalizing and publishing sensitive data.
- This pilot study illustrates how the optimal data generalization policy depends on the attacker model employed, and how strategic analysis can inform the publication decision process.

2 Related work

Recent work has proposed using game theory to model attacker behavior in a variety of security-related applications. In network security, Xu and Lee (18) use game theory to model the network of botnet attackers and defenders for analyzing the performance of a proposed defense system and guiding its design. Kunreuther and Heal (12) define a generic model of *interdependent security games*, where players make individual decisions to invest in security measures, but the resultant security risks depend on investments by all the players. Kearns and Ortiz (9) develop algorithms tailored to solving such games, and apply them to a scenario about the airline adoption of baggage screening technology (8). Perhaps the most prominent example is the ARMOR system deployed at Los Angeles International Airport (16), which models security resource scheduling and terrorist attack decisions as a Stackelberg game. Computational advances in game solving are facilitating the application of this approach to increasingly complex models (10).

A second relevant body of research addresses the problem of modeling information sharing activities and the associated incentives and disincentives in these multiagent systems. Kleinberg et al. (11) examine different information-exchange scenarios and measure the participants' willingness to share information using solution concepts of the coalition games. Agrawal and Terzi (1) introduce a database-related information sharing scenario, in which private database owners reveal information to others in order to improve their query-answer capability.

Economics has become an increasingly important tool for information security analysis, as attackers have become increasingly motivated by financial profits over the years (5). *This has led to research relying on the observation that economic incentives play a significant role in the strategies of attackers and potential victims.* For example, the study by Grossklags et al. (6) focuses on economic outcomes in modeling security investment decision-making by potential victims to protect themselves against malicious Internet attacks.

3 Data generalization background

Consider a data set D that a publisher would like to make available to the public. The publisher applies some data generalization method A to D, obtaining a generalized version D_A, which it publishes instead of D in order to protect the privacy of people whose information is contained in the data set. Figure 1 illustrates an example original data set and its generalized version. As explained above, despite generalization, privacy attackers may be able to derive sensitive information from D_A if they possess sufficient background knowledge (3; 15; 14).

We denote by t some *target individual* whose *sensitive value* $\sigma_t \in \Sigma_t$ is of interest to attackers. Chen et al. (3) propose to classify background knowledge regarding a particular target t into three categories of facts, represented by sets L, K, and M. Each fact is a stylized ground expression. L comprises information about sensitive values $\sigma'_t \neq \sigma_t$ that target t does not have, for instance "Alan does not have flu". K is a set of facts about sensitive values for other individuals $t' \neq t$, for example "Bob has flu". Facts in M specify the relationships between t and other individuals, such as "if Erin has AIDS then Alan has AIDS".

Given this classification, the tuple $B = (L, K, M)$ fully describes an attacker's background knowledge and thus indirectly specifies her ability to successfully *breach* a published data set. We say that D_A has been breached if an attacker can deduce the target's sensitive value σ_t.[1]

For many applications, it may be advantageous to adopt a more abstract and compact specification of background knowledge, rather than enumerating it explicitly. Chen et al. (3) propose a summary representation that replaces the specific instances with counts of the number of facts in the respective categories. In this scheme, background knowledge $B = (L, K, M)$ is summarized by the tuple of quantities $b = (|L|, |K|, |M|)$. This abstraction relaxes the requirement to reason about instance-specific knowledge of attackers, and is exponentially more compact. Although it discards instance-specific information, the summary still enables a designer to reason about the degree of generalization required to thwart breach of the data set in the worst case. Given our examination of small examples in this study, however, we retain the full specification of background knowledge, B, for the remainder of this paper.

4 Information sharing among attackers

We examine a network of n attackers who seek to discover the target individual t's sensitive value σ_t in the data set D_A. These attackers may exchange background information with one another prior to launching their attacks, in order to improve their prospects for compromising D_A. The attacker faces a fundamental tradeoff in its incentives for sharing information:

– Acquiring relevant background facts generally improves the ability of an individual attacker to breach the target data set, which in turn generates value for the attacker.

– Revealing relevant information also improves the likelihood that other attackers will successfully breach the data set. As more attackers succeed, the value of the breached information typically declines for each attacker. For example, the price an attacker could obtain by selling the sensitive information would decrease to the extent it is commonly available.

Each attacker i starts with some prior knowledge, $B_i = (L_i, K_i, M_i)$. From the perspective of the database publisher and other attackers, the background knowledge of attacker i is uncertain, drawn from some distribution β, which can be modeled using various approaches (13).

4.1 Information sharing mechanism

We describe a simple mechanism by which the attackers share information. Although in practice we cannot mandate the process whereby attackers will coordinate in this way, defining some specific process is necessary to frame the strategic environment in which the attackers operate. The sharing mechanism we assume relies on a principle of reciprocity to induce mutually beneficial sharing. That is, one attacker provides information to a neighbor on the attacker network only to the extent that this neighbor provides information in return. Specifically, the number of facts in each category transferred between two attackers is the same in each direction. For simplicity, we also assume that information exchanged among the attackers is accurate; that is, attackers do not distort information that they share with others.

The basic decision made by each attacker is which facts to offer to share. That is, given prior knowledge $B_i = (L_i, K_i, M_i)$, the set of available information-sharing actions S_i for attacker i comprises all $s_i = (s_{l,i}, s_{k,i}, s_{m,i})$ such that $s_{l,i} \subseteq L_i$, $s_{k,i} \subseteq K_i$, and $s_{m,i} \subseteq M_i$. Given the category L sharing offers $s_{l,i}$ and $s_{l,j}$ of two neighboring attackers, the number of facts shared in that category is therefore $\min(|s_{l,i}|, |s_{l,j}|)$. Category K and M sharing operates identically. The sharing mechanism thus determines the quantities of facts to be shared for each pair of connected attackers, for each category. In each case, when the number of facts to be shared is fewer than the number offered by one party, the subset of offered facts actually transmitted to the other is selected randomly.

4.2 Attacker utility

Our model of attackers' utility presumes their primary objective is to discover the target individual's sensitive information. Let r_t be the reward obtained from discovering

[1] The concept of breach could be treated more generally. Past work has sought to model an attacker's uncertainty about sensitive facts (for example, using a distribution over possible worlds (3; 15)), and then defined the idea of breach incorporating this uncertainty. For example, we might instead say that D_A has been breached if an attacker can determine σ_t with certainty exceeding some threshold c. However, in the interest of simplicity, we fix $c = 1$.

(e.g., by selling) the sensitive information σ_t. The more attackers who have this piece of information, the less valuable it is to each attacker. As a result, the reward each attacker receives decreases with the number of attackers successfully compromising the target data set. Specifically, if there are μ successful attackers, we assume that each attacker who obtains t's sensitive value receives reward $\frac{r_t}{\mu^2}$. According to this utility function, a successful attacker's reward deteriorates faster as μ increases.

Attackers need to make decisions about how much information they would like to share with others in order to maximize their rewards. Since we consider scenarios with only one target, without loss of generality we can set $r_t = 1$.

Given a *strategy profile* (sharing decision for each attacker) $s = (s_1, \ldots, s_n)$, we can calculate the amount of knowledge each attacker obtains from sharing information. From this information and the specifications of the generalized database and distribution of prior knowledge, we can evaluate each attacker's prospects for compromising the target database, and consequently their expected reward, or utility. The utility to attacker i playing strategy s_i when other agents play their strategies collectively denoted s_{-i} is given by $u_i(s_i, s_{-i})$.

Example 1. *Consider the scenario specified in Figure 1. Three privacy attackers X, Y, and Z would like to know Alan's disease, denoted as* $Disease[Alan]$*. X knows* $Disease[Alan] \neq$ *cancer and* $Disease[Dana] =$ *cancer. Y knows* $Disease[Bob] =$ *flu and* $Disease[Carol] =$ *flu. Z knows if* $Disease[Erin] =$ *AIDS then* $Disease[Alan] =$ *AIDS, and* $Disease[Erin] \neq$ *cancer. Suppose that X wants to share* $Disease[Dana] =$ *cancer and Y wants to share that* $Disease[Bob] =$ *flu. After sharing information with X, Y now knows* $Disease[Dana] =$ *cancer, in addition to her initial background knowledge. Y therefore can infer* $Disease[Alan]$ *and consequently collect a reward of* 1 *if she is the only attacker capable of discovering his disease. If X, Y, and Z succeed in discovering* $Disease[Alan]$*, each would then collect a reward of* $\frac{1}{9}$ *instead.*

4.3 Database publisher

In privacy-preserving data publishing, the database publisher typically strives to strike a balance between protecting individual privacy and maintaining the published data's value (minimizing *information loss*) when choosing her generalization strategy (3; 15; 14). We incorporate both information loss and privacy breach risk when computing the publisher's utility u_d.

We denote by s_d the publisher's strategy for anonymizing the released data set. Formally, s_d fully describes the resulting generalized data set, which we denote D_{s_d}. Thus s_d can be of different formats, depending on the chosen generalization method. In our example in Figure 1, the publisher's data generalization action that transforms that original data set to the generalized data set can be fully specified by $s_d = (s_{d,1}, \ldots, s_{d,|D|}) = (1, 1, 2, 2, 2)$. In this particular representation, $s_{d,i} = s_{d,j}$ for $i, j \in [1, |D|]$ indicates that the two records i and j are "generalized" so that in D_{s_d} they are indistinguishable based on other non-sensitive attributes.

We first quantify the generalization-induced information loss of the generalized data set D_{s_d}, given the publisher's action s_d. For simplicity, out of many previously proposed measures of information loss, we adopt a variation of the "discernibility penalty" proposed by Bayardo and Agrawal (2). For each record e in generalized data set D_{sd}, we define the *equivalence class* $\pi(e, D_{sd})$, which is the set of records in D_{sd} that are indistinguishable from e on quasi-identifier attributes due to generalization. The intuition is to assign each record a penalty based on the size of its equivalence class. Thus, the information loss is quantified as

$$il(s_d) = \frac{1}{Z_D} \sum_{e \in D_{s_d}} |\pi(e, D_{s_d})|,$$

where Z_D is the largest information loss possible for any data set of D's size, and thus is constant for a fixed-size data set. This normalization factor allows us to discount the effect of the data set's size on our measure of information loss.

The second factor in the publisher's utility is the prospect for data privacy breach. We capture this in a random variable br, whose probability distribution depends on the strategies of attackers as well as the publisher. The variable takes value one if the sensitive data is breached, zero otherwise.

We formulate the publisher's payoff $u_d(s_d, s)$ such that it is normalized on [0,1], decreasing with privacy breach and information loss. There are many possible ways to integrate these factors in an overall utility function. The simplest is to linearly combine information loss and privacy breach, weighted by parameter w:

$$u_d(s_d, s) = 1 - [w \times il(s_d) + (1 - w) \times br(s_d, s)]. \quad (1)$$

4.4 Privacy breach

Suppose that the database publisher chooses action s_d, the attackers' initial background knowledge is $\mathbf{B} = (B_1, \ldots, B_n)$, and their strategy profile is s. As described in Section 4.1, s determines the attackers' resulting posterior knowledge, collectively denoted as $\mathbf{B}' = (B'_1, \ldots, B'_n)$. In order to calculate their final reward, we need estimate the likelihood that each can breach the data set D_{s_d} given their posterior knowledge $\mathbf{B}'$.

For each attacker i, with posterior knowledge B'_i, we can reason logically about the sensitive values i can eliminate when attempting to deduce t's sensitive value from D_{s_d}. The payoff $u_i(s, s_d)$ to this attacker is calculated as described previously. Applying this reasoning to all attackers, we can calculate the number μ that are successful for any configuration of posterior knowledge among the attackers. For this configuration, we then conclude $br(s_d, s) = 1$ if $\mu > 0$ and 0 otherwise.

Given a distribution β over attackers' prior background knowledge $\mathbf{B}$, and a profile of attacker strategies s, we can further compute a distribution over attackers' posterior background knowledge $\mathbf{B}'$. These elements are therefore sufficient to calculate expected utilities for all agents (publisher and attackers), using the definitions specified above.

5 Game-theoretic modeling

We model the strategic environment with n privacy attackers plus the database publisher d as a game, employing the strategy sets and utility functions defined above. The game plays out in two stages:

1. The database publisher first chooses her action s_d and publishes the data set D_{s_d}.
2. The attackers observe the publisher's action. They then choose their actions s, exchange background knowledge, attack the data set, and collect reward if they succeed.

Because the publisher moves first, we can characterize her problem as optimizing the database design, subject to the outcome of the *information-sharing subgame* played among the attackers conditional on this design. We thus focus on defining and analyzing this attacker subgame.

5.1 Information-sharing subgame

Technically, the information-sharing game among the attackers is a game of incomplete information, with information structure defined by the distribution β over prior background knowledge. Each agent's strategy in the incomplete-information game is a mapping from its own type (assignment of prior knowledge) to a sharing offer. Here we simplify the model structure by translating to normal form, explicitly constructing the payoff for every combination of attacker strategies.[2]

Recall our subgame is conditioned on the publisher's action s_d selected in the first stage. A given s_d and the distribution β of background facts among privacy attackers defines the expected payoff for any profile of attacker strategies. We can calculate these payoffs by Monte Carlo sampling, given a budget of H samples. To estimate the expected payoff of attacker strategy profile s:

1. Draw a background knowledge configuration $\mathbf{B} = (B_1, \ldots, B_n)$, according to the distribution β.
2. Calculate the distribution of privacy breach events given s_d, $\mathbf{B}$, and s, based on the sharing mechanism described in Section 4.1.
3. Tally the expected payoffs u_i for each attacker as well as the expected value of publisher's privacy breach br based on the results for this configuration.
4. Repeat steps 1–3 H times.
5. Average over the sampled u_i and br values to construct estimated expected values.

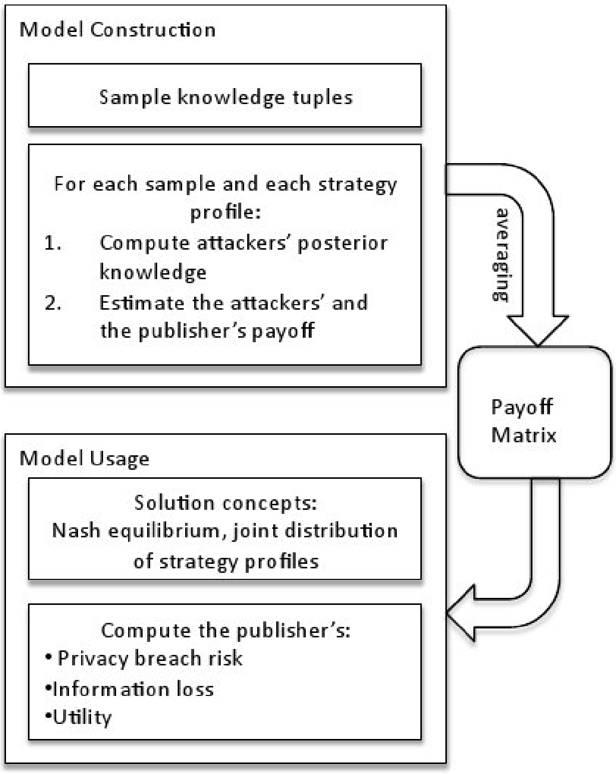

Model Construction
Sample knowledge tuples
For each sample and each strategy profile:
1. Compute attackers' posterior knowledge
2. Estimate the attackers' and the publisher's payoff
averaging
Payoff Matrix
Model Usage
Solution concepts:
Nash equilibrium, joint distribution of strategy profiles
Compute the publisher's:
• Privacy breach risk
•Information loss
•Utility

Figure 2: Overview of the strategic model of privacy attackers and database publisher.

We can construct the complete expected payoff matrix of the game by repeating the above procedure for each attacker strategy profile s and each database publisher's action s_d. In practice, we will not be able to do so exhaustively, but instead would focus on a salient subset of strategy combinations and induce a game model that best cap-

Given a subgame form constructed as specified above, we are interested in identifying the Nash equilibria (NE).

Definition 1. *A strategy profile s^* is a Nash equilibrium if no unilateral deviation in strategy by any single player is beneficial for that player given the others' designated strategies. That is, $\forall i, s_i' \in S_i.\ u_i(s_i^*, s_{-i}^*) \geq u_i(s_i', s_{-i}^*)$.*

If all agents play pure (non-probabilistic) strategies in s^*, then s^* is a *pure-strategy NE* (PSNE).

Definition 2. *Player i's* regret, *$\epsilon_i(s)$, represents the maximum gain in payoff i can obtain through unilaterally reconsidering its own strategy s_i given others' strategies s_{-i}.*

$$\epsilon_i(s) = \max_{s_i' \in S_i} u_i(s_i', s_{-i}) - u_i(s_i, s_{-i}).$$

A profile's regret $\epsilon(s)$ is defined as

$$\epsilon(s) = \max_i \epsilon_i(s).$$

[2] In practice, this will generally entail restrictions on the flexibility of attacker strategies, particularly in how they are conditioned on the realization of prior background knowledge.

By these definitions, if profile s^* is an NE, strategy s_i^* is player i's *best response* to others' play s_{-i}^* and therefore induces zero regret ($\epsilon_i(s^*) = 0$). All else equal, profiles with zero (or small) regret are considered more likely to be played by rational agents, as high-regret profiles offer some agent a large incentive to deviate. Thus, a database publisher may wish to choose a design s_d that performs well when attackers follow equilibrium strategies conditional on that design. Figure 2 summarizes the game-theoretic modeling and analysis process as applied to our data privacy attack scenario.

6 Illustrative example and analysis

In this section we present a toy example illustrating how our game model can be used to analyze a privacy-preserving publishing scenario.

6.1 Example

The original data set D for our example comprises the records in Figure 1(a), plus the set of records specified in Figure 3.

Name	Age	Gender	Zipcode	Disease
...	...	...	...	...
David	24	M	13344	heart
Daniel	32	M	13455	allergy
Frank	24	M	12334	AIDS
Grace	40	F	12445	cancer
Heather	45	F	13445	allergy

Figure 3: Additional records appended to Figure 1(a) to define the original data set D for the example.

There are three attackers ($n = 3$) in this example who are interested in identifying Alan's disease. Moreover, it is common knowledge that each person's disease can be either heart, allergy, AIDS, cancer, or flu. Although this would never be the case for realistic data sets, our toy example is sufficiently small that we can exactly account for all possible background knowledge instances in all three categories. In this case, there are four instances of type L, and nine each of types K and M. We further restrict that each attacker initially starts with only one instance of each category, which means $|L_i| = |K_i| = |M_i| = 1$. The distribution of prior background knowledge β draws a fact in each category with equal probability for each attacker.

Given this configuration of prior knowledge, an attacker needs to decide whether or not to share her available fact for each knowledge category. We assume that attackers make this decision unconditional on the particular fact drawn for the respective categories, which results in a total of eight possible strategies. For example, one possible strategy is to share one's L and K facts, but not the M fact.

The database publisher's strategy s_d can be represented by a ten-element array of equivalence-class indices, following the format described in Section 4.4. The strategy specifies to which class each record belongs as a result of the publisher's data generalization method.

Since there are too many possible publisher actions (168,440 even for this small data set) to evaluate them all, we identified a select set of ten candidate strategies, spread out in the design space. We deliberately selected these ten candidates from a set of more than a hundred designs sampled from the publisher's strategy space, to ensure that information loss is spread relatively uniformly over the possible range. For each design, we constructed the corresponding normal-form information-sharing subgame, using the procedure detailed in Section 5.1. Our Monte Carlo budget was $H = 5000$, a sufficient number of samples to render negligible the variance in expected payoff calculations for the attackers.

For each profile of attacker strategies, we record the attacker payoffs as well as the probability of privacy breach. From this we can calculate database publisher's utility using Equation (1). We set the publisher's tradeoff weight $w = 0.25$, implying that the publisher values lowering the probability of privacy breach by a given increment three times as much as lowering information loss by that same increment on the specified scale.

6.2 Empirical results

For each publisher strategy, we evaluate the outcome achieved under three different assumptions about attackers' behavior:

Scenario	Assumption
No	No attackers share any information.
NE	Attackers play a PSNE profile.
All	All attackers share all information.

The No scenario is a best-case assumption: attackers are unable or unwilling to share information, for whatever reason, thus they attack based only on their individual information. All is the worst-case scenario for the publisher. Under NE, the attackers are treated as rational strategic players, predicted to play an equilibrium profile of the information-sharing subgame. In general these subgames may have multiple equilibria. Our analysis identifies all the PSNE, and defines the NE scenario as an equiprobable selection among these.

Figure 4 presents the information loss and expected privacy breach for each of the ten selected publisher strategies, under each of the attacker behavior assumptions No, NE, and All. Since information loss does not depend on the attackers' actions, a given publisher action is represented by three points at the same y-axis level, associated with the respective attackers' behaviors. The separation of these points on the x-axis confirms that attackers' behavior in equilibrium is generally different from all-or-none information sharing.

Inspection of Figure 4 allows us to identify and rule out the *dominated* publisher actions, that is, any s_d that

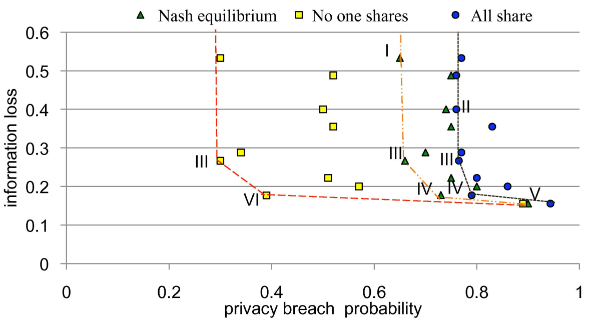


Figure 4: Expected privacy breach and information loss under various generalization actions and attacker behaviors.

is worse on both information loss and privacy breach than some other available publisher strategy or convex combination of strategies, under the same assumption on attacker behavior. Accordingly, in the figure we draw piecewise-linear curves for each attacker scenario, connecting the frontier of non-dominated publisher strategies. Given any weight parameter for publisher utility (1), the optimal generalization design (among the ten evaluated here) lies on this non-dominated frontier. We label the non-dominated actions with roman numerals.

As expected, generalization actions that induce greater information loss generally partition the original data set into fewer groups and/or generalize more records in the same group. For instance, action I partitions the original ten records into two groups of three and seven, whereas action V divides them into four smaller groups.

The distinction in composition and shape among the frontiers for the three behavior scenarios confirms the possibility that the publisher's optimal choice will be different under the respective assumptions. For instance, a publisher that pessimistically assumes that all attackers share information (All) may pick action II. However, this strategy is dominated under the NE or No assumptions.

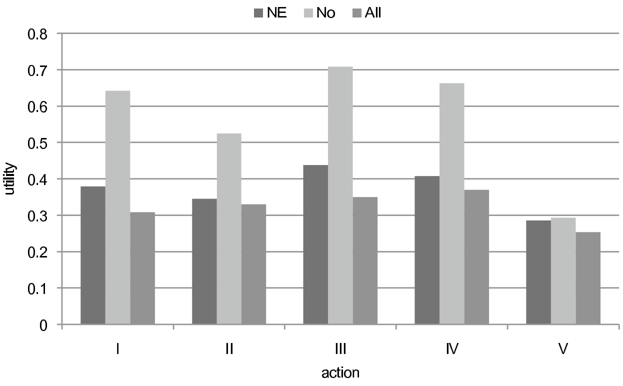


Figure 5: Database publisher's utility ($w = 0.25$) under different generalization actions and scenarios for attackers' behavior.

Given a particular weight for trading off information loss and privacy risk, we can identify the publisher's optimal choices. Figure 5 plots the publisher's utility for its non-dominated actions, at tradeoff weight $w = 0.25$, under each of the attacker behavior scenarios. This chart reveals that the worst-case assumption (All) that all share everything indeed leads to choosing action IV, which is suboptimal under the NE model.

Like in most multiagent-system models, the solutions generated from our models are sensitive to the choice of attackers' utility function described in Section 4.2. In particular, higher powers of μ in the attacker's utility function (i.e., reward dropping off faster than the square of number of successful attackers) may lead to overoptimistic estimates of the risk of privacy breach. Further, the choice of utility function can vary considerably with specific data publication scenarios.

7 Conclusions and future work

Past research in privacy-preserving data publishing has demonstrated the importance of accounting for an attacker's background knowledge. A variety of generalization tools have been developed, but at a minimum these still require the database publisher to know the amount of background knowledge available to attackers (3; 15). The presence of multiple attackers with capabilities for pooling background knowledge significantly magnifies this uncertainty, absent a model of how attackers will actually share information.

This paper initiates a game-theoretic study of privacy attackers as a knowledge-sharing network. Rather than simply guessing about attackers' information-sharing behavior, we propose a grounded framework for reasoning about attackers' interactions, which in turn assists the data publisher in choosing a generalized data set to publish. Our empirical study demonstrates that attacker incentives (and their resulting behavior) can influence the database publisher's optimal strategy.

Whereas this paper illustrates the importance of reasoning about attackers' incentives when choosing a data publishing strategy, our initial models by no means cover all attack scenarios. Future work should refine these models based on behavioral observations to enrich the data publisher's limited information about attackers' knowledge and behavior.

In addition, representing the full content of attackers' background knowledge as we did for this initial study will not remain feasible as we scale the resulting model to larger networks of attackers. Thus, it is also important to adopt a more compact representation of background knowledge, such as the quantified summaries of background knowledge proposed by Chen et al. (3) and Martin et al. (15).

Game-theoretic analysis may provide useful grounds for predicting attacker behavior, but it is by no means the only source of evidence. Attackers may not be perfectly rational, or their information and incentives may not be accurately captured by the model. Graphical multiagent models (GMMs) are designed to support integration of game-theoretic and other sources of knowledge about multiagent behavior (4). Like other graphical models, GMMs also take advantage of locality in agent interactions (e.g., structure in the information-sharing network), and provide a compact representation for efficient computation of joint distributions over agent behavior.

Finally, we are interested in applying a similar framework to study privacy protection mechanisms other than generalization (e.g., input and output perturbation techniques for statistical databases).

References

[1] Rakesh Agrawal and Evimaria Terzi. On honesty in sovereign information sharing. In *Tenth International Conference on Extending Database Technology*, pages 240–256, Munich, 2006.

[2] Roberto J. Bayardo and Rakesh Agrawal. Data privacy through optimal k-anonymization. In *Twenty-first International Conference on Data Engineering*, pages 217–228, Tokyo, 2005.

[3] Bee-Chung Chen, Kristen LeFevre, and Raghu Ramakrishnan. Privacy skyline: Privacy with multidimensional adversarial knowledge. In *Thirty-Third International Conference on Very Large Data Bases*, pages 770–781, Vienna, 2007.

[4] Quang Duong, Michael P. Wellman, and Satinder Singh. Knowledge combination in graphical multiagent models. In *Twenty-Fourth Conference on Uncertainty in Artificial Intelligence*, pages 153–160, Helsinki, 2008.

[5] Jason Franklin, Vern Paxson, Adrian Perrig, and Stefan Savage. An inquiry into the nature and causes of the wealth of Internet miscreants. In *Fourteenth ACM Conference on Computer and Communications Security*, Alexandria, VA, 2007.

[6] Jens Grossklags, Nicolas Christin, and John Chuang. Secure or insure?: A game-theoretic analysis of information security games. In *Seventeenth International Conference on World Wide Web*, pages 209–218, Beijing, 2008.

[7] Patrick R. Jordan and Michael P. Wellman. Generalization risk minimization in empirical game models. In *Eighth International Conference on Autonomous Agents and Multi-Agent Systems*, pages 553–560, Budapest, 2009.

[8] Michael Kearns. Economics, computer science, and policy. *Issues in Science and Technology*, 21(2):37–47, 2005.

[9] Michael J. Kearns and Luis E. Ortiz. Algorithms for interdependent security games. *Advances in Neural Information Processing Systems*, 16, 2004.

[10] Christopher Kiekintveld, Manish Jain, Jason Tsai, James Pita, Fernando Ordóñez, and Milind Tambe. Computing optimal randomized resource allocations for massive security games. In *Eighth International Conference on Autonomous Agents and Multi-Agent Systems*, pages 689–696, Budapest, 2009.

[11] Jon Kleinberg, Christos H. Papadimitriou, and Prabhaka Raghavan. On the value of private information. In *Eighth Conference on Theoretical Aspects of Rationality and Knowledge*, pages 249–257. San Francisco, 2001.

[12] Howard Kunreuther and Geoffrey Heal. Interdependent security. *Journal of Risk and Uncertainty*, 26: 231–249, 2003.

[13] Tiancheng Li, Ninghui Li, and Jian Zhang. Modeling and integrating background knowledge in data anonymization. In *Twenty-Fifth International Conference on Data Engineering*, pages 57–66, Shanghai, 2009.

[14] Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam. l-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data*, 1 (1), 2007.

[15] David J. Martin, Daniel Kifer, Ashwin Machanavajjhala, Johannes Gehrke, and Joseph Y. Halpern. Worst-case background knowledge in privacy. In *Twenty-Third International Conference on Data Engineering*, pages 126–135, Istanbul, 2007.

[16] James Pita, Manish Jain, Janusz Marecki, Fernando Ordóñez, Christopher Portway, Milind Tambe, Craig Western, Praveen Paruchuri, and Sarit Kraus. Deployed ARMOR protection: The application of a game theoretic model for security at the Los Angeles International Airport. In *Seventh International Conference on Autonomous Agents and Multiagent Systems*, pages 125–132, Estoril, Portugal, 2008.

[17] Latanya Sweeney. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty Fuzziness and Knowledge-based Systems*, 10(5):557–570, 2002.

[18] Jun Xu and Wooyoung Lee. Sustaining availability of web services under distributed denial of service attacks. *IEEE Transactions on Computers*, 52(2):195–208, 2003.

Planning to Discover and Counteract Attacks

Tatiana Kichkaylo, Tatyana Ryutov, Michael D. Orosz and Robert Neches
Information Sciences Institute
University of Southern California,
Marina del Rey, CA 90292, USA

Keywords: planning, attack recognition, intrusion detection

Received: September 4, 2009

A major function of a security analyst is to analyze collected intelligence looking for plans, associated events, or other evidence that may identify an adversary's intent. Armed with this knowledge, the analyst then develops potential responses (e.g., countermeasures) to deter the discovered plan or plans, weighs their strengths and weaknesses (e.g., collateral damage) and then makes a recommendation for action. Unfortunately, the collected intelligence is typically sparse and it is not possible for the analyst to initially discover the adversary's specific intent. Under these circumstances, the analyst is forced to look at the range of possible plans/actions an adversary may take. The full range of potential attack scenarios is too rich to generate manually. Its complexity also bars direct analysis and evaluation of the potential impact of alternative actions and countermeasures. To address these issues, we are developing a set of tools that exhibit the following features/capabilities:

- *Using available partial plan segments (referred to as snippets), construct multiple feasible scenarios/pathways that an adversary may take to reach an identifiable end goal*
- *Provide visual tools for exploring sets of possible scenarios under various observables, importance, and likelihood conditions, helping the analyst generate information probes, actions and countermeasures*
- *Compare the potential impact of alternative data probes, actions and countermeasures on an adversary's actions by assessing their discrimination/attack mitigation potential and possible side-effects*
- *Automatically suggest potential data probes, actions and countermeasures based on partial understanding of the adversary's plan and given observable activity*

These tools can provide decision support for many different domains, including terrorist activity recognition and network intrusion detection.

Povzetek: Opisan je nabor orodij za napovedovanje napadov.

1 Introduction

The problem of detection, prevention and/or response to attacks is common to multiple domains. In the cyber community, a goal of an attacker may be to gain access to protected resources or to disrupt service to legitimate users. In counter-terrorism, an attack manifests itself as actions in the physical world. In robotic soccer, an attack is a sequence of actions by a team aimed at outmaneuvering the opponents and scoring goals.

Attacks can be classified along several dimensions. Different techniques and tools for attack detection and prevention/counteraction are applicable to different regions of this multi-dimensional space.

- *Attacks may be fast or slow.* Slow attacks, such as terrorism in the physical world, leave sufficient time for human analysts to respond to alerts and warnings. In the case of some cyber-attacks (e.g., Denial of Service) a reaction is required in real time, which often leaves no room for human operators or complex reasoning algorithms.
- *Vulnerabilities may or may not be known in advance.* In the case of existing physical infrastructures or legacy systems it may be impossible to eliminate all vulnerabilities. However, for many of these legacy systems and physical infrastructures, it is possible to identify possible attack strategies for known weak points and design systems that detect such attacks. In the case of unknown vulnerabilities these external detection systems need to monitor the general health of the legacy system and dynamically devise detection and response strategies if something abnormal is detected.
- *An attack may appear either as an expected behavior of other agents, or an anomaly.* In soccer, one may assume that any action of the opposing team is a part of an attack. In the cyber-world, most actions are part of some legitimate user activity. Therefore, it is desirable

to minimize the overhead of monitoring and limit disruption of legitimate activity.

In this paper we discuss a set of tools and approaches for identifying vulnerabilities, detecting potential attacks based on observables, and devising detection and/or countermeasures aimed at minimizing disruption of normal operations. We discuss the potential of such tools for human analysts for detection of slow attacks. In addition, we describe an initial prototype system developed for the Intelligence Advanced Research Projects Activity (IARPA) sponsored Proactive Intelligence (PAINT) project. The system generates both benign and nefarious pathways (i.e., plans that an opponent or adversary may take to achieve an objective) based on an initial goal and a collection of partial plan segments (which we call *snippets*). The implemented system was targeted for the counter-terrorism domain. We also present network intrusion detection examples from as a new potential application for our system.

2 The problem

Recognizing attack plans is a key activity of security analysts. Plan recognition has been a research area in artificial intelligence (AI) for decades. In AI, plan recognition is a process of inferring the goals of an agent from observations of the agent's activities. In security applications (e.g., intrusion detection), the plan recognition process is concerned with adversary recognition, where attackers try to avoid or interfere with recognition process and can take deliberate actions to hide their actions and intentions.

The assumptions used in traditional plan recognition [3] [5] are not valid in the security-related adversary recognition domain. In some domains (e.g., RoboCup soccer [16] , computer games [4] , and military simulations [15]) the adversary and its high level goals are known. In security domains (e.g., computer security, information warfare, and antiterrorism), the adversary tries to hide its actions and identity. Additionally, its real intentions are not always clearly identifiable.

The challenges in adversary plan recognition and response in security domain include the following [10] [11] :

Uncertainty and incompleteness. In some cases, the functional limitations of security sensors or a less-than optimum deployment pattern may increase uncertainty by hindering our ability to observe all attacker activities and steps. Moreover, we may have an incomplete knowledge of the possible attack plans.

Partially ordered plans. Often attackers' plans are flexible in the ordering of the plan's steps; therefore we must be able to recognize the multiple possible instantiation orderings created by these plans.

Multiple concurrent goals. Attackers can have multiple dynamic attack plans. For example, a hacker might be interested in stealing sensitive data as well as using computers to launch attacks against other targets.

Actions used for multiple effects. Often a single action can be used for multiple effects. For example, scanning of a domain can be used both for planning a DoS (Denial of Service) attack as well as to identify the web server that a hacker wants to deface.

Misleading behavior. In addition to actions directly contributing to achieving a goal, an attacker can take actions to mislead plan recognition, or to exploit some of its weaknesses.

Multiple weighed hypotheses. Ranking the possibilities is often more helpful than providing a single (possibly incorrect) explanation for observed activities. Say one observes scanning activity. While this action indicates a hacker is interested in a network, the observation of the action itself provides very little evidence about the hacker's intent. Rather than giving just one of the many equally likely answers, it is much more helpful to report the relative likelihood of each of the possibilities.

Automated vs. human-in-the-loop operation. In security applications, the purpose of adversarial plan recognition is to predict possible attacks in order to generate effective countermeasures. In many current human-in-the-loop operations, real-time detection and response is not achievable. The resulting delays could enable an actual attack. As [20] [6] have pointed out, in such applications, an automated attack recognition and reaction system avoids these delays and can stop an attack before the damage is done. Automated systems, however, can produce negative outcomes since some response actions (e.g., changes to firewall rules) can negatively affect legitimate users. Hence, the application of automated techniques is limited to well-known attacks (e.g., signature-based intrusion detection) and targeted countermeasures.

Side-effects of probes and countermeasures. Detection of stealthy attacks, such as malicious insider covert activity and terrorist preparations, can be complex and may require probes (i.e., actions designed to engender a response likely to provide additional useful information). To prevent a situation where a probe or a response action causes more damage than the actual attack, a system must allow analysts to reason about the likelihood and severity of an attack as well as about the effects of candidate alternative probe/response mechanisms.

Existing techniques for predicting adversary actions, include game theoretic approaches and game playing, adversarial planning, and pattern recognition. These provide partial/limited solutions at best. Data mining approaches have been used to find information that helps to detect attacks; however they suffer from a high false alarm rate and do not help analysts connect separate events. In order to minimize damage, new algorithms and tools for security analysts are needed to enable them to further analyze and correlate attack scenarios, make accurate situational assessments and quickly execute appropriate responses.

3 Modelling alternatives

The algorithmic core of our toolkit is a constraint-based planning system [18] . The system core maintains alternative plans consisting of tokens. A token may represent an action or a state. Both past events (observed or hidden) and future ones (plan projection) are part of a plan. Tokens contain variables representing temporal (start, end, duration) and resource properties (actor, equipment) of the action/state. Variables can be used as arguments to constraints. Constraints define tuples of values their arguments can take [8] . Constraints enforce partial temporal ordering and/or resource dependencies between tokens.

A plan is seeded with a set of tokens representing high-level goals, target states, and/or observed events. The planning module then refines the seeded plans according to snippets. Snippets are similar to HTN methods [9] . A snippet captures a modification of a plan as an implication of the form “if a certain configuration of tokens is a part of the plan, then the following configuration of tokens is also a part of the plan”. Snippets thus can represent expansion of high-level goals into lower-level actions and states. They also can enforce causality (e.g., if an attacker knows a password, then an action of stealing the password should have preceded the attack). If multiple snippets are applicable to a given situation (e.g., there are multiple ways to achieve the same goal), the planner will create alternative plans using each of the snippets. In the current system implementation, the user has sole responsibility for generating attack snippets. A possible direction for future research is to extract snippets from execution traces using machine learning techniques.

Consider the following example. Figure 1 shows a snippet describing the use of the *lpr attack*[1] to achieve a state when an attacker gets access to a protected file. This snippet says that a possible way to explain the presence of a **secret printed** state in a plan is to add to the plan a partially ordered sequence of actions implementing the attack. Lines on the figure represent temporal constraints. Note that the snippet imposes only a partial ordering on its steps, thus allowing for generation of partially ordered plans. Additional resource constraints (not shown) declare that all variables marked as **f** (or **s**) refer to the same file. Constraint propagation helps to limit the set of possibilities and to define windows of interest. For example, suppose an analyst has an estimate for the earliest time a printed copy of a file *secret.txt* was available to an attacker. Propagation of this file name and time point will limit the set of **ln –s** commands that would be considered as a possible part of the attack. Note also that application of snippets may reuse existing tokens in the plan. For example, **block ppt** command might have been issued by a different user for a legitimate reason and hijacked by the attacker.

In addition to describing goal/sub-goal relationships, snippets can describe necessary effects of actions (e.g., "if a port scan detection tool is installed and there is a port scan in progress, an alert will be generated").

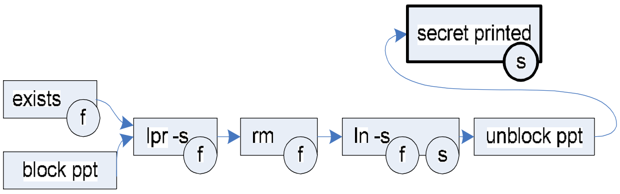


Figure 1: lpr attack snippet.

Representing domain knowledge in the form of snippets allows the system to mix-and-match various steps, making it possible to discover new combinations that could constitute possible new attacks. Further, by using automated planning algorithms instead of having human analysts manually compose scenarios, our approach provides better coverage of both attacks and normal operations. This in turn leads to higher-quality analysis of vulnerabilities, possible attacks and side-effects of countermeasures. The potential downside of this approach is that an operator could be overwhelmed with too many different plans to consider. However, we will discuss techniques which ameliorate that risk by reducing and filtering the set of all *possible* plans down to a more manageable set of *plausible* plans.

We now describe how this algorithmic core may be used in end-user tools for security analysis.

4 Discovering potential plans of attack

One way to discover potential vulnerabilities in a system is to think like an attacker. Our tools allow the user to specify multiple high-level goals, both for attacks and for (typically) benign activities. In the network intrusion detection domain, the goals of an attacker can be *denial of service* (e.g., taking down a web site) *unauthorized access* to sensitive information (e.g., password or credit card numbers) and *unauthorized modification* of data (e.g., web site defacement). To stage an attack, multiple objectives may need to be achieved (e.g., first steal a web site password, and then deface the site).

Once the goals/objectives have been established/defined, the planning system then builds a set of plans in the form of partially ordered sets of executable actions and/or observable states that achieve these goals. Since multiple goals are considered together, actions are reused where possible. For example, in the data network domain, installation of Trojan software can be used to 1) hide an attacker break in to avoid early detection, 2) capture key strokes for stealing passwords, and/or 3) install backdoors that enable an attacker to use the compromised system for a DDoS attack.

The resulting set of plans is similar to a set of attack trees [21] , where each path through an attack tree represents a unique attack with overall attacker's objective placed in the root. A snippet can be seen as node (or a subset of nodes) in an attack tree. An attack tree representation requires explicit chronological orderings of "exploits" (nodes). Since our approach

[1] To carry out the attack, an attacker requests printing of file *doc.txt* that he is authorized to access. User rights are checked by lpr command, access is granted, and the request is put into the printer queue. Then, before the printing actually starts, the attacker removes the printed file and replaces it by a link to file *secret.txt* he is not allowed to access. As a result, the latter file will eventually be printed and the attacker will be able to get the sensitive information.

supports snippet reuse and captures partial orderings of plans, it offers a more compact representation. Furthermore, attack trees do not facilitate adequate visualization support for analyzing different attack plans. After a set of plans has been constructed, our scheme enables a user/analyst to both explore individual plans and run multiple analyses on the set of plans as a whole. The rest of this section illustrates some of the ways this can be done. The screenshots are taken from several prototype tools built as part of the IARPA PAINT (ProActive INTelligence) program.

4.1 Exploring details of a plan and comparing individual plans

Note that because of reuse, an action may serve as a sub-goal supporting multiple goals. Each token has a set of variables, including a description of the start and end of the action, resources involved, and the agent performing the action. For example, to achieve the goal "find vulnerable system to install a Trojan", several actions are needed:

- Step 1: Scan for active IP addresses, open ports, operating systems (OS) and any applications running.
- Step 2: Create a report.
- Step 3: Determine the patch level of the OS or applications.
- Step 4: Attempt to exploit the vulnerability.

Scanners may either be malicious or friendly. Friendly scanners usually stop at step 2 and occasionally step 3 but never go to step 4.

Constraints, such as temporal ordering and resource restrictions, and semantic relations, connect the variables. In the "Trojan" example above, the temporal constraints are: Step1 before Step2 before Step3 before Step4.

The user can browse details of each plan to discover why each action is added to the plan and which potential goals each action is contributing to. In addition, the user can compare individual plans with each other or explore common features among sets of plans.

Figure 2 shows a graphical interface for visually comparing two plans. Circles and squares represent high-level and atomic actions respectively. Each rhombus represents a constraint. Blue shapes describe elements common to both plans. Green and white shapes show elements present in one plan but not the other. Such comparisons help the analyst identify differentiating points between attacks and benign activity, as well as potential conditions for generating alarms and warnings.

The table interface (Figure 3) is used for comparing group of plans. The user selects a set of plans from the tree of generated plans. The table interface distributes different states and actions in the plans into separate columns. Each row describes one plan. This arrangement allows users to see common and distinguishing

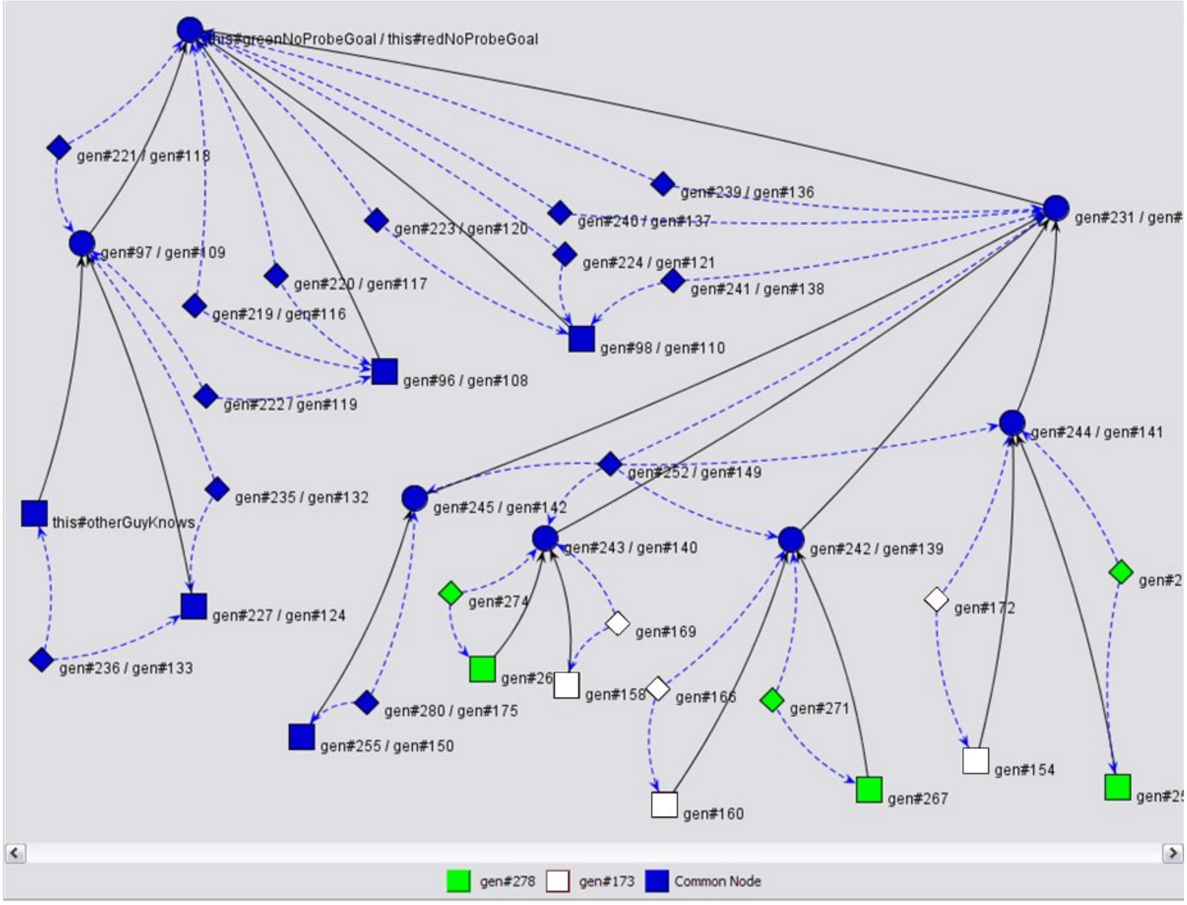


Figure 2: Plan comparison interface.

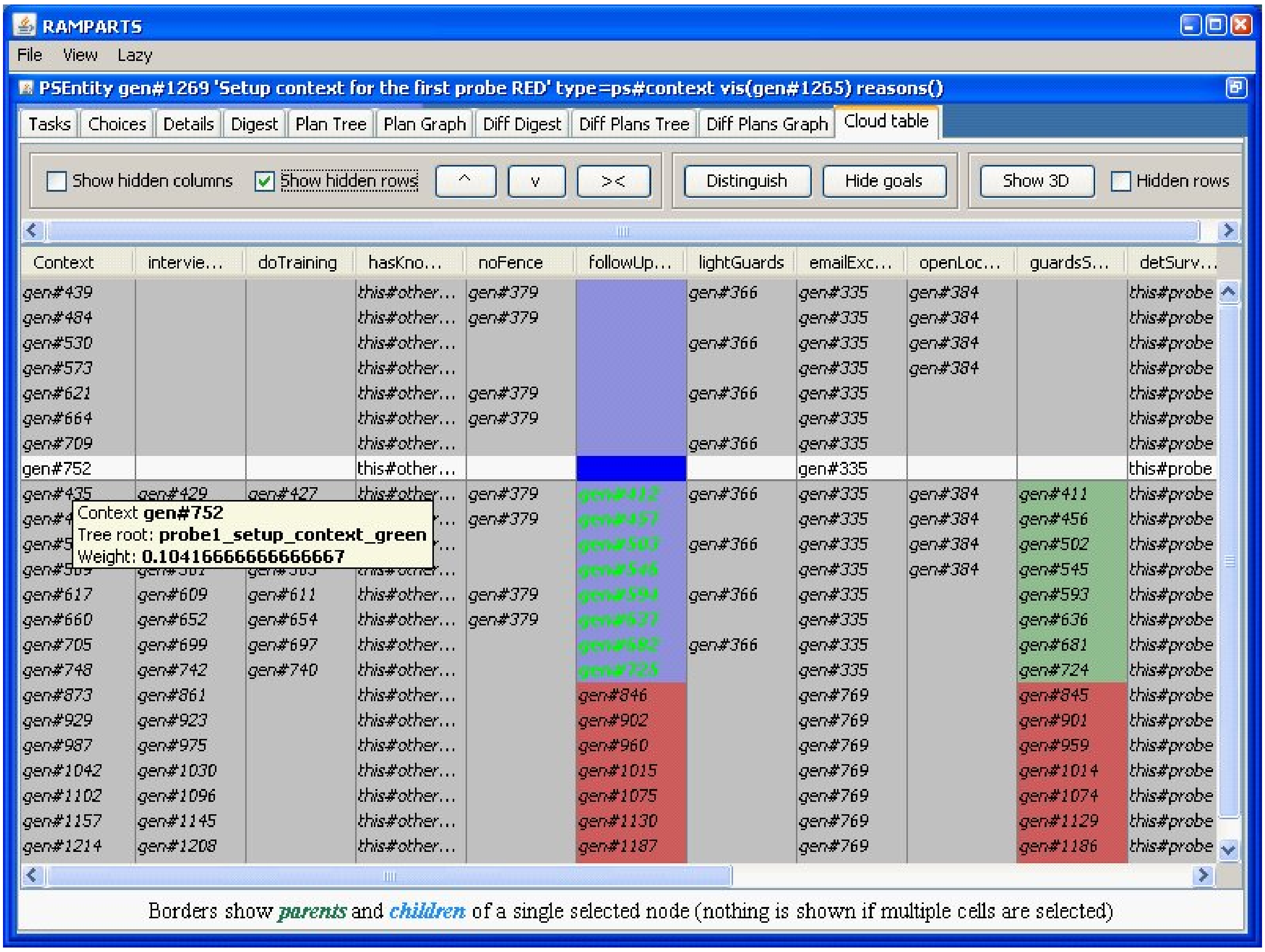
components of plans. The table can be sorted by any column, for example, to separate plans which contain a particular undesirable action from those that do not.

Any column of the table can be declared observable or hidden. The user can then select a set of rows and use the user interface (the Distinguish button) to determine if the selected plans can be distinguished from those not selected. This function allows one to quickly determine if a given set of observations is sufficient to determine the intent of the adversary regardless of which particular plan in the set is followed. Distinguishable plans are greyed out. In Figure 3, all plans are distinguishable given the chosen set of observables except the scenario identified as **gen#752**.

The graphical version of the group comparison interface (Figure 4) shows the goal-subgoal structure of the plans and visualizes the relative frequency of different actions in plans and presence of individual actions in distinguishable and non-distinguishable plans. Where this approach is helpful is in situations where the set of generated (and possible) plans can be divided into a *benign* group and a *nefarious* group and there exists a certain event (i.e., a node) which, if observable, clearly distinguishes the direction (benign vs. nefarious) the potential attacker is going regardless of the remaining nodes/events/steps in each plan. In such a situation, if a certain event is observed that clearly indicates the potential adversary is implementing one of the benign plans, there is no need to continue monitoring for additional intelligence/observables. On the other hand, if a certain event is observed that clearly indicates that the potential adversary is implementing one of nefarious plans, then further analysis is required to determine the path being taken in order to implement appropriate countermeasures.

4.2 Plan recognition

In addition to building plans from goals to observables, our approach could also be used bottom up (i.e., from observables to goals). This process provides alternative feasible explanations of observed activity (note: this has not yet been implemented, however). For instance, if we observe the actions depicted in Figure 1, we may suspect an unauthorized file access scenario exploiting the **lpr** vulnerability. The list of actions indicates an attack if a user who executes the commands does not have access to the file *secret.txt*. Otherwise, this may indicate unusual, but non-malicious activity.



PSEntity gen#1269 'Setup context for the first probe RED' type=ps#context vis(gen#1265) reasons()

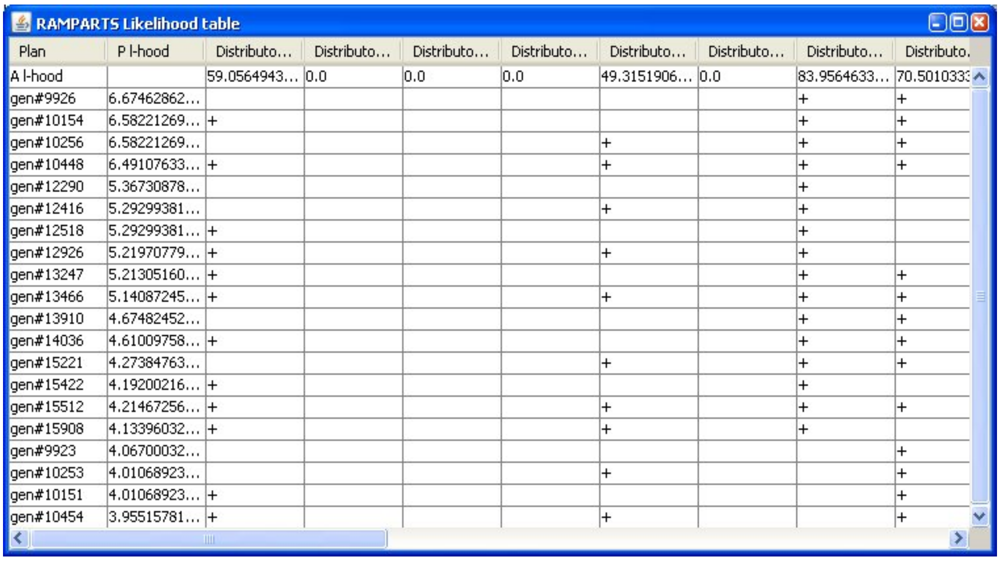
□ Show hidden columns ☑ Show hidden rows □ Hidden rows

Context	intervie…	doTraining	hasKno…	noFence	followUp…	lightGuards	emailExc…	openLoc…	guardsS…	detSurv…
gen#439			this#other…	gen#379		gen#366	gen#335	gen#384		this#probe
gen#484			this#other…	gen#379			gen#335	gen#384		this#probe
gen#530			this#other…			gen#366	gen#335	gen#384		this#probe
gen#573			this#other…				gen#335	gen#384		this#probe
gen#621			this#other…	gen#379		gen#366	gen#335			this#probe
gen#664			this#other…	gen#379			gen#335			this#probe
gen#709			this#other…			gen#366	gen#335			this#probe
gen#752			this#other…				gen#335			this#probe
gen#435	gen#429	gen#427	this#other…	gen#379	gen#412	gen#366	gen#335	gen#384	gen#411	this#probe
gen#4[illegible]	[illegible]	[illegible]	[illegible]	gen#379	gen#457		gen#335	gen#384	gen#456	this#probe
gen#5[illegible]	[illegible]	[illegible]	[illegible]		gen#503	gen#366	gen#335	gen#384	gen#502	this#probe
gen#509	gen#501	gen#503	this#other…		gen#546		gen#335	gen#384	gen#545	this#probe
gen#617	gen#609	gen#611	this#other…	gen#379	gen#594	gen#366	gen#335		gen#593	this#probe
gen#660	gen#652	gen#654	this#other…	gen#379	gen#637		gen#335		gen#636	this#probe
gen#705	gen#699	gen#697	this#other…		gen#682	gen#366	gen#335		gen#681	this#probe
gen#748	gen#742	gen#740	this#other…		gen#725		gen#335		gen#724	this#probe
gen#873	gen#861		this#other…		gen#846		gen#769		gen#845	this#probe
gen#929	gen#923		this#other…		gen#902		gen#769		gen#901	this#probe
gen#987	gen#975		this#other…		gen#960		gen#769		gen#959	this#probe
gen#1042	gen#1030		this#other…		gen#1015		gen#769		gen#1014	this#probe
gen#1102	gen#1096		this#other…		gen#1075		gen#769		gen#1074	this#probe
gen#1157	gen#1145		this#other…		gen#1130		gen#769		gen#1129	this#probe
gen#1214	gen#1208		this#other…		gen#1187		gen#769		gen#1186	this#probe

Context **gen#752**
Tree root: **probe1_setup_context_green**
Weight: **0.10416666666666667**

Borders show *parents* and *children* of a single selected node (nothing is shown if multiple cells are selected)

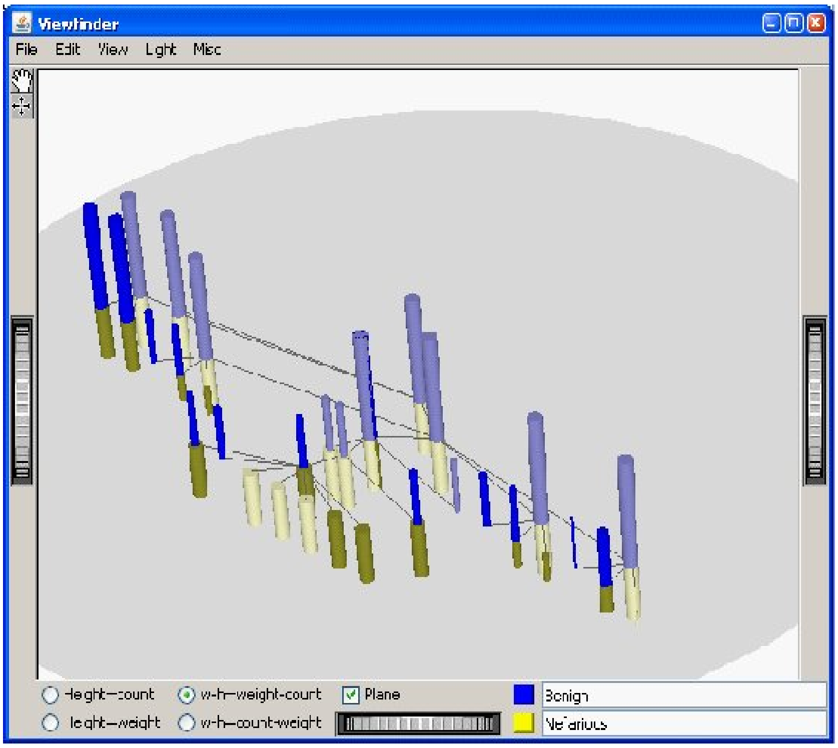
Figure 3: Table-based interface for comparing sets of plans



RAMPARTS Likelihood table

Plan	P l-hood	Distributo...	Distributo...	Distributo...	Distributo...	Distributo...	Distributo...	Distributo...	Distributo.
A l-hood		59.0564943...	0.0	0.0	0.0	49.3151906...	0.0	83.9564633...	70.501033
gen#9926	6.67462862...							+	+
gen#10154	6.58221269...	+						+	+
gen#10256	6.58221269...					+		+	+
gen#10448	6.49107633...	+				+		+	+
gen#12290	5.36730878...							+	
gen#12416	5.29299381...					+		+	
gen#12518	5.29299381...	+						+	
gen#12926	5.21970779...	+				+		+	
gen#13247	5.21305160...	+						+	+
gen#13466	5.14087245...	+				+		+	+
gen#13910	4.67482452...							+	+
gen#14036	4.61009758...	+						+	+
gen#15221	4.27384763...					+		+	+
gen#15422	4.19200216...	+						+	
gen#15512	4.21467256...	+				+		+	+
gen#15908	4.13396032...	+				+		+	
gen#9923	4.06700032...								+
gen#10253	4.01068923...					+			+
gen#10151	4.01068923...	+							+
gen#10454	3.95515781...	+				+			+

Figure 5: Interface for weighting plans and actions



○ Height–count ● w-h-weight-count ☑ Plane 🟦 Benign

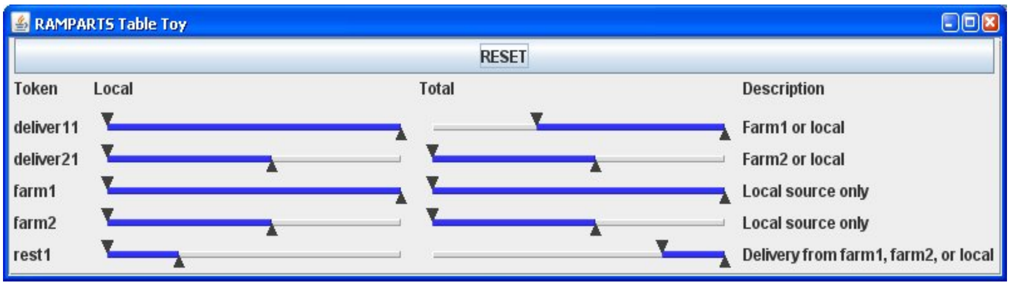
○ Height–weight ○ w-h-count-weight 🟨 Nefarious

Figure 4: Graphical interface for comparing sets of plans

4.3 Weighting plans and actions

Typically, there are multiple paths that achieve the same result. In these cases, some plans are more preferable (and/or likely) than others. Figure 5 shows a user interface that allows the analyst to adjust weights of individual plans and/or actions. This tool can be used to identify correlations between actions and states. For example, this feature may be useful in cases where an event has occurred unobserved, but a related event is observable. For instance, we may not know that our system was compromised and is being used as "zombie" to stage a DDoS attack against other host, but we can observe unusual traffic indicating the presence of a hacker DDoS tool (e.g., trinoo or TFN).

The weight assigned to an action or observable state (specified in the columns) is the likelihood of it to actually happening, given the set of plans of which it is a part and likelihoods of these plans. The weight of a plan is the likelihood of it being carried out as opposed to alternative plans. The initial weights of plans may be assigned uniformly or using a heuristic function. For example, due to relative likelihood, the same attack plans against a server for an online banking service, might be


RAMPARTS Table Toy

RESET

Token	Local	Total	Description
deliver11			Farm1 or local
deliver21			Farm2 or local
farm1			Local source only
farm2			Local source only
rest1			Delivery from farm1, farm2, or local

Figure 6: Slider interface for exploring likelihoods.

assigned higher values than they would be against a personal home server.

Users can adjust weights of plans and actions to simulate various "what-if" scenarios. Once a single value is modified, the implications are propagated to other values. For example, declaring that a particular observable has been detected will increase the likelihood of plans containing it and decrease likelihood of plans that do not contain it. This, in turn, affects likelihoods of other actions and observables constituting these plans.

4.4 Likelihood propagation within a plan

The logical structure of tokens, constraints, and goal-sub-goal links can be leveraged to reason about the likelihood of violations within each plan. The user interface presented in Figure 6 provides access to such reasoning. This interface also allows the user to treat likelihoods qualitatively, rather than quantitatively. In practice, single numbers describing likelihoods are both hard to obtain and hard to understand. Our user interface instead represents the range of likelihoods from 0 (definitely false) to 1 (definitely true), using double-headed sliders. For example, a range of [0, 1] means "no information", while [0, 0.3] would mean "unlikely, but not yet completely ruled out".

Figure 6 illustrates a simple scenario, where a restaurant buys produce from two farms and delivers using two independent companies. The sliders describe the likelihood of food contamination at each stage of this transportation network. The **Local** column describes introduction of a contaminant at the given location, and the **Total** column corresponds to the presence of the contaminant resulting from all upstream sources. This interface allows the user to see the pruning effect of probes. For example, suppose food contamination has been detected at the restaurant (hence **rest1**'s **Total** slider set to the right), but local contamination at the site has been ruled out (hence **rest1**'s **Local** slider set to the left). These two observations do not give us any additional information. If we further observe that the produce delivered from **farm2** is clean (hence **deliver21**'s slide in the **Total** column being set to the left), the system can deduce that both **farm2** and the delivery company are clean. It thus follows that the contaminant was introduced somewhere along the second transportation chain, although it is not possible to say whether it happened at **farm1** or during transportation (which is why the corresponding sliders show the [0, 1] interval).

Stopping attacks with minimum collateral damage

The analytical features described in the previous section can be combined to generate more powerful tools for a given domain. One goal of such tools is development of countermeasures that minimize collateral damage.

False positives are a major problem in alert generation. One way to decrease them is to perform diagnostic actions upon initial detection of a potential problem. The reaction to these alert-triggered diagnostics can, in many cases, provide a clue as to whether or not the alert in question is indeed part of an attack. This allows adjusting the confidence in the original alert, thereby reducing false positives.

By analysing multiple ways to execute an attack and contrasting them with multiple ways to perform benign activities, our tools help to identify potential points for alert generation and intervention. For example, these tools may help an analyst identify an action (i.e., a probe) that could force a potential adversary to (unknowingly) proceed down a pathway that produces a predicted benign outcome rather than proceed down a pathway that produces a predicted nefarious outcome.

One challenge for this approach is selection of an action for diagnosis or intervention. Often, such actions may affect not only an attack but benign activity as well. For example, detection of SYN flooding attack may use active probing [23] that collects data on the delay between the server and the client. While timely and reliable, this intervention imposes processing and storage overhead. Techniques already described can be reapplied to address this class of problems.

Earlier we described tools that help an analyst analyze sets of attack plans and contrast them with benign activities. To do so, a scenario is seeded with a set of goals, and the system generates various potential instantiation of the scenario. This same technique can also be used to analyse potential effects of diagnostic actions or countermeasures. The new action is simply added to the description of the scenario. The resulting set of plans can then be contrasted with the one obtained without the diagnostic action/countermeasure.

This approach can be used to analyze effects of response strategies in addition to individual actions. Multiple actions can be added to the scenario. Further, by introducing a new snippet instead of a fixed action, the analyst can model actions triggered by certain activity. As a result, such actions will occur in some realizations of the scenario but not in others.

5 Related work

Huang et. al. [16] developed a technique for automated plan recognition in the field of RoboCup simulation soccer games. For each agent representing a player in the game, they translated the actions observed from various adversaries (consecutive or discrete multivariable streams) into behavior queues using prediction and backfill techniques. After populating an agent's behavior queue, frequent and interesting behavior sequences were identified using a statistical dependency test. These sequences were then retrieved and transformed into formalized plans. Finally the plans were refined as multi-agent teams adopted them. [13] applied Hidden Markov Models (HMMs) for recognizing opponent behaviours in RoboCup soccer simulations. HMM states corresponded to decomposed robot behavior. Uncertainty in recognizing behaviour was represented as probabilistic transitions between the states. [19] adopted case-based reasoning (CBR) for opponent modelling and planning players' strategies in RoboCup competitions. Solutions to problems were found by reusing solutions to similar problems encountered in the past.

None of these techniques are directly applicable to the security domain. In every case, their plan recognition process depends on the observations of opponent players, position of ball and gates, and the game state at a particular moment. In security domain, we may not know who the adversary is, what its goal is, and whether the adversary exists (observed activity can be legitimate).

Attacker plan recognition [10] [7] [22] in the network security domain largely concentrates on correlation of observed actions and alerts produced by intrusion detection systems. [10] presented a probabilistic model of plan recognition for recognizing and predicting the intentions of the agents based on the construction of execution traces from raw security alerts. This method requires a library of fully predefined attack plans and lacks support for reasoning about deceptive actions by an adversary. [7] proposed a method for detecting various steps of an intrusion scenario, casting it as a planning activity based on a declarative description of actions, goals, and plans. The method does not, however, provide additional information to distinguish between more vs. less plausible scenarios. As we noted earlier, this is a very important issue because the number of possible scenarios can be quite large. [1] extends the previous approach by providing the ability to rank possible scenarios. [22] proposed a graph-based technique to correlate isolated attack scenarios derived from low-level alerts. Attack trees define attack plan libraries used to correlate isolated alert sets that are converted into causal networks with assigned probability distributions to evaluate the likelihood of attack goals and predict future attacks.

None of these systems provide visual tools for an analyst to explore sets of possible scenarios under various observables, levels of importance (or priorities), and likelihood conditions. These aids are essential for helping analysts generate probes and countermeasures.

[14] proposed a method to analyse and test threats posed by malicious insiders. They used AI planning to automatically generate courses of action an adversary could choose in subverting the system. The analyst can then use this information to evaluate the vulnerability of a system to attacks, and to select the most reasonable defensive measures. There is no notion of uncertainty or likelihood in the generated plans, and no support for comparative analysis of several plans to achieve a given goal. [17] presented an application of plan recognition techniques to support analysts in processing national security alerts by automatically identifying the hostile intent behind them. The system needs a complete library of manually-generated attack templates, a daunting requirement.

6 RAMPARTS Prototype 1.0

As noted earlier, we have developed an initial prototype implementation of a number of the features and capabilities described in this paper. The Risk Analyses and Models of Plans of Attack for Recognizing Terrorist Schemes (RAMPARTS) project was funded by IARPA as part of the ProActive INTelligence (PAINT) program. The objective of this effort was to demonstrate an initial proof-of-concept by developing supporting infrastructure and implementing a subset of capabilities.

Based on an initial set of goals and a set of plan snippets (generated by subject matter experts), the RAMPARTS prototype (1.0) generates and visually displays possible plans (both nefarious and benign) that a potential adversary/opponent might follow. The RAMPARTS toolkit also allows the user/analyst to explore the plans to help determine which key actions/events – if observed – could be used to help the analyst predict whether the potential adversary is going down a nefarious or benign pathway without actually knowing which exact pathway is being taken.

The next step (to be implemented in prototype 2.0) is to determine which "probe" or "probes" (active or passive) to implement to possibly cause the specified event/action to be observed or to cause (or at least attempt to cause) the potential adversary to go down a benign pathway (ideally, without their knowledge). In addition, we plan to use the DHS and NSF funded DETER [2] infrastructure for conducting experiments in computer security, as a test bed for further development of the project. Further, we plan to test the next prototype on a number of port security scenarios as part of the DHS sponsored USC Center for Risk and Economic Analysis of Terrorism Events (CREATE) PortSec (Port Security) project.

Acknowledgement

This research was partially supported by the Intelligence Advanced Research Projects Activity (IARPA) under grant number FA8750-07-2-0177. However, all opinions, findings, and conclusions or recommendations in this document are those of the authors and do not necessarily reflect the views of IARPA.

References

[1] S. Benferhat, F. Autrel et F. Cuppens (2003). Enhanced Correlation in an Intrusion Detection Process. *In Proceedings of Second International Workshop Mathematical Methods, Models and Architectures for Computer Networks Security*.

[2] T. Benzel, R. Braden, D. Kim, A. Joseph, C. Neuman, R. Ostrenga, S. Schwab, K. Sklower (2007). Design, Deployment, and Use of the DETER Testbed, *In Proceedings of the DETER Community Workshop on Cyber Security Experimentation and Test.*

[3] J. Blythe (1999). Decision-Theoretic Planning. *AI Magazine*, 20(2).

[4] M. Buro & T. Furtak (2003), RTS Games as Test-Bed for Real-Time Research, *Invited Paper at the Workshop on Game AI*, JCIS

[5] CALO (2003). Cognitive agent that learns and organizes, *http://calo.sri.com.*

[6] C. A. Carver, J. M. D. Hill, J. R. Surdu, and U. W. Pooch (2000), A Methodology for using Intelligent Agents to provide Automated Intrusion Response, *Proceedings of the IEEE Systems, Man, and Cybernetics Information Assurance and Security Workshop*.

[7] F. Cuppens, F. Autrel, A. Mi`ege and S. Benferhat (2002). Recognizing Malicious Intention in an Intrusion Detection Process. *Soft Computing Systems - Design, Management and Applications*, volume 87, 806–817.

[8] R. Dechter (2003). Constraint Processing. *Morgan Kaufmann Publishers Inc*.

[9] K. Erol, J. Hendler, and D. Nau (2004). UMCP: A sound and complete procedure for hierarchical task-network planning. *Proceedings of AIPS*.

[10] C. W. Geib and R. P. Goldman (2001). Plan Recognition in Intrusion Detection Systems. *In Proceedings of the Second DARPA Information Survivability Conference and Exposition*.

[11] Geib, C., Goldman, R. (2002), Requirements for Plan Recognition in Network Security Systems, *Proceedings of the Recent Advances in Intrusion Detection conference.*

[12] Frank M, Frans V (2003). A formal description of tactical recognition [J]. *Information Fusion*, 4(1): 47-61.

[13] K. Han and M. Veloso (1999). Automated robot behavior recognition applied to robotic soccer. *Proceedings of the Workshop on Team Behaviors and Plan Recognition*, 47 –52.

[14] S. Harp, J. Gohde (2005), Thomas Haigh, M. Boddy Automated Vulnerability Analysis Using AI Planning, 2005 *AAAI Spring Symposium on AI for Homeland Security.*

[15] C. Heinze, S. Goss, and A. Pearce (1999). Plan Recognition in Military Simulation: Incorporating Machine Learning with Intelligent Agents. *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence*, Workshop on Team Behaviour and Plan Recognition, pages 53-63Author (year). *Title of the book*. Publisher.

[16] Z. Huang, Y. Yang and X. Chen (2003). An approach to plan recognition and retrieval for multi-agent systems. *Proceedings of AORC.*

[17] Jarvis, P.; Lunt, T. F.; Myers, K. L (2004). Identifying terrorist activity with AI plan recognition technology. *National Conference on Artificial Intelligence, AAAI Press.*

[18] Kichkaylo, T., van Buskirk, C., Singh, S., Neema, H., Orosz, M., and Neches (2007), R. Mixed-Initiative Planning for Space Exploration Missions, *Workshop on Moving Planning and Scheduling Systems into the Real World*.

[19] C Marling, M Tomko, M Gillen, D Alexander, D (2003). Case-based reasoning for planning and world modeling in the robocup small size league, IJCAI *Workshop on issues in designing physical agents.*

[20] P. A. Porras and P. G. Neumann (1997), EMERALD: Event Monitoring Enabling Responses to Anomalous Live Disturbances, *Proceedings of the National Information Systems Security Conference*, pp. 353-365.

[21] Schneier, B., “Attack Trees.” Dr Dobbs Journal. December 1999.

[22] Qin X. and Lee W. (2004), Attack Plan Recognition and Prediction Using Causal Networks, *ACSAC-O4*, 370-379.

[23] B. Xiao, W. Chen, Y. He, E. H.-M. Sha (2005). An active Detecting Method Against SYN Flooding Attack, icpads, vol. 1, pp.709-715, *11th International Conference on Parallel and Distributed Systems* (ICPADS'05).

Improving Morphosyntactic Tagging of Slovene Language through Meta-tagging

Jan Rupnik, Miha Grčar and Tomaž Erjavec
Jožef Stefan Institute, Jamova cesta 39, Ljubljana
E-mail: {jan.rupnik, miha.grcar, tomaz.erjavec}@ijs.si, http://kt.ijs.si

Keywords: PoS tagging, meta-tagger, Slavic languages, FidaPLUS, JOS corpus, machine learning, Orange, decision trees, CN2 rules, Naive Bayes

Received: July 12, 2009

Part-of-speech (PoS) or, better, morphosyntactic tagging is the process of assigning morphosyntactic categories to words in a text, an important pre-processing step for most human language technology applications. PoS-tagging of Slovene texts is a challenging task since the size of the tagset is over one thousand tags (as opposed to English, where the size is typically around sixty) and the state-of-the-art tagging accuracy is still below levels desired. The paper describes an experiment aimed at improving tagging accuracy for Slovene, by combining the outputs of two taggers – a proprietary rule-based tagger developed by the Amebis HLT company, and TnT, a tri-gram HMM tagger, trained on a hand-annotated corpus of Slovene. The two taggers have comparable accuracy, but there are many cases where, if the predictions of the two taggers differ, one of the two does assign the correct tag. We investigate training a classifier on top of the outputs of both taggers that predicts which of the two taggers is correct. We experiment with selecting different classification algorithms and constructing different feature sets for training and show that some cases yield a meta-tagger with a significant increase in accuracy compared to that of either tagger in isolation.

Povzetek: V članku je opisano označevanja slovenskih besedil z združevanjem Amebisovega označevalnika in označevalnika TnT.

1 Introduction

Morphosyntactic tagging, also known as part-of-speech tagging or word-class syntactic tagging is a process in which each word appearing in a text is assigned an unambiguous tag, describing the morphosyntactic properties of the word token. Such tagging is the basic pre-processing step for a number of applications or more advanced analysis steps, such as syntactic parsing. Morphosyntactic tagging is, in general, composed of two parts: the program first assigns, on the basis of a morphological lexicon all the possible tags that a word form can be associated with (morphological look-up), and then chooses the most likely tag on the basis of the context in which the word form appears in the text (disambiguation). For words not appearing in the lexicon, various taggers either ignore them or employ heuristics to guess at their tag.

Unlike English, morphologically richer Slavic languages such as Czech (Hajič and Hladka, 1998) or Slovene typically distinguish more than a thousand morphosyntactic tags. In the multilingual MULTEXT-East specification (Erjavec, 2004) almost 2,000 tags (morphosyntactic descriptions, MSDs) are defined for Slovene. MSDs are represented as compact strings, with positionally coded attribute values, so they effectively serve as shorthand notations for feature-structures. For example, the MSD `Agufpa` expands to `Category = Adjective, Type = general, Degree = undefined, Gender = feminine, Number = plural, Case = accusative`.

Having such a large number of tags makes assigning the correct one to each word token a much more challenging task than it is e.g. for English. The problem for Slovene has been exacerbated by the lack of large and available validated tagged corpora, which could serve as training sets for statistical taggers.

Recently, new annotated language resources have become available for Slovene. FidaPLUS[1] (Arhar & Gorjanc, 2007) is a 600 million word monolingual reference corpus automatically annotated with MULTEXT-East MSDs by the Slovene HLT company Amebis[2]. But while FidaPLUS is freely available for research via a Web concordancer, it is not generally available as a dataset. In order to remedy the lack of publicly available annotated corpora for HLT research on Slovene, the JOS project (Erjavec and Krek, 2008) is making available two corpora under the Creative Commons license. Both contain texts sampled from FidaPLUS, with the smaller jos100k containing 100,000 words with fully validated morphosyntactic annotations, and the larger, jos1M having 1 million words, and partially hand validated annotations – project resources preclude fully validating the latter.

Previous experiments (Erjavec et al., 2000) showed that from various publicly accessible taggers the best

[1] http://www.fidaplus.net/

[2] http://www.amebis.si/

results were achieved by TnT (Brants, 2000). TnT is a Hidden Markov Model tri-gram tagger, which also implements an unknown-word guessing module. It is fast in training and tagging, and is able to accommodate the large tagset used by Slovene.

Having the validated jos100k at our disposal, we experimented with training TnT and seeing how its errors compare to the ones assigned by the Amebis tagger. It turned out that the two taggers are comparable in accuracy, but make different mistakes. This gave us a method of selecting the words that should be manually corrected in jos1M – only those tokens where the annotations between the taggers differ were selected for manual inspection. This approach concentrated on validating the words where state-of-the-art taggers are still able to make correct decisions, at the price of ignoring cases where both taggers predict the same but incorrect tag, i.e. the truly difficult cases.

Having several automatically tags for each word also offers the possibility of combining their outputs in order to increase accuracy, say, over the whole FidaPLUS corpus. Experiments in combining PoS taggers have been attempted before, using various learning strategies, and for various languages, e.g. voting, stacking, etc. for Swedish (Sjöbergh, 2003) or multi-agent systems for Arabic (Othmane Zribi et al., 2006). An experiment, more similar to ours, is reported in Spoustová et al. (2007) for Czech, also using a rich positional tagset, where several stochastic taggers are combined with a rule based one; the rule based tagger is used predominantly as a pre-disambiguation step, to filter out unacceptable tags from the ambiguity classes of the tokens.

This paper presents a similar experiment, which, however, uses only two independent taggers therefore precluding combination methods such as voting or pipelining. But as in the Czech case, we also need to deal with a very large and positionally encoded tagset.

The rest of this paper is structured as follows: Section 2 presents the dataset used in the experiments, Section 3 explains the methods used to combine the output of the taggers, Sections 4 and 5 give the results of experiments on the jos100k and jos1M corpora with different methods and features, and Section 6 gives the conclusions and directions for further work.

2 Dataset

The dataset used in the first set of experiments is based on the jos100k corpus; the corpus contains samples from almost 250 texts from FidaPLUS, cca. 1,600 paragraphs or 6,000 sentences. The corpus has just over 100,000 word tokens, and, including punctuation, 120,000 tokens. jos100k contains only manually validated MSDs, of which 1,064 different ones appear in the corpus.

For the dataset we added MSDs assigned by Amebis and TnT to the manually assigned ones. Two sentences from the dataset are given in Figure 1. Annotations marking texts and paragraphs have been discarded and end of sentence is marked by an empty line. Punctuation is tagged with itself.

Prišlo	Vmep-sn	Vmep-sn	Vmep-sn
je	Va-r3s-n	Va-r3s-n	Va-r3s-n
do	Sg	Sg	Sg
prerivanja	Ncnsg	Ncnsg	Ncnsg
in	Cc	Cc	Cc
umrla	Vmep-sf	Vmep-sf	Vmep-sf
je	Va-r3s-n	Va-r3s-n	Va-r3s-n
.	.	.	.
Tega	Pd-nsg	Pd-msa	Pd-msg
se	Px------c	Px------c	Px------c
sploh	Q	Q	Q
nisem	Va-r1s-y	Va-r1s-y	Va-r1s-y
zavedel	Vmep-sm	Vmep-sm	Vmep-sm
.	.	.	.

Figure 1: Example stretch of the corpus dataset ("*Prišlo je do prerivanja in umrla je. Tega se sploh nisem zavedel.*"). First column is the word-form, second the gold standard manually assigned tag, third the one assigned by TnT, and the fourth by Amebis. Note the first word of the second sentence, where both taggers make a mistake.

The source FidaPLUS corpus also contains, for each word token, all possible MSDs that could be assigned to it, i.e. its ambiguity class. Based on this information, we computed the average per-word MSD ambiguity which turns out to be 3.13 for the jos100k corpus. So, on the average, a tagger needs to choose the correct MSD tag between three possibilities. Note that disambiguation is only possible for known words.

2.1 Amebis MSDs

The Amebis MSDs were taken from the source FidaPLUS corpus; as mentioned, the Amebis tagger is largely a rule-based one, although with heuristics and quantitative biases. The tagger uses a large lexicon, leaving only 2% of the word tokens in jos100k unknown. Amebis doesn't tag these words, and they have all been given a distinguished PoS/MSD "unknown". Furthermore, FidaPLUS is annotated according to the MULTEXT-East specification, while the JOS corpus uses a modification, based on, but different from the MULTEXT-East/FidaPLUS one. Differences concern reordering of attribute positions, changes in allowed values, etc., as well as lexical assignment. For the most part an information-preserving conversion is possible, but for MSDs (attributes) of some lexical items only heuristics can be used for the conversion. Taking into account that all Amebis "unknowns" are by definition wrong, as all words are manually annotated with specific MSDs, and that a certain number of errors is introduced by the tagset mapping, Amebis obtains 87.9% accuracy on all tokens (incl. punctuation) in the dataset.

2.2 TnT MSDs

The TnT tagger was trained on the dataset itself, using 10-fold cross-tagging. The dataset was split into 10 parts, with 9 folds used for training, and the remaining fold tagged with the resulting model, and this process repeated for all 10 folds. As the lexical stock of jos100k is small, the tagging model used a backup lexicon which was extracted from the FidaPLUS corpus and its annotations. In other words, tri-gram statistics and lexicon containing uni-gram statistics of word-forms (their ambiguity classes) of frequent words were learned from jos100k, while less frequent words obtained their ambiguity classes from MSDs assigned by the Amebis tagger. Given such a tagging set-up, the obtained accuracy over the all dataset tokens (incl. punctuation) for TnT is 88.7%, slightly better than Amebis; but TnT has the advantage of learning how to correctly tag at least some unknown words (such as those marked as "foreign", i.e. tokens in spans of non-Slovene text), as well as having less problems with tagset conversion. Nevertheless, on the dataset it performs better than Amebis, so the TnT accuracy can be taken to constitute the baseline for the experiment.

2.3 Error comparison

Table 1 compares the errors made by the taggers against the gold standard. The first line gives the complete size of the corpus in words. The second gives the number of correct MSD assignment to word tokens for TnT (86.6% per-word accuracy), and the third for Amebis (85.7%). The fourth line covers cases where both taggers predict the correct MSD, for 78% of the words.

Lines 5 and 6 cover cases where one tagger correctly predicts the tag, while the other makes a mistake. These two lines cover a significant portion (2/3) of all the errors, so if such mistakes can be eliminated by deciding which tagger made the correct choice, the gains in accuracy are considerable.

The last two lines indicate upper bounds on the gains achieved by concentrating on choosing the correct tag. Line 7 gives cases where both taggers agree, but on an incorrect tag (3.2%), and line 8 the number of cases where both are wrong, but in different ways (2.4%); the upper bound on combination accuracy is thus 94.3%.

Let us look at two typical examples of cases 7 and 8. An example of both taggers being wrong, but agreeing on the assigned tag is exemplified in the fragment *"ni mogoče povedati" (it is not possible to tell)* where *"mogoče"* should be an adverb but both taggers assign it an adjectival tag. An example of both taggers being wrong in different ways is the fragment *"ni priporočene/Adj zgornje/Adj mejne/Adj vrednosti/Adj" (there is no recommended upper bound value)*. The correct tag for the noun is *Ncfsg*, i.e. feminine singular genitive, the genitive being determined by the (long distance) dependency on "*ni"*. The Amebis tagger correctly predicts this tag, while TnT makes a mistake, and assigns to the noun the plural accusative. As adjectives must agree with the noun in gender, number and case, the three adjectives preceding the noun must also be tagged as feminine singular genitive. Here both taggers are wrong: while TnT correctly posits the agreement between the noun and adjectives, all the adjective tags are wrong, due to the noun being incorrectly tagged. Amebis, on the other hand, does not pick up the agreement, and tags all three adjectives as masculine ones.

	Words	Gold	Amebis	TnT	Gloss
1	100,003	MSD1			Words in dataset
2	86,623	MSD1		MSD1	TnT tagger correct
3	85,718	MSD1	MSD1		Amebis tagger correct
4	78,018	MSD1	MSD1	MSD1	Both taggers correct
5	7,700	MSD1	MSD1	MSD2	Amebis correct, TnT error
6	8,605	MSD1	MSD2	MSD1	Amebis error, TnT correct
7	3,232	MSD1	MSD2	MSD2	Both wrong, and identical
8	2,448	MSD1	MSD2	MSD3	Both wrong, and different

Table 1: Comparison of tagging accuracy of Amebis and TnT over the 100k dataset.

3 Combining the taggers

As mentioned, our meta-tagger is built on top of two taggers, the Amebis rule-based tagger and TnT. The sole task of the meta-tagger is to decide which tag to consider correct. The meta-tagger is implemented as a classifier which, if the two underlying taggers disagree, classifies the case into one of the two classes indicating which of the two taggers is more likely to be correct. To train the classifier, we needed two things: a way to describe a case with a set of features, and a classification algorithm. The following section describes the feature construction process and the subsequent section the classification algorithms we tried out for this task.

3.1 Feature construction

To be able to train the classifier we needed to describe each case with a set of features. We decided to keep our meta-tagger relatively simple and to construct features solely out of tags predicted by the underlying taggers. Alternatively, we could compute content features as well (such as *n*-grams, prefixes, and suffixes) as it is the case with the SVM-based taggers such as SVMTool (Giménez & Márquez, 2004).

For training and testing we used the dataset discussed in Section 2, with each word assigned three tags: the correct tag (assigned manually), the tag assigned by TnT, and the tag assigned by the Amebis tagger. Each of these three tags can be decomposed into 15 attributes such as the part-of-speech category, type, gender, number, and so on. For a given tag, not all attribute values are set, therefore the data is sparse in this sense (e.g. the value of gender and number for prepositions is "undefined").

The attributes of the tags assigned by the two taggers (but not those of the manually assigned tags) were directly used as features for training. In addition, we constructed features that indicate whether the two taggers agree on a particular attribute value or not (the so called agreement features). The example was labeled according to the tagger which correctly tagged the word (the label was thus either TnT or Amebis). Note that we built a training feature vector only when the two taggers disagreed and one of them was correct (if none of the taggers was correct, we were unable to label the feature vector). The entire feature construction process is illustrated in Figure 2.

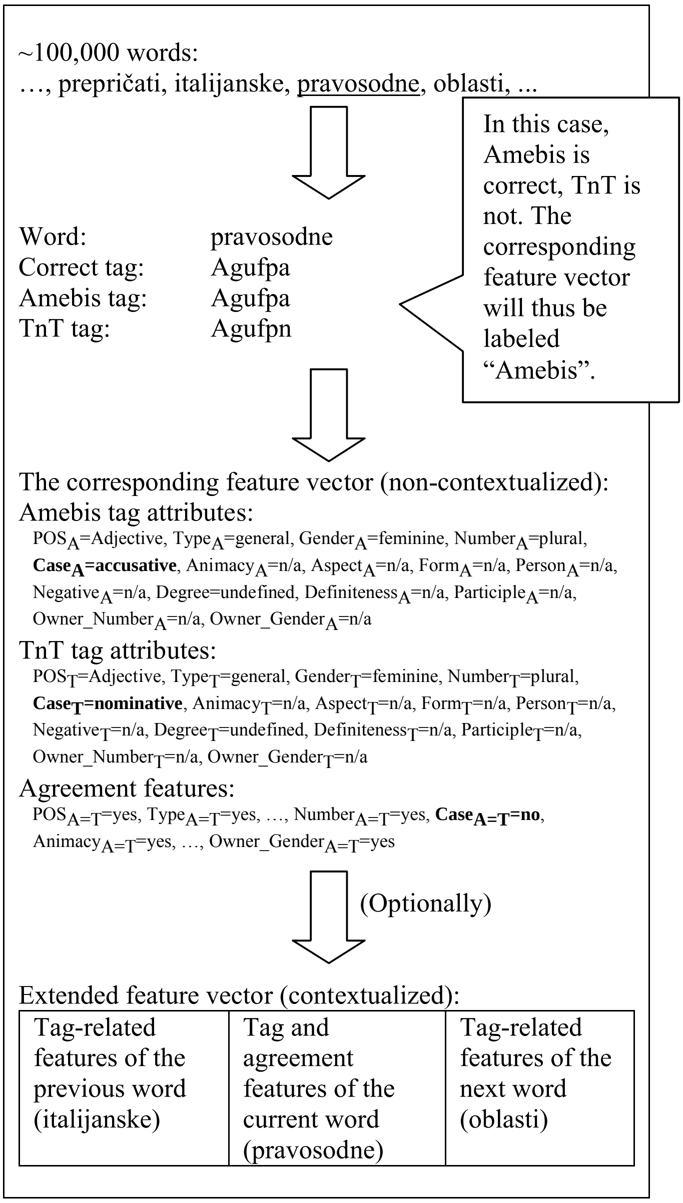
For the first set of experiments we used the tag attributes and agreement features of the current word to construct a feature vector (termed non-contextualized features in Figure 2). In the second set of experiments, on the other hand, we also added tag features (from both, TnT and Amebis) from the previous and the next word (termed contextualized features in Figure 2). It is also important to mention that we ran a set of experiments where we excluded punctuation from the text and a set of experiments where each different type of punctuation was treated as a separate part-of-speech category (e.g. POS$_T$=,) with all the other attributes set to "not applicable". Each of these settings gave slightly different results. The results are discussed in Section 4 in more detail.

3.2 Learning algorithms

We experimented with three different classification algorithms: the Naive Bayes classifier, CN2 rule-induction algorithm, and C4.5 decision tree building algorithm. In this section, we briefly describe each of them.

The **Naive Bayes (NB)** classifier is a probabilistic classifier based on Bayes' theorem.[3] It naively assumes a strong independence of features. Furthermore, it is a black box classifier in the sense that its decisions are not easily explainable.

CN2 is an if-then rule-induction algorithm (Clark & Niblett, 1989). It is a covering algorithm meaning that each new rule covers a set of examples which are thus removed from the dataset. Unlike the Naive Bayes classifier, the trained model (i.e. a set of induced rules) provides an explanation for a decision (i.e. an if-then rule that was taken into account when classifying the example). Looking at the induced rules, it is also possible to read, understand, and also verify the knowledge that was discovered in the training set.



~100,000 words:
…, prepričati, italijanske, pravosodne, oblasti, ...

In this case, Amebis is correct, TnT is not. The corresponding feature vector will thus be labeled "Amebis".

Word: pravosodne
Correct tag: Agufpa
Amebis tag: Agufpa
TnT tag: Agufpn

The corresponding feature vector (non-contextualized):

Amebis tag attributes:
POS$_A$=Adjective, Type$_A$=general, Gender$_A$=feminine, Number$_A$=plural, **Case$_A$=accusative**, Animacy$_A$=n/a, Aspect$_A$=n/a, Form$_A$=n/a, Person$_A$=n/a, Negative$_A$=n/a, Degree=undefined, Definiteness$_A$=n/a, Participle$_A$=n/a, Owner_Number$_A$=n/a, Owner_Gender$_A$=n/a

TnT tag attributes:
POS$_T$=Adjective, Type$_T$=general, Gender$_T$=feminine, Number$_T$=plural, **Case$_T$=nominative**, Animacy$_T$=n/a, Aspect$_T$=n/a, Form$_T$=n/a, Person$_T$=n/a, Negative$_T$=n/a, Degree$_T$=undefined, Definiteness$_T$=n/a, Participle$_T$=n/a, Owner_Number$_T$=n/a, Owner_Gender$_T$=n/a

Agreement features:
POS$_{A=T}$=yes, Type$_{A=T}$=yes, …, Number$_{A=T}$=yes, **Case$_{A=T}$=no**, Animacy$_{A=T}$=yes, …, Owner_Gender$_{A=T}$=yes

(Optionally)

Extended feature vector (contextualized):

Tag-related features of the previous word (italijanske)	Tag and agreement features of the current word (pravosodne)	Tag-related features of the next word (oblasti)

Figure 2: The feature construction process.

C4.5 is an algorithm for building decision trees; it is based on information entropy[4] (Quinlan, 1993). C4.5 uses the fact that each attribute of the data can be used to make a decision that splits the data into smaller subsets. It examines the normalized information gain (difference in entropy) that results from choosing an attribute for splitting the data. The attribute with the highest normalized information gain is the one used to make the decision. This process is repeated several times on smaller and smaller subsets of data. Similarly to CN2 (the rule-induction algorithm), C4.5 builds glass box models. Unlike its predecessor, the ID3 algorithm, C4.5 knows how to handle data with missing values (i.e. sparse data) and prunes the tree by cutting off branches that do not contribute to the classification accuracy.

[3] c.f. http://en.wikipedia.org/wiki/Naive_Bayes_classifier

[4] c.f. http://en.wikipedia.org/wiki/C4.5_algorithm

4 Experiments

In this section, we present tagging accuracies of the meta-tagger for different combinations of feature sets and underlying classification models. The size of the set of examples for training and testing is 16,305 and consists of 8,605 cases where TnT tagger predicted the correct tag and Amebis tagger did not and 7,700 cases where Amebis was correct and TnT was not. All experiments were conducted with the Orange data mining tool (Demšar et al., 2004). 5-fold cross validation method was used to evaluate the tagging accuracy of the meta-tagger in all experimental scenarios. We first discuss two baseline models for the meta-tagger, after that we define several different feature sets, then continue with the description of non-contextualized models and end the section with models that incorporate context features.

4.1 Baselines

The first baseline is the majority classifier which always predicts that TnT tagger is correct. This classifier achieves the accuracy of 52.8%.

The second baseline model is a Naive Bayes model trained on only one feature: Amebis MSD. This is a very simple model, since to classify a new example (with only one feature f, that is the Amebis MSD), all one needs to do is count the number of cases with MSD equal to f where Amebis was correct and the number of cases with MSD equal to f where Amebis was incorrect (*P(x = f, y = amebis-correct)* and *P(x = f, y = amebis-incorrect)*) and predict the class (amebis-correct or amebis-incorrect) with the higher count. This model achieves the accuracy of 70.95% (approx. 18% higher than the first baseline).

Let us consider two examples. Assume that there were 200 cases where Amebis predicted the tag Pd-nsg, and it was correct in 150 of these cases (this means the TnT was correct in the remaining 50 cases). This means that P(Amebis-predicts: Pd-nsg, Amebis-correct) = 0.75. In this case the meta-tagger would always predict the tag Pd-nsg if Amebis predicted it as well.

Now, if we assumed that Amebis was correct in 80 of 200 cases, P(Amebis-predicts: Pd-nsg, Amebis-correct) = 0.4), then the meta-tagger would always predict the tag predicted by TnT, given that Amebis predicted Pd-nsg (the evidence in the training data tells us not to trust the Amebis tagger, since the probability of it being correct is less than 0.5).

4.2 Feature sets

We will now describe the features for the non-contextualized models. The first set of features for the non-contextualized models are the so called FULL features; they only include full Amebis MSD and full TnT MSD (two features). The second set of features called DEC is a decomposition of the FULL features as described in Section 3.1 (45 features: 15 Amebis features, 15 TnT features, 15 Agreement features). The third set of features, BASIC, is a subset of DEC features, where we only take the features corresponding to Category, Type, Gender, Number and Case into account (10 features: 5 for Amebis and 5 for TnT). The final set of features, ALL, is a union of FULL and DEC (47 features).

Feature sets for contextualized models (with and without punctuation) are extensions of non-contextualized feature sets, where the features of examples surrounding our training example are added (see Section 3.1). The context features (i.e. the features of the previous and next word) are the same ones as that of the current word except for the Agreement features which are only computed for the current word (in the DEC feature set we thus keep only 15 Agreement features: the ones of the current word).

Features ALL, when contextualized, include six features for MDS tags (Amebis-Prev, Amebis, Amebis-Next, TnT-Prev, TnT, TnT-Next), 45 for Amebis tag features (3 × 15 features), 45 for TnT tag features and 15 Agreement features, which sums up to 111 features.

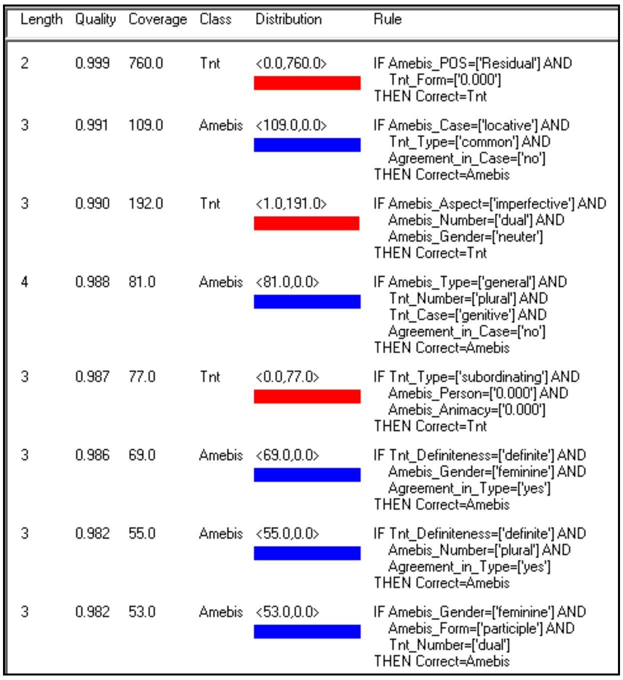
4.3 Non-contextualized models

Experiments with features that do not take context into account (Table 2) show that C4.5 is the most robust classifier with respect to different feature sets and that it can achieve the highest accuracy. We can also observe that tag features are not very suitable for the Naive Bayes classifier because the conditional independence assumptions are too strongly violated.

Feature set / Classifier	FULL	DEC	BASIC	ALL
NB	73.90	67.55	67.50	69.65
C4.5	73.51	**74.70**	74.23	73.59
CN2	60.61	72.57	71.68	70.90

Table 2: Non-contextualized models (accuracy in %). Feature sets FULL, DEC, BASIC and ALL are explained in Section 4.2.

Even though the CN2 algorithm results in slightly lower accuracy it can prove useful since the rules that it produces are easy to interpret and thus discover the strengths and weaknesses of the TnT and Amebis classifiers (see Figure 3).



Length	Quality	Coverage	Class	Distribution	Rule

2	0.999	760.0	Tnt	<0.0,760.0>
				🟥
					IF Amebis_POS=['Residual'] AND
					Tnt_Form=['0.000']
					THEN Correct=Tnt

3	0.991	109.0	Amebis	<109.0,0.0>
				🟦
					IF Amebis_Case=['locative'] AND
					Tnt_Type=['common'] AND
					Agreement_in_Case=['no']
					THEN Correct=Amebis

3	0.990	192.0	Tnt	<1.0,191.0>
				🟥
					IF Amebis_Aspect=['imperfective'] AND
					Amebis_Number=['dual'] AND
					Amebis_Gender=['neuter']
					THEN Correct=Tnt

4	0.988	81.0	Amebis	<81.0,0.0> 🟦	IF Amebis_Type=['general'] AND Tnt_Number=['plural'] AND Tnt_Case=['genitive'] AND Agreement_in_Case=['no'] THEN Correct=Amebis

3	0.987	77.0	Tnt	<0.0,77.0>
				🟥
					IF Tnt_Type=['subordinating'] AND
					Amebis_Person=['0.000'] AND
					Amebis_Animacy=['0.000']
					THEN Correct=Tnt

3	0.986	69.0	Amebis	<69.0,0.0>
				🟦
					IF Tnt_Definiteness=['definite'] AND
					Amebis_Gender=['feminine'] AND
					Agreement_in_Type=['yes']
					THEN Correct=Amebis

3	0.982	55.0	Amebis	<55.0,0.0>
				🟦
					IF Tnt_Definiteness=['definite'] AND
					Amebis_Number=['plural'] AND
					Agreement_in_Type=['yes']
					THEN Correct=Amebis

3	0.982	53.0	Amebis	<53.0,0.0>
				🟦
					IF Amebis_Gender=['feminine'] AND
					Amebis_Form=['participle'] AND
					Tnt_Number=['dual']
					THEN Correct=Amebis

Figure 3: List of rules discovered by CN2 in Orange. Rules are ordered by their quality which is a function of rule coverage and rule accuracy. The second rule, for example, tells us that if Amebis predicted locative case and TnT predicted some other case and TnT predicted common type, then the meta-tagger should predict the same tag as Amebis. The first rule, IF Amebis_POS=['Residual'] AND TnT_Form=['0.000'] THEN Correct = TnT, covers the examples mentioned in Section 2.1, where Amebis predicts POS tag "unknown" (by definition incorrect). The rule says that in such case, TnT is always correct, which is what is expected.

4.4 Context and punctuation

When comparing the results of experiments with context, we notice that taking punctuation into account (see Section 3.1) is beneficial in almost all cases (see Tables 3 and 4). This can be explained by the fact that ignoring punctuation can yield unintuitive context tags, for instance the sequence of tags T1, T2, T3, where T1 is the last word of a sentence, T2 the first word and T3 the second word of the next sentence.

We notice that C4.5 can best benefit from extra contextual features, whereas the performance of the other algorithms does not change notably.

Feature set / Classifier	FULL	DEC	BASIC	ALL
NB	73.10	68.29	67.96	70.55
C4.5	73.10	78.51	**79.23**	76.72
CN2	62.16	73.26	72.75	72.29

Table 3: Context without punctuation (accuracy in %).

Feature set / Classifier	FULL	DEC	BASIC	ALL
NB	73.44	68.32	68.14	70.53
C4.5	74.18	78.91	**79.73**	77.68
CN2	62.23	74.27	72.82	73.01

Table 4: Context with punctuation (accuracy in %).

5 Large-scale experiment

In addition to the experiments on the jos100k corpus, we also performed a large-scale experiment on a larger subset of FidaPLUS, the jos1M corpus, consisting of 1,000,017 word tokens (without punctuation). The corpus was first tagged by both taggers (i.e. Amebis and TnT). Amebis is a rule-based tagger and does not require training, TnT, on the other hand, was trained on the complete jos100k corpus. Then, if (and only if) the two taggers disagreed on a particular word token, the token was manually validated. Consequently, we are unable to determine cases when both taggers are correct or agree on an incorrect tag. Dataset statistics (analogous to the ones in Table 1) are given in Table 5.

	Words	Gold	Amebis	TnT	Gloss
1	1,000,017				Words in dataset
2	809,897		MSD1	MSD1	Both taggers agree
3	75,378	MSD1	MSD1	MSD2	Amebis correct, TnT error
4	88,657	MSD1	MSD2	MSD1	Amebis error, TnT correct
5	26,085	MSD1	MSD2	MSD3	Both wrong, and different

Table 5: The jos1M corpus statistics.

5.1 Experimental setting

We confronted Naive Bayes with C4.5 (building CN2 rules was computationally too expensive). We experimented with all defined feature sets: FULL, DEC, BASIC, and ALL, with and without context. Punctuation was included in the contextualized cases. For some reason, the C4.5 algorithm was unable to handle feature sets FULL and ALL when contextualized. We speculate that the implementation in Orange does not manage memory efficiently when it comes to attributes with 1000+ different values. The results of the experiments are presented in the following section.

5.2 Results

In this section, we present tables analogous to the ones in Section 4. We show how the algorithms perform under different feature sets. As already said, we do not show results for the CN2 algorithm and for C4.5 under certain conditions (denoted with "N/A"). The results fully support our observations on the smaller jos100k corpus and are presented in Tables 6 and 7. Note also that the second baseline yields 72.39% accuracy on the jos1M corpus.

Feature set / Classifier	FULL	DEC	BASIC	ALL
NB	73.93	66.85	66.67	69.81
C4.5	76.45	**76.56**	76.29	76.49

Table 6: The jos1M corpus – non-contextualized models (accuracy in %).

Feature set / Classifier	FULL	DEC	BASIC	ALL
NB	73.74	67.59	67.86	70.28
C4.5	N/A	**84.18**	84.01	N/A

Table 7: The jos1M corpus – context and punctuation (accuracy in %).

6 Conclusions

The paper presents a meta-tagger built on top of two taggers, namely the TnT HMM-based tagger and the Amebis rule-based tagger. The purpose of the meta-tagger is to decide which tag to take into account if the two taggers disagree in a particular case.

The experimental results show that the two taggers are quite orthogonal since very little information is needed to get a significant increase in performance from the first baseline.

Furthermore, using context can improve the performance of some models and taking punctuation into account when constructing context features is better than ignoring it. C4.5 with context and punctuation features achieves the highest accuracy, 79.73% on jos100k and 84.18% on jos1M, which results in a meta-tagger with significantly higher accuracy than Amebis tagger or TnT tagger. The overall accuracies are given in Figure 4. Note that the first baseline is equal to the TnT overall accuracy.

There are roughly 5% cases in which both taggers assign an incorrect tag. By using the technique discussed in this paper (i.e. rule inference), it would be possible to learn under which conditions the two taggers are both mistaken and thus alert the user about such tags.

Furthermore, it would be possible to apply our technique on a per-attribute basis. We would be able to predict incomplete tags, i.e. tags with some attributes missing, where the missing attributes would be those most likely predicted falsely by both taggers. This would be very useful as guidance for human taggers preparing the JOS corpus. The missing attributes would have to be entered manually; the rest would only need to be validated.

Tagging on a per-attribute basis and looking at cases in which both taggers predict an incorrect tag will be the focus of our future research. In addition, we will consider including more taggers into the system. The main idea is to develop taggers, specialized to handle cases in which the two currently used taggers are not successful.

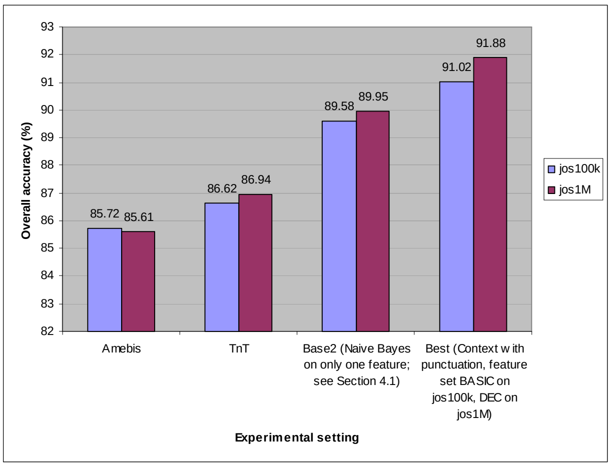


Figure 4: The overall accuracies (%). We can see that our meta-tagger exhibits around 4%–5% overall improvement over the two underlying taggers (i.e. TnT and Amebis). For computing the accuracies on the jos1M corpus, we needed to estimate the number of cases where the two taggers agreed on a correct tag. Looking at the statistics of the jos100k corpus (Table 1), we can see that the taggers are correct in 96.4% of the cases where they agree on the tag. Therefore, we computed the required number as 96.4% of 809,897 which is 780,740.71.

Acknowledgements

The work described in this paper was supported in part by grant ARRS J2-9180 "Jezikoslovno označevanje slovenskega jezika: metode in viri" and EU 6FP-033917 SMART "Statistical Multilingual Analysis for Retrieval and Translation".

References

[1] Arhar, Š. and Gorjanc, V. (2007). Korpus FidaPLUS: nova generacija slovenskega referenčnega korpusa. Jezik in slovstvo, 52(2): 95–110.

[2] Brants T. (2000). TnT – A Statistical Part-of-Speech Tagger. In Proceedings of the Sixth Applied Natural Language Processing Conference ANLP-2000, 224–231.

[3] Clark, P. and Niblett, T. (1989). The CN2 Induction Algorithm. Machine Learning, 3(4): 261–283.

[4] Demšar J., Zupan B. and Leban G. (2004). Orange: From Experimental Machine Learning to Interactive Data Mining. White Paper (www.ailab.si/orange), Faculty of Computer and Information Science, University of Ljubljana.

[5] Erjavec, T., Džeroski, S. and Zavrel, J. (2000). Morphosyntactic Tagging of Slovene: Evaluating PoS Taggers and Tagsets. In Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'2000). ELRA, Paris.

[6] Erjavec, T. (2004). MULTEXT-East Version 3: Multilingual Morphosyntactic Specifications, Lexicons and Corpora. In Proceedings of the Fourth International Conference on Language Resources and Evaluation, LREC 2004, 1535–1538.

[7] Erjavec, T. and Krek, S. (2008). The JOS morphosyntactically tagged corpus of Slovene. In Proceedings of the Sixth International Conference on Language Resources and Evaluation, LREC 2008.

[8] Giménez, J. and Márquez, L. (2004). SVMTool: A General POS Tagger Generator Based on Support Vector Machines. In Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04).

[9] Hajič, J. and Hladka, B. (1998). Tagging Inflective Languages: Prediction of Morphological Categories for a Rich, Structured Tagset. COLING-ACL'98. ACL.

[10] Quinlan, J.R. (1993). C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, Inc.

[11] Sjöbergh, J. (2003). Combining POS-taggers for improved accuracy on Swedish text. In NoDaLiDa 2003, 14th Nordic Conference on Computational Linguistics. Reykjavik.

[12] Spoustová, D., Hajič, J., Votrubec, J., Krbec, P. and Květoň, P. (2007). The Best of Two Worlds: Cooperation of Statistical and Rule-Based Taggers for Czech. Proceedings of the Workshop on Balto-Slavonic Natural Language Processing. June 2007. Prague, Czech Republic. Association for Computational Linguistics.

[13] Zribi, C.B.O., Torjmen, A. and Ahmed, M.B. (2006). An Efficient Multi-agent System Combining POS-Taggers for Arabic Texts. In Computational Linguistics and Intelligent Text Processing. LNCS Volume 3878/2006, Springer.

A Study of Analysing IT Digital Coping Strategies

Yao-Ming Chu, Li-Ling Hsu and Jung-Tsung Yang
Department of Industrial Technology Education, NKNU
No.62, Shenjhong Rd., Yanchao Township, Kaohsiung County 824, Taiwan
E-mail: lingo@mail2.cy.edu.tw, googo.hr@gmail.com

Keywords: digital divide, SWOT analysis, information technology (IT), information literacy

Received: July 8, 2008

In the present study, we explore the analyses of coping strategy for IT digital divide on the elementary campus. The questionnaire was distributed by stratified random sampling. The cases included 39 schools and 150 teachers. The study adopted the questionnaire of "SWOT Strategic Analysis for Digital Divide" to investigate the digital learning environment on campus and to understand teachers' opinions on reducing the digital divide. The main finding after analyzing the score statistically are as follows:

1. The emphasis degree of the principle's promotion for digital learning, high – 38.7%, medium—55.3%, low – 6%.

2. The rate of schools that had computers in every class was 48.7%, network 76%, computers and network 46%.

3. Information-credit-taken teachers and qualified teachers in information subject: teachers from urban schools occupy the highest rate; teachers from general schools the second, and teachers from remote schools get the lowest rate.

4. Most of the teachers agreed on reducing digital divide on campus.

Povzetek: Prispevek analizira rezultate analize anketiranja 150 učiteljev na temo digitalne ločnice.

1 Introduction

The growth of the digital environment is double-edged. On the one hand it integrates information and on the other hand it creates the secret worry of the digital divide. In recent years, much attention has been paid to the IT digital divide internationally. The possible impacts on society have been discussed widely and it has been concluded that generally the digital divide is an obstacle to the development of civilisation. Therefore, the termination of the status quo for unbalanced IT development between regions, groups, and individuals has become a common consensus.

Following land, labour, and capital, information has become an important factor in this knowledge economy era. In the information society, those who can promptly master and gather information will be competitive [5]. How though, will the development of information technology (IT) influence the social equality of wealth and justice? The optimistic view holds that the use of IT benefits the balance of accessing information, as people can accumulate resources and promote their situation by accessing IT to obtain important information. However, the pessimistic view holds that IT enlarges the inequality between the rich and the poor; thus the rich become richer and the poor become poorer. To create opportunities for social equality and fair competition, the most efficient method is to provide a fair IT educational environment and advocate the universality of accessing IT to reduce the digital divide and promote circulation on different levels.

USA was the first country that systematically observed the digital divide. The National Telecommunications and Information Administration (NTIA) in the USA have continuously delivered reports into the digital divide (Falling through the Net) since 1997[12]. It was discovered that opportunities to access IT were differentiated, according to people's income, race, educational background, and region of residence. The differences seem to be widening. Recently most countries have also discovered that the problems of the digital divide may create a wall for disadvantaged minorities from attending social activities. To protect and promote fair information access opportunities and social justice, digital divide on campus is an important issue that needs to be carefully considered.

2 Review of literature

IT is defined as any computer-based tool that people use to interpret information and carry out the information processing needs of an organization [6]. IT has paved the way for an information society sans frontiers to have easy access to information and communication, also connects the machine environment with human applications, and has emerged as a force for global connectivity. Therefore there is a fair claim in the common statement, “IT has radically changed the lives of millions of people [19].” People that do not have IT access are in danger of exclusion from participation in the knowledge-based global economy. Because IT can directly contribute to human capabilities and support economic growth through the productivity gains that it generates.

2.1 Impacts of the digital divide on society

Does the digital create gaps or opportunities? Bill Clinton, the former American president, made a groundbreaking speech pointing out that the Internet removes barriers between countries and cultures to bring people closer together and create opportunities. However, it will be a tragedy if the use of the Internet creates new barriers because of its unavailability for some people where it was intended to remove those barriers.

Nevertheless, there is a disparity in the spread of IT across the world between the developed and the developing nations. There were 232 million Internet users in developed countries, as opposed to only 83 million Internet users in developing countries. There were 77 million registered online computers in the United States, 6 million in Japan, 5 million in Canada. In contrast, there were less than 10 registered online computers in Bangladesh, Angola, Chad, and Iraq and none in Burundi, Benin, and Syria. In terms of access to personal computers (PCs), there are 70 PCs for every 1000 people in the world. There are 3 PCs for every 10 people in developed countries, 7.5 PCs per 10,000 people in Sub-Saharan Africa, 2.9 PCs per 1000 people in South Asia, and 0.7 PCs out of 1000 people in Mali. There is an estimated 56Gbps bandwidth between the United States and Europe and 18Gbps of bandwidth between the United States and Asia. In contrast, there is only 0.2Gbps between Africa and Europe and 0.5Gbps between Africa and the United States [2].

Digital divide can be categorized as ([7],[15],[21]):

1. Global digital divide: This is the first divide wherein the Internet users account for only 6% of world population and 85% of them are in the developed countries where 90% of the Internet hosts are located.
2. Regional digital divide: Within Asia, the personal computer (PC) penetration is 0.58% in Indian (Asia is at 3.24% and world average is at 7.96%). The current Internet subscriber base is only 0.4% in Indian, in sharp contrast to Asian countries as Korea with 58, Malaysia with 11 and China with 2%.
3. National digital divide: Within nation, there is an urban–rural digital divide; within urban, there is educated–uneducated digital divide; amongst educated there is rich–poor digital divide.

The digital divide in education is built on disparities in investment in education as a whole. While European countries spend 6.77% of Gross National Product on education, South Asian and East Asian countries spend only 2.94% to 3.51%, respectively (Bridges.org). As a consequence of the global divide, students in poorer countries have less access to digital content, and lack competitiveness for participation in the knowledge-based global economy.

Generally speaking, there are two directions for reducing the digital divide: the positive one is to promote social justice; and the passive one is to avoid social instability caused by unbalanced access opportunities to IT.

Reference	Definition
NTIA（1997）	The gap of “Have” and “Have-Not” for IT, such as computer, internet and the ability of using them [12].
OECD（2001）	The differences presented by different social economic environments and different Internet access activities for each person, household, enterprise, and region [16].
Norris（2001）	Global divide as the divergence of ICT access between industrialized and developing nations. The social divide refers to the gap between information rich and information poor within the same nation. The democratic divide separates those people who use digital technology and information to participate in public life, and those who do not [15].
Wolff and MacKinnon（2002）	Inequalities exist in the degrees to which populations can access and use ICTs. This inequality is called the digital divide [22].
Siriginidi（2005）	Differences based on race, gender, geography, economic status and physical ability; in access to information, the Internet and other information technologies; in skills, knowledge and ability to use information and other technologies.
S.F. Tseng（2002）	The use and development of IT may be different because of gender, race, class and region of residence. Therefore, the opportunities to get access to IT are differentiated [20].

Table 1: Summary of researchers’ definitions of digital divide

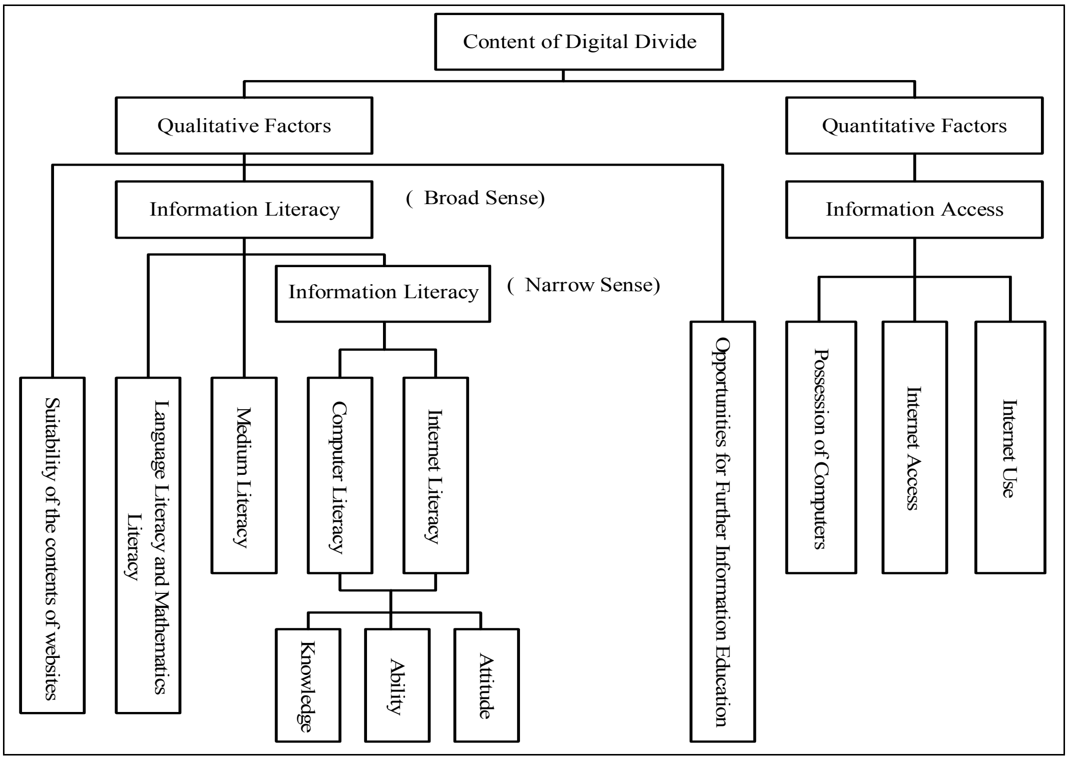


Figure 1: The content of digital divide

2.2 Definition and content of digital divide

The phrase 'Digital Divide' was first seen in the report Falling through the Net-New Information of Digital Divide in 1997 and Falling through the Net-The Definition of Digital Divide in 1999 delivered by the NTIA. The reports claim that information tools, such as computers and the Internet, have a crucial influence on individual economic achievements and career development in an information society. The PC ownership and the ability of using them will dominate the gaps between the rich and the poor.

There are two parts in the digital divide; the first is to analyse the different rates of people who have or access the Internet. The second is to investigate computer use to compare people's information literacy. There will also be differences when considering the digital divide if thinking is centered on people alone. The American Children Education Organisation claims that users may have problems accessing information on the Internet because of the lack of local network connections, information literacy, language obstacles, and cultural diversities when reading the websites. Furthermore, obstacles of accessing the Internet for residents in low income communities, the readability of website content, and the friendliness of surfing software are also possible factors that will influence the digital divide.

Researchers have amended the viewpoints provided by S.F.Tseng(2002), Siriginidi(2005), and integrated McClure's four factors of information literacy: traditional language and mathematics literacy, medium literacy, information literacy, and internet literacy, as well as the "suitability of the contents of websites" in a qualitative dimension (see Fig. 1).

Researchers have integrated the contents of the digital divide and defined it as: in a digital information society, there are differences in the opportunities of accessing IT, user ability, and the suitability of selecting the contents according to individual social attributes, such as differences in gender, race, household income, class, and region of residence. The opportunities for accessing IT include the possession of computer equipment, opportunities for internet connection, and the conditions of using the internet, etc. The abilities required for using IT include information literacy and information technique literacy.

2.3 Influential factors on the digital divide on campus

Researchers have concluded from relevant documents that the main factors that influence the digital divide on campus are race, geographical region, personal factors (gender, age), family factors (the education degree of the parents, household income, profession, social and economic position, household IT equipment, and attendance and recognition of householders), school factors (IT equipment, maintenance, quality of information education, internet quality, emphasis degree), teacher factors (teachers' information ability, learning attitude, and opportunities of education training), and communities and government factors (policy and internet cost), etc. Among those factors, the region of residence, gender, age, educational degree of the parents, household income, social and economic position of householders, household IT equipment, and school IT environment are considered to be most significant.

To sum up, the differences of different regions of residence, individual family factors, school equipment, and the teacher's literacy will create a digital divide. These are all included in the questionnaire for this study as research variables.

2.4 Theory of SWOT strategy

A derivative of the Harvard policy model, also referred to as the "design school model" [11], the SWOT approach seeks to address the question of strategy formation from a two-fold perspective: from an external appraisal (of threats and opportunities in an environment) and from an internal appraisal (of strengths and weaknesses in an organization). SWOT generates lists, or inventories, of strengths, weaknesses, opportunities, and threats. Organizations use these inventories to generate strategies that fit their particular anticipated situation, their capabilities and objectives ([1],[3],[17]).

The actions to be undertaken that can be deduced from the four elements of SWOT are: building on strengths, eliminating weaknesses, exploiting opportunities, and mitigating the effect of threats [4].

The major analysis tool for strategy management is a SWOT analysis. This also applies to school organisations. In a changing society, how to look for and identify the strengths and weaknesses, and how to examine the external environmental opportunities and threats, are questions worthy of investigating, and also an important basis for a strategy approach to solving IT digital divides on campus.

3 Method

3.1 Research structure

Figure 2 is the structure of this study. According to the documents, this study uses two directional predictors, including population variables (age, teaching seniority, teacher classification, IT-relevant experience, etc.), and e-learning environment variables (region of schools, school classification, school scale, IT equipment, and emphasis of the degree of the principle's promotion of digital learning). The dependent variables include four SWOT strategic directions for the digital divide: strengths, weaknesses, opportunities, and threats, which include four strategic analyses: school administration, teacher, IT equipment, and government policy.

3.2 Subject

The research subjects for this study are teachers from public elementary schools in Kaohsiung City and Kaohsiung County in Taiwan. The questionnaire was distributed by a stratified random sampling technique. The teacher population was classified according to the location and scale (large, medium, and small) of their schools. Data were randomly selected at a rate proportionate to the amount of schools in the same group. There were one administrative teacher, one computer teacher, and two tutoring teachers in the class filling in the questionnaire in each sampling school. The sample size was 39 schools and 156 teachers, in which 152 questionnaires were retrieved. Among them, 150 were valid and the validity rate of retrieved questionnaires was 96%. Table 2 and 3 show the demographic characteristics of the samples and the current climate of digital learning

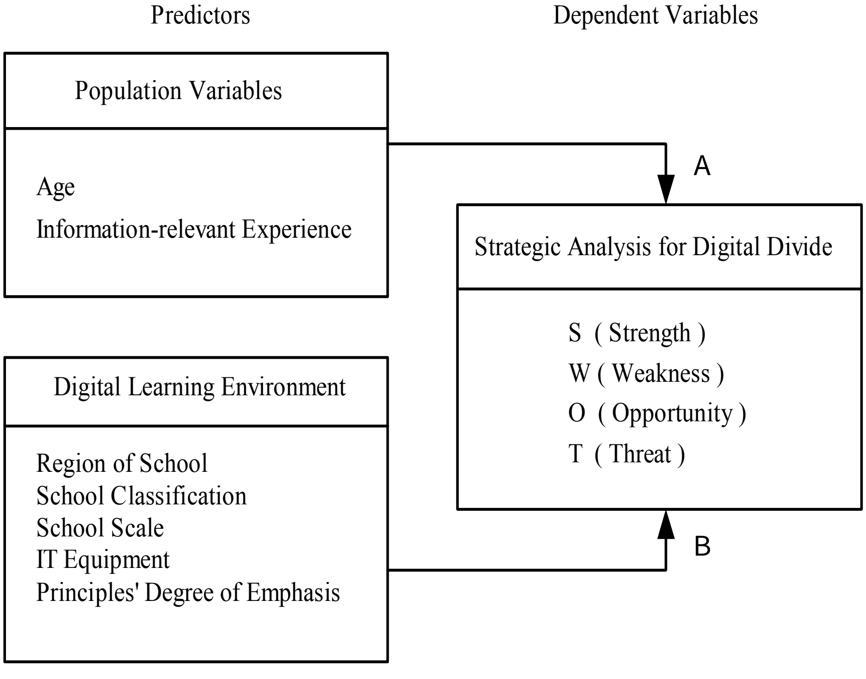


Figure 2: The structure of this study

3.3 Instrument reliability and validity

The questionnaire, “SWOT Strategic Analysis for Digital Divide” was designed for this study. There are four subscales in this questionnaire, including internal strengths (15 questions), internal weaknesses (16 questions), external opportunities (16 questions), and external threats (14 questions) of the digital divide on campus. Likert’s 5-point-scale was adopted in this study (1=strongly disagree; 5=strongly agree).

To increase the external validity and internal validity of this questionnaire, five academic experts and three teachers from elementary schools were invited to review the content. Furthermore, 65 valid questionnaires were chosen randomly for pre-testing. In the ‘Item Analysis’, the results of the third question and the thirteenth one at the ‘Threat’ part were deleted because their CR value did not reach the level of significance. The internal consistency (α) for each scale are: ‘Strengths’.91, ‘Weaknesses’.83, ‘Opportunities’.84, ‘Threatens’.84. Furthermore, the internal relationship between each part was significant. The correlation degrees with amounts were: .773, .791, .854, and .850.

4 Results

4.1 Analysis of teachers’ recognition towards the strategic analysis for the digital divide on campus

Teachers’ Recognition in this questionnaire is shown in Table 4. The mean was 4.12 which is closer to the degree of ‘strongly agree’.

Variables	Number of Subjects / Percentage											
Age	Under 30 years old 44			31-40 years old 74			41-50 years old 28			Over 50 years old 4		
Information Relevant Experiences	Attended information workshops at the school			Attended information workshops outside the school			Taken information credit courses at universities or colleges, including distance learning			Qualified in information subjects		
	33			65			37			15		
Vs School Classification (percentage)	urban 12.7	general 6	remote 3.3	urban 18.7	general 10.7	remote 14	urban 12	general 8.7	remote 4	urban 6.7	general 2.7	remote 0.7

Table 2: Summary of the demographic characteristics in the samples

Variables	Number of Subjects / Percentage				
Region of School	Kaohsiung City 59		Kaohsiung County 91		
School Classification	Urban school 75	General school (township) 42	Remote school 33		
School Scale	Under 6 classes 30	7-18 classes 22	19-36 classes 36	37-60 classes 26	Over 61 classes 36
Principles’ Degree of Emphasis	High 58 (38.7%)	Medium 83 (55.3%)	Low 9 (6%)		
Having Computers in each Class	yes 73（48.7%）		no 77（51.3%）		
Having Computers and the Internet in each Class	yes 69（46%）		no 81（54%）		

Table 3: Analysis of the current digital learning environment

Part	M	SD	Min	Max	N of questions	Mean of each question
Strength	64.25	6.45	44	75	15	4.28
Weakness	60.23	7.17	45	78	16	3.76
Opportunity	68.89	6.91	51	80	16	4.30
Threat	49.93	5.54	36	60	12	4.16
Four Parts					59	4.12

Table 4: Summary of teachers' recognition in the strategic analysis for the digital divide

At the 'Strength' part, data were ranked according to teachers' degree of recognition as shown in Table 5. In Question S01, the mean was 4.46. It shows that teachers expressed a highly positive attitude towards the promotion of students' information literacy and the reduction of the digital divide, if schools perform information education. The mean was 4.45 in S02 and S06. It shows that teachers held a highly positive attitude towards information education, teachers' further education in IT, and that using the internet as a teaching assistance tool will promote digital learning at school. (S01: Schools should perform information education to promote students' information literacy. S02: Schools should encourage teachers' further education of information ability. S06: Schools should equip the Internet in each class as a teaching assistance.)

Generally speaking, the recognition mean of the question analysis was between 4.01 and 4.46. It shows that subjects agreed on reducing the digital divide on campus at the 'Strength' part.

Question	S01	S02	S06	S05	S03	S07	S14	S10	S15	S04	S12	S11	S08	S13	S09
M	4.46	4.45	4.45	4.40	4.38	4.37	4.30	4.28	4.25	4.23	4.21	4.19	4.18	4.11	4.01
Rank	1	2	2	3	4	5	6	7	8	9	10	11	12	13	14

Table 5: Summary of teachers' recognition in the strategic analysis for the digital divide at the 'Strength' part

At the 'Weakness' part, data were ranked according to the teachers' degree of recognition as shown in Table 6. The highest mean, 4.44, was for Question W06. The second highest was for Question W07. This shows that there should be complete IT equipment available if a good digital learning environment is to be established, and government should constantly budget for subsidies to maintain justice for digital learning. Therefore, there would not be serious digital divide caused by the different locations of schools, or the different social and economic backgrounds of students. (W06: Insufficient financial support for IT equipment maintenance will influence digital learning on campus. W07: Insufficient quantities of computers and computer classrooms will influence digital learning on campus.)

Question	W06	W07	W04	W14	W02	W05	W08	W16	W11	W12	W13	W09	W15	W03	W10	W01
M	4.44	4.13	3.89	3.85	3.79	3.78	3.78	3.74	3.73	3.71	3.71	3.69	3.69	3.51	3.49	3.29
Rank	1	2	3	4	5	6	6	7	8	9	9	10	10	11	12	13

Table 6: Summary of teachers' recognition in the strategic analysis for the digital divide at the 'Weakness' part

At the 'Opportunity' part, data were ranked according to the degree of teachers' recognition as shown in Table 7. The mean for Question O05 was the highest at 4.56. Teachers considered that internet use should be more popular; however, current national internet use is still expensive. This provides a burden for students with lower household incomes. In order to achieve a fair access to IT, market competition should be promoted to reduce the internet price. The mean for O07 is 4.49. It shows that teachers recognise that there should be diverse digital teaching software for the diversification of teaching as an assistance tool for teaching. The planning of a digital book reservation library would provide this service. The mean for O06 was 4.48, and the mean for O04 was 4.47. It shows that teachers consider that an increase of IT equipment and accessing opportunities to internet resources would promote students' digital learning skills and knowledge. The mean for O02 was 3.83. This means that teachers considered that the sufficiency of IT resources on campus was more important than subsidies for individuals and families. (O05: Promote market competition to reduce the price of the internet. O07: Establish a digital book reservation library to provide software for teachers and students. O06: Provide internet equipment in public libraries. O04: Reduce IT equipment prices to increase opportunities for accessing IT. O02: The government subsidises the expense of accessing the internet and computers for those of low household income.)

Question	O05	O07	O06	O04	O08	O14	O13	O11	O03	O09	O10	O16	O01	O12	O15	O02
M	4.56	4.49	4.48	4.47	4.46	4.39	4.37	4.34	4.29	4.29	4.28	4.27	4.26	4.08	4.03	3.83
Rank	1	2	3	4	5	6	7	8	9	9	10	11	12	13	14	15

Table 7: Summary of teachers' recognition in the strategic analysis for the digital divide at the 'Opportunity' part.

At the 'Threat' part, data were ranked according to the degree of teachers' recognition, as shown in Table 8. The mean for Question T01 was 4.47 and the mean for T02 was 4.30. This shows that in the teachers' opinion, the social and economic status of households and the educational degree of the householders are influential factors on the digital divide. The mean for T05 was 4.33, and the mean for T06 was 4.32. Teachers considered that there will be a digital divide between schools because of different school scales and different resources for IT equipment and its maintenance. The mean for T03 was 4.21, and the mean for T09 was 4.19. In 1998, each school was subsidised for at least one IT classroom by the government, however, the equipment is no longer usable and maintenance is unavailable. A lack of subsidies from the government for renewing IT equipment has created a digital divide because of the different resources available for schools. (T01: Students from families of a lower social and economic status will have less experience in accessing IT. T02: Students with parents of a lower education background will have less information literacy. T05: IT equipment is different according to the school scales.T06: Budgets for maintenance are different according to the school scales. T03: Government funds are not sufficient for constantly subsidising hardware, software, and maintenance. T09: Though the government emphasises the digital learning policy, budgets for promoting the policy are not sufficient.)

Question	T01	T05	T06	T02	T03	T09	T10	T04	T11	T12	T07	T08
M	4.47	4.33	4.32	4.30	4.21	4.19	4.17	4.14	3.98	3.96	3.94	3.92
Rank	1	2	3	4	5	6	7	8	9	10	11	12

Table 8: Summary of teachers' recognition in the strategic analysis for the digital divide at the 'Threat' part

5 Discussion and conclusions

5.1 Discovery of this study

1.The degree of emphasis of the principles' promotion of digital learning: medium degree of emphasis occupies the highest rate while a lower degree of emphasis gets the lowest rate.

2.Half of the subjects' schools have computers in each class; less than half have computers and internet access.

3.Information-credit-taken teachers and qualified teachers in information subject: teachers from urban schools occupy the highest rate; teachers from general schools the second, and teachers from remote schools get the lowest rate.

4.The teachers' degree of recognition is high on the four parts of the SWOT strategic analysis for the digital divide: strength, weakness, opportunity, and threat.

5.2 Strategic analysis of reducing the digital divide on campus at school administration part

1. Schools should promote information education to advance students' information literacy.
2. Schools should promote teachers' information literacy and encourage teachers' further information education.
3. Schools should provide students with computer use in their free time.
4. Schools should make systematic IT promotion plans to provide research opportunities for teachers to advance their abilities of using IT.
5. Schools should have a complete plan for their IT equipment to provide more opportunities for accessing information for both students and parents.
6. It relies on the government to solve the problems of changing the IT equipment and maintaining it.

5.3 Strategic analysis of reducing the digital divide on the campus at school teachers part

1. Teachers should have the ability of integrating IT into teaching and encourage students to learn with internet resources to promote students' positive attitude.
2. Schools should have sufficient professional IT teachers for proper management of IT equipment and promotion of information education.
3. Too much workload for IT teachers and the circulation of teachers to remote schools will influence the promotion of IT education

5.4 Strategic analysis for reducing the digital divide on campus at school IT equipment part

1. Schools should equip computers and internet access in each class for teachers to integrate IT into teaching.
2. It should be emphasised that differences in IT equipment and the construction of broadband between schools will severely influence opportunities of digital access on campus.
3. The government should subsidise IT equipment for schools and teachers, to promote teaching integrated with IT.
4. Public libraries should provide internet access to increase IT access opportunities for students from disadvantaged backgrounds.

5.5 Strategic analysis for reducing the digital divide on campus at government policy part

1. The government should allocate a budget for improving the digital environment on campus to subsidise software, hardware, and maintenance.
2. Effects of reducing the digital divide between students are limited because of insufficient hours for computer courses.
3. The Ministry of Education should establish a resources exchange centre to provide hardware and software for learning resources, and online resources as a channel for sharing experiences.
4. The government should promote market competition to reduce the prices of the IT equipment and the Internet, to promote opportunities to access IT.
5. The government should encourage universities, colleges, and civil organisations to attend IT education to train local IT technical staff in remote areas.

6 Suggestions

This study should have enlarged the sampling area and increase the sampled nations to expand the suitability of the research. However, due to the financial and time constraints of this study, this was not possible.

The penetration rate of computers in Taiwan is among the highest in the world because of its special geographical environment and IT resources. However, there is still a digital divide on campus, which calls for further research on this issue. It would also be hoped that the questionnaire results in this study might serve as a research guide for future researchers.

References

[1] Bourgeois (1996) L.J. III, Strategic Management: From Concept to Implementation, The Dryden Press, Fort Worth.

[2] Bridges.org. （2001） Spanning the digital divide: understanding and tackling the issues. Retrieved July 12, 2003 from the Bridges.org website: http://www.bridges.org/spanning/report.html .

[3] David, F. (1997) Strategic Management, sixth ed. Prentice-Hall, Upper Saddle River, NJ.

[4] Dealtry, T.R.(1992) "Dynamic SWOT analysis", Developer´s Guide, Dynamic SWOT Associates, Birmingham.

[5] Drucker, P. F. (2002) Managing in the next society. St. Martin Press.

[6] Haag, Cummings, & Dawkins (1998) Haag S., Cummings M., Dawkins J., Management information systems for the information age, McGraw-Hill, Boston.

[7] ITU （2003） World telecommunication indicators. Available from: http://www.itu.int/ITU-D/ict/statistics/at_glance/Internet00.pdf

[8] Jayajit Chakraborty and M. Martin Bosman （2002） Race and Society, Volume 5, Issue 2, Pages 163-177

[9] Ma. Mercedes T. Rodrigo （2005） International Journal of Educational Development, Volume 25, Issue 1, Pages 53-68

[10] [10] McClure, C. R. (1994) Network Literacy: A Role for Libraries Information Technology and Libraries.13 （2） ,p116-125.

[11] Mintzberg, H. （1994） The Rise and Fall of Strategic Planning, Prentice-Hall, New York.

[12] National Telecommunications and Information Administration （1997） “Falling Through the Net II: New Data on the Digital Divide.” Retrieved February 7, 2004, from the WWW： http://-www.ntia.doc.gov/ntiahome/net2/falling.html

[13] National Telecommunications and Information Administration （1999） “Falling Through the Net: Defining the digital divide”. Retrieved February 12, 2004, from the World Wide Web： http://-www.ntia.doc.gov/ntiahome/fttn99/contents.html

[14] National Telecommunications and Information Administration （2002） “A Nation Online: How Americans Are Expanding Their Use of the Internet,” Retrieved February 27,2004,from the World Wide Web： http://www.ntia.doc.gov/-ntiahome/dn/html/anationonline2.html

[15] Norris, P. （2001） Digital divide? Civic engagement, information poverty & Internet in

democratic societies. Retrieved July 11, 2003 from the Harvard University's John F. Kennedy School of Government web site: http://ksghome.harvard.edu/~.pnorris.shorenstein.ksg/Book1.htm

[16] OECD（2001）Closing the gap：Securing benefits for all from education and training,in“Education Policy Analysis”.

[17] Pearce, J.A. II, Robinson R.B. Jr.（1997） Strategic Management. Formulation, Implementation, and Control, sixth ed. Irwin, Chicago.

[18] Robert Kozma, Ray McGhee, Edys Quellmalz and Dan Zalles（2004）International Journal of Educational Development, Volume 24, Issue 4, Pages 361-381

[19] Shivanthi（2004）The International Information & Library Review, Volume 36, Issue 4, Pages 319-327

[20] S.F. Tseng（2002）Research of the digital divide in Taiwan. Research, Development and Evaluation Commission, Executive Yuan.

[21] Telecom Regulatory Authority of India（2004） Telecom Regulatory Authority of India, 2004. Consultation paper on growth of telecom services in rural India: The way forward, Paper No. 16/2004. Available from: http://www.trai.gov.in/27octcon.htm

[22] Wolff and MacKinnon（2002）Wolff, L., MacKinnon, S., 2002. What is the digital divide? Retrieved July 12, 2003 from the TechKnowLogia.org web site: http://www.techknowlogia.org/

[23] Yifei Sun and Hongyang Wang（2005）Journal of Rural Studies, Volume 21, Issue 2, Pages 247-258

Appendix

Questionnaire of the “SWOT Strategic Analysis for Digital Divide”

1. Strength, (S)

Please express your opinion about the ‘Strength’ of promoting digital learning at school and reducing the digital divide

S01: Schools should perform information education to promote students’ information literacy.

S02: Schools should encourage teachers’ further education of information ability.

S03: Teachers should encourage students to use the Internet in their learning.

S04: Schools should have enough human and financial resources for IT equipment maintenance.

S05: Schools should equip computers in each class to integrate IT into teaching.

S06: Schools should equip the Internet in each class as a teaching assistance.

S07: Schools should have sufficient professional IT teachers to promote IT education.

S08: Administrative staff should emphasise the promotion of digital learning at school.

S09: Students have computers and access to the Internet at home.

S10: Schools should continue promoting the training of teachers’ information literacy.

S11: Teachers should have the ability to integrate IT into teaching.

S12: Schools should integrate IT into teaching to educate students’ positive attitude toward computer learning.

S13: Schools should provide circulated teaching software.

S14: Teachers should teach students to pay attention to information moral principles.

S15: Teachers should encourage students’ cooperation to promote their IT ability.

2. Weakness, (W)

Please express your opinion about the ‘Weakness’ in promoting digital learning at your school and reducing the digital divide

W01: Lack of firewall and antivirus software on schools’ internet systems will affect internet teaching quality.

W02: Teachers have fewer opportunities for further education in information at remote schools.

W03: The circulation of teachers will influence schools' promotion of information education.

W04: Insufficient numbers of qualified teachers in information-related subjects will affect IT education at school.

W05: Internet speed at school will affect teaching quality.

W06: Insufficient financial support for IT equipment maintenance will influence digital learning on campus.

W07: Insufficient quantities of computers and computer classrooms will influence digital learning on campus.

W08: Schools' not able to provide access to computers for students in their free time will affect students' opportunities for IT access.

W09: Insufficient opportunities for teachers' computer-related research and study will affect their ability of digital teaching.

W10: Elder teachers will find it difficult to cooperate in digital learning.

W11: Insufficient quantity of computer-assisted teaching software will affect the digital learning.

W12: A few hours of computer courses will limit the effect of reducing the digital divide between students.

W13: The lack of a systematic plan will mean difficulties for advocating teachers' abilities of using IT.

W14: Insufficient numbers of professional IT teachers will create difficulties in the ability of using IT.

W15: Lack of a proper plan for IT equipment will affect the outcome of teaching assistance.

W16: Lack of professional IT management will affect the digital learning environment on campus.

3. Opportunity, (O)

Please express your opinion about the 'Opportunities' of promoting digital learning at your school and reducing the digital divide

O01: Training local IT professionals for remote schools will promote digital learning.

O02: The government subsidise the expense of accessing the internet and computers for low household income.

O03: The government supports communal computer learning courses to establish lifelong learning channels.

O04: Reduce IT equipment prices to increase opportunities for accessing IT.

O05: Promote market competition to reduce the price of the internet.

O06: Provide internet equipment in public libraries.

O07: Establish a digital book reservation library to provide software for teachers and students.

O08: Establish a resources database in public libraries to support school teaching.

O09: Adopt a fan-shaped mode in teacher training of IT ability. Train seeded teachers first and they will promote IT ability.

O10: Hold practical workshops to help teachers constantly update with new knowledge and promote their ideas of teaching.

O11: The Ministry of Education should establish a resource exchange centre to provide software, hardware and internet learning resources for experience sharing.

O12: The 'Learning Fueling Station' at the Ministry of Education can promote digital learning.

O13: Encourage teachers to buy laptops with governmental subsidies to help integrate IT into teaching.

O14: Buy overhead projectors for each class-group with governmental subsidies to promote teaching integrated with IT.

O15: Develop the characteristics of seeded schools and they will promote communal schools.

O16: Encourage universities, colleges, and non-governmental organisations to attend IT education courses to encourage its promotion.

4. Threat, (T)

Please express your opinion about the 'Threats' in promoting digital learning at your school and reducing the digital divide

T01: Students from families of a lower social and economic status will have less experience in accessing IT.

T02: Students' with parents of a lower education background will have less information literacy.

T03: Government funds are not sufficient for constantly subsidising hardware, software, and maintenance.

T04: Maintenance factories for IT equipment are not sufficient in remote areas.

T05: IT equipment is different according to the school scales.

T06: Budgets for maintenance are different according to the school scales.

T07: The Ministry of Education does not integrate indexical courses according to students' information learning ability in different phases.

T08: Education units do not provide various channels for further education for teacher training in teaching and combining IT.

T09: Though the government emphasises the digital learning policy, budgets for promoting the policiy are not sufficient.

T10: A Lack of widespread broadband construction influences the development of digital learning.

T11: IT seeded teachers are overloaded, as their teaching hours are not reduced.

T12: Current practice of IT education is based on the mode which operates from the higher level to the lower level. However, another mode should be considered which operates from the lower to the higher level.

Piecemeal Journey to 'HALCYON' World of Pervasive Computing: From Past Progress to Future Challenges

Rolly Seth
Amity University, Lucknow, India,
E-mail: rolly.seth@gmail.com

Rishi Kapoor
Indian Institute of Information technology, Allahabad, India,
E-mail: rkapoor.rishi@gmail.com

Hameed Al-Qaheri
College of Business Administration, Kuwait University,
E-mail: alqaheri@cba.edu.kw

Sugata Sanyal
Tata Institute of Fundamental Research, India,
E-mail: sanyal@tifr.res.in

Keywords: halcyon model, pervasive computing, nonchalant, connected, hidden, calm technology, privacy.

Received: October 1, 2009

Although 'Halcyon' means serene environment which pervasive computing aims at, we have tried to present a different interpretation of this word. Through our approach, we look at it in context of achieving future 'calm technology'. The paper gives a general overview of the state of pervasive computing today, proposes the 'HALCYON Model' and outlines the 'social' challenges faced by system designers.

Povzetek: Prispevek podaja pregled računalništva po meri ljudi in posebej model HALYCON.

1 Introduction

All the technological advancements in the world are the outcome of chimera of a handful that was seen much before. Technology has progressed by leaps and bounds in the past few years. As in the dream of Mark Weiser, 'Ubiquitous computing' [1] has started laying its foundation in today's real world, the future with pervasive computing assumes a pastoral world where an individual will not be distracted by plethora of information surrounding him. This paper encompasses our representation of this 'HALCYON' world and how will its piecemeal journey progress.

By saying the future world to be 'Halcyon', we mean a peaceful technology world, where although at any given instant of time, plenty of information will surround an individual, yet it will sit calmly without interfering with one's work or trying to grab its attention. A person unaware of its presence will perform daily tasks without the need to carry several devices wherever the person goes. Information will always be at his side and will come to his aid as and when the situation demands, like a 'guardian angel'. This might sound confusing as surfeit information always distracts users, but this is the charm of pervasive computing "Everywhere, Anytime yet ulterior".

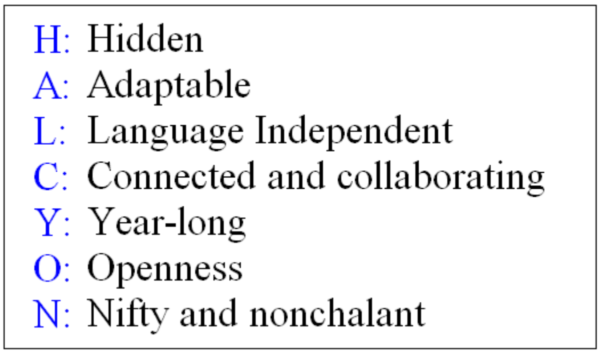
To realize this, it requires a shift in view from a sender's need to that of the receiver's. For example, our mobile phone rings whenever the sender wants to talk to us, irrespective of our preoccupancy. Although we have the option of picking up the call or disconnecting it, yet it asks for our attention at that instant only and thereby interrupting us. Thus, this perception is far from the realization of the halcyon world. We require technologies like e-mail that will be present at the receiver's end but only be seen when the user is at ease and wants to use it.

This paper gives a general overview of the current state of pervasive computing in section 2, proposes the 'HALCYON Model' in sections 4 and 5 and outlines the 'social'challenges faced by system designers in section 6.

2 Background

The clear indication for pervasive world to come into existence is through the use of "asynchronous communication". The reason why today's world cannot be categorized into that of 'calm technology' [2] is that a sender usually requires immediate attention of the receiver. Both the sender and the receiver are required to be present and free at the same instant, which is not always the case. In 'Halcyon' world, information is only provided to the user when he wants it.

A vital consideration while developing such a world requires each device to perform varied functions corresponding to different environments. In other words, devices should be made ‘environment specific’ and not function specific. This is in contrast to today’s world where a user’s mind quickly maps to one particular device, being subjected to a particular requirement. For instance, to measure human temperature, one would look for a thermometer and so on. All this boils down to the use of ‘augmented functionality’, that is add-on functions to be performed apart from the basic ones.



H: Hidden
A: Adaptable
L: Language Independent
C: Connected and collaborating
Y: Year-long
O: Openness
N: Nifty and nonchalant

Figure 1: meaning of HALCYON

We conceptualize the meaning of ‘HALCYON’ as shown in Figure 1:

Hidden: The first and foremost requirement of a peaceful/halcyon world would be that devices and their working be concealed from the outer world. By this, we do not mean intrusiveness into one’s life. It would be of non-intruding type, thus safeguarding the privacy and safety of an individual. This is a major issue against fast pace progress towards a ‘calm technology’ world as to how such surplus amount of information will be kept hidden and administered so that it is not misused? For this to happen, the invisible front-end is to be supported by complicated back-end servers.

Adaptable: It should be easily adaptable to the environment or be context-aware. It should have the flexibility to re-configure itself. It governs the display properties of the information. For example, if someone calls the user and he is busy driving a car, then the environment will sense that and divert the call to another device, an answering machine or so. A premature example of it would be the popularity of call diversion to voice messages in today’s era, yet it requires more refinement to adapt to the bigger vision. It requires ‘diversification’, that is to choose a device from a set of many others that best fits the requirement instead of any predefined diversion route.

Language Independent: By language independent, we do not mean non-standardization of platforms. On the other hand, we mean that halcyon world should not be restricted to the use of only one mode of communication, which is through use of words. It should involve other modalities too. The environment will decide which one of them should be used according to the situation. For example, the use of visual representation and colour changes are the most common options used today as in datafountain [3], Auraorb [4] , flashbag [5] etc.

Connected and collaborating: This is the major role after being ‘hidden’ that devices have to perform. Physical distance would not be a barrier to connectivity. No doubt mobile phones offer this paradigm but for one to stay connected even at remote places, they need to answer a call and talk or reply via SMS etc. But, in contrast, in the HALCYON world, one does not need to disturb the other person in order to stay connected with him. Consider for instance, that the wall of our room changes colour according to our work and design patterns which depict our mood. A blue wall with flowers on it shows that we are in a cheerful mood. Some ambient interfaces like hello-wall, forecast umbrella and visual calendar have shown a promising step towards this development. Through connectivity, one could even be able to control the operations of their home appliances from remote places.

Connectivity would also require mutual collaboration between the devices. A message will be automatically transferred from one device to another without the use of human intervention. Consider one such example of an alarm clock, which apart from waking up an individual will also request the kitchen to make tea or breakfast for a user and kitchen in turn will look at user’s past history, interest, health report and other details to choose the best breakfast to be made. It will also check with the health report if the user is ill or not. If yes, then it will instruct back to the alarm clock so as not to ring and wake the user as he is not well and waits till he gets up. If his reports are fine then kitchen will prepare breakfast by collaborating with kitchen appliances by the time the individual gets ready to go to the office.

Year-long: Back–end servers will need to be run continuously, gathering 24X7 data from different inputs and using this data for future use. For continuous usage various constraints come in, like enormous power requirements for day/night running. One solution available for continuous use is the usage of RFID (radio frequency identification) [6] tag/reader used in wearable computing. Polaris [7] being one such example.

Openness: It means open access to data or material resources. Any one should be able to extract information quickly from the ‘open ocean’ of data and use it for own good. The biggest reason for Linux platform’s fast popularity was it being available as open source to one and all, where anyone can use and change to meet one’s requirements. For this to be possible a large ‘repository’ needs to be maintained. One such approach is PIE [8], (Personal Information Everywhere). It will be a boon in various ways. Consider the case of healthcare, where information stored in the past would be used to cure an individual at present and provide medications whenever required. Many wearable and ambient displays are made for healthcare purposes which show blood pressure, sugar level and according to past history, warn the user in case of any danger, but that is just the beginning. Openness of data will also help in removal of problems like traffic congestion and will automatically command the driver to take another route in case of traffic congestion on a certain route. The hindrance to this paradigm is obviously how to prevent the sensitive information from falling in wrong hands? Balancing between full and legitimate usage is an issue yet to be tackled with. Also storage of such surfeit information still remains a question and ‘Memex’ [9] being one such proposed concept. There are further serious limitations on memory usage.

Nifty and nonchalant: Although it would be very complex making such an environment, it must be easy to use and understand. It should naturally blend with the surroundings so that an individual does not feel disturbed by its presence. It must also be nifty to understand the changes in the surroundings and “dynamically” adapt to them. Some companies like Ericsson and Electrolux are developing an intelligent refrigerator [10] to detect shortage of any food item and automatically order it from the supplier. To some extent, intelligence has already been inbuilt into today’s washing machines that have the power to detect water level needed, soap requirements etc and automatically rinse the clothes and then switch off the power supply. One needs to develop such adroit machines that would sense a danger and then automatically work to revert it. Although natural disaster sensing machines have been developed, their functions do not extend beyond displaying warning signals. For a halcyon world to be developed, ‘integration and synchronization’ of many tasks should take place. Like a fire alarm on sensing fire in the house will automatically dial and call the fire control department and before the rescue team arrives, will try to extinguish the fire by itself by turning on the sprinkler valves.

3 Related work

Pervasive Computing is the envisioned fifth generation of computers which has evolved through a piecemeal journey of seventy years when the first generation of computers came into existence in 1940s. Since then each leap to the next level brought many drastic changes for computers. And so this is expected from this new wave of computing. Among profuse metamorphosis in various areas in each era, we would like to give a special mention to:

1->Size,
2->Memory (Data storage) Capacity,
3-> language for interaction and
4-> Input / Output Display

These became the benchmarks for assessing the technological advancements of each era.

The first yardstick of evaluating development had been reduction in size. The first generation computers used to occupy the whole room which UNIVAC and ENIAC belong to. Then up to the fourth generation, a number of alterations were done through the use of Integrated circuits. A large number of components of the computer could be placed on a single Silicon chip, the size of a finger tip. Moore's law still holding well, what could be prophesied is that the size would become so small that it will touch atomic encode information. Also the much awaited use of 'Nanotechnology' provides flexibility of doing things that could not be thought of earlier. The new promising 'Nokia Morph' [11] makes use of this nanotechnology to provide flexible, transparent, context aware mobile devices.

The second yardstick is information storage which has always been a concern with ever increasing requirements. Actual size of devices is decreasing, yet, there needs to be steady increase in the memory size. The initial concept in this field was the 'Memex', put forward by Vannevar Bush [12], about a large place to store and modify data as per user demand, which laid the foundation of continuous progress in this area. Initially, magnetic drums were used to quench the demands. Then came technologies like floppy disk, magnetic tape etc for secondary data storage. Presently, semiconductor memories are much more prevalent. However, this change is also an interim solution. In order to hold much more data again, nanotechnology is going to play an important role. However, several other software concepts are coming into picture, not only for storing more data but to store what we call 'Intelligent' data. 'Conceptnet3' [13] is one of such ideas which adds common sense knowledge to the data being stored. Thus, there has been a shift in functionality now, as to what else a memory could do apart from storing data, thereby trying to match it with human brain.

The third benchmark is the language required in order to interact with the computer and get its work done. The transition of language has been judged from the ease of use for the user in commanding the computer. Each generation has shown an increasing level of abstraction from earlier 0/1 binary machine code and mnemonics in assembly language to today's English like sentences used in High Level Language. Still the need for a 'natural language' is felt where one does not have to memorize the instruction set and rules to be followed while directing machine to perform a certain task. This is what has emerged as Artificial intelligence languages for fifth generation languages such as LISP and PROLOG [14]. LISP, due to its inbuilt structure takes less effort and less memory space to do a similar task as compared to ADA or any other language. However, it is still far apart to achieving a natural language tag which could be used for pervasive computing.

The last yardstick but of paramount nature in context to pervasive era is the use of Input /Output devices chosen for gathering data from environment and displaying information in the environment. The initial generations relied heavily on use of punched cards, printouts. Then came GUI (graphic user interface) and mouse for I/O. Mark Weiser described such devices namely Tabs, Pads and Boards in his Pervasive Computing vision. Some new era devices making use of it are Ambient Trolley [15], Aura Orbs, Infocanvas [16] etc. Nowadays, apart from taking the information from words, other modalities have also been explored. What is emerging as the unanimous choice for making pervasive I/O devices is the use of RFID (Radio Frequency Identification) where tags attached to different objects send radio frequency in all directions (each tag is associated with unique frequency) and accessed by readers who want it at that instant. It implements the pervasive concept that information will be floating around the user any time and he will access it whenever he feels the need of it. Another gadget being widely used nowadays is 'Phidgets' [17], [18] that can be attached to the USB port of a PC and acts as a building block for providing sensing and controlling operation from the PC. It is built with a powerful Application Programming Interface (API) which provides the user abstraction from the underlying working. It can sense various functions such as light, distance, humidity, motion, pressure, touch, voltage etc and can be interfaced from a wide list of platforms like Java, C, C++, MATLAB etc.

It is worth noticing here that with each wave of new technology more stress is given on visualization. The most recent archetype of it being Hello-wall [19], Data fountain and History table cloth [20]. These make use of daily life objects to display information. The most exciting point of these inventions is that the information is visible through them 24X7. Yet these technologies do not interfere with one’s daily routine to grasp their attention. Data fountain will display the real time comparative currency rates of different countries by means of the height that water level rises to, in each of the fountains (assigned to each currency). Similarly, Info Canvas provides information with the help of paintings; such as, number of people in the painting will represent the traffic conditions on the road; Sky colour of the painting will show weather condition in the real world etc.

As each era progresses from one to another, it marks the beginning of some new modalities. Pervasive Computing aims at not restricting itself to only one label. Any mode will be chosen from one of the several others that are available, as and when required. Thus, with today's technology it tries to explore every possible mode of providing information - be it voice, colour, area, picture or sound.

4 System overview

4.1 Pervasive computing: Other side

Although everybody expects Pervasive computing to be a boon for everyone, it has a graver side of it too. As they say ‘everything comes at a price and so it has’. In this paper, we also try to unfold this other side of the coin. The effects of pervasive computing can already be seen in the social life of us.

Primarily, artificial intelligence will result in automatically cutting down physical activity of an individual, who will now not get up and walk much, in order to perform the daily routine tasks. We assume that devices will be integrated to pass on a request from one to another by themselves. This will, in turn, have an effect on biological metabolism which requires certain daily dose of physical activities for its proper working. A computer animated movie ‘Wall-E’ (2008) also tries to portray this issue where people have become boneless as they do not even need to walk in order to get their work done (as some machine does it for you, right from brushing your teeth).

Huge amount of e-waste [21], [22] would be generated, and disposing it could become a problem.

Circumlocutory, the radiations from it will affect the human body resulting in various obscure diseases or defects.

Social interaction will subside as people will not feel the need to talk to one another to know their well being (as that information will be automatically sensed) or to get some work done. We human beings, as social animals will thereby have drastic changes in our behavior, who will then consider computers as better companions than Homo sapiens.

Dependency on computers would increase, as humans will feel helpless in performing any task without the availability of machines.

Indirect effect of decline in human-to-human interaction will increase aloofness and probably more chances of suicide rates as then physiological counseling of one (provided by talking to family, friends etc) will be less and surfeit of information will even guide different ways to commit suicide.

Risk rates can become enormously large as a small mistake in commanding a machine or interpretation by a device can lead to severe damage. This may result in a series of many others that are collaboratively working. Thus, a small aberration will take the scale of much bigger mishaps.

Last, but of paramount importance is the concern for privacy, as anyone and everyone’s sensitive data will be floating all around. An individual would be easily tracked or others would be able to see which activity you are currently involved in. Similar to the problem of e-waste, there will be issue of spamming. Our email in-boxes are already filled with those. It is not difficult to imagine what will happen when these seamless machines will be monitoring each of our activities.

Also consider for example project Aura [23]. Carnegie Mellon University conceived of this project as an example as a grand ubiquitous computing project, aiming to have a large scale computing system, demonstrating this concept. This one covers wearable, handheld and other type computers. Pervasive computing system keeps track of user location, behaviour and habits. It constantly tunes itself on basis of above data. Though this information is necessary for successful system deployment, it is a serious threat to user’s privacy. It is important to question our design principles of the “pervasive computing system” so as to strike a balance between the “invisible machine” and loss of user’s privacy.

4.2 Pervasive computing: Benefits

However, apart from all these shortcomings, following benefits will always be provided to one and all through pervasive computing: In Halcyon world we can imagine a world with easy access to information resulting in time saved to obtain that information. Consequently, travel costs required for getting the work done will decrease.

A transparent society [25] would be created where anyone will not be able to provide anyone with wrong information. Your family and friends will know about your well-being without disturbing you again and again by calling or messaging. A huge number of accidents could be dodged at the correct time thereby cutting down the enormous premature death rates due to mishaps. Traffic congestion problems could be avoided. Help will be at your doorstep as and when required to assist you in case of emergency.

One does not have to search for the appropriate information required for his use out of the plethora of information. Smart devices will automatically find it for you. Things will be linked together working collaboratively thereby eliminating the need to feed output of one machine to another for further processing. Their intelligent system will self regulate that.

5 Methodology proposed

Since the idea of pervasive computing was first proposed, much has been written on it but still there is a need for a single unified approach for integrating all pervasive computing. As can be seen in figure 2, the whole model is conceptualized by using a 2-channel grid technology. The two channel system will aid in duplex communication where one channel will allow anyone to components. Through this paper we will try to propose a 'HALCYON MODEL' in order to pace up the process of continuously transmit information and other one being for simultaneous reception. These two channels will cover the entire globe using grid system.

An individual will then send or obtain information from any of these channels without disturbing anyone in the process.

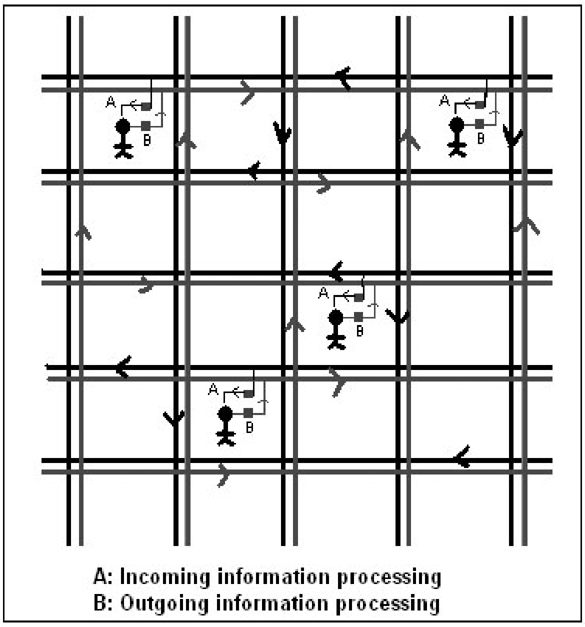

A
B
A
B
A
B
A
B
A: Incoming information processing
B: Outgoing information processing

Figure 2: HALCYON Model of pervasive world

The prime focus of this model is on the two small boxes labelled A and B in figure 2. These two processing subsystems will include all the main considerations required for calm technology world to come into existence. Figure 3 shows an enlarged version of box A which will come in between an individual and the entire grid of continuous flowing information. Whenever this individual will consciously or unconsciously need any information to his aid, he will extract information from this free flowing ocean. Before it is made accessible to him, it will be needed to be processed inside subsystem B to check whether he is an authorized user. As can be seen in figure 3, seven levels are to be passed after raw information is read from the grid. Let us see each one separately.

LEVEL I - SECURITY CHECK LEVEL: This level is to address the most important issue hindering the fast pace development of pervasive world. Here an individual identity will be checked if he is the legitimate user to have access to the specific information or not? One of the many approaches could be through the use of RFID tags/ readers system where each individual would be provided with a RFID tag. One must note that with the free flowing information in the grid, a list of authentic users also run along with it to check that the validity of an individual with that list. It will therefore provide privacy in a way that even though the information would be flowing around everyone, only appropriate users will be able to use it. If the user stands justified obtaining information, the information is directly sent to Level VI.

LEVEL II- MODALITY AND DEVICE DECISION ACCORDING TO URGENCY: The second level considers the fact that in HALCYON world, the things have to be context aware or adaptable to the environment. In order to make this possible flexibility needs to be provided regarding the selection of device to be used for providing the information. In this level firstly urgency level and environment is checked to know how quickly the data is to be provided to the user so that it will aid him at right time. If user is busy now and it is found inappropriate at the moment to provide the information to him, it is sent to a queue to be passed on to someone else. Otherwise, different modalities are chosen according to the priority levels like vision, sound, smell, feel etc. If the information is not important for the moment then it is also sent to a queue and stored for future use, so as not to disturb the user at the moment

It is worth mentioning here that it purely depends on the kind of information as to which path to be followed. It does not solely depend on urgency level sent. For example, fire at home (urgent) could be tackled by someone else if the user is currently busy in a meeting. But urgent information like a chance of accident with a truck while you are driving is to be provided at the same instant. Thus, you can see that although both pieces of information were urgent, yet their processing was different depending on user environment. If the information to be provided to the user is important at the

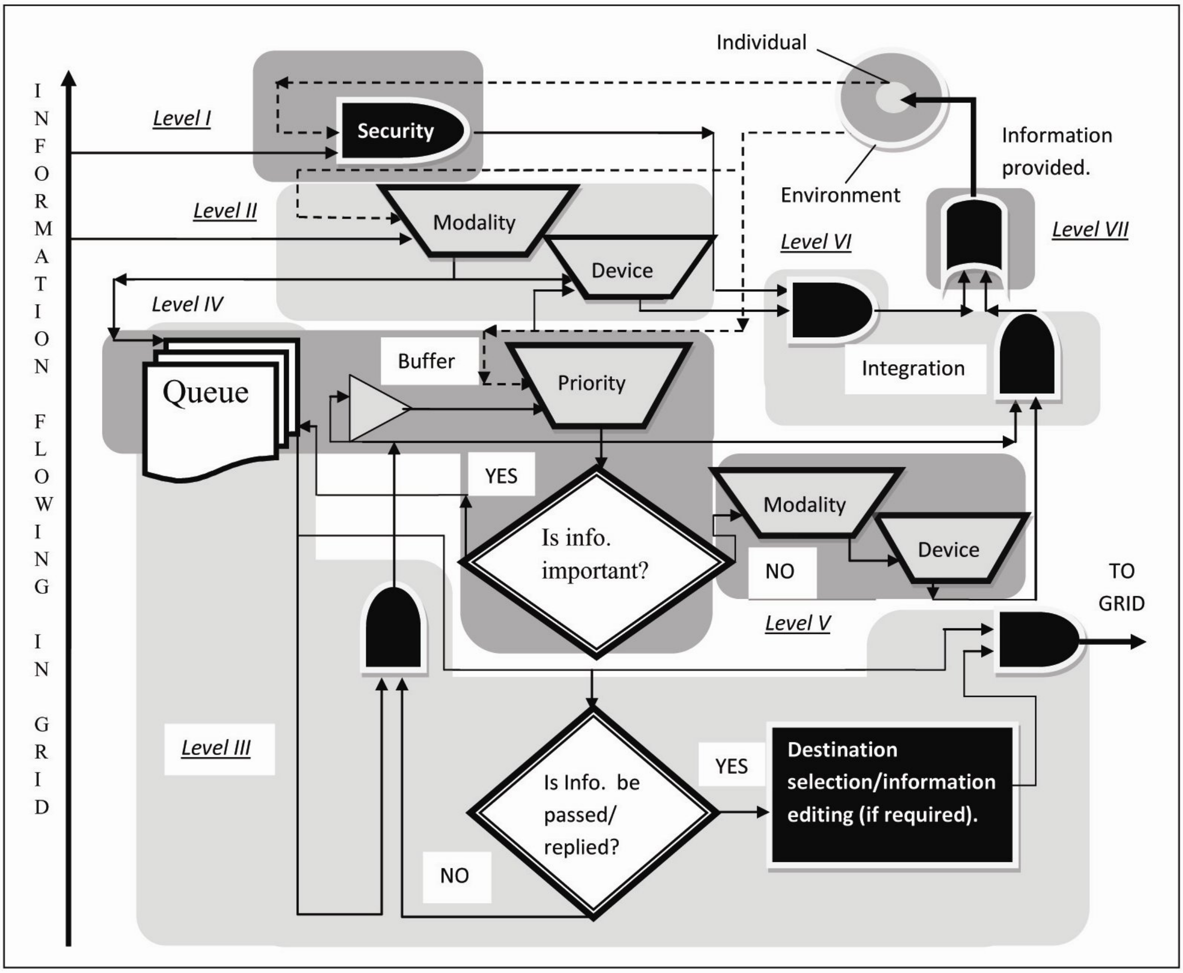


Figure3: System 'A' of HALCYON model

given moment only, then one of the modalities like sound, vision, touch etc is chosen, whatever is more appropriate at that instant. After the choice of modality is done, this and environment data are taken as selection lines in order to decide the appropriate device available with the user like cell phone, speaker etc in his environment, to pass on the information. This accounts for the fact that one has the option to provide information only through available devices with or around the user.

For example, consider you are driving a car towards north and a truck is coming from the right lane. There is high probability of a collision between the two vehicles. According to urgency level of decision, it is decided that the information of utmost importance is to be provided to the user at that instant. Considering the scenario, the appropriate modality will be voice. From the environment, it is found that speaker is the best device available with the user in order to draw his immediate attention. The user can then take immediate action to prevent the same.

LEVEL III- DECISION TO SEND INFORMATION TO OTHER SYSTEMS/ REPLY BACK FOR FURTHER OPERATION: At several occasion it becomes difficult when systems/devices demand your attention in order to send their output to the next one for further operations to be done. Consider a case when you are busy in a very important meeting in your office and your house catches fire. Although it is very urgent, you can not give immediate attention to it. Now, a possible solution offered in the pervasive world is that accessing your preoccupancy does not disturb you. On the other hand, it automatically gives authority to the fire extinguisher to switch on until the fire has been extinguished. At the same instant it must also call fire department if it senses that this fire is of massive scale which can not be put out by extinguisher alone. Thus, a decision is made at this level to know if the individual did not receive that information since it is not urgent or user is very busy in some other work. In this case, the first option is found correct and information progresses to Level IV else information is sent to the grid with proper authority list of users and urgency level. If the information is to be passed on, firstly the information is deciphered to find out the appropriate users where this information is to be sent, depending upon some predefined set of criteria. For example, in this fire case, fire department is found to be the appropriate receiver. Also, reply is given according to some predefined rules.

LEVEL IV- UNIMPORTANT INFORMATION IN QUEUE RECHECKED FOR PRIORITY AFTER A CERTAIN TIME DELAY: In case the information is not required at that instant, the user is not disturbed and thereby sent to a queue, which is unique for every individual. In case there is no hurry to send that information to other system/user then after a certain time delay, again the data stored is checked in “First In First Out” (FIFO) order to decide for the fact if it still is of any value to be provided at that instant or not? If not, then data is sent back to the queue again. Else it is sent to level V for further processing. One must note that the queue portion is common for both Level III and Level IV usage.

LEVEL V- CHOICE OF MODALITY AND DEVICE OF DELAYED INFORMATION ACCORDING TO URGENCY: If the data is important for the individual and this is to be known now, not after certain time, the same process as explained in Level II is repeated. Only difference being here is that now the option for sending data back to queue is not provided as now the information is of value only for that instant. Actually one must have seen the familiarity of these blocks with the digital world where such symbol represents a multiplexer. Thereby only one output is chosen among a plethora of many others depending upon one or more select lines.

LEVEL VI- INTEGRATION OF INFORMATION WITH CONTEXT AWARE DEVICE SELECTED: This level includes two parts. First being merging of the information with the device selected from those available at that instant only. Second also performs the same operation but it is kind of delayed version of the same coming from the queue.

LEVEL VII- INFORMATION MADE AVAILABLE TO THE INDIVIDUAL: This being the final level, here the information is coming from either of the two separately integrated portions and is made accessible to the individual at that instant only. Thus, one can see that this 7 levels subsystem A could to be installed between the free flowing information grid and an individual / device in order to transform this world into a calm one.

The proposed framework so far focuses only on one part of the whole picture, that is, how the information will be accessed by someone. The second part encompasses how it will be sent to the grid. Although, in part A, some information needs to be sent to the grid, it is for the sole purpose of linking two processes or operations. In other words, there is just an intermediary information link. On the other side, the subsystem B will depict how the information will be delivered if it directly originates from the user /device with no predecessor. However, one must note that operation wise the two portions are similar. Figure 4 shows the subsystem B of how the originated information is sent to the grid. It is quite simple in approach compared to that of A. Here the only fact to be kept in mind is that the information can not be simply sent to the grid of free flowing information as then various security and destination issues will pop up. Thus, in order to prevent that from happening along with the information, three other details are also sent along with it. One is the identity of real destinations, only where the information is to be made accessible, identity of the sender and the urgency level with which it is to be provided at the target. Figure 4 accounts for these facts only.

Maybe one should not overlook the fact that the focus has been shifted to the receiver rather than sender which demands more processing at receiver end. One could reason out that generally, everyone who sends the information will send it on high priority, to be processed at that time only. But, the model, being receiver centric, data is sent without giving preference to its urgency. Thus, if the user is busy, the information is sent to a queue where it is either sent to some other user, to be processed immediately or checked after some time if individual is free and it is necessary to provide the information. One application of it could be to stop numerous advertisements that demand your immediate attention.

Overall, it must be kept in mind that all these levels are only broad classifications not restricting to only one operation each level. Also, any level can be skipped, depending upon the requirement.

It should be kept in mind that these two subsystems namely A and B will be ulterior and continuously running in the backend without one’s knowledge.

In order to give a better picture of the model proposed let us consider an example. Here we have tried explaining part A by taking the fire instance. Imagine one house has caught fire. Now, as shown in the figure it immediately sends information to the person X (to whom this home belongs) by embedding it in the free flowing channel. The information sent contains four parts- Message: ‘Fire at home’, sender, receiver identity, and urgency level. This information sent reaches the person X but before being accessed by the user, it passes through subsystem A. The processes happening at each level have been explained alongside. It is to be again mentioned that all levels have not been covered in it as there was no need for those in this scenario. At level I, firstly security check is done if the given user is person X or not? Since if he is found to be the legitimate user, level II is then followed where it is initially checked if he is free or not? If free then appropriate modality and device is chosen for providing information. In this case as seen from the figure 5, the user is busy in a meeting and thus not free. Thus, this information is sent to the queue to self manage the things, which transfers it to level III. At level III it is checked if information is to be passed on to other systems by providing authority to them. Since, fire at home is of utmost priority it is passed on to fire station (after deciphering the message that there is fire at home, so fire department is to be called according to predefined actions). Apart from it, it also replies back to home to use fire extinguisher

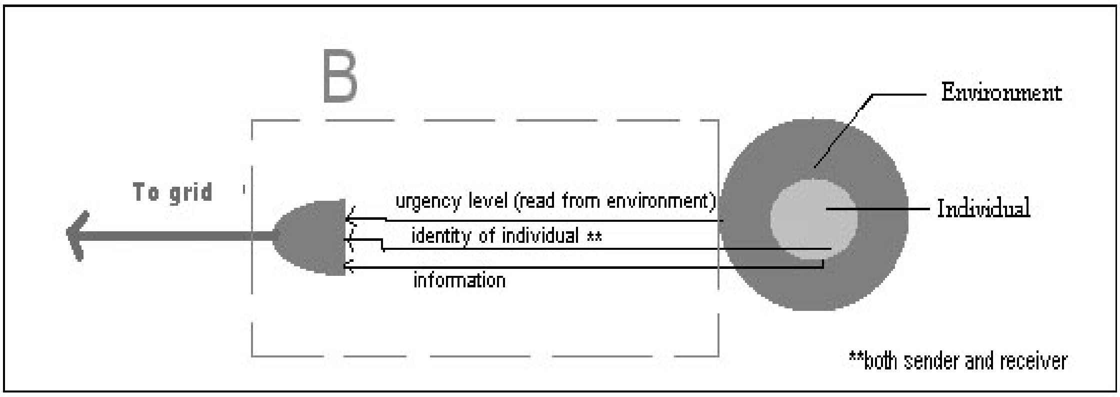
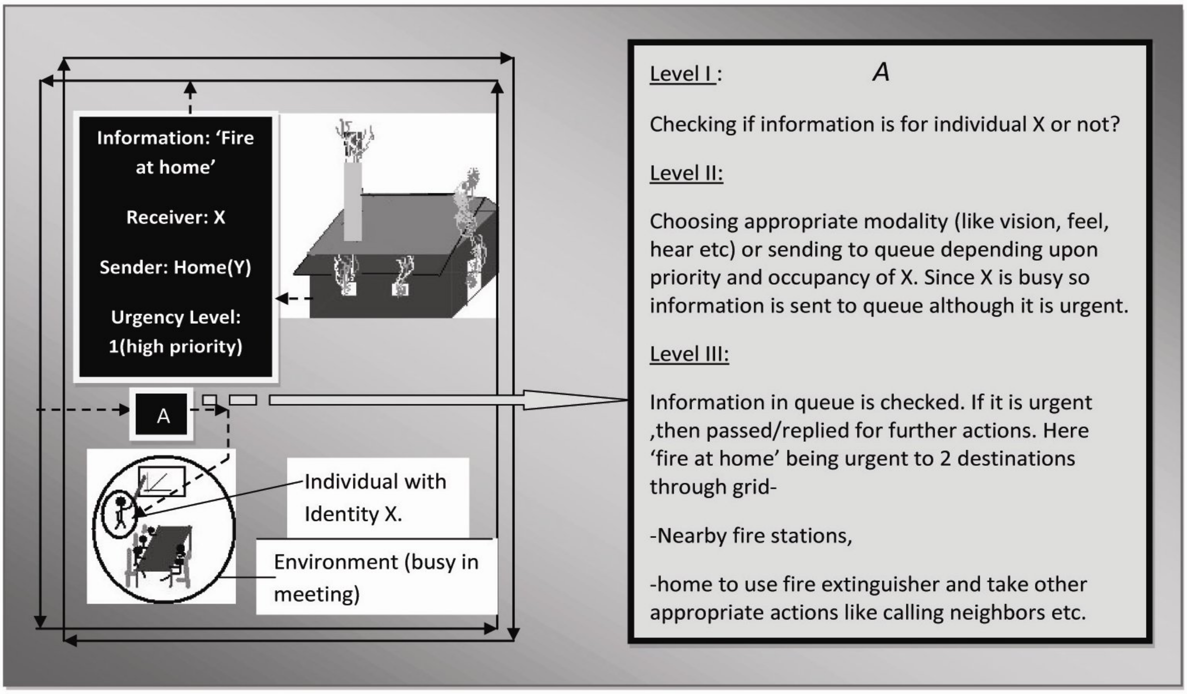

B
To grid
urgency level (read from environment)
identity of individual **
information
Environment
Individual
**both sender and receiver

Figure 4: Subsystem 'B' of HALCYON model


Information: 'Fire at home'
Receiver: X
Sender: Home(Y)
Urgency Level: 1(high priority)
A
Individual with Identity X.
Environment (busy in meeting)

Level I : A

Checking if information is for individual X or not?

Level II:

Choosing appropriate modality (like vision, feel, hear etc) or sending to queue depending upon priority and occupancy of X. Since X is busy so information is sent to queue although it is urgent.

Level III:

Information in queue is checked. If it is urgent ,then passed/replied for further actions. Here 'fire at home' being urgent to 2 destinations through grid-

-Nearby fire stations,

-home to use fire extinguisher and take other appropriate actions like calling neighbors etc.

Figure 5: A sample case of halcyon model.

in the meanwhile, or call neighbours. In case it is difficult to decipher the message by the system, it can be passed to some other individual, for example secretary or some other relative.

6 Issues and future challenges

The major issue that still needs to be figured out in adopting this model is to discover a technique through which such huge amount of information could be stored, which is exclusive for each individual (as shown in figure 3, where each individual is provided with his unshared Queue).

As 'HALCYON' will be for the whole world, the whole world should unanimously agree upon a common standard such as Sync ML, XML or similar proposed standards [27]. Other challenge is of finding a universal standard for the communication and the flow of this information in the grid. Although many proposed solutions are available, everybody has not agreed upon a single one method. In case two systems co-exist like Bluetooth and 802.11 [26], they will result in interference, thereby affecting proper working of each other.

Another concern is of finding a way for spectrum allocation so as to provide easy access and identity to every device/individual as spectrum shortage will be a process considering the mass scale of such operation. An idea proposed is to keep monthly subscription for everyone who only can access the portion of spectrum provided he has paid their monthly charges like for an internet connection. Through this, unnecessary users could be removed easily but it requires vast amount of flexibility in the system which could also be upgraded, whenever required.

Another concern is of power consumption. Things like batteries, solar cells are options but they surely can not quench the unending demand for power and thus deemed impractical.

Last but not the least is the issue for proving mobility for such systems by concealing itself in the world. What could be thought for giving a real shape to this model is through the use of something like satellites which will keep track of you without even your knowledge. Consider the example of GPS (Global Positioning System) as an example.

The integration of both complex back-end and front-end is not an easy task and requires ‘decentralization’ where everyone will be working in a distributed network but with synchronization between each other. Devices like Jini [28] and Universal Plug and Play (UPnP) [29] are paving the way for this third wave of communication. This next wave of communication will have major effect on environment and thus utmost care needs to be given to see that nature is not affected with this creation. Thus, in a nutshell, one must aim at the bigger vision when trying to make a “Halycon Home” for us and through this paper we have tried to achieve the same.

7 Conclusion

Every technology is coming into existence to harvest into a boon for the populace, yet it is in the aftermath of its implementation in the world that the darker side of it also comes into picture. One then plans to eliminate it by launching several ‘version series’ in the market, one after another. In this paper we tried to address drawbacks with the positive side, so as to make a wave of hitting the right target at the first attempt only. Can we do that? Surely, we can by every individual doing his small part to make a ‘better world’. We are in no way against this third era of technology. What we just want is to make a world with no shortcomings, so as not to repent later on. Through this paper we also tried to explain achieving the same by proposing a HALYCON framework and by bringing forth all the points under one roof.

References

[1] Mark Weiser (1991) Computer for the 21st Century http://nano.xerox.com/hypertext/weiser/SciDraft3.html

[2] Mark Weiser and John Seely Brown (1996). Coming Age of Calm Technology, Xerox PARC http://www.johnseelybrown.com/calmtech.pdf

[3] Data Fountain, Money translated to water, http://www.koert.com/work/datafountain/

[4] Mark Altosaar, Roel Vertegaal, Changuk Sohan, Daniel Chang: Auraorb (2006): Social Notification Appliance, Human Media Labs. http://interruptions.net/literature/Altosaar-CHI06-p381-altosaar.pdf

[5] Martin Tomitsch (2007) Towards a Taxonomy for Ambient Information Systems, Pervasive’07 Workshop http://deco.inso.tuwien.ac.at/fileadmin/user_upload/Pervasive07-WS-AIS-Taxonomy.pdf

[6] Information Technology Association of America (ITAA), Radio Frequency Identification. RFID ...Coming of age (2004) http://www.itaa.org/rfid/docs/rfid.pdf

[7] Polaris Networks, http://www.polarisnetworks.net/datasheet/pervcompressrealease230109.pdf

[8] Boaz Carmeli, Benjamin Cohen, Alan J. Wecker (2000) Proceedings of the eleventh ACM on Hypertext and hypermedia http://portal.acm.org/citation.cfm?id=336296.336502

[9] Vannevar Bush (1945) As we may Think, Atlantic Monthly. http://www.ps.uni-sb.de/~duchier/pub/-vbush/vbush-all.shtml

[10] John Thackara (2001), Pervasive Computing, Receiver Magazine, http://www.vodafone.com/-flash/receiver/05/articles/pdf/01.pdf

[11] The Morph Concept. http://www.nokia.com/about-nokia/research/demos/the-morph-concept

[12] Gene Michael Stover (2005), Notes about Vannevar Bush’s As We May Think http://cybertiggyr.com/nmemex/nmemex.pdf

[13] Kenneth Arnold: ConceptNet3, MIT Media Lab, 2007. http://conceptnet.media.mit.edu/

[14] F. Hattori, K. Kushima, T. Wasano: A comparison of Lisp, Prolog, and Ada programming productivity in AI area, 1985. http://portal.acm.org/citation.cfm?id=319655

[15] John Stasko, Myungcheol Doo, Brian Dorn, Christopher Plaue (2007). Explorations and Experiences with Ambient Information Systems, Pervasive’07 Workshop. http://www.cc.gatech.edu/~john.stasko/papers/pervasive07-sys.pdf

[16] John Stasko, Chris Paule, Zach Pausman: The InfoCanvas- Information Art, Aware Home, Georgia Tech., http://www.awarehome.gatech.edu/-projects/The_InfoCanvas.pdf

[17] Peter Beens: An Overview of Phidgets – Low Cost USB, Interfacing, Proceedings of 7th Annual ACSE Conference York University-November, 2005 http://wiki.acse.net/images/b/b6/Phidgets_Peter_Beens_ACSE2005.ppt

[18] Phidgets: Programming Manual, http://www.phidgets.com/documentation/Programming_Manual.pdf

[19] Thorsten Prante et al (2003), HelloWall- Beyond Ambient Displays , 5th International Conference on Ubiquitous Computing, USA. http://www.ipsi.fraunhofer.de/ambiente/paper/2003/prante-hello.wall_ubicomp03-withCopyright-letter.pdf

[20] Sara Routarinne, Johan Redstrom (2007), Domestication as Design Intervention http://www.johan.redstrom.se/papers/domestication.pdf

[21] The e-waste problem. http://www.greenpeace.org/-international/campaigns/toxics/electronics/the-e-waste-problem

[22] E-waste- An Indian Perspective. http://www.assocham.org/events/recent/event_64/An_Indian_Perspective_by_Toxic_Links.ppt

[23] Project Aura, http://www-2.cs.cmu.edu/~aura/

[24] David A. Cieslikows, Naomi J. Halewood, Kaoru Kimura and Christine Zhen-Wei Quang (2009): Key Trends in ICT Development, Information and Communication for Development. http://siteresources.worldbank.org/EXTIC4D/Resources/5870635242066347456/IC4D_2009_Key_Trends_in_ICT_Deelopment.pdf

[25] David Brin: Transparent Society http://www.usemod.com/cgi-bin/mb.pl?TransparentSociety

[26] Cheryl Ajluni (2009) Can Bluetooth and 802.11b co-exist? http://www.carl-chapman.com/articles/-archives/coexist.pdf

[27] Pervasive Computing: The Mobile World, Uwe Hansmann, Springer Professional Computing, pg 395 http://books.google.com/books?-id=8yyAbiMPOF0C&printsec=frontcover&dq=pervasive+computing&ei=DBOdSqOcKaCCkASI8PWdAQ#v=onepage&q=&f=false

[28] Introduction to Jini, Wikipedia. http://www.jini.org/wiki/Category:Introduction_to_Jini

[29] Universal Plug and Play, Wikipedia http://en.wikipedia.org/wiki/Universal_Plug_and_Play

Separating Interleaved HTTP Sessions Using a Stochastic Model

Marko Poženel, Viljan Mahnič and Matjaž Kukar
Faculty of Computer and Information Science
University of Ljubljana
Tržaška cesta 25, SI-1000 Ljubljana, Slovenia
E-mail: {marko.pozenel, viljan.mahnic, matjaz.kukar} @fri.uni-lj.si
Corresponding author: matjaz.kukar@fri.uni-lj.si

Keywords: Markov model, clickstream analysis, Web user behavior

Received: October 30, 2009

We describe a novel method for interleaved HTTP session reconstruction based on first order Markov model. Interleaved session is generated by a user who is concurrently browsing a web site in two or more web sessions (browser windows). In order to assure data quality for subsequent phases in analyzing user's browsing behavior, such sessions need to be separated in advance. We propose a separating process based on trained first order Markov chains. We develop a testing method based on various measures of reconstructed sessions similarity to original ones. We evaluate the developed method on two real world clickstream data sources: a web shop and a university student records information system. Preliminary results show that the proposed method performs well.

Povzetek: V članku predstavljamo metodo za razpletanje prepletenih HTTP sej s pomočjo markovskega modela.

1 Introduction

In the past decades World Wide Web (WWW) has become one of the main sources of information. It has enabled unprecedented exchange of data between different parties. Companies need web sites to reach customers and sell their products, institutions furnish information about their services, individuals can effectively access various services over Internet. With the growing number of web pages and documents, web sites are coping with stronger competition. It is difficult to attract new customers and retain the existing ones. Under such circumstances only the web sites that understand the needs of their customers will prevail. Analysing users' behavior has become an important part of web page data analysis. *Clickstream* data represent the main data source for the analysis of user behavior (11). A sequence of clicks that a user makes while browsing through a website is called a clickstream. Analysis of web data such as clickstreams entails certain problems with availability and quality of data (7).

Data about behaviour of web site visitors have become one of the most important sources of information in most web-aware companies. They play an important part in daily transactions and important business decisions. It is essential to get reliable data analyses, which require both appropriate methods and data. The quality of the the patterns discovered in data analysis depends on the quality of the data on which data mining is performed. A *user session* is represented by one visit of a user to a web site. For better web usage mining results we need reliable sessions. Clickstream data from a normal website are noisy, page events are often not explicitly linked to page requests. The preprocessing phase is therefore prone to errors. Although many methods for sessions reconstruction have been devised (1; 13), reliable session reconstruction still remains a challenge.

Especially really interested and capable users often browse the same web site with multiple browser windows opened. In each web browser they perform actions to complete a certain task. Typically, users switch between browsing tasks so that they work on a task only for a certain time period. Even if only one user is currently active, we actually have concurrent sessions, each for one web browser window (i.e. task). In a web server log file all concurrent sessions will be seen as a single long session. We call such sessions *interleaved sessions*. They cannot be easily separated without some kind of context help. Such sessions have negative effect on data quality so we have to deal with the issue. We have three choices: (i) neglect the problem, (ii) simply abandon such sessions, (iii) try to separate them. The first choice is bad for data quality since such sessions can affect web usage analysis results. If we abandon such sessions we also abandon useful knowledge about web site usage. Such sessions are usually generated by advanced users whose behaviour colud be potentially extremly valuable to us. Therefore we decided to develop a method for separating interleaved sessions.

We present a novel approach for session separation using a trained first-order Markov model to facilitate session separation. To the very best of our knowledge, the Markov approach has not been used for this purpose before. Actually, the interleaved session problem has been largely neglected in Web mining, with the only exception being Viermetz et al. (14) who use an entirely different approach based on building a clicktree. This clicktree contains all possible paths a user could have taken through a website map. While their approach is dedicated to better understanding of actual user behavior, our approach is focused on separation process. Based on training first-order Markov model on validated (clean) sessions, our approach is very effective in deinterleaving process (with linear complexity). We introduce a special purpose methodology for evaluation of separation process, evaluate our method on clickstreams from different sources, and present preliminary results.

2 Methods

2.1 Clickstream

In order to attract more visitors to our web site we have to know who our visitors are, what they do on our site, and what they would like to be changed. A great aid in achieving this goal is clickstream data. Clickstream is a sequence of clicks or pages visited as a visitor explores a particular Web site. Clickstream data are often large, inadequately structured, and show incomplete picture of users' activity. For example, server side log data do not involve browser and e.g., network caching ('Back' browser actions or requesting pages in intermediate server's cache) (7).

Clickstream data needs to be gathered, preprocessed and cleaned prior to the analysis. This step depends on the type and the quality of data. Work done in this phase affects the quality of results of further analyses.

The basic form of clickstream data from a Web server is stateless – no session identifier is logged. This is the consequence of the fact that the HTTP protocol is stateless. Each line in the log file shows an isolated resource retrieval event, but does not provide a link to other events in a user session. Since we are interested in all user actions in a certain period of time, we have to gather all individual events in a user session. The process is called *sessionization*. Without some context help it is hard or impossible to reliably identify complete user session. Berendt et al. (1) report that these sessionization tools are based on heuristic rules and assumptions about the site's usage and are therefore prone to errors.

2.2 Discrete Markov models for clickstream analysis

Markov chain is defined as follows. We have *a set of states* $S = \{s_1, s_2, ..., s_N\}$, where N denotes the number of states. The process starts in one of the states and moves forward from one state to another at regularly spaced discrete times. For example, the chain is currently in the state s_i and it moves next to s_j with the *transition probability* p_{ij}. The starting state is defined by a probability distribution. We denote the steps in which the process changes states as $t = 1, 2, ...n$ and the state at time t as q_t. Associated with each state is a set of transition probabilities p_{ij}, where

$$p_{ij} = P(s_i \to s_j) = P(q_t = s_j | q_{t-1} = s_i) \qquad (1)$$

that is, given the present state, the future and the past states are independant. This paper focuses on time-homogenous Markov chains, in which

$$\forall t : P(q_{t+1} = s_i | q_t = s_j) = P(q_t = s_i | q_{t-1} = s_j) \quad (2)$$

for all t, meaning that the transition probabilities do not change with time. We restrict our discussion to Markov chains defined on a finite state-space. The probability of transition between states in a single step can be written as *transition probability matrix* T:

$$T = \begin{bmatrix} p_{11} & \cdots & p_{1N} \\ \vdots & \ddots & \vdots \\ p_{N1} & \cdots & p_{NN} \end{bmatrix}, \quad \sum_j p_{ij} = 1 \qquad (3)$$

The final parameter of a Markov chain is the *starting state*, which can either be a predefined fixed state or can be chosen from a probability distribution on a set of states given in the form of a probability vector π,

$$\pi = (\pi_1, \pi_2, \pi_3, \ldots, \pi_N) \qquad (4)$$

where π_i denotes the probability that state s_i is initial and N denotes number of states.

$$\pi_i = P(q_1 = s_i), 1 \leq i \leq N \qquad (5)$$

Given a sequence of states $(q_1, q_2, \ldots, q_k)$ we can calculate the probability of the sequence by multiplying the probability of the initial state $P(q_1)$ with the probability of transitions to the successive state as follows:

$$P(q_1, q_2, \ldots, q_k) = P(q_1) \cdot \prod_{i=2}^{k} P(q_{i-1} \to q_i) \qquad (6)$$

In the first-order Markov chain the next step depends only on current state. If the step depends on the current and the previous state, we get a somewhat more complicated second-order Markov model. Its states correspond to all possible pairs of actions that can be performed in a sequence. We can generalize this approach to the Kth-order Markov model, which computes the predictions by looking at the last K actions performed by the user, leading to a state-space that contains all possible sequences of K actions (6).

2.3 Related work

Data pre-processing is important part of web usage analysis since it requires large amount of time and affects the results of analyses. This problem motivated researchers to develop new methods for pre-processing.

Colley et al. (3) proposes a series of steps for data preprocessing for web usage mining. These include data cleaning, user identification, session identification and data formatting. Zhang et al. (13) improved statistical-based time oriented heuristics for the reconstruction of user sessions. They used statistical analysis and usage mining techniques to improve time-oriented heuristics. Ting et al. (11) developed the Pattern Restore Method (PRM) algorithm, which attempts to reconstruct missing server-side clickstream data based on referring site information and the Website's link structure. Berendt et al. (1) used the web site structure to reconstruct incomplete sessions.

Markov models have also been used in the clickstream analysis area. Many approaches have been proposed. In (2) the authors primarily focus on visualization aspects of website navigation patterns. Model-based clustering (using finite mixtures of Markov models) is used to assign users to clusters. Sarukkai (10) employs Markov chains to both predict the most likely sites that a user will visit next and generate tours (sequences of websites) that a user might be interested in according to his or her current browsing history. The model can be continuously updated with data provided by new users of the websites it covers.

In (6) authors look at the ways of reducing the state-space complexity of higher order Markov models, while retaining their high coverage. This is achieved by first building a full model from some of the training data, then pruning it with the rest. The results show that these methods can greatly reduce the state-space complexity while generally improving its accuracy. Ypma and Haskes (12) expanded the work done by Cadez and Heckerman by using mixtures of Hidden Markov models. This enabled them to process the dataset without first grouping actual URI requests into page categories. Their work shows that even without artificially categorized webpages, a mixture of HMMs will generate classes of pages with similar characteristics.

2.4 Separating interleaved sessions with Markov model

The process of separating interleaved sessions is one of the phases in data pre-processing. First, clickstream data has to be cleaned and sessionized. We refer to sessions, that have been restored without deficiencies, as *clean* sessions. Durring the sessionization process we detect interleaved sessions which we cannot separate at that time either by using some background knowledge, or by applying a pre-trained Markov model (MM). Interleaved sessions are separated from clean sessions and are additionally processed. The separation process is based on stochastic methods which have been used to solve some other issues related to cliskstream. Because of generality and simplicity we decided to use first-order Markov model. We build a Markov model and train it with data from clean sessions. Training proceeds as follows. If there is a transition $s_i \rightarrow s_j$ in training data, the frequency counter n_{ij} is incremented by one. We can use last pre-processing clean sessions or clean sessions from last few pre-processings. Trained markov model is then used to separate interleaved sessions. In case of more than two interleaved sessions only the first one is considered as clean, and the second one is submitted to further separation. This results in more reliable pre-processed user behavior data. The last step in a analysis is evaluation of separated sessions with several methods.

For separating interleaved sessions we use a trained first-order MM. We utilize site map data as background knowledge. Site map consists of links between pages that are explicitly connected with hyperlinks. A link between pages S_1 and S_2 in a site map means higher prior probability of transition between these two pages than if there were no link in a site map. When we train the MM we also use the web site map. Based on links between page sites we calculate initial transition probability between pages $p_{ij}^{(0)}$, where i, j denotes source and target state. Formula for calculating $p_{ij}^{(0)}$:

$$p_{ij}^{(0)} = \frac{1 - P_A^{(ij)}(N - n_t)}{n_t}, n_t \geq 1 \qquad (7)$$

where j denotes all states that are connected to state i, N denotes number of states, n_t number of outgoing links from state i[1] and $P_A^{(ij)} = 1/N^2$ an uninformed probability of transition between any two states. If there is no connection between i and j, probability $P_A^{(ij)}$ is assigned. Parameter $P_A^{(ij)}$ determines the prior probability of transition between arbitrary two pages in the site map.

Let each session be represented as sequence of pages $S = \{q_1, q_2, \dots q_n\}$ where n denotes length of session. q_1 denotes the entry page and q_n the last page the user visited in this session. For a transition from $q_{i-1} = s_j$ to $q_i = s_k$, training data site map data can be combined with *m-estimate* (5):

$$P(s_j \rightarrow s_k) = p_{jk} = \frac{(n_{jk} + m p_{jk}^{(0)})}{n_j + m} \qquad (8)$$

where n_{jk} denotes number of transitions from state j to k, which we got from training data. n_j is number of visits of state j. m denotes the weight which presents the ratio between prior (web site map) and posterior knowledge. p_{jk}^0 denotes transition probability based on web site map. Parameter m represents the importance rate of prior knowledge. The higher the m is, the more important the prior knowledge is. If $m = 0$, then we completly neglect the meaning of prior knowledge. In that case m-estimate converts to relative frequency $p_{jk} = n_{jk}/n_j$.

2.5 The separation process

Separating interleaved session is based on a fact that a transition between sites $q_i \rightarrow q_{i+1}$ is more likely to belong to one of the consisting sessions. If we have interleaved session $S_p = [q_1, q_2, \dots, q_n]$ that consists of two clean sessions length n_1 and n_2, where $n_1 + n_2 = n$. The number of

[1] We assume that there is always the reflective transition from s_i to s_i, so n_t is always greater than 0.

possible different separations is $C = \binom{n_1+n_2}{n_1} = \binom{n_1+n_2}{n_2}$. Let us say that the last page of the first session that we already managed to separate is S_{1i}. Similarly for the second session we denote the last page as S_{2i}. For each page S_i in an unprocessed interleaved session, we check what is the transition probability from last page of separated session to current page S_i. If $P(S_{1i} \rightarrow S_i) > P(S_{2i} \rightarrow S_i)$ we add page S_i to the first separated session, otherwise to the second one. Until both of the separated sessions get the first element (entry page), we have to check whether S_i is an entry page for second session. Separating process can be seen on Figure 1.

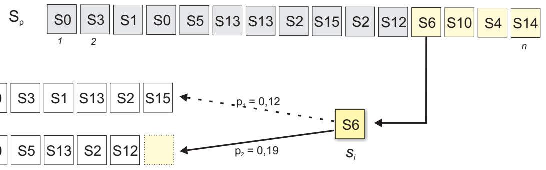


Figure 1: Figure shows simple process of separating interleaved session.

2.6 Evaluation of separating process

Separated sessions needs to be evaluated to see how successful our method was. Each session is represented as a sequence of pages. Evaluating quality of separated sessions can be viewed as evaluating their similarity. Determining the similarity between sequences is one of the basic tasks in machine translation as well as in computational biology (8). Basically, two sequences are more similar if they have more symbols in common and the symbols' order is similar. There are many methods of measuring similarity between two sequences. We use several more or less strict methods based on: perfect match, Levenshtein distance, longest common subsequence (LCS) and weighted longest common subsequence (9).

Perfect match is a simple method where only sequences that perfectly match contribute to the end result.

Alternative approach to measure sequence similarity is based on sequence distance, named *edit distance*. The distance between two sequences is defined as the smallest sum of edit operations' costs that transforms one sequence to another. If we have only three edit operations: inserting, deleting and swapping symbols, and all have the cost of 1, we get *Levenshtein distance*.

A sequence $Z = [z_1, z_2, ..., z_n]$ is a subsequence of another sequence of sequence $X = [x_1, x_2, ...x_m]$ if there exists a strict increasing sequence $i_1, i_2, ...i_k$ in X such that for all $j = 1, 2, ..., k$ we have $x_{ij} = z_j$ (4). If we have sequences X and Y, the longest common subsequence of X and Y is a common subsequence with the maximum length. The longer the common subsequence, the more two sessions are similar to each other. One advantage of LCS is that it does not require consecutive matches but in-sequence matches that reflect level element order as n-grams. Deficiency of LCS is that it only counts the main in-sequence elements. Other common subsequences are not reflected in a result (8). We estimated these methods are apropriate for evaluation of separating process.

We can improve LCS method to differentiate LCS in relation to other elements in the sequence. Chin et al. (9) called this method *weighted LCS* (WLCS). They also propose the use of *F-measure* to estimate the similarity between two sequences X of length m and Y of length n. We decided to use F-measure for presenting end results.

3 Materials

3.1 Synthetic data

First we created a test environment that is similar to real one but is not as complex. We checked what is the average HTTP session length on a local web server. For testing we fixed the number of Web pages to 30. We created an artificial web site map that represented links with higher probability. According to the site map we generated a number of sessions that were used for MM training data, and some of them for creating interleaved sessions. After training MM, we applied the process for separating interleaved sessions and verified the results. About 48% of interleaved sessions were separated 100% correctly, which encouraged us to proceed to real data.

3.2 Real-world data

We applied the interleaved session separating process on two real clickstream sources. The first clickstream originates from log files of university student records information system. It has been used by 16 member institutions. It has approximately 300 different pages. Each state in MM corresponds to an individual page. Typical user paths are well defined. Users have to be logged on in order to use the system. Sometimes they are logged on with different user roles at the same time, and this creates interleaved sessions. Since users have to be logged on we can always determine the session entry point. The Web server log files use the basic CLF format. Clickstream data was taken for 4 months of use, which resulted in 150.000 user sessions.

The second clickstream source is taken from a web shop, which is considerably different from the student records information system. Users do not have to sign in (except for buying items), it has many more users and many more pages. We had to cut down number of states of Markov model in order to efficiently use it. Every state of our Markov model represents a group of pages, not an individual page. We transformed the web shop pages to 900 states. Session entry point can be almost any page, which makes separating interleaved sessions harder. The Web shop site map has plenty of links between pages. In fact only few pages are not linked with all others. The web shop generates about 10.000 user sessions a day.

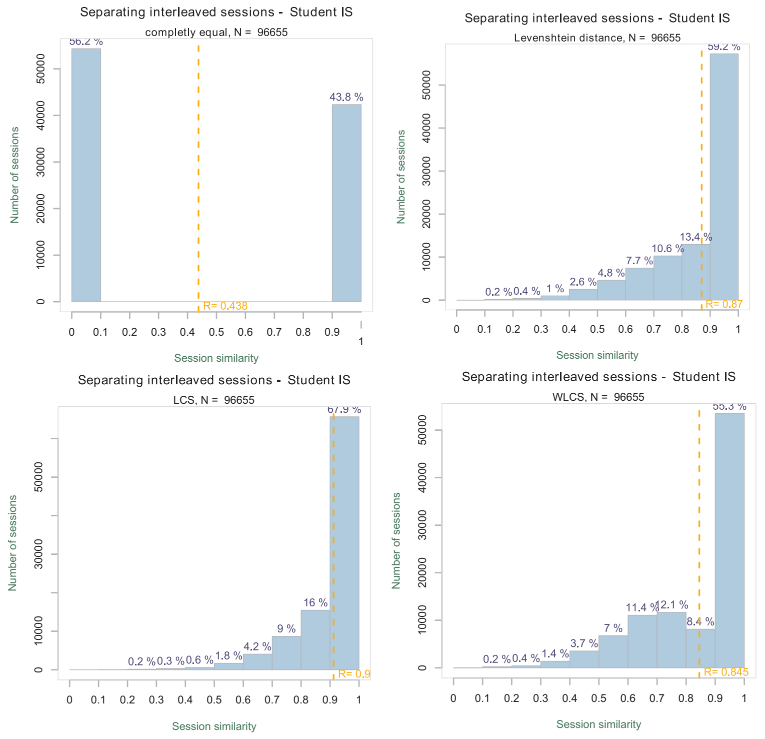


Figure 2: Results of separating for Student records IS clickstream. R denotes weighted average of session similarities.

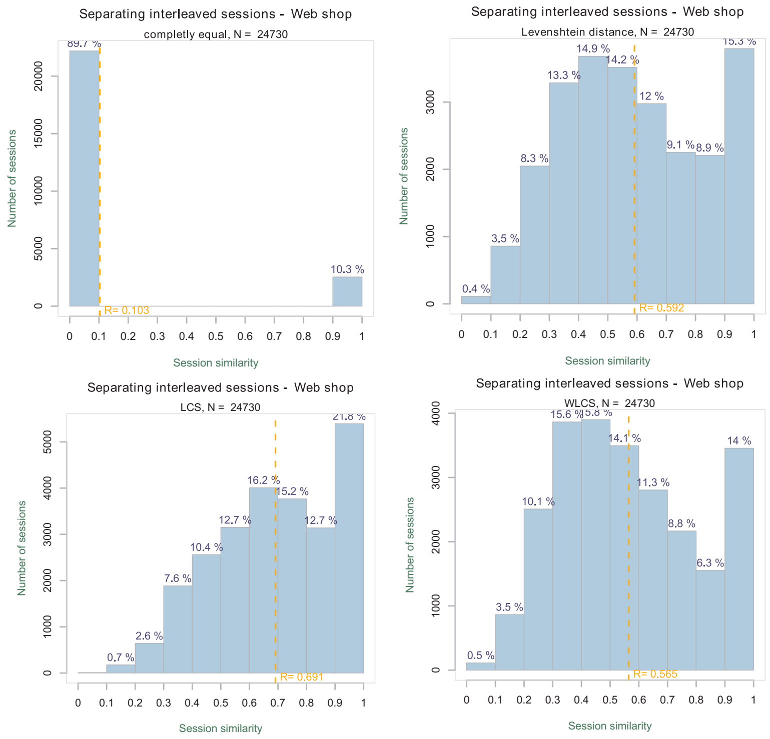


Figure 3: Results of separating for Web shop clickstream. R denotes weighted average of session similarities.

For both clickstreams we took the same steps as with artificially generated data. Initial clean sessions, used for learning, were generated during the sessionization process of clickstream data. During the sessionization we applied all the neccessary steps in order to remove noisy data. We analysed what a typical user session looks like and removed all sessions that did not meet the rules (e.g. too short or too long sessions). 70% of clean sessions were used as a training set for MM, and the rest were used to generate interleaved sessions in order to evaluate the separation process. After separating interleaved sessions we evaluated results with evaluation methods that we presented earlier.

4 Results

In Figures 2 and 3 we can see graphs for evaluation methods and source of clickstream. Each graph corresponds to one evaluation method. The X axis shows intervals for F-measure based similarity and the Y axis shows number of sessions that fall in that interval. Figure 2 reports results for student IS clickstream. 96655 interleaved sessions have been created and separated. On the first graph we see that 43% sessions have been separated 100% correctly (session sequence similarity = 1). This result is much better in comparison with Web shop. Other three graphs on at Figure 2 depict how well the sessions have been separated according to evaluation method. LCS and WLCS graphs show that majority of sessions are more than 50% similar to the original ones.

If we look at Figure 3 we see results for Web shop. 24730 interleaved sessions have been created and separated. Looking at the first graph in that Figure, one sees how many sessions have been separated 100% correctly. For web shop this percentage is a little more than 10%, which is quite low. However even 10% is better than throwing away all interleaved sessions. One of the reasons is that grouping pages together affects the results. Since the site map is larger, there may be numerous user paths, what also affects the results. User can enter the web shop at almost any page, so it is harder to detect where the second session in interleaved session starts. Results on a graph that show LCS seem better, since LCS is a less strict method of evaluation than WLCS.

5 Conclusion

We propose a new method for improving the quality of clickstream data in pre-processing phase that is based on a first-order Markov model. To the very best of our knowledge, the Markov approach has not been used for this purpose before. Proposed method is very effective in deinterleaving sessions (linear complexity). We present the motivation that led us to implementation and have applied method on two real data clickstreams. The presented results show that in certain cases method gives promising results. We analysed the domain and detected possible causes of worse results. In order to minimize method deficiencies we plan to work on the issues we presented. First we have to improve the method for detecting interleaved session starting pages. We are also planning to use second-order Markov model and Hidden Markov Model (HMM) for separating process.

References

[1] Bettina Berendt, Bamshad Mobasher, Miki Nakagawa, and Myra Spiliopoulou. The impact of site structure and user environment on session reconstruction in web usage analysis. In *WEBKDD - KDD Workshop on Web Mining and Web Usage Analysis*, pages 159–179, 2002.

[2] Igor Cadez, David Heckerman, Christopher Meek, Padhraic Smyth, and Steven White. Model-based clustering and visualization of navigation patterns on a web site. *Data Min. Knowl. Discov.*, 7(4):399–424, 2003.

[3] Robert Cooley, Bamshad Mobasher, and Jaideep Srivastava. Data preparation for mining world wide web browsing patterns. *Knowledge and Information Systems*, 1:5–32, 1999.

[4] T. H. Cormen, C. E. Leiserson, and R. L. Rivest. *Introduction to Algorithms.* The MIT Press and McGraw-Hill Book Company, 1989.

[5] S. Džerovski, B. Cestnik, and I. Petrovski. Using the m-estimate in rule induction. *J. Comput. Inf. Technol.*, 1(1):37–46, 1993.

[6] Mukund Deshpande and George Karypis. Selective markov models for predicting web page accesses. *ACM Trans. Interet Technol.*, 4(2):163–184, 2004.

[7] Ron Kohavi. Mining e-commerce data: The good, the bad, and the ugly. In Foster Provost and Ramakrishnan Srikant, editors, *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 8–13, 2001.

[8] G. Leusch, N. Ueffing, and H. Ney. A novel string-to-string distance measurewith applications to machine translation evaluation. In *In Proceedings of MT Summit IX*, pages 240–247, 2003.

[9] C-Y. Lin and F. J. Och. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *ACL '04: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, page 605, Morristown, NJ, USA, 2004. Association for Computational Linguistics.

[10] R. R. Sarukkai. Link prediction and path analysis using markov chains. In *Proceedings of the 9th international World Wide Web conference on Computer networks : the international journal of computer and telecommunications netowrking*, pages 377–386, Amsterdam, The Netherlands, The Netherlands, 2000. North-Holland Publishing Co.

[11] I-Hsien Ting, Chris Kimble, and Daniel Kudenko. A pattern restore method for restoring missing patterns in server side clickstream data. *Lecture Notes in Computer Science*, 3399:501–512, March 2005.

[12] Alexander Ypma and Tom Heskes. Automatic categorization of web pages and user clustering with mixtures of hidden markov models. In *WEBKDD*, pages 35–49, 2002.

[13] J. Zhang and A.A. Ghorbani. The reconstruction of user sessions from a server log using improved time-oriented heuristics. In *Communication Networks and Services Research, 2004. Proceedings. Second Annual Conference on*, pages 315–322, May 2004.

[14] M. Viermetz, C. Stolz, C. Gedov, and M. Skubacz. Relevance and Impact of Tabbed Browsing Behavior on Web Usage Mining. In *Web Intelligence, 2006. WI 2006, 2006. Proceedings. IEEE/WIC/ACM International Conference on*, pages 262–269, Dec 2006.

A Simulation Study on the Impact of Mobility Models on the Network Connectivity, Hop Count and Lifetime of Routes for Ad hoc Networks

Natarajan Meghanathan
Department of Computer Science
Jackson State University
Jackson, MS 39217, USA
E-mail: natarajan.meghanathan@jsums.edu

Keywords: mobile ad hoc networks, vehicular ad hoc networks, mobility models, route lifetime, hop count, connectivity, simulation

Received: September 23, 2009

The high-level contribution of this paper is a simulation-based analysis of the network connectivity, hop count and lifetime of the routes determined for ad hoc networks under the following four mobility models: Random Waypoint model, Gauss-Markov model, City Section model and the Manhattan model. Two kinds of routes are determined: routes with the minimum hop count and routes with the longest lifetime. Extensive simulations have been conducted for different conditions of network density and node mobility for each of the four mobility models and also for different values of the degree of randomness parameter for the Gauss-Markov mobility model. We arrive at rankings of the mobility models with respect to network connectivity, hop count of minimum hop routes, lifetime of minimum hop routes, lifetime of stable routes and the hop count of stable routes. We also observe a route lifetime–hop count tradeoff for all the four mobility models. The general trend of the results is: the more realistic and constrained is a mobility model, the larger is the number of hops in the minimum hop routes and smaller is the lifetime of the stable routes determined under the mobility model.

Povzetek: Opisana je analiza mrež s štirimi modeli mobilnosti: Random Waypoint, Gauss-Markov, City Section in Manhattan.

1 Introduction

A mobile ad hoc network (MANET) is a dynamically distributed system of mobile wireless nodes. The network bandwidth is limited and the transmitted signals are prone to interference and collision as the medium is shared. Since the nodes operate with limited battery power, the transmission range of a node is often limited. As a result, multi-hop routing is a common feature in MANETs. As the nodes move, there is unlikely to be a single fixed route throughout the duration of a source-destination session. MANET routing protocols proposed in the literature can be of two types [1]: proactive and reactive. Proactive routing protocols tend to predetermine routes between any pair of nodes in the network through periodic exchange of route-information bearing control packets among the nodes in the network. Reactive routing protocols use a broadcast query-reply cycle to determine routes between a pair of nodes only when required. In dynamic environments, reactive routing has been preferred over proactive routing as the on-demand routing protocols incur a relatively lower control overhead [2].

Vehicular ad hoc networks (VANETs) are one of the most promising application areas of MANETs. VANET communication is normally accomplished through special electronic devices placed inside each vehicle so that an ad hoc network of the vehicles is formed on the road. A vehicle equipped with a VANET device should be able to receive and relay messages to other VANET-device equipped vehicles in its neighborhood. VANET applications can be broadly classified into two categories: safety applications and comfort applications [3]. An example of a safety application is on-board active safety systems to assist drivers with information (like accidents, road surface conditions, intersections, highway entries and etc) about the road ahead. Comfort applications are those applications that can provide non-critical services like weather information, gas station or restaurant locations, mobile e-commerce, Internet access, music downloads and etc.

VANETs resemble MANETs with respect to the dynamically and rapidly changing network topologies due to fast moving vehicles. However, the mobility of the vehicles is normally constrained by predefined roads and speed limitations. Mobility of the vehicles is also affected due to traffic congestion in the roads and the traffic control mechanisms (like stop signs and traffic lights). Route stability is an important design criterion to be considered in the design of MANET and VANET routing protocols. The routing protocols should be able to dynamically adapt to the rapidly changing network topologies while taking into consideration the layout of the roads. The commonly used route discovery approach of flooding the Route-Request (RREQ) packets can easily lead to congestion in the network and also consume the battery charge of the nodes. Frequent route changes can also result in out-of-order packet delivery, causing high jitter in multi-media, real-time applications. For safety applications, it is better to route all the critical data packets through the same path so that the receiver can reassemble the packets and get a consistent view of the network condition.

In an earlier work [4], we proposed an algorithm called *OptPathTrans* to determine the sequence of stable routes between a given source-destination pair over the duration of a communication session. Given the complete knowledge of the future topology changes over the entire duration of the communication session between a source *s* and destination *d*, algorithm *OptPathTrans* operates as follows: Whenever an *s-d* path is required at a time instant *t*, choose the longest-living *s-d* path from *t*. The above strategy is repeated over the duration of the *s-d* session. The sequence of such longest living stable paths is called the Stable Mobile Path (SMP). The performance of algorithm *OptPathTrans* has been largely studied under the commonly used Random Waypoint (RWP) mobility model [5] for MANETs. Note that throughout this paper, we use the terms ‘path’ and ‘route’ interchangeably. They mean the same. Similarly, the terms ‘vehicle’ and ‘node’ are used interchangeably, but mean the same.

In this paper, we study the performance of algorithm *OptPathTrans* with respect to the commonly used City Section mobility model [6] and the Manhattan mobility model [7] for VANETs, in addition to the Random Waypoint mobility model and the Gauss-Markov mobility model [8] used in MANETs. Most of the simulation studies in MANETs use the RWP model as the node mobility model. Even though the RWP model is easy to simulate, it has some unrealistic assumptions about node movement [9]: sharp turns and sudden stop. Sharp turns occur whenever a node changes its direction after travelling for a random amount of time and sudden stops occur when the node decides to stop at a particular time instant and changes directions. During a direction change, the speed chosen by a node is totally independent of the previous speed. The Gauss-Markov mobility model proposed by Liang and Haas [8] is more realistic compared to the RWP model. It eliminates the twin problems of sharp turns and sudden stops by considering the past speed and direction to influence the future speed and direction. But, the Gauss-Markov mobility model has been very rarely used in MANET simulation studies, mainly due to the relatively larger complexity involved in simulating it compared to the Random Waypoint model.

We use the *OptPathTrans* algorithm to compute the sequence of stable paths (the Stable Mobile Path) and the *Dijkstra* algorithm [10] to compute the sequence of minimum hop paths (called the Minimum Hop Mobile Path). For different conditions of network density and node mobility considered, we compute the Minimum Hop Mobile Path and the Stable Mobile Path and rank the four mobility models (Random Waypoint mobility model, Gauss-Markov mobility model, City Section mobility model and the Manhattan mobility model) with respect to three critical metrics: (i) network connectivity (ii) path hop count and (iii) route lifetime. In all of the simulations, we observe a route lifetime – hop count tradeoff for each of the four mobility models considered. Routes with longer lifetime have larger hop count and routes with smaller hop count have shorter lifetime. We could not find any such comprehensive evaluation study in the literature about the impact of the four mobility models on the network connectivity, lifetime of stable paths and the hop count of minimum hop paths.

The rest of the paper is organized as follows: Section 2 briefly discusses the working of the Random Waypoint, City Section, Manhattan and the Gauss-Markov mobility models and also discusses the simulation methodology adopted for each model. Section 3 describes related work with regards to studying the impact of mobility models on the performance of the routing protocols. Section 4 provides an overview of the *OptPathTrans* algorithm used to determine the sequence of long-living stable paths in ad hoc networks. Section 5 illustrates the simulation results and compares the four mobility models with respect to the results obtained for network connectivity, route lifetime and hop count. Section 6 summarizes the results of the simulations and Section 7 concludes the paper.

2 Mobility models and their simulation methodology

In this section, we first provide a brief overview of the four mobility models studied in this paper. All the four mobility models assume the network is confined within fixed boundaries. The Random Waypoint mobility model assumes that the nodes can move anywhere within a network region. Under the Gauss-Markov mobility model, a node periodically updates its speed and direction of movement with a certain degree of Gaussian randomness and based on the past speed and direction of movement. The City Section and the Manhattan mobility models assume the network to be divided into grids: square blocks of identical block length. The network for these two VANET mobility models is thus basically composed of a number of horizontal and vertical streets. Each street has two lanes, one for each direction (north/ south direction for vertical streets, east/ west direction for horizontal streets). A node is allowed to move only along the grids of horizontal and vertical streets. For all the four mobility models, the mobility profile for each node, spanning the entire simulation time period, is created offline. The mobility profiles of the nodes are then input to the routing algorithm.

2.1 Random waypoint mobility model

Initially, the nodes are assumed to be placed at random locations in the network. The movement of each node is independent of the other nodes in the network.

The mobility of a particular node is described as follows: The node chooses a random target location to move. The velocity with which the node moves to this chosen location is uniform-randomly selected from the interval [v_{min},…,v_{max}]. The node moves in a straight line (in a particular direction) to the chosen location with the chosen velocity. After reaching the target location, the node may stop there for a certain time called the pause time. The node then continues to choose another target location and moves to that location with a new velocity chosen again from the interval [v_{min},…,v_{max}]. The selection of each target location and a velocity to move to that location is independent of the current node location and the velocity with which the node reached that location.

Simulation Methodology: The nodes are uniform-randomly distributed throughout the network. The mobility profile for a node is updated each time the node changes its direction and randomly chooses a new target location to move. The mobility profile for a node *i* comprises of a sequence of tuples, each of which contain the following information: $[t_i^a, t_i^b,\ \ (x_i^a, y_i^a),\ (x_i^b, y_i^b), v_i^{a-b}]$ where t_i^a is the time instant that node *i* is at location (x_i^a, y_i^a), changes its direction to move to a randomly selected location (x_i^b, y_i^b) with a randomly chosen velocity $v_i^{a-b} \in [v_{\min}, \ldots, v_{\max}]$ and reaches the chosen location at time instant t_i^b. The above process is independently repeated for each node until the simulation time period. We assume the pause time for a node to be zero in all of our simulations. Whenever the routing algorithm needs to compute a graph at a particular time instant, say t^s, we determine the location of each node in the network using the mobility profile for the node. For example, to determine the location of node *i* at time instant t^s, we index into the sequence of tuples constituting the mobility profile of node *i* and choose the entry whose two time instants t_i^a, t_i^b are such that $t_i^a \le t^s \le t_i^b$. The location (x_i^s, y_i^s) of node *i* at time instant t^s is basically given by:

$$x_i^s = \left[\frac{t_i^s - t_i^a}{t_i^b - t_i^a} * x_i^b\right] + \left[\frac{t_i^b - t_i^s}{t_i^b - t_i^a} * x_i^a\right]$$

$$y_i^s = \left[\frac{t_i^s - t_i^q}{t_i^b - t_i^a} * y_i^b\right] + \left[\frac{t_i^b - t_i^s}{t_i^b - t_i^a} * y_i^a\right].$$

2.2 City section mobility model

Initially, the nodes are assumed to be randomly placed in the street intersections. Each street (i.e., one side of a square block) is assumed to have a particular speed limit. Based on this speed limit and the block length, one can determine the time it would take to move in the street. Each node placed at a particular street intersection chooses a random target street intersection to move. The node then moves to the chosen street intersection on a path that will incur the least amount of travel time. If two or more paths incur the least amount of travel time, the tie is broken arbitrarily. After reaching the targeted street intersection, the node may stay there for a pause time and then again choose a random target street intersection to move. This procedure is repeated independently by each node.

Simulation Methodology: The nodes are uniform-randomly distributed on the street intersections. More than one node may be placed at a street intersection. The mobility profile for a node is updated each time the node moves from a street intersection to a randomly selected street intersection through a path that will incur the minimum number of street intersections. The mobility profile for a node *i* comprises of a sequence of tuples, each of which contain the following information: $[t_i^a, t_i^b,\ \ p_i^{a-b}(x_i^a, y_i^a), (x_i^b, y_i^b), v_i]$ where t_i^a is the time instant that node *i* is at street intersection (x_i^a, y_i^a) and chooses to move to a randomly selected street intersection (x_i^b, y_i^b) with a velocity v_i through the shortest path p_i^{a-b} that will have the minimum number of street intersections between (x_i^a, y_i^a) and (x_i^b, y_i^b). Node *i* reaches (x_i^b, y_i^b) at time instant t_i^b. The above process is independently repeated for each node until the simulation time period. We assume the pause time for a node to be zero in all of our simulations.

To determine the location of node *i* at time instant t^s, we index into the sequence of tuples constituting the mobility profile of node *i* and choose the entry whose two time instants t_i^a, t_i^b are such that $t_i^a \le t^s \le t_i^b$. The location (x_i^s, y_i^s) of node *i* then at time instant t^s is determined as follows: Let the shortest path p_i^{a-b} between street intersections (x_i^a, y_i^a) and (x_i^b, y_i^b) be represented as (x_i^a, y_i^a), (x_i^{k1}, y_i^{k1}), (x_i^{k2}, y_i^{k2}), ……, (x_i^{kh}, y_i^{kh}),(x_i^b, y_i^b); where *h* is the number of intermediate street intersections; k_1, k_2, k_3, ………..k_h are the intermediate street intersections constituting the shortest path p_i^{a-b} and t_i^{k1}, t_i^{k2}, ……, t_i^{kh} represent the time instants the node *i* is at these intermediate street intersections respectively. We find the two time instants t_i^{kl} and t_i^{kl+1} such that $t_i^{kl} \le t^s \le t_i^{kl+1}$ where $1 \le l \le h$. The location (x_i^s, y_i^s) of node *i* then at time instant t^s is basically given by:

$$x_i^s = \left[\frac{t_i^s - t_i^{kl}}{t_i^{kl+1} - t_i^{kl}} * x_i^{kl+1}\right] + \left[\frac{t_i^{kl+1} - t_i^s}{t_i^{kl+1} - t_i^{kl}} * x_i^{kl}\right]$$

$$y_i^s = \left[\frac{t_i^s - t_i^{kl}}{t_i^{kl+1} - t_i^{kl}} * y_i^{kl+1}\right] + \left[\frac{t_i^{kl+1} - t_i^s}{t_i^{kl+1} - t_i^{kl}} * y_i^{kl}\right].$$

2.3 Manhattan mobility model

Initially, the nodes are assumed to be randomly placed in the street intersections. The movement of a node is decided one street at a time. To start with, each node has equal chance (i.e., probability) of choosing any of the streets leading from its initial location. After a node begins to move in the chosen direction and reaches the next street intersection, the subsequent street in which the node will move is chosen probabilistically. If a node can continue to move in the same direction or can also change directions, then the node has 50% chance of continuing in the same direction, 25% chance of turning to the east/north and 25% chance of turning to the west/south, depending on the direction of the previous movement. If a node has only two options (this occurs when the node is in one of the four bounding streets of the network), then the node has an equal (50%) chance of exploring either of the two options. If a node has only one option to move (this occurs when the node reaches any of the four corners of the network), then the node has no other choice except to explore that option.

Simulation Methodology: The mobility profile for a node is updated each time the node moves from a street intersection to an adjacent street intersection. The movement is decided as follows: Let (x_i^a, y_i^a) be the street intersection of node *i* at time instant t_i^a. Let $SI - set_i^a$ be the set of all neighbouring street intersections of (x_i^a, y_i^a). If there exists only one entry in $SI - set_i^a$, say $SI - set_i^a$ = [(x_i^E, y_i^E)], then the adjacent street intersection (x_i^b, y_i^b) to which node *i* moves at time instant t_i^b is basically (x_i^E, y_i^E). If there are two possible candidate adjacent street intersections in $SI - set_i^a$, say $SI - set_i^a$ = [(x_i^E, y_i^E), (x_i^F, y_i^F)], then we generate a random number r_i^a from 0 to 1. If r_i^a < 0.5, then we assign (x_i^b, y_i^b) = (x_i^E, y_i^E); otherwise, we set (x_i^b, y_i^b) = (x_i^F, y_i^F). If there are three possible candidate adjacent street intersections in $SI - set_i^a$, say $SI - set_i^a$ = [(x_i^E, y_i^E), (x_i^F, y_i^F), (x_i^G, y_i^G)], where let (x_i^E, y_i^E) be the street intersection that is in the same axis as that of (x_i^a, y_i^a) and let (x_i^F, y_i^F) and (x_i^G, y_i^G) be the street intersections that are not in the same axis as that of (x_i^a, y_i^a). We generate a random number r_i^a from 0 to 1. If r_i^a < 0.5, then we assign (x_i^b, y_i^b) = (x_i^E, y_i^E); if 0.5 ≤ r_i^a < 0.75, we assign (x_i^b, y_i^b) = (x_i^F, y_i^F) and for r_i^a ≥ 0.75, we set (x_i^b, y_i^b) = (x_i^G, y_i^G). We then add the tuple $[t_i^a, t_i^b,\ (x_i^a, y_i^a),\ (x_i^b, y_i^b), v_i]$ to the mobility profile of node *i*, where v_i is the velocity with which node *i* moves from (x_i^a, y_i^a) to (x_i^b, y_i^b). The above process is independently repeated for each node until the simulation time period. We assume the pause time for a node to be zero in all of our simulations.

To determine the location of node *i* at time instant t^s, we index into the sequence of tuples constituting the mobility profile of node *i* and choose the entry whose two time instants t_i^a, t_i^b are such that $t_i^a \le t^s \le t_i^b$. The location (x_i^s, y_i^s) of node *i* at t^s is basically given by:

$$x_i^s = \left[\frac{t_i^s - t_i^a}{t_i^b - t_i^a} * x_i^b\right] + \left[\frac{t_i^b - t_i^s}{t_i^b - t_i^a} * x_i^a\right]$$

$$y_i^s = \left[\frac{t_i^s - t_i^a}{t_i^b - t_i^a} * y_i^b\right] + \left[\frac{t_i^b - t_i^s}{t_i^b - t_i^a} * y_i^a\right].$$

2.4 Gauss-Markov mobility model

Initially, the nodes are placed at random locations in the network. The movement of a node is independent of the other nodes in the network. Each node *i* is assigned a mean speed, $\overline{S_i}$, and mean direction, $\overline{\Theta_i}$ of movement. For every constant time period, a node calculates the speed and direction of movement based on the speed and direction during the previous time period, along with a certain degree of randomness incorporated in the calculation. The node is assumed to move with the calculated speed and in the calculated direction during every fixed time period. For a particular time instant, t_i^{a+1}, the speed and direction of a node *i* is calculated as follows:

$$S_i^{a+1} = \alpha * S_i^a + (1-\alpha) * \overline{S_i} + \sqrt{1-\alpha^2} * S^G\left(t_i^a\right)$$

$$\Theta_i^{a+1} = \alpha * \Theta_i^a + (1-\alpha) * \overline{\Theta_i} + \sqrt{1-\alpha^2} * \Theta^G\left(t_i^a\right)$$

The parameter α (0 ≤ α ≤ 1) is used to incorporate the degree of randomness while calculating the speed and direction of movement for a time period. The degree of randomness decreases as we increase the value of α from 0 to 1. When α is closer to 0, the degree of randomness is high, which may result in sharper turns. When α is closer to 1, the speed and direction during the previous time period are given more importance (i.e., the model is more temporally dependent) and the node prefers to move in a speed and direction closer to what it has been using so far. Thus, the movement of a node gets more linear as the value of α approaches unity. The terms $S^G(t_i^a)$ and

$\Theta^G(t_i^a)$ are random variables chosen independently by each node from a Gaussian distribution with mean 0 and standard deviation 1. If (x_i^a, y_i^a) are the co-ordinates of node *i* at time instant t_i^a, the co-ordinates (x_i^{a+1}, y_i^{a+1}) of the node at time instant t_i^{a+1} are given by:

$$X_i^{a+1} = X_i^a + [S_i^a * \cos(\Theta_i^a)]$$
$$Y_i^{a+1} = Y_i^a + [S_i^a * \sin(\Theta_i^a)]$$

3 Related work

In addition to the mobility models described in Section 2, we now discuss few other relevant mobility models that have also been proposed for ad hoc networks. These include the Random Walk model [6], Random Direction model [6], Random Trip model [13], Freeway model [7] and the Random Point Group (RPG) model [6].

3.1 Review of other mobility models

The Random Walk model is a slight variation of the Random Waypoint model: a node moves in a randomly chosen direction and speed until the network boundary is reached. After reaching the network boundary, the node chooses another random direction and speed to move. The Random Trip mobility model is also a slight variation of the Random Waypoint mobility model: at a trip transition instant, a node selects a random direction, trip duration and speed and moves in the chosen direction with the chosen speed for the chosen trip duration. If the node reaches the boundary of the network during the trip, the node is reflected. After the expiration of the chosen trip duration, another set of random direction, trip duration and speed values is chosen and the movement is continued. In the Random Waypoint, Random Walk and Random Trip mobility models, the random speed selected by a node is chosen from a pre-specified range. In the Random Direction model, a node moves in a randomly chosen direction and travels to the border of the simulation area in that direction. As soon as the network boundary is reached, the node stops for a certain period of time and then moves by choosing another angular direction (between 0 and 180 degrees).

The Freeway model can be used to emulate the movement of nodes in a freeway. A freeway is assumed to comprise of one or more lanes, in both directions. The movement of a node is however restricted to a particular lane of the freeway and there can be no lane changes. The velocity of a node *i* at time instant *t*+1, V_i^{t+1} is dependent on the velocity of the node at time instant *t*, V_i^t and the maximum possible acceleration/deceleration for a unit time, a_i^t. If node *i* travels behind node *j* and the distance between them at time instant *t* is within the Safety Distance (SD), then $V_i^t < V_j^t$. Thus, the Freeway model has high spatial and temporal dependence.

The RPG mobility model works as follows: Nodes move as a group with each group having a group leader (a logical centre for the group) whose movement determines the group's mobility pattern. Because of this property, the mobility pattern of the nodes under the RPG model is expected to have high spatial dependence. Initially, each group member is assumed to be uniform-randomly distributed in the neighbourhood of the group leader. For every time instant, the speed and direction of a node is derived by randomly deviating from that of the group leader.

3.2 Literature review

In [14], the authors study the impact of the three Randomness-based mobility models (Random Waypoint, Random Direction and the Random Walk models) on the minimum-hop based Ad hoc On-demand Distance Vector (AODV) [15] routing protocol. Simulation results show that AODV incurred the lowest hop count for its routes when simulated under the Random Waypoint mobility model. The Random Direction mobility model generated the maximum routing overhead (ratio of the number of control packets sent to the number of data packets sent) and the Random Waypoint model generated the minimum routing overhead. The Random Waypoint mobility model also yielded the maximum packet delivery ratio and throughput.

In [16], the authors compare the performance of the proactive Destination Sequenced Distance Vector (DSDV) routing protocol [17] and the on-demand Dynamic Source Routing (DSR) [18] and the AODV routing protocols with respect to the Random Waypoint and Random Trip mobility models. The end-to-end delay per data packet for all the three routing protocols was found to be lower when simulated with the Random Waypoint model compared to the Random Trip model.

In [19], the authors compare the performance of DSR and DSDV protocols under the Random Waypoint model, Random Point Group model, Freeway model and the Manhattan model. The throughput obtained with DSR was greater than that obtained with DSDV under all the four mobility models. For smaller node velocities, DSR yielded a higher throughput under the Freeway mobility model. For moderate and higher node velocities, DSR yielded a higher throughput under the Random Waypoint model, closely followed by the Manhattan model. As the mobility of the nodes increased, the throughput of both DSR and DSDV decreased drastically under the Freeway model.

In [20], the authors study the impact of the Random Waypoint model, the RPG mobility model and its variants on the performance of DSDV, AODV and DSR. DSR had the highest packet delivery ratio, followed by DSDV and then AODV. Routes under the RPG model are more likely to exist for a longer time as nodes are in the close vicinity of each other. Hence, the routing protocols incurred a lower routing overhead as well as a larger throughput with the RPG model compared to the Random Waypoint model. The packet delivery ratio of the routing protocols was strongly influenced by the distribution of the nodes within the network. There were erratic variations of the packet delivery ratios of the routing protocols with respect to the RPG model and its variants; whereas, there was less variation in the packet delivery ratios of the routing protocols under the Random Waypoint model. Under the Random Waypoint model, nodes are more likely to be homogeneously distributed throughout the network all the time.

In [21], the authors compare the performance of DSR, DSDV and the AODV under the Random Waypoint model, Freeway model, Manhattan model and the RPG model. Each of the three routing protocols achieved the highest throughput and the least overhead with the RPG model and incurred high overhead and low throughput with both the Freeway and Manhattan models. The relative rankings of the routing protocols with respect to different performance metrics varies with the mobility model used.

3.3 Motivation for our research

All of the above work study the performance impact of the mobility models on the minimum-hop based routing protocols. None of the work studied the performance impact of the mobility models on the stability-based routing algorithms and protocols. As a first step in this direction, in this paper, we study the impact of the Random Waypoint, City Section model, Manhattan and the Gauss Markov mobility models on the *OptPathTrans* algorithm (algorithm to find the sequence of most stable paths in an ad hoc network) and compare the performance with that of the algorithm to find the sequence of minimum hop paths. The next step in our research would be to study the impact of the four mobility models on the stable path MANET routing protocols such as the Flow-Oriented Routing Protocol (FORP) [22], Associativity Based Routing (ABR) protocol [23] and the Route-lifetime Assessment Based Routing Protocol (RABR) [24] and compare their performance with that of the minimum-hop based routing protocols such as DSR, AODV and DSDV.

4 Algorithm to determine the optimal number of path transitions

In this section, we briefly review the *OptPathTrans* algorithm, recently proposed by us in [4], to determine the optimal number of path transitions in ad hoc networks. The algorithm uses the notions of mobile graph to record the sequence of network topology changes and mobile path to record the sequence of paths in a mobile graph.

4.1 Mobile graph

A mobile graph [11] is defined as the sequence $G_M = G_1G_2 \ldots G_T$ of static graphs that represents the network topology changes over some time scale *T*. In the simplest case, the mobile graph $G_M = G_1G_2 \ldots G_T$ can be extended by a new instantaneous graph G_{T+1} to a longer sequence $G_M = G_1G_2 \ldots G_T\, G_{T+1}$, where G_{T+1} captures a link change (either a link comes up or goes down). But such an approach has very poor scalability. In this paper, we sample the network topology periodically for every 0.25 seconds, which could, in reality, be the instants of data packet origination at the source.

4.2 Mobile path

A *mobile path* [11], defined for a source-destination (*s-d*) pair, in a mobile graph $G_M = G_1G_2 \ldots G_T$ is the sequence of paths $P_M = P_1P_2 \ldots P_T$, where P_i is a static path between the same *s-d* pair in $G_i = (V_i, E_i)$, V_i is the set of vertices and E_i is the set of edges connecting these vertices at time instant t_i. That is, each static path P_i can be represented as the sequence of vertices $v_0v_1 \ldots v_l$, such that $v_0 = s$ and $v_l = d$ and $(v_{j-1},v_j) \in E_i$ for $j = 1,2, \ldots, l$. The timescale of *T* normally corresponds to the duration of a session between *s* and *d*.

The Stable Mobile Path (SMP) for a given mobile graph and *s-d* pair is the sequence of static *s-d* paths such that the number of route transitions (change from one static *s-d* path to another) is as minimum as possible. In other words, the constituent static paths of an SMP have the longest possible route lifetime. A Minimum Hop Mobile Path (MHMP) for a given mobile graph and *s-d* pair is the sequence of minimum hop static *s-d* paths. The SMP for an *s-d* pair on a given mobile graph is determined by using algorithm *OptPathTrans*. The MHMP for an *s-d* pair on a given mobile graph is determined by repeatedly running the minimum path weight Dijkstra algorithm on the static graphs. We follow the Least Overhead Routing Approach (LORA) [12] for ad hoc networks. Accordingly, a minimum hop *s-d* path determined by running *Dijkstra* algorithm on a static graph G_i is assumed to be used in the subsequent static graphs G_{i+1}, G_{i+2}, …., as long as the path exists in these static graphs.

4.3 Algorithm *OptPathTrans*

Algorithm *OptPathTrans* (pseudo code given in Figure 1) operates on the following greedy strategy: Whenever a path is required, select a path that will exist for the longest time. Let $G_M = G_1G_2 \ldots G_T$ be the mobile graph generated by sampling the network topology at regular instants t_1, t_2, …, t_T of an *s-d* session. When an *s-d* path is required at sampling time instant t_i, the strategy is to find a mobile sub graph $G(i, j) = G_i \cap G_{i+1} \cap \ldots \cap G_j$ such that there exists at least one *s-d* path in $G(i, j)$ and no *s-d* path exists in $G(i, j+1)$. A minimum hop *s-d* path in $G(i, j)$ is selected. Such a path exists in each of the static graphs G_i, G_{i+1}, …, G_j. If sampling instant $t_{j+1} \leq t_T$, the above procedure is repeated by finding the *s-d* path that can survive for the maximum amount of time since t_{j+1}. A sequence of such maximum lifetime static *s-d* paths over the timescale of a mobile graph G_M forms the stabile mobile *s-d* path in G_M. The run-time complexity of the algorithm is $O(n^2T)$, where *n* is the number of nodes in the network and *T* is the number of static graphs in a mobile graph (*T* is thus a measure of the timescale of the network communication session between the source *s* and destination *d*).

```
Input: $G_M = G_1 G_2 \dots G_T$, source $s$, destination $d$
Output: $P_S$          // Stable-Mobile-Path
Auxiliary Variables: $i, j$
Initialization: $i$=1; $j$=1; $P_S = \Phi$

Begin OptPathTrans

while ($i \le T$) do
    Find a mobile graph $G(i, j) = G_i \cap G_{i+1} \cap \dots \cap G_j$ such that there exists at least one s-d path in $G(i, j)$ and {no s-d path exists in $G(i, j+1)$ or $j = T$}
    $P_S = P_S$ U {minimum hop s-d path in $G(i, j)$ }
    $i = j + 1$
end while

return $P_S$

End OptPathTrans
```

Figure 1: Pseudo code for algorithm *OptPathTrans.*

5 Simulations

The network dimensions are 1000m x 1000m. The Random Waypoint and Gauss-Markov mobility models work by assuming an open network field without any grid constraints. For the City Section and Manhattan mobility models, we assume the network is divided into grids: square blocks of length (side) 100m. The network for the two VANET mobility models is thus basically composed of a number of horizontal and vertical streets. Each street has two lanes, one for each direction (north and south direction for vertical streets, east and west direction for horizontal streets). A node is allowed to move only along the grids of horizontal and vertical streets. The wireless transmission range of a node is 250m. The network density is varied by performing the simulations with 25 (low), 50 (moderate) and 75 (high) nodes. The node velocity values used for each of the four mobility models are 5 m/s (about 10 miles per hour), 15 m/s (about 35 miles per hour) and 30 m/s (about 65 miles per hour), representing levels of low, moderate and high node mobility respectively. For the Gauss-Markov mobility model, these velocity values represent the mean speed $\overline{S}$ for a node and for the other three mobility models, these values represent the velocity of every node in the network. Note that, in the case of the Random Waypoint model, for a given mobility level, we let all the nodes in the network to move in the same velocity by letting $v_{min} = v_{max}$. The pause time is 0 seconds; so, all the nodes are constantly in motion.

In the case of the Gauss-Markov mobility model, each node is assigned a random value for the mean direction of movement, $\overline{\Theta}$, chosen from the range [0…360˚]. When a node travels beyond the boundaries of the simulation field, the mean direction of movement of the node is forced to flip 180 degrees so that the node can remain within the boundary of the simulation field. The constant time period for updating the speed and direction of movement of the nodes under the Gauss-Markov mobility model is 1 second.

We obtain a centralized view of the network topology by generating mobility trace files for 1000 seconds under each of the four mobility models. The network topology is sampled for every 0.25 seconds to generate the static graphs and the mobile graph. Two nodes *a* and *b* are assumed to have a bi-directional link at time *t*, if the Euclidean distance between them at time *t* (derived using the locations of the nodes from the mobility trace file) is less than or equal to the wireless transmission range of the nodes. Each data point in Figures 2 through 12 is an average computed over 5 mobility trace files and 20 randomly selected *s-d* pairs from each of the mobility trace files. The starting time of each *s-d* session is uniformly distributed between 1 to 20 seconds.

5.1 Performance metrics

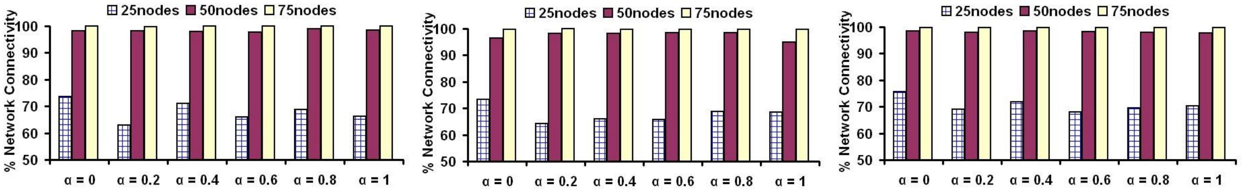
The following performance metrics are evaluated:

- *Percentage Network Connectivity*: Percentage network connectivity indicates the probability of finding an *s-d* path between any two nodes in the network for a given network density and level of node mobility. Measured over all the *s-d* sessions, this metric is the ratio of the number of static graphs in which there is an *s-d* path to the total number of static graphs in the mobile graph.
- *Average Route Lifetime*: The average route lifetime is the average of the lifetime of all the static paths of an *s-d* session, averaged over all the *s-d* sessions.
- *Average Hop Count*: The average hop count is the time averaged hop count of a mobile path for an *s-d* session, averaged over all the *s-d* sessions. The time averaged hop count for an *s-d* session is measured as the sum of the products of the number of hops for the static *s-d* paths and the corresponding lifetime of the static *s-d* paths divided by the number of static graphs in which there existed a static *s-d* path. For example, if a mobile path comprises of a 2-hop static path p_1, a 3-hop static path p_2, and a 2-hop static path p_3, existing in static graphs 1-2, 3-5 and 6-10 respectively, then the time-averaged hop count of the mobile path would be (2*2 + 3*3 + 2*5) / 10 = 2.3.

Note that for a given condition of network density and node mobility, the values reported for the performance metrics under the Gauss-Markov mobility model in figures 7 through 12 are the average of the performance metric values obtained for different values of α.

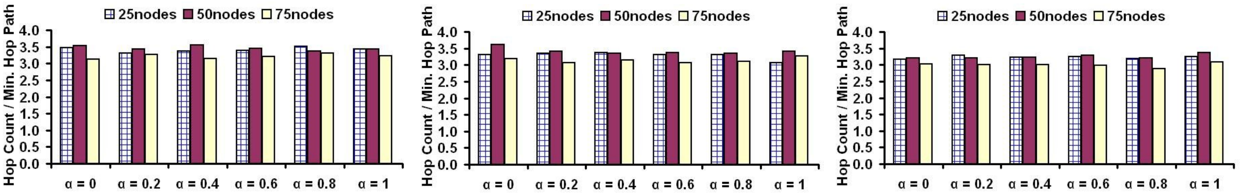
5.2 Impact of the degree of randomness parameter in the Gauss-Markov mobility model

In moderate and high-density networks, there is no significant influence of parameter α on network connectivity. As we add more nodes, randomness or the absence of it, has no significant impact on network connectivity. The network connectivity obtained in low-density networks seems to slightly depend on the value of α used. As seen in Figure 2, the highest network connectivity for low-density networks is obtained when α = 0, corresponding to the scenario in which the mobility of the nodes is totally random and not dependent on the past history. As nodes move randomly without restricting to a fixed path, the connectivity of the nodes can be guaranteed for a longer time, especially in low-density networks. Otherwise, if nodes move on a linear path for a long time, it is likely that they fall out of the range of each other after a while, thus reducing the connectivity of any two nodes in the network. Nevertheless, the standard



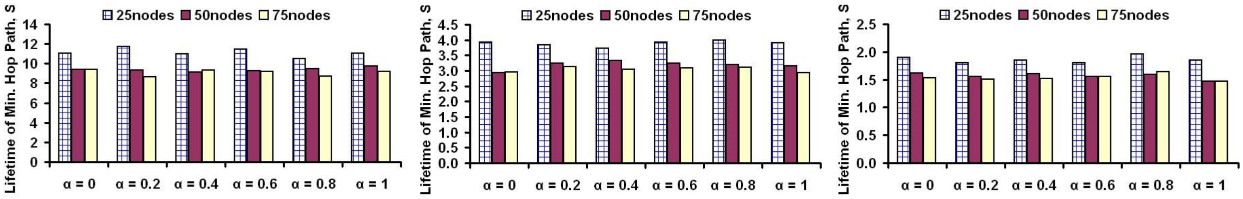
2.1: Node Velocity = 5m/s **2.2:** Node Velocity = 15m/s **2.3:** Node Velocity = 30m/s

Figure 2: Gauss-Markov Model: Percentage Network Connectivity Vs Degree of Randomness



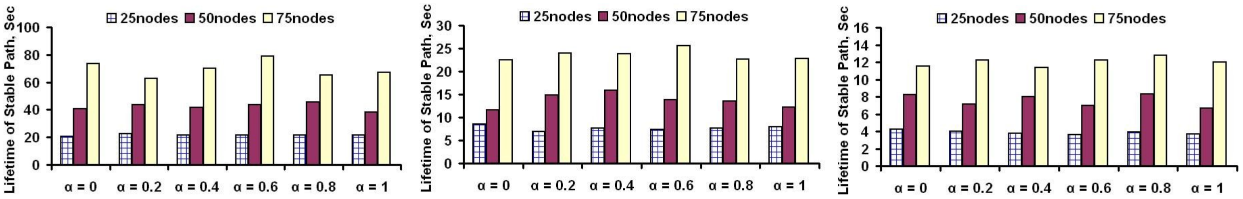
3.1: Node Velocity = 5m/s **3.2:** Node Velocity = 15m/s **3.3:** Node Velocity = 30m/s

Figure 3: Gauss-Markov Model: Average Hop Count per Minimum Hop Path Vs Degree of Randomness.



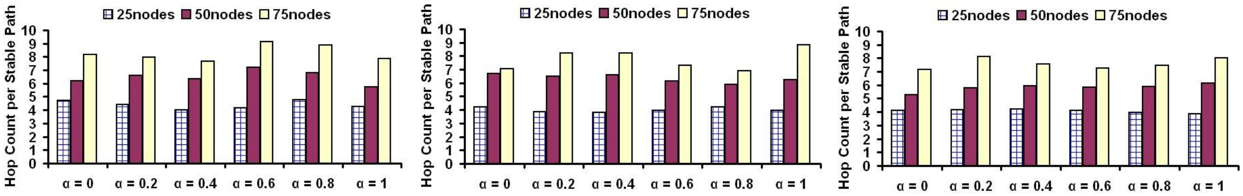
4.1: Node Velocity = 5m/s **4.2:** Node Velocity = 15m/s **4.3:** Node Velocity = 30m/s

Figure 4: Gauss-Markov Model: Average Lifetime per Minimum Hop Path Vs Degree of Randomness.



5.1: Node Velocity = 5m/s **5.2:** Node Velocity = 15m/s **5.3:** Node Velocity = 30m/s

Figure 5: Gauss-Markov Model: Average Lifetime per Stable Path Vs Degree of Randomness.



6.1: Node Velocity = 5m/s **6.2:** Node Velocity = 15m/s **6.3:** Node Velocity = 30m/s

Figure 6: Gauss-Markov Model: Average Hop Count per Stable Path Vs Degree of Randomness.

deviation of the network connectivity in low-density networks is not more than 5% of the mean network connectivity obtained for different α values.

With respect to the influence of the degree of randomness parameter α on the hop count of the minimum hop paths, we observe that the maximum difference in the minimum hop count of the paths obtained for different values of α for a given condition of node mobility and network density is 0.3, while the average hop count of the paths under the Gauss-Markov model is in the range of 3.0 to 3.6. In more statistical terms, the standard deviation of the minimum hop counts is not even 3% of the average of the minimum hop counts obtained for different values of α. So, we can very well conclude that the degree of randomness in the Gauss-Markov model does not significantly influence the hop count of the minimum hop paths in the network. We also observe that the lifetime of the minimum hop paths do not significantly differ from each other for different values of α. Similar to the observation made for the hop count, the standard deviation of the lifetime of the minimum hop paths is not even 4% of the average of the lifetime of the minimum hop paths obtained for different values of α. The above observations are consistent with the simulation results obtained in [9], wherein it has been concluded that the degree of randomness parameter in the Gauss-Markov model has no significant influence on the throughput and end-to-end delay per data packet.

With respect to the influence of the degree of randomness parameter α on the lifetime of the stable paths, we observe that it is possible to determine long-living routes by adopting larger intermediate values of α (i.e., α values of 0.6 to 0.8), when the model is considered to give higher weight to the speed and direction of movement in the previous time period while determining the speed and direction of movement for the subsequent time period. But, there is still a degree of randomness associated with the determination of speed and direction (note that we are not letting α to approach 1). Such a small level of randomness is required to occasionally deflect nodes from their linear path so that the nodes continue to remain as neighbours with a larger probability and do not move far away from each other. If the nodes strictly take a linear path, it is possible for them to move away from each other relatively sooner. A certain degree of randomness is required for the nodes to get deflected and move in the vicinity of each other for relatively little longer. From Figures 6.1 through 6.3, we can observe that the lifetime of the stable routes determined with α value of 0.6 can be sometimes about 20% more than the lifetime of the stable routes determined with α value of 1.0. For lower values of α, the degree of randomness increases and the number of link changes increases significantly. Nevertheless, the lifetime of stable routes determined for lower values of α is not significantly lower than those obtained for α values in the range of 0.6 to 0.8 and is almost close to the lifetime values obtained when α is unity. The standard deviation of the lifetimes of stable routes is still below 10% of the mean of the lifetime of the stable routes computed over the different values of α for a particular condition of network density and node mobility.

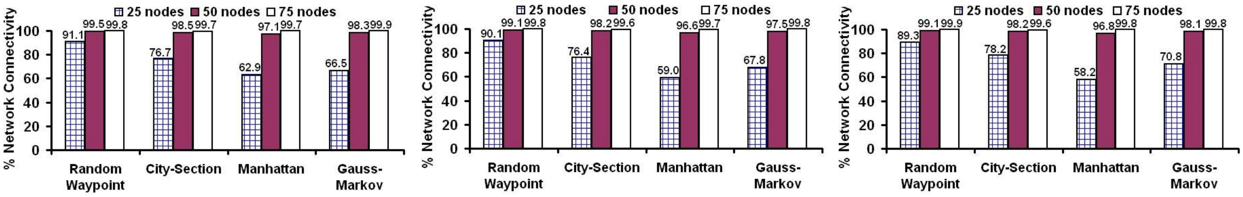
Similarly, the standard deviation of the hop counts of stable routes is within 10% of the mean of the hop counts of stable routes computed over different α values for a given network density and node mobility.

The above observations justify the usage of the average of the values of a performance metric for different values of α as a measure of the performance under the Gauss-Markov mobility model in figures 7 through 12 that compare the performance under the different mobility models.

5.3 Network connectivity

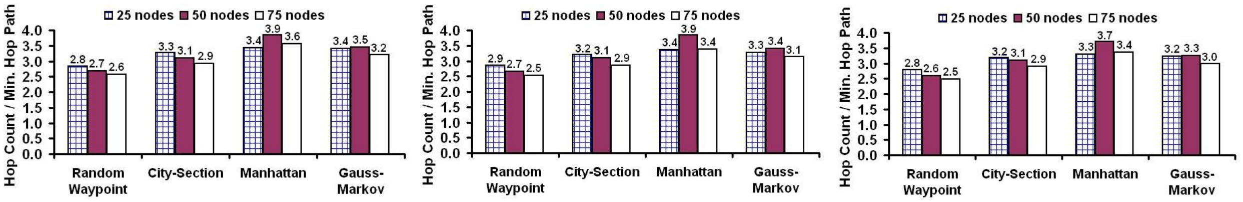
The Random Waypoint mobility model provided the maximum network connectivity for any combination of network density and node mobility. In low-density and moderate-density networks, for all levels of node mobility, the RWP model is the only mobility model to provide a connectivity of at least 90% and 99% respectively. In high-density networks, for any level of node mobility, each of the four mobility models provided connectivity of at least 99.5%. The following ranking can be observed for the four mobility models in decreasing order of network connectivity in low and moderate density networks: Random Waypoint model, City Section model, Gauss-Markov model and the Manhattan model. Note that the network connectivity values plotted in Figures 7.1 through 7.3 is an average of those obtained for the Minimum Hop Mobile Path and Stable Mobile Path.

In low density networks, the lower network connectivity obtained with the two VANET mobility models can be attributed to the constrained motion of the nodes only along the streets of the network. In the case of the Manhattan mobility model, the probabilistic nature of direction selection after reaching each street intersection is also a reason behind the lowest network connectivity observed for this mobility model among all the four mobility models. The number of nodes distributed in the streets of the network may not be sufficient enough to connect any pair of source-destination nodes all the time. In the case of the Gauss-Markov mobility model, the direction of movement of the nodes is restricted close to the initially assigned mean direction of movement. Note that, we assign each node a mean direction of movement randomly chosen from [0…360˚]. But, still, when there are few nodes in the network, the restricted movement of the nodes close to the mean direction of movement is a limiting factor for network connectivity. As we increase the number of nodes in the network, both the VANET mobility models and the Gauss-Markov mobility model demonstrate a significant increase in network connectivity, for all levels of node mobility. This illustrates the fact that the randomness associated with the mobility models assures that any pair of nodes will remain connected, provided we have at least a reasonably larger number of nodes (like moderate density networks), irrespective of the different levels of node mobility.



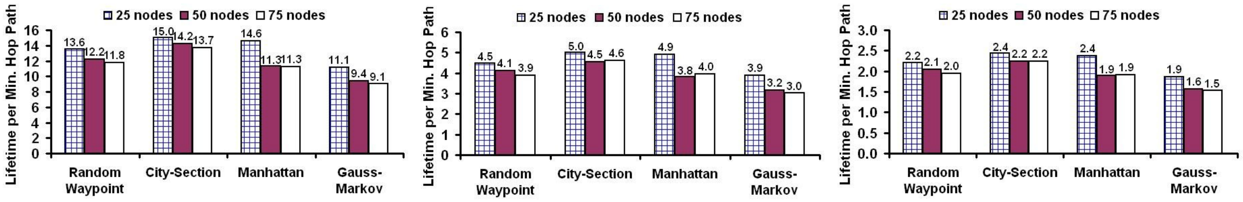
7.1: Node Velocity = 5m/s **7.2:** Node Velocity = 15m/s **7.3:** Node Velocity = 30m/s

Figure 7: Percentage Network Connectivity under the Different Mobility Models.



8.1: Node Velocity = 5m/s **8.2:** Node Velocity = 15m/s **8.3:** Node Velocity = 30m/s

Figure 8: Average Hop Count per Minimum Hop Path under the Different Mobility Models.



9.1: Node Velocity = 5m/s **9.2:** Node Velocity = 15m/s **9.3:** Node Velocity = 30m/s

Figure 9: Average Lifetime per Minimum Hop Path under the Different Mobility Models.

5.4 Minimum hop mobile path

We now discuss the time averaged hop count per minimum hop path (refer Figure 8) and the average route lifetime (refer Figure 9) of the minimum hop paths determined as the constituent paths of the Minimum Hop Mobile Path under the four mobility models.

5.4.1 Average hop count per minimum hop path

The average hop count per minimum hop path determined for the two VANET mobility models and the Gauss-Markov model is considerably larger than the hop count per minimum hop path determined for the Random Waypoint mobility model. The relatively larger hop count can be attributed to the constrained mobility of the nodes under these three mobility models. The minimum hop paths in the street networks are most likely not to exist on or close to the straight line between source and destination nodes. Similarly, due to the temporal dependency associated with the Gauss-Markov mobility model, one cannot always find minimum hop paths lying on a straight line connecting the source and destination nodes. Based on our observations in Figures 8.1 through 8.3, we can arrive at the following ranking of the four mobility models in the increasing order of the magnitude of the hop count for the minimum hop paths: Random Waypoint model, City Section model, Gauss-Markov model and the Manhattan model. This ranking holds good for all the simulated conditions of network density and levels of node mobility.

For a given node velocity, the average hop count per minimum hop path under the City Section mobility model, Gauss-Markov mobility model and the Manhattan mobility model is respectively about 14%, 17% and 19% more than that incurred for the Random Waypoint mobility model in low-density networks. In moderate and high-density networks, the average hop count per minimum hop path under the City Section, Gauss-Markov and Manhattan mobility models is respectively about 18%, 25% and 40% more than that incurred for the Random Waypoint mobility model. We also observe that with increase in network density, the average hop count per minimum hop path for the Random Waypoint mobility model, City Section mobility model and the Gauss-Markov mobility model decreases (by a factor of 5%-10%). On the other hand, with increase in network density, the average hop count per minimum hop path for the Manhattan mobility model remained the same or even sometime increases up to 14%. This can be attributed to the significant increase in the network connectivity for the Manhattan mobility model with increase in network density, but at the cost of increase in hop count. Given the extremely constrained and random nature of node movement under the Manhattan mobility model, in order to connect the source and destination, more intermediate nodes have to be accommodated in the source-destination paths.

5.4.2 Average lifetime per minimum hop path

When we aim for minimum hop count in the paths and determine the Minimum Hop Mobile Path by repeated application of the Dijkstra algorithm on the static graphs, we observe that the minimum hop paths determined under the City Section mobility model are relative more stable (i.e., have a larger route lifetime) compared to the minimum hop paths determined under the other three mobility models. We even observe that the minimum hop paths determined under the Manhattan mobility model in low-density networks are relatively more stable than those determined under the Random Waypoint and the Gauss-Markov mobility models. In low-density networks, the average lifetime of the minimum hop paths determined under the Manhattan mobility model is only about 3% less than the average lifetime of the minimum paths determined under the City Section mobility model. But, as we increase the network density, the number of hops in the minimum hop paths determined under the Manhattan mobility model increased to provide better network connectivity. However, such minimum hop paths have been observed to be relatively unstable.

For a given level of node mobility, in low-density networks, the average lifetime of the minimum hop routes determined under the Manhattan model, Random Waypoint model and the Gauss-Markov model is about 3%, 10% and 25% less than the average lifetime of minimum hop routes determined under the City Section mobility model. But, as we increase the network density, we observe that the magnitude of this difference increases further. For a given level of node mobility, in moderate and high-density networks, the average lifetime of the minimum hop routes determined under the Manhattan model, Random Waypoint model and the Gauss-Markov model is about 10%-15%, 10%-15% and 30%-34% less than the average lifetime of minimum hop routes determined under the City Section mobility model. The relatively poor lifetime of minimum hop routes determined under the Gauss-Markov mobility model can be attributed to the temporal dependency of the nodes in choosing their direction of movement. When we attempt to optimize the number of hops, it may be possible to obtain paths with lower hop count. But, such paths may have links whose constituent nodes are on the verge of moving away from each other, travelling in different directions.

We also observe that with increase in network density, the average lifetime of the minimum hop routes decreases for all the four mobility models. This can be attributed to the decrease in the hop count of the minimum hop routes as we increase the number of nodes in the network. It gets possible to reduce the number of hops by choosing the intermediate nodes from a larger pool of available nodes. But, the physical distance between the constituent nodes of the links in such minimum hop paths tends to be close to or even more than 80% of the transmission range of the nodes at the time of route selection itself. Such routes are bound not to last longer. Thus, with a small reduction in the minimum hop count of the paths, we incur a significant reduction in the lifetime of the minimum hop routes. Note that for the Manhattan mobility model, even though the hop count of the minimum hop routes increases with increase in network density, the lifetime of the minimum hop routes decreases with increase in network density. In the case of the Manhattan mobility model, the network connectivity was very low in low-density networks. So, as we increased the network density, the network connectivity improved but more intermediate nodes have to be added, even if we aim for minimum hop paths.

The following ranking can be assigned for the four mobility models in the decreasing order of lifetime of minimum hop routes for the different network densities, irrespective of the level of node mobility:

- *Low-density networks*: City Section model, Manhattan model, Random Waypoint model, Gauss-Markov model
- *Moderate and High-density networks*: City Section model, Random Waypoint model, Manhattan model and Gauss-Markov model

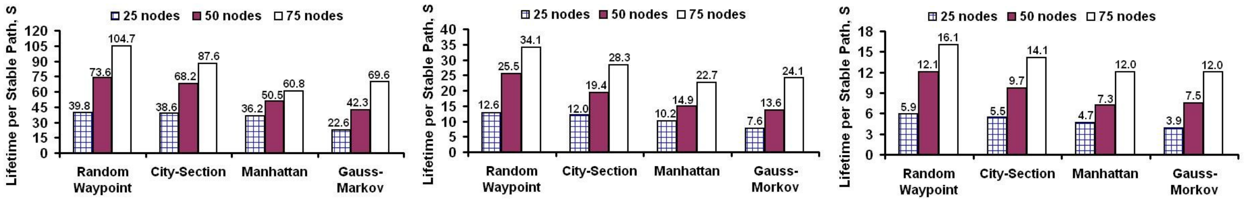
5.5 Stable mobile path

We now discuss the average route lifetime (refer Figure 10) per stable path and the time averaged hop count per stable path (refer Figure 11) as we determine the constituent stable paths of the Stable Mobile Path under all the four mobility models.

5.5.1 Average lifetime per stable path

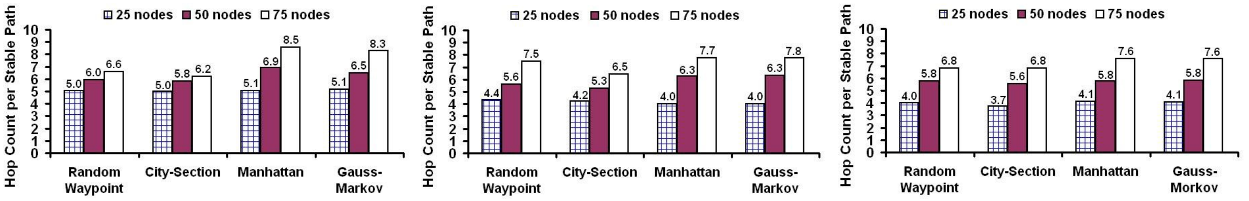
For all the four mobility models, algorithm *OptPathTrans* was able to determine stable paths by adding more intermediate nodes to the path such that the average physical distance between the constituent nodes of the links in the stable path is only about 60%-70% of the transmission range of the nodes. Thus, the long-living routes are determined at the cost of an increase in the hop count. This is a tradeoff that cannot be avoided. We also observe that for each of the four mobility models, for a given level of node mobility, the lifetime of the stable routes increased as we increase the network density. This is because algorithm *OptPathTrans* gets a larger pool of nodes to select from by looking ahead at the future. Nevertheless, we do see an increase in the hop count of the stable routes in moderate and high-density networks vis-à-vis low-density networks, an observation again vindicating the tradeoff between route lifetime and hop count in ad hoc networks.

Overall, the lifetime of stable routes obtained under the Random Waypoint model was the largest in all of the simulated scenarios. This can be attributed to the unconstrained mobility of the nodes and algorithm *OptPathTrans* faces no restrictions in choosing the intermediate nodes that form the stable paths. With the other three mobility models, due to the constrained mobility of the nodes in one way or the other, algorithm *OptPathTrans* faces restrictions in choosing the intermediate nodes for the stable paths. As a result, the stable paths determined under these mobility models with restricted node movement have a relatively lower



10.1: Node Velocity = 5m/s **10.2:** Node Velocity = 15m/s **10.3:** Node Velocity = 30m/s

Figure 10: Average Lifetime per Stable Path under the Different Mobility Models.



11.1: Node Velocity = 5m/s **11.2:** Node Velocity = 15m/s **11.3:** Node Velocity = 30m/s

Figure 11: Average Hop Count per Stable Path under the Different Mobility Models

lifetime when compared to those incurred under the RWP model.

The following rankings can be assigned for the four mobility models in the decreasing order of the lifetime of the stable paths for the different network density and node mobility scenarios:

- *High network density, High node mobility*: Random Waypoint model, City Section model, Gauss-Markov model and Manhattan model
- *All other conditions of network density and node mobility*: Random Waypoint model, City Section model, Manhattan model and Gauss-Markov model

In high-density networks with high level of node mobility, the Gauss-Markov model yielded stable routes with a slightly larger lifetime than the Manhattan mobility model. This can be reasoned as follows: In high-density networks, due to the temporal dependency on the direction of movement of the nodes, it is possible to find stable paths involving nodes that are travelling in the same direction or at least not moving away from each other. With the Manhattan model, it is less likely that the algorithm *OptPathTrans* can find nodes travelling in the same direction as the direction of movement of the nodes is probabilistically decided at each street intersection.

As we increase network density from low to high, the two VANET mobility models experience a relatively smaller increase in the lifetime of stable routes compared to the Random Waypoint and Gauss-Markov mobility models. This can be attributed to the constrained mobility of the nodes in the street networks. As discussed above, the Gauss-Markov model makes the most use of increase in the number of nodes in the network, as it becomes increasingly possible to find nodes travelling in a similar direction. Since the nodes moving under Gauss-Markov model are temporally dependent on the previous direction of movement and the mean direction of movement, algorithm *OptPathTrans* has good chance of finding a stable path that will involve nodes travelling in a similar direction and that the link between the constituent nodes of such a path will exist for a relatively longer time.

5.5.2 Average hop count per stable path

As discussed in Section 4.5.1, algorithm *OptPathTrans* determines paths with a relatively longer lifetime than the minimum hop paths, but the hop count of such stable paths is larger than the minimum hop count possible for the same operating conditions of network density and node mobility. Such a lifetime-hop count tradeoff exists for all the four mobility models.

The hop count of the stable routes determined under the City Section mobility model is the smallest for most of the operating scenarios. The hop count of the stable routes determined under the Random Waypoint mobility model is larger than those determined under the City Section mobility model by at most 10%-15%. But as we observed in Section 4.5.1, algorithm *OptPathTrans* was able to find long-living routes under the Random Waypoint model. The lifetime of such stable routes can be as large as 30% compared to those incurred with the City Section mobility model. This observation again illustrates the tradeoff between lifetime and hop count. The hop count of the stable routes determined with the Gauss-Markov and the Manhattan mobility models are almost the same for most of the operating scenarios, with the Manhattan mobility model having a slightly larger hop count in networks with low node mobility. But the hop count of the stable routes determined under these two models is considerably larger than those determined under the City Section and the Random Waypoint mobility models.

As we increase the level of node mobility, the lifetime of the links decreases, forcing algorithm *OptPathTrans* to determine stable routes with a relatively smaller lifetime compared to those determined in networks of low node mobility. The hop count of the stable routes determined under conditions of high node mobility does mostly get smaller compared to those determined under conditions of low node mobility.

Nevertheless, as we increase the level of node mobility from 5 m/s to 30 m/s, for all of the four mobility models, the reduction in the hop count of stable paths is more in low-density networks compared to that incurred in high-density networks. This is because, as we increase the level of node mobility, the rate of decrease in the lifetime of stable routes in low-density networks is relatively larger than the rate of decrease of the lifetime of stable routes in high-density networks under each of the four mobility models.

5.6 Route lifetime – hop count tradeoff

We observe a tradeoff between the objectives of optimizing the route lifetime and the hop count per path for all the four mobility models. Both of these performance metrics cannot be optimized at the same time. For a given simulation condition of network density and node mobility, the average hop count of a Minimum Hop Mobile Path is smaller than the average hop count of a Stable Mobile Path; the average route lifetime of a Stable Mobile Path is more than the average route lifetime of a Minimum Hop Mobile Path. We capture the route lifetime-hop count tradeoff in terms of the lifetime ratio and the hop count ratio. The lifetime ratio is the ratio of the average lifetime per stable path in the Stable Mobile Path to that of the average lifetime per minimum hop path in the Minimum Hop Mobile Path. The hop count ratio is the ratio of the average hop count per stable path in the Stable Mobile Path to that of the average hop count per minimum hop path in the Minimum Hop Mobile Path. The range of values of the lifetime ratios and hop count ratios observed for the four mobility models in low and high-density networks is illustrated in Figures 12.1 and 12.2.

One can observe from the figures that largest values for both the lifetime and hop count ratios are incurred with the Random Waypoint mobility model. This illustrates that under the Random Waypoint mobility model, it is possible to determine stable paths with relatively a very longer lifetime compared to the lifetime of the minimum hop paths, but there is a corresponding increase in the hop count of the stable paths. Nevertheless, the increase in the hop count ratio is not proportional and is sub-linear compared to the increase in the lifetime ratio. For example, the lifetime ratio for RWP model is in the range of 2.6-2.9 and 8.2-8.9 in low and high density networks respectively, where as the hop count ratio for the RWP model is in the range of 1.4-1.75 and 2.5-2.75 in low and high-density networks respectively.

The lifetime ratio for the Gauss-Markov mobility model is the lowest of all the four mobility models in low-density networks. But, in high-density networks, we observe that the lifetime ratio for the Gauss-Markov mobility model is above that of the City Section and Manhattan mobility models. This supports our earlier reasoning that in high-density networks, algorithm *OptPathTrans* manages to find more nodes travelling in the same direction under the Gauss-Markov mobility model and hence can find long-living stable paths involving those nodes. Of course, there is a corresponding increase in the hop count ratio for the Gauss-Markov mobility model in high-density networks. The increase in the lifetime ratio for the City Section and Manhattan mobility models is relatively modest and this can be attributed to the constrained mobility of the nodes in the street networks. Similar to the RWP model, the increase in the hop count ratio for the other three mobility models is also sub-linear compared to the increase in the lifetime ratio.

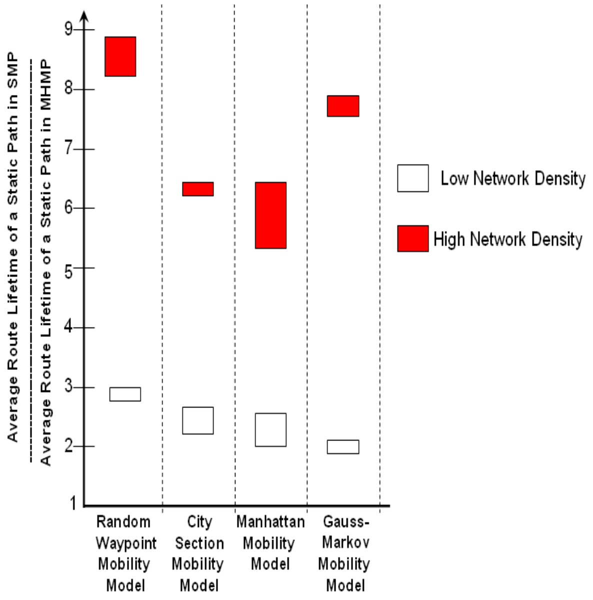


Figure 12.1: Lifetime Ratio.

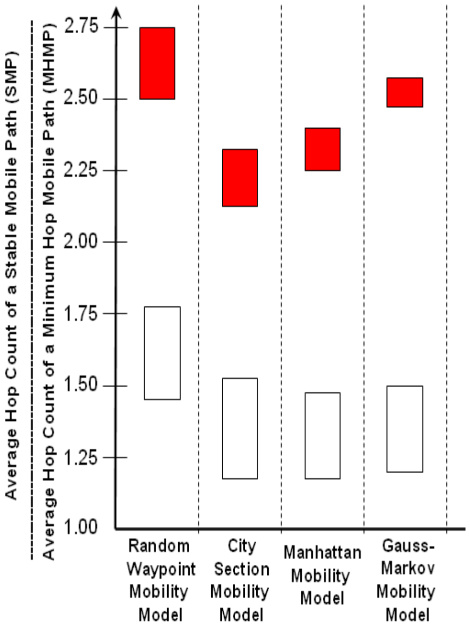


Figure 12.2: Hop Count Ratio.

Figure 12: Route Lifetime – Hop Count Tradeoff.

6 Summary of results

The performance of the Gauss-Markov mobility model for different values of the degree of randomness parameter α can be summarized as follows: In moderate and high-density networks, there is no significant influence of parameter α on network connectivity. The highest network connectivity for low-density networks is obtained when α = 0, corresponding to the scenario in which the mobility of the nodes is totally random and not dependent on the past history. The degree of randomness does not significantly influence the hop count and lifetime of the minimum hop paths in the network. We also observe that one can determine long-living routes by adopting larger intermediate values of α (i.e., α values of 0.6 to 0.8).

The following ranking can be observed for the four mobility models in the decreasing order of network connectivity in low and moderate density networks: Random Waypoint model, City Section model, Gauss-Markov model and the Manhattan model. In high-density networks, for any level of node mobility, all the four mobility models provided a connectivity of at least 99.5%. As we increase the number of nodes in the network, both the VANET mobility models and the Gauss-Markov mobility model demonstrate a significant increase in network connectivity, for all levels of node mobility.

The minimum hop paths in the street networks are most likely not to exist on or close to the straight line between the source and destination nodes. Similarly, due to the temporal dependency associated with the Gauss-Markov mobility model, one cannot find minimum hop paths lying on a straight line connecting the source and destination nodes. For any condition of network density and node mobility, we can arrive at the following ranking for the four mobility models in the increasing order of the magnitude of the hop count for the minimum hop paths: Random Waypoint model, City Section model, Gauss-Markov model and the Manhattan model.

The minimum hop paths determined under the City Section mobility model are relatively more stable (i.e., have a larger route lifetime) compared to the minimum hop paths determined under the other three mobility models. The following rankings can be assigned for the four mobility models in the decreasing order of lifetime of minimum hop routes for the different network densities, irrespective of the level of node mobility: (i) Low-density networks – City Section model, Manhattan model, Random Waypoint model, Gauss-Markov model and (ii) Moderate and High-density networks – City Section model, Random Waypoint model, Manhattan model and Gauss-Markov model.

The lifetime of stable routes obtained under the Random Waypoint mobility model was the largest in all of the simulated scenarios. The following rankings can be assigned for the four mobility models in the decreasing order of the lifetime of the stable paths for the different network density and node mobility scenarios: (i) High network density, High node mobility – Random Waypoint model, City Section model, Gauss-Markov model and Manhattan model (ii) All other conditions of network density and node mobility – Random Waypoint model, City Section model, Manhattan model and Gauss-Markov model.

The hop count of the stable routes determined under the City Section mobility model is the smallest for most of the operating scenarios. The hop count of the stable routes determined under the Random Waypoint mobility model is larger than those determined under the City Section mobility model by at most 10%-15%. The hop count of the stable routes determined under the Manhattan and Gauss-Markov mobility models are closer to each other, but are considerably larger than those determined under the City Section and the Random Waypoint mobility models.

As we aim for stable routes, the increase in the hop count of the paths is only sub-linear to the increase in the lifetime compared with that of the minimum hop paths. The Random Waypoint model provided the maximum increase in the path lifetime as well as the maximum increase in the hop count vis-à-vis the minimum hop paths. The City Section and Manhattan mobility models provided only a modest increase in the path lifetime and also incurred only a correspondingly modest increase in the hop count compared with that of the minimum hop paths. The Gauss-Markov mobility model too provided only a modest increase in the path lifetime and hop count in low-density networks. But, as we increase the network density, the improvement in the path lifetime is significantly high accompanied by a reasonably larger increase in the hop count of the stable paths vis-à-vis the minimum hop paths.

7 Conclusions

In conclusion, the general trend is: the more realistic is a mobility model, the larger is the number of hops in the minimum hop routes and smaller is the lifetime of stable routes determined under the mobility model. The Random Waypoint model yielded the lowest hop count for minimum-hop routes and the largest lifetime for stable routes. On the other hand, more realistic mobility models such as the Gauss-Markov model and the Manhattan model yield a relatively larger number of hops for minimum-hop routes and a relatively smaller lifetime for stable routes.

We observe a tradeoff between the objectives of optimizing the route lifetime and the hop count per path for all the four mobility models. Both of these performance metrics cannot be optimized at the same time. For a given simulation condition of network density and node mobility, the average hop count of a Minimum Hop Mobile Path is smaller than the average hop count of a Stable Mobile Path; the average route lifetime of a Stable Mobile Path is more than the average route lifetime of a Minimum Hop Mobile Path.

References

[1] C. Siva Ram Murthy and B. S. Manoj, "Routing Protocols for Ad Hoc Wireless Networks," *Ad Hoc*

Wireless Networks: Architectures and Protocols, Chapter 6, Prentice Hall, June 2004.

[2] J. Broch, D. A. Maltz, D. B. Johnson, Y. C. Hu and J. Jetcheva, "A Performance Comparison of Multi-Hop Wireless Ad Hoc Network Routing Protocols," *Proceedings of the 4th Annual ACM/IEEE Conference on Mobile Computing and Networking*, pp. 85-97, Oct. 1998.

[3] T. Taleb, M. Ochi, A. Jamalipour, K. Nei and Y. Nemoto, "An Efficient Vehicle-Heading Based Routing Protocol for VANET Networks," *Proceedings of the IEEE International Wireless Communications and Networking Conference*, pp. 2199-2204, April 2006.

[4] N. Meghanathan and A. Farago, "On the Stability of Paths, Steiner Trees and Connected Dominating Sets in Mobile Ad Hoc Networks," *Elsevier Ad Hoc Networks*, Vol. 6, No. 5, pp. 744 – 769, July 2008.

[5] C. Bettstetter, H. Hartenstein and X. Perez-Costa, "Stochastic Properties of the Random-Way Point Mobility Model," *Wireless Networks*, pp. 555 – 567, Vol. 10, No. 5, September 2004.

[6] T. Camp, J. Boleng and V. Davies, "A Survey of Mobility Models for Ad Hoc Network Research," *Wireless Communication and Mobile Computing*, Vol. 2, No. 5, pp. 483-502, September 2002.

[7] F. Bai, N. Sadagopan and A. Helmy, "IMPORTANT: A Framework to Systematically Analyze the Impact of Mobility on Performance of Routing Protocols for Ad hoc Networks," *Proceedings of the IEEE International Conference on Computer Communications*, pp. 825-835, March-April, 2003.

[8] B. Liang and Z. Haas, "Predictive Distance-based Mobility Management for PCS Networks," *Proceedings of the IEEE International Conference on Computer Communications*, Vol. 3, pp. 1377-1384, March 1999.

[9] J. Ariyakhajorn, P. Wannawilai and C. Sathitwiriyawong, "A Comparative Study on Random Waypoint and Gauss-Markov Mobility Models in the Performance Evaluation of MANET," *Proceedings of the International Symposium on Communications and Information Technologies*, pp. 894-899, September 2006.

[10] T. H. Cormen, C. E. Leiserson, R. L. Rivest and C. Stein, "Single-Source Shortest Paths," *Introduction to Algorithms*, 2nd Edition, Chapter 24, MIT Press, 2001.

[11] A. Farago and V. R. Syrotiuk, "MERIT: A Scalable Approach for Protocol Assessment," *Mobile Networks and Applications*, Vol. 8, No. 5, pp. 567 – 577, October 2003.

[12] M. Abolhasan, T. Wysocki and E. Dutkiewicz, "A Review of Routing Protocols for Mobile Ad hoc Networks," *Elsevier Ad Hoc Networks*, Vol. 2, No. 1, pp. 1-22, January 2004.

[13] S. P. Chaudhri, J. Y. L. Boudec, M. Vojnovic, "Perfect Simulations for Random Trip Mobility Models," *Proceedings of the 38th Annual Symposium on Simulation*, pp. 72-79, San Diego, USA, April 2005.

[14] M. I. M. Saad and Z. A. Zukarnain, "Performance Analysis of Random-based Mobility Models in MANET Routing Protocol," *European Journal of Scientific Research*, Vol. 32, No. 4, pp. 444-454, 2009.

[15] C. E. Perkins and E. M. Royer, "Ad hoc On-Demand Distance Vector Routing," *Proceedings of the 2nd IEEE Workshop on Mobile Computing Systems and Applications*, pp. 90-100, February 1999.

[16] S. A. Kulkarni and G. R. Rao, "Mobility Model Perspectives for Scalability and Routing Protocol Performances in Wireless Ad hoc Networks," *Proceedings of the First International Conference on Emerging Trends in Engineering and Technology*, pp. 176 – 181, July 2008.

[17] C. E. Perkins and P. Bhagwat, "Highly Dynamic Destination Sequenced Distance Vector Routing for Mobile Computers," *ACM SIGCOMM*, pp. 234 – 244, October 1994.

[18] D. B. Johnson, D. A. Maltz, and J. Broch, "DSR: The Dynamic Source Routing Protocol for Multi-hop Wireless Ad hoc Networks," *Ad hoc Networking*, edited by Charles E. Perkins, Chapter 5, pp. 139-172, Addison-Wesley, 2001.

[19] B. Divecha, A. Abraham, C. Grosan and S. Sanyal, "Impact of Node Mobility on MANET Routing Protocols Models," *Journal of Digital Information Management*, Vol. 4, No. 1, pp. 19 – 23, February 2007.

[20] G. Ravikiran and S. Singh, "Impact of Mobility Models on the Performance of Routing Protocols in Ad hoc Wireless Networks," *Proceedings of the 59th IEEE Vehicular Technology Conference* (*VTC 2004-Spring*), Vol. 4, pp. 2185 – 2189, May 2004.

[21] F. Bai, N. Sadagopan and A. Helmy, "The IMPORTANT Framework for Analyzing the Impact of Mobility on Performance of Routing Protocols for Ad hoc Networks," *Elsevier Ad hoc Networks*, Vol. 1, No. 4, pp. 383 – 403, November 2003.

[22] W. Su, S-J. Lee and M. Gerla, "Mobility Prediction and Routing in Ad hoc Wireless Networks," *International Journal of Network Management*, Vol. 11, No. 1, pp. 3-30, 2001.

[23] C-K. Toh, "Associativity-based Routing for Ad hoc Mobile Networks," *IEEE Personal Communications*, Vol. 4, No. 2, pp. 103-109, 1997.

[24] S. Agarwal, A. Ahuja, J. P. Singh and R. Shorey, "Route Lifetime Assessment Based Routing Protocol for Mobile Ad hoc Networks," *Proceedings of the IEEE International Conference on Communications*, pp. 1697-1701, June 2000.

Enhanced Relevance-Based Approach for Network Control

Aneel Rahim, Fahad bin Muhaya, Zeeshan Shafi Khan
Prince Muqrin Chair for IT Security,
King Saud University, Kingdom of Saudi Arabia
E-mail: aneelrahim@ksu.edu.sa, fmuhaya@ksu.edu.sa

M.A. Ansari, Muhammad Sher
International Islamic University, Islamabad, Pakistan
E-mail: mnsr.alam@gmail.com, m.sher@iiu.edu.pk

Keywords: vehicular ad hoc networks, broadcast, 802.11e

Received: March 9, 2009

Simple flooding, probabilistic approach, area-based scheme, knowledge-based approach and Multi hop Vehicular broadcast is not suitable for VANETs scenario because of its dynamic nature. Relevance scheme is proposed to disseminate the relevant message for sharing in VANETs and discards the redundant messages from the network and improves the over all performance of network. The relevance-based approach does not provide network control and it only broadcast user traffic. This paper presents an improvement in mathematical model to consider the network control. Simulations using NS-2 show that proposed mathematical model consider the network control and improve the global benefit.

Povzetek: Predstavljen je matematičen model izboljšanega nadzora in upravljanja z mrežo.

1 Introduction

Broadcast is the main building block of mobile applications and routing protocols in mobile adhoc networks [1]. Adhoc network is infrastructure less temporarily network, which is mainly used for disaster area and battle field. [2] Vehicular Ad-Hoc Networks is some how different from it in term of battery and mobility.

VANET is the collection of vehicles that communicate with each other from time to time and require no base station, no router for their communication. They can share information either directly or through intermediate nodes [3].

Mostly in VANETs, vehicles are interested in the same kind of information for example information about any accident, road block and weather situation of particular route [4]. So broadcast is the only best option for communication in VANETs.

In Mobile adhoc network a lot of work has been done for broadcast schemes but these existing techniques doses not perform well in VANETs. Simple flooding, probabilistic approach, area-based scheme, knowledge-based approach and Multi hop Vehicular broadcast are not suitable for VANETs scenario. As Collision, Contention and redundant messages [7] are the shortcoming of simple flooding. Probabilistic approach try to solve the redundant message and works fine in dense network but it performance degrades in sparse network. Area-based and knowledge-based approaches also not perform very well because of the dynamic nature of VANETs. Multi hop Vehicular broadcast [6] have Scalability problem. These schemes also ignore the relevance of information and inject the surplus information in network. Relevance approach is proposed to differentiate between high and low priority traffic and improve the performance of network by discarding the redundant messages from the network. The relevance-based approach also has one problem that it does not provide network control and it broadcast only user traffic.

This paper presents an enhancement in mathematical model of relevance-based approach to overcome this problem and global benefit of the network is enhance by adding the network control.

This paper is organized as follows: In section 2, previous work is described. In section 3, enhanced mathematical model is proposed. In section 4, simulation study and results are shown. Lastly in section 5 conclusions is given.

2 Related work

In this section, we will discuss the basic techniques for broadcast i.e. simple flooding, probabilistic approach, area-based scheme and knowledge-based approach. But these techniques can't work fine in VANETS because of dynamic nature of the network. After that we discuss the relevance-based approach that is designed specially for VANETs. We describe its properties, methodology and implementation.

2.1 Previous broadcast approaches

Simple flooding: approach to perform broadcast is by flooding. In this method, a vehicle sends a message to all of its neighbors and its neighbors in return send message to its neighbors. This process continues until all the vehicles get the same message.

Probabilistic scheme: the message is broadcast with some fixed probability. In dense network, due to share coverage only few nodes can do rebroadcast to save network resources. [9]

Area-based scheme: a node calculates the additional coverage area on bases of received redundant messages. If a node achieve sufficient additional coverage area with broadcast then it will rebroadcast else not. [9]

Neighbor knowledge: every node maintains neighbor node information. With help of this information a node decide to rebroadcast a massage or not. To get neighbor information each node has to exchange periodic Hello packets with its neighbor nodes. [10]

Existing broadcast techniques like simple flooding have shortcoming such as redundant rebroadcasts, collision, contention, and probabilistic approach works like simple flooding in spare network.

The performance of neighbor knowledge method depends upon the exchange of hello packet. If the nodes exchange hello message with short interval it will cause contention and collision. If the interval is large its performance degrades due to mobility.

2.2 Relevance-based approach

When two vehicles are in the same VANETs for only a short duration due to high mobility and both the vehicles have too much information in its buffer that they want to exchange with each other. So it is not possible to share all information. They select only important and relevant message for sharing by using relevance-based approach.

Properties

Altruism, application-oriented information differentiation and controlled unfairness are some basic characteristics of relevance-based approach [5] [6] [11]. Altruism means nodes are not selfish and malicious. They forward the information to increase the global benefit regardless of their own benefit. Application-oriented information differentiation means that existing techniques depend on packet specific data but now we get the application oriented data to remove the redundant and surplus information. Controlled unfairness means message are forwarded according to their priority rather than the time they spent in queue.

Methodology

The relevance-based approach is consisting of two steps. First is to calculate the importance of message using the information from three contexts (vehicle context (v), message context (m), information context (i)). Second is to forward the messages according to their relevance value [6].

Implementation

The cross layer design is used to implement the relevance-based approach. Relevance of each message is calculated at application layer and that value is attached to message header before passing it to link layer. Benefit-based extension change the functionality of interface queue and medium access control and forward messages according to their priority by getting information from application layer through interlayer communication [5] [6].

Relevance-based approach can also be implement through 802.11e protocol [8] but it is not suitable due following shortcoming. Firstly the four queues of 802.11e do not give internal resorting of the packets in a packet queue. Packets are inserted into one of the four different priority queues according to their relevance value but for dequeuing it ignore the relevance value and follow only FIFO principle. Secondly they are no mechanism to assign a priority to a given packet. Sort packets into four queues are harmful, because data packets of different relevance value are inserted into the same queue. Thirdly the performance of global benefit decreases because packets of less importance more often get the medium than the high relevance value due to no internal contention of four queues [5].

3 Proposed mathematical model for relevance-based approach

The mathematical model that relevance-based approach used to calculate Message Benefit is given below.

Message Benefit =

$$\frac{1}{\Sigma_{i=1}^{N} a_i} * \sum_{i=1}^{N} a_i * b_i(m, v, i)[6]$$

Message (m), Vehicle (v), and Information (i) context parameters are used to compute a message benefit for every message. Message context includes message age, last transmission, last reception etc. Vehicle context includes speed, road position and connectivity. Information context includes distance, impact and interest etc. Application dependent function bi is used to compute N parameters. The N parameters are then weighted with application dependent factors a_i. At the end all parameters are added and divided by the sum of all a_i. The message benefit value lies between 0 and 1.

Global Benefit = sum of local benefit of all vehicles

The mathematical model that relevance-based approach uses does not consider the network control it only consider user traffic. So its global benefit can be improved by improving the mathematical model by including network traffic as well.

First we will divide the network traffic into different categories. After that priority will be set for each type of traffic and new parameter will be introduce with the existing model that will represents the network traffic. The value of new parameter added will depend on type of traffic.

Basically we have two type of traffic i.e. user traffic and network traffic. User traffic is assign value 0 and network traffic is divided into three categories i.e. operational level, maintenance level and administrative level traffic. We assign the values to network traffic according to their importance e.g. Operational level traffic is assign a value one, then administrative level traffic has value two and lastly maintenance level traffic has three value. We assign values to user and network traffic from 0 to 3.So it is easy to handle them by using 802.11e protocols.

Enhanced Message Benefit =

$$\frac{1}{\Sigma_{i=1}^{N} a_i} * \sum_{i=1}^{N} a_i * b_i(m,v,i) + \sum_{i=1}^{N} p_i \quad (1)$$

a) $\sum_{i=1}^{N} Pi$ =0 if it is user traffic

Where as

$\sum_{i=1}^{N} Pi$ =1 for Operational level network problem

$\sum_{i=1}^{N} Pi$ =2 for Administrative level

$\sum_{i=1}^{N} Pi$ =3 for Maintenance level

b) Message Benefit = $\sum_{i=1}^{N} Pi$ (for Network Traffic only)

If $0 > \sum_{i=1}^{N} Pi \leq 3$ Then

$$\frac{1}{\Sigma_{i=1}^{N} a_i} * \sum_{i=1}^{N} a_i * b_i(m,v,i) = 0$$

802.11e protocol has four queues for data transmission at MAC layer and q0 has greater preference than q1 and q1 has greater preference than q2 and so on. We forward the user and network traffic in to queues according to their values. In existing message benefit calculation they are not considering the network control traffic and they have the range of 0 to 1 but in our proposed message benefit we have range from 0 to 3 is only to handle for 802.11e queues and the for the calculation of global benefit we divide enhance message benefit by three so that it values lies between 0 and 1.

4 Simulation study and results

In order to validate our proposed mathematical model, we compare its performance with existing relevance-based approach. We used NS-2, a network simulator, to simulate the behavior of broadcast schemes under VANETs scenarios.

We use Manhattan Mobility Model and traffic is generated by Generic Mobility Simulation Framework [12].We consider an area of 3000m x3000m with vehicles moving at a speed of 72Km/hr to 108 Km/hr.

4.1 Global benefit with relevance-based approach

Global Benefit (GB) is sum of all local benefits of vehicles during the simulation. Figure 1 shows the global benefit that can be achieved by using relevance-based approach. In existing mechanism there is no parameter for network traffic and no priority is assigned to network traffic so only user traffic getting more and more bandwidth than network traffic as its priority is set higher in existing mechanism.

Relevance-based approach consider only user traffic and ignore network traffic. So its global benefit can be improved by improving the mathematical model. We now evaluate the performance of relevance-based approach by adding the network control parameter in the existing formula. Figure 2 shows the global benefit with

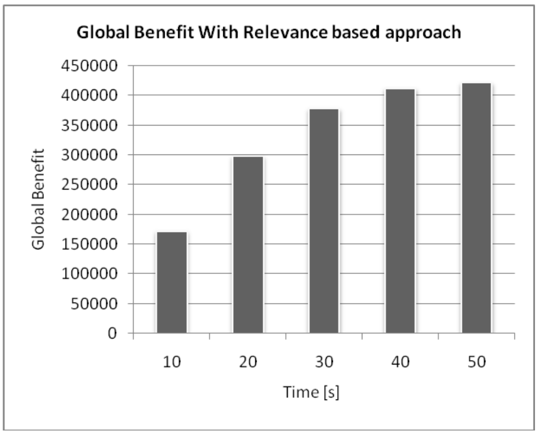


Figure 1: Relevance Approach.

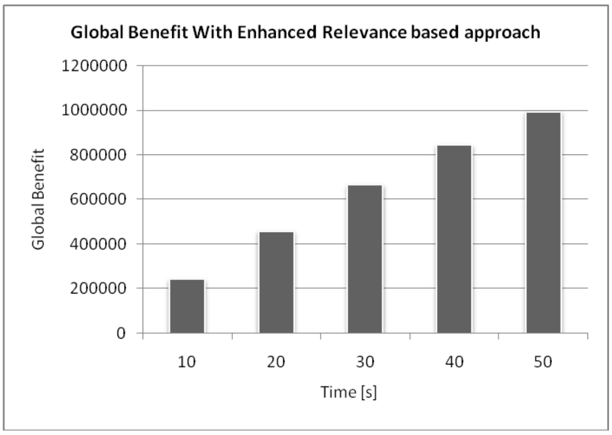


Figure 2: Enhanced Relevance Approach.

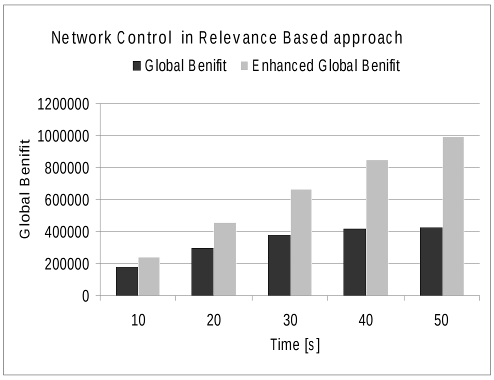


Figure 3: Comparison of GB and EGB

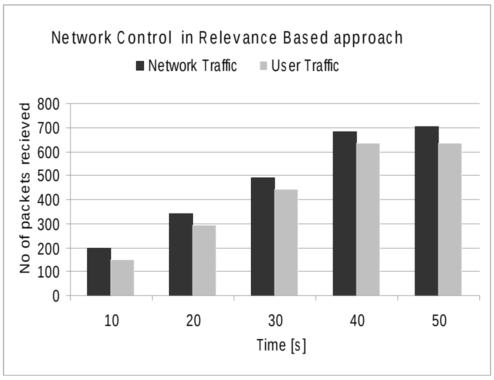


Figure 4: Network Control

enhanced relevance-based approach. It is clear from figure 1 and 2 that global benefit is improved by using enhanced relevance-based approach because in figure 2 network control traffic set higher priority and get more bandwidth than user traffic. So lower priority traffic cannot get more bandwidth than higher priority traffic as it happens in the existing scenarios. That's why the global benefit is improved by adding the network parameter in relevance-based approach.

4.2 Performance evaluation of network traffic and user traffic

In this study we have fifty vehicles, moving with a speed of speed 20 to 30 m/s and simulation time is 50 seconds. Figure 3 shows the behavior of network traffic and user traffic using enhanced relevance-based approach.

Figure 3 shows that network messages have higher priority and it get more medium than user traffic. At time 10, 20, 30, 40, 50 sec, 148, 294, 441, 635 and 636 messages for user traffic and 196, 343, 491, 684 and 706 packets for network traffic are received. It is clear from the simulation that high priority traffic (network traffic) gets more medium than user traffic and the overall benefit of network is higher when we consider the network parameter in the message benefit.

5 Conclusion

Relevance-based approach is proposed for VANETs to give the safety and high priority traffic more bandwidth. But the network parameter is missing in existing message benefit formula. So the proposed approach enhance the global benefit by adding the network parameter in relevance-based approach for network control traffic and simulation shows that global benefit is improved by using enhanced relevance-based approach as higher priority traffic get more medium than lower bandwidth traffic.

References

[1] C. Ellis, H. Miranda, F. Taiani (2009) Count on me: Lightweight Ad-Hoc Broadcasting in Heterogeneous Topologies, *ACM M-MPAC.*

[2] S. Kumar, R. K. Rathy, D Pandey (2009) Design of an Ad-hoc Network Model for Disaster Recovery Scenario Using Various Routing Protocols*, , ACM ICAC3.*

[3] A. Rahim, M. Yasin, I Ahmad, Z. S. Khan, M. Sher (2009) Relevance Based Approach with Virtual Queue Using 802.11e protocol for Vehicular Adhoc Networks*, IEEE-IC4*.

[4] A Rahim, I. Ahmad, Z. S. Khan, M. Sher, M. Shoaib, A. Javed, R. Mahmood (2009) A COMPARATIVE STUDY OF MOBILE AND VANETs, *International Journal of Recent Trends in Engineering*, *Vol 2, No. 4.*

[5]]S. Eichler, C. Schroth, T. Kosch, M. Strassberger (2006) Strategies for context-adaptive message dissemination in vehicular ad hoc networks, *Second International Workshop on Vehicle-to- Vehicle Communications*.

[6] T. Kosch, C. J. Adler, S. Eichler, C. Schroth, M. Strassberger (2006) the scalability problem of vehicular ad hoc networks and how to solve it, *IEEE Wireless Communications*

[7] S-Y. Ni, Y-C Tseng, Y-S Chen, J-P Sheu (1999) The broadcast storm problem in mobile ad hoc networks, *Proc. ACM MobiCom '99*.

[8] T. Osafune, L. Lin, M. Lenardi (2006) Multi-Hop Vehicular Broadcast, *6th International Conference on ITS Telecommunications*.

[9] B. Williams,T. Camp (2002) Comparison of Broadcasting Techniques for Mobile Ad Hoc Networks, *ACM MOBIHOC*.

[10] J. Yoo, H-ryeol Gil, C-kwon Kim (2003) INK: Implicit Neighbor Knowledge Routing in Ad Hoc Networks, IEEE.

[11] C. Schroth, R. Eigner, S. Eichler, M. Strassberger (2006) A Framework for Network Utility Maximization in VANETs, *ACM, International Conference on Mobile Computing and Networking*.

[12] R. Baumann, F. Legendre, P. Sommer (2008) Generic Mobility Simulation Framework (GMSF) *ACM,MobilityModels'08*.

Fast Scalar Multiplications on Hyperelliptic Curve Cryptosystems

Lin You
School of Communication Engineering, Hangzhou Dianzi University, Hangzhou 310018, China
mryoulin@gmail.com

Jiwen Zeng
Department of Mathematics, Xiamen University, Xiamen 361005, China
jwzeng@xmu.edu.cn

Keywords: Hyperelliptic Curve cryptosystems, scalar multiplications, Frobenious Endomorphism, Frobenious Expansion, Euclidean length

Received: August 14, 2008

Scalar multiplication is the key operation in hyperelliptic curve cryptosystem. By making use of Euclidean lengths of algebraic integral numbers in a related algebraic integer ring, we discuss the Frobenius expansions of algebraic numbers, theoretically and experimentally show that the multiplier in a scalar multiplication can be reduced and converted into a Frobenius expansion of length approximate to the field extension degree, and then propose an efficient scalar multiplication algorithm. Our method is an extension of the results given by Müller, Smart and Günther et al. If some (optimal) normal basis is employed, then, for some hyperelliptic curves over finite fields, our method will make the computations of scalar multiplications be lessened about fifty-five percent compared with the signed binary method.

Povzetek: Predstavljena je metoda pohitrenega skalarnega množenja.

1 Introduction

Elliptic curve cryptosystems (ECC) have now widely been studied and applied in e-commerce, e-government and other secure communications. The practical advantages of ECC is that it can be realized with much smaller parameters compared to the conventional discrete logarithms based cryptosystems or RSA but with the same levels of security. This advantage is especially important in the environments with limited processing power, storage space and bandwidth.

As a natural generalization of elliptic curve cryptosystems, the hyperelliptic curve cryptosystem (HECC) was first proposed by Koblitz (1; 2). In a hyperelliptic curve cryptosystem, the rational point group of an elliptic curve, is replaced by the Jacobian group of a hyperelliptic curve , and its security is based on the discrete logarithm on this Jacobian group, that is, based on the hyperelliptic curve discrete logarithm problems(HECDLP). Since the order of the Jacobian group can be constructed much large over a small base field in HECC, HECC has gotten much attention in cryptography, a lot of work has been done to study the group structures and operations on the Jacobian groups.

Let q be a power of some prime and $\mathbb{F}_q$ be the finite field of q elements. A hyperelliptic curve C of genus g over $\mathbb{F}_q$ is defined by the equation

$$v^2 + h(u)v = f(u), \qquad (1)$$

where $h(u)$, $f(u) \in \mathbb{F}_q$ with $\deg_u(h) \leq g$ and $\deg_u(f) = 2g+1$, and there is no solution $(u,v) \in \overline{\mathbb{F}}_q \times \overline{\mathbb{F}}_q$ which simultaneously satisfy the equation $v^2 + h(u)v = f(u)$ and the partial derivate equations $2v + h(u) = 0$ and $h'(u)v - f'(u) = 0$. If the characteristic of $\mathbb{F}_q$ is odd, then the curve (1) is isomorphic to a hyperelliptic curve with the corresponding $h(u)$ equal to 0.

A *divisor* D on C over $\mathbb{F}_q$ is defined as a finite formal sum of rational $\overline{\mathbb{F}}_q$-points $D = \sum m_i P_i$ on C with its *degree* defined as the integer $\sum m_i$. The Jacobian group $\mathbb{J}_C(\mathbb{F}_q)$ of the curve C over $\mathbb{F}_q$ is an Abelian group composed of reduced divisors on C. Every element or reduced divisor D in $\mathbb{J}_C(\mathbb{F}_q)$ can be uniquely expressed by a pair of polynomials $< a(u), b(u) >$ with the properties

$$\begin{cases} \deg_u b(u) < \deg_u a(u) \leq g \\ b(u)^2 + h(u)b(u) - f(u) = 0 \bmod a(u) \end{cases}, \qquad (2)$$

where $a(u)$, $b(u) \in \mathbb{F}_q[u]$. Generally, $a(u)$ is a monic polynomial of degree g and $b(u)$ is a polynomial of degree $g-1$ with a overwhelming probability. The zero element of $\mathbb{J}_C(\mathbb{F}_q)$ can be expressed as $< 1, 0 >$.

In practical hyperelliptic curve cryptosystems, the vital computation that dominates the whole running time is *scalar multiplication*, that is, the computation of the repeated divisor adding

$$\underbrace{D + D + \cdots + D}_{m}$$

for a given divisor $D \in \mathbb{J}(\mathbb{F}_{q^n})$ and a positive integer $m \geq 1$, which is denoted as mD.

Such as in the hyperelliptic curve Diffie-Hellman key exchange protocol(HECDH), suppose Alice and Bob wish to generate their shared secret key for their secure communication, then they do the followings:

- First they agree on a positive integer n and a hyperelliptic curve C over a finite field $\mathbb{F}_q$, and also a divisor $D \in \mathbb{J}(\mathbb{F}_{q^n})$.
- Alice randomly chooses an positive integer m_A that is smaller than $\sharp\mathbb{J}_C(\mathbb{F}_{q^n})$, and then compute the scalar multiplication $D_A = m_A D$ and send D_A to Bob.
- Bob similarly chooses an positive integer m_B, compute $D_B = m_B D$ and send D_B to Alice.
- Alice and Bob compute the scalar multiplications $D_{A,B} = m_A D_B$ and $D_{B,A} = m_B D_A$,respectively.
- Since $D_{A,B} = m_A D_B = m_A(m_B D) = (m_A m_B)D = (m_B m_A)D = D_{B,A}$, Alice and Bob get their shared secret key $D_{A,B}$.
- Using this shared secret key $D_{A,B}$ and some symmetric cryptographic algorithm of their choice, Alice and Bob can communicate securely.

As the above shown, each of Alice and Bob compute two scalar multiplications and the scalar multiplication is the unique operation that involved in HECDH. Also in the hyperelliptic curve digital signature algorithm(HECDSA), it takes three dominating scalar multiplications except for some simple field operations.

A natural algorithm to compute the scalar multiplication mD is (signed) binary method. In (6; 7), Müller and Smart employed Frobenius automorphism to compute point scalar multiplications on elliptic curves over small fields of characteristic even or odd, respectively. In (8), Günther et al employed Frobenius automorphism to compute scalar multiplications on two hyperelliptic curves of genus 2. Their ideas are based on the two facts: One is that, for a point or divisor D, computing $\phi(D)$ is much faster than doubling D, and the other is that every $\mathbb{Z}[\tau]$-integer can be represented as Frobenius expansion or τ-adic expansion of finite lengths, where τ is a root of $P(T)$. In this paper, we will extend their methods to compute scalar multiplications on hyperelliptic curves of general genus.

The remainder of this paper is organized as follows: In Section 2, we briefly describe the Frobenius endomorphism on Jacobian groups of hyperelliptic curves over finite fields and a lemma contributed to Weil's theorem((5)), and in this section, we also introduce the Euclidean length in the algebraic integral ring $\mathbb{Z}[\tau]$ with τ a root of some hyperelliptic curve's characteristic polynomial. In Section 3, we discuss the lengths of τ-adic expansions of algebraic integral numbers in $\mathbb{Z}[\tau]$ and obtain an upper bound for them. In Section 4, we study the cyclic τ-adic expansions and the optimization of the τ-expansions's lengths. The τ-expansion's length of any algebraic integral number is optimized in Section 5, An efficient scalar multiplication algorithm is proposed in Section 6, and the last section gives the conclusion.

2 Frobenius endomorphism over Jacobian groups of hyperelliptic curves

The Frobenius map ϕ of $\overline{\mathbb{F}}_q$ is defined as the map $x \longmapsto x^q$ for $x \in \overline{\mathbb{F}}_q$. Naturally, ϕ induces an endomorphism ϕ_J of $\mathbb{J}_C(\mathbb{F}_{q^n})$ as follows:

$$\begin{tikzcd}
\mathbb{J}_C(\mathbb{F}_{q^n}) \arrow[r, "\phi_J"] & \mathbb{J}_C(\mathbb{F}_{q^n}) \\
<\sum_{i=0}^{g} a_i x^i, \sum_{j=0}^{g-1} b_j x^j > \arrow[r, mapsto, "\phi_J"] & <\sum_{i=0}^{g} a_i^q x^i, \sum_{j=0}^{g-1} b_j^q x^j >,
\end{tikzcd}$$

where $D =< a(u), b(u) >=< \sum_{i=0}^{g} a_i x^i, \sum_{j=0}^{g-1} b_j x^j >$ is a reduced divisor or an element of $\mathbb{J}_C(\mathbb{F}_{q^n})$ with $a_i, b_j \in \mathbb{F}_{q^n}$.

For convenience, ϕ_J is also denoted by ϕ.

Lemma 1((5)) *For any positive integer r, let M_r denote the number of rational points of the hyperelliptic curve C defined by Equation* (1) *over $\mathbb{F}_{q^r}$ and $\sharp\mathbb{J}_C(\mathbb{F}_{q^r})$ denotes the order of the Jacobian group $\mathbb{J}_C(\mathbb{F}_q)$. Then*

1. *The zeta-function $Z(t)$ has the expression*

$$Z(t) = \exp(\sum_{n=1}^{\infty} \frac{M_n}{n} t^n) = \frac{L(t)}{(1-t)(1-qt)},$$

where $L(t)$ is an integral coefficient polynomial of degree $2g$.

2. *Let*

$$P(T) = t^{2g} L(1/T) = \prod_{i=1}^{2g} (T - \tau_i),$$

then $|\tau_i| = \sqrt{q}$, and the roots come in complex conjugate pairs such that there exists an ordering with $\tau_{i+g} = \bar{\tau}_i$, and hence, $\tau_{i+g}\tau_i = q$.

3. *$P(T)$ is the characteristic polynomial of Frobenius endomorphism ϕ and $P(T)$ is an integral coefficient polynomial of the following form*

$$\begin{array}{c} P(T) = T^{2g} + a_1 T^{2g-1} + a_2 T^{2g-2} + \cdots + \\ +a_g T^g + q a_{g-1} T^{g-1} + \cdots + q^{g-1} a_1 T + q^g \end{array}. \qquad (3)$$

4. *Let $a_0 = 1$, then for $1 \le i \le g$*

$$\begin{array}{rcl} i a_i & = & (M_i - q^i - 1)a_0 + (M_{i-1} - q^{i-1} - 1)a_1 \\ & & + \cdots + (M_1 - q - 1)a_{i-1} \end{array}.$$

5. *For any positive integer n,*

$$\sharp\mathbb{J}_C(\mathbb{F}_{q^n}) = \prod_{i=1}^{2g} (1 - \tau_i^n).$$

For cryptographic purposes, in order to resist all possible attacks on the HECDLP, such as Pollardąŕs rho algorithm((3)) and Pohlig-Hellman algorithm((4) or their improved versions, it is most desirable that $\sharp\mathbb{J}_C(\mathbb{F}_{q^n})$ have a large prime integer factor, or to the best, $\sharp\mathbb{J}_C(\mathbb{F}_{q^n})$ is by itself a large prime or almost large prime. For the best possibility, the necessary condition is that $P(T)$ is irreducible. Hence, $P(T)$ is assured to be irreducible here.

3 Euclidean lengths in the algebraic integral ring $\mathbb{Z}[\tau]$

Let C be a hyperelliptic curve of genus g over $\mathbb{F}_q$ with the characteristic polynomial (3). Let τ be a root of $P(T)$. Then, since $P(T)$ is irreducible, every element ξ in $\mathbb{Z}[\tau]$ can be uniquely expressed as the form

$$x_0 + x_1\tau + \cdots + x_{2g-1}\tau^{2g-1}.$$

Let $\tau = \tau_1,\ \tau_2,\ \cdots,\ \tau_g$ be the g roots of $P(T)$ which are not conjugate each other. Then, we can define a positive number $N(\xi)$ corresponding to ξ as the following

$$N(\xi) = \sqrt{|\sum_{i=0}^{2g-1} x_i\tau_1^i|^2 + \cdots + |\sum_{i=0}^{2g-1} x_i\tau_g^i|^2},$$

where $|x|$ denotes the complex absolute value of x. $N(\xi)$ is often called the Euclidean length of ξ.

It is clear that $N(\xi\eta) \le N(\xi)N(\eta)$ and $N(\xi+\eta) \le N(\xi) + N(\eta)$ hold for any $\xi,\ \eta \in \mathbb{Z}[\tau]$. And $N(\xi)^2$ is a positive definite quadratic form in the variables $x_0,\ x_1,\ \cdots,\ x_{2g-1}$, with the coefficients being integer polynomials of $P(T)$'s coefficients $a_i (1 \le i \le g)$.

For $g = 1$ and $\xi = x_0 + x_1\tau$, we have

$$N(\xi)^2 = x_0^2 - a_1x_0x_1 + qx_1^2.$$

For $g = 2$ and $\xi = x_0 + x_1\tau + x_2\tau^2 + x_3\tau^3$, we have

$$\begin{array}{ll} N(\xi)^2 & = 2x_0^2 - a_1x_0x_1 + (a_1^2 - 2a_2)x_0x_2 \\ & -(a_1^3 - 3(a_1a_2 - a_1q))x_0x_3 + 2qx_1^2 - a_1qx_1x_2 \\ & +(a_1^2 - 2a_2)qx_1x_3 + 2q^2x_2^2 - a_1q^2x_2x_3 + 2q^3x_3^2 \end{array}.$$

In general, let $S_i = \sum_{j=1}^{g}(\tau_j^i + \bar{\tau}_j^i)$, $X = (x_0, x_1, \cdots, x_{2g-1})$, and let

$$A = \begin{pmatrix} g & S_1/2 & S_2/2 & \cdots & S_{2g-1}/2 \\ S_1/2 & qg & qS_1/2 & \cdots & qS_{2g-2}/2 \\ S_2/2 & qS_1/2 & q^2g & \cdots & q^2S_{2g-3}/2 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ S_{2g-1}/2 & qS_{2g-2}/2 & q^2S_{2g-3}/2 & \cdots & q^{2g-1}g \end{pmatrix}$$

Then we can easily prove

$$N(\xi)^2 = XAX^T,$$

where S_i can be computed by the following Newton's formula:

$$S_i + a_1S_{i-1} + a_2S_{i-2} + \cdots + a_{i-1}S_1 + ia_i = 0$$

with $a_0 = 1$ and $a_j = a_{2g-j}q^{j-g}$ for $j > g$.

4 Convert m into τ-adic expansion

Similar to Lemma 1 in (6), we have

Lemma 2 (**Division With Remainder** in $\mathbb{Z}[\tau]$) *Let $m \in \mathbb{Z}[\tau]$, then there exists a unique pair of elements m' and r such that*

$$m = m'\tau + r \qquad (4)$$

with $m' \in \mathbb{Z}[\tau]$, $r \in \{-\lceil q^g/2\rceil + 1, \cdots, \lfloor q^g/2\rfloor\}$.

Theorem 1 *Let $m \in \mathbb{Z}[\tau]$, then m can be uniquely represented as a τ-adic expansion*

$$m = \sum_{i=0}^{k-1} r_i\tau^i + m'\tau^k,\ \ r_i \in \{-\lceil q^g/2\rceil + 1, \cdots, \lfloor q^g/2\rfloor\}.$$

If $k \ge 2\log_q \frac{2(\sqrt{q}-1)N(m)}{\sqrt{g}}$, then $N(m') < \frac{q^g\sqrt{g}}{2(\sqrt{q}-1)}$.

Proof Iterate the Division With Remainder in $\mathbb{Z}[\tau]$ for $m_0 = m$, then we have

$$m_i = m_{i+1}\tau + r_i,\ \ r_i \in \{-\lceil q^g/2\rceil + 1, \cdots, \lfloor q^g/2\rfloor\}.$$

Hence,

$$m_0 = \sum_{i=0}^{j-1} r_i\tau^i + m_j\tau^j.$$

Apply triangle inequality for Euclidean length in $m_i = m_{i+1}\tau + r_i$, and we will get

$$N(m_j) < \frac{N(m_0)}{\sqrt{q}^j} + \frac{\sqrt{g}(\lfloor q^g/2\rfloor)}{\sqrt{q}-1}.$$

Hence, if $\frac{N(m_0)}{\sqrt{q}^j} \le \frac{\sqrt{g}}{2(\sqrt{q}-1)}$ or

$$j \ge 2\log_q \frac{2(\sqrt{q}-1)N(m_0)}{\sqrt{g}},$$

then

$$N(m_j) < \frac{\sqrt{g}(\lfloor q^g/2\rfloor + 1/2)}{\sqrt{q}-1} = \frac{q^g\sqrt{g}}{2(\sqrt{q}-1)}.$$

Hence, for $k = \lceil 2\log_q \frac{2(\sqrt{q}-1)N(m)}{\sqrt{g}}\rceil + 1$ and $m' = m_k$, we have $N(m') < \frac{q^g\sqrt{g}}{2(\sqrt{q}-1)}$. □

Lemma 3 $a_1 \le 2\lfloor 2\sqrt{q}\rfloor$. *And if $a_2 = 0$, then*

$$|a_1| < \sqrt{q}.$$

Proof $a_1 \le 2\lfloor 2\sqrt{q}\rfloor$ holds obviously since every root of $P(T)$ has the complex absolute value $\sqrt{q}$. If $a_2 = 0$, then $|a_1| < \sqrt{q}$ follows the facts $M_2 - q^2 - 1 + a_1^2 = 0$ and $M_2 > 1$.

Lemma 4 *If C is a hyperelliptic curve of genus 2 with the irreducible characteristic polynomial $P(T) = T^4 + a_1T^3 + a_2T^2 + a_1qT + q^2$, then $a_1^2 - 4a_2 + 8q$ is non-square and*

$$2|a_1|\sqrt{q} - 2q < a_2 < a_1^2/4 + 2q.$$

Proof If $a_1^2 - 4a_2 + 8q$ is square, then $P(T) = (T^2 + \frac{1}{2}(a_1 \pm \sqrt{a_1^2 - 4a_2 + 8q})T + q)(T^2 + \frac{1}{2}(a_1 \mp \sqrt{a_1^2 - 4a_2 + 8q})T + q)$, which contradicts our hypothesis that $P(T)$ is irreducible.

Due to (9), we have $\lceil 2|a_1|\sqrt{q}-2q\rceil \leq a_2 \leq \lfloor a_1^2/4+2q\rfloor$, and it follows $2|a_1|\sqrt{q}-2q < a_2 < a_1^2/4+2q$. □

Theorem 2 *Let C be a hyperelliptic curve of genus g and its irreducible characteristic polynomial $P(T)$ have a root τ. Let $R=\{-\lceil q^g/2\rceil+1,\cdots,\lfloor q^g/2\rfloor\}$, and a τ-adic expansion means a τ-polynomial with the coefficients belong to R. Let $\xi \in \mathbb{Z}[\tau]$ with*

$$N(\xi) < \frac{q^g\sqrt{g}}{2(\sqrt{q}-1)}. \tag{5}$$

1. *If $g=2$, $a_1=a_2=0$, then for every positive integer m, m has a τ-adic expansion of length about $\frac{1}{2}\lfloor \log_q m\rfloor$. (But, in this case, the curves are supersingular and not suitable for cryptosystems).*
2. *If $g=2$, and only one of a_1 and a_2 equal to 0, then ξ has a τ-adic expansion of length at most 5.*
3. *If $g=2$, and none of a_1 and a_2 equals to 0, then ξ has a τ-adic expansion of length at most 8.*
4. *If $g \geq 3$, then ξ has a τ-adic expansion of length $l \leq 2g+4$.*

Proof Suppose $\xi \in \mathbb{Z}[\tau]$ satisfying the inequality (5), then

$$N(\xi)^2 < \frac{gq^{2g}}{4(\sqrt{q}-1)^2}.$$

1. Since $q^2+\tau^4=0$ or $q^2=-\tau^4$, it obviously follows that m has a τ-adic expansion of length about $\frac{1}{2}\lfloor \log_q m\rfloor$.

2. Suppose $a_1=0$. Then $|a_2|<2q$, and it follows $4q^2-a_2^2 \geq 4q-1$. Hence,

$$\begin{aligned}
&N(\xi)^2\\
&=2(x_0^2-a_2x_0x_2+qx_1^2-a_2qx_1x_3+q^2x_2^2+q^3x_3^2)\\
&=2((x_0-a_2/2x_2)^2+(q^2-a_2^2/4)x_2^2+q(x_1-a_2/2x_3)^2\\
&\quad+q(q^2-a_2^2/4)x_3^2)\\
&=2((1-\tfrac{a_2^2}{4q^2})x_0^2+q^2(x_2-\tfrac{a_2}{2q^2}x_0)^2+(q-\tfrac{a_2^2}{4q})x_1^2\\
&\quad+q^3(x_3-\tfrac{a_2}{2q^2}x_1)^2)\\
&<\tfrac{q^4}{2(\sqrt{q}-1)^2}
\end{aligned} \tag{6}$$

Thus,

$$\left\{\begin{array}{rcl}
|1-\frac{a_2^2}{4q^2}|^{1/2}|x_0| & < & \frac{q^2}{2(\sqrt{q}-1)}\\
|1-\frac{a_2^2}{4q^2}|^{1/2}|x_1| & < & \frac{q^2}{2(q-\sqrt{q})}\\
|q^2-a_2^2/4|^{1/2}|x_2| & < & \frac{q^2}{2(\sqrt{q}-1)}\\
|q^2-a_2^2/4|^{1/2}|x_3| & < & \frac{q^2}{2(q-\sqrt{q})}
\end{array}\right., \tag{7}$$

and so,

$$|x_0| < \tfrac{q^3}{(\sqrt{q}-1)\sqrt{4q-1}},\ |x_1| < \tfrac{q^3}{(q-\sqrt{q})\sqrt{4q-1}},$$
$$|x_2| < \tfrac{q^2}{(\sqrt{q}-1)\sqrt{4q-1}},\ |x_3| < \tfrac{q^2}{(q-\sqrt{q})\sqrt{4q-1}}.$$

a) If $q \geq 4$, then $|x_0| \leq q^2$ and $|x_1| \leq q^2/2$. Hence, if $x_0 > q^2/2$(similar for $x_0 < -q^2/2$), then from (6), we have $(2q^2x_2-a_2x_0)^2 < q^6/(\sqrt{q}-1)^2-q^4(4q-1)/4$, and so $|x_2-a_2| \leq |x_2-\frac{a_2}{2q^2}x_0|+|\frac{a_2}{2q^2}x_0-a_2| < \sqrt{4q^2-(4q-1)(\sqrt{q}-1)^2/(4(\sqrt{q}-1))+3(2q-1)/4} \leq q^2/2$. Thus

$$\begin{aligned}
\xi &= x_0+x_1\tau+x_2\tau^2+x_3\tau^3\\
&=(x_0-q^2)+x_1\tau+(x_2-a_2)\tau^2+x_3\tau^3-\tau^4
\end{aligned}$$

is a τ-adic expansion of length at most 5 .

b) If $q \geq 4$ and $|x_0| \leq q^2/2$, then ξ is itself a τ-adic expansion of length 4 .

c) If $q=2$, then $a_2=\pm 1$. Hence, from (7), we have $|x_0| \leq 4$, $|x_1| \leq 3$, $|x_2| \leq 2$ and $|x_3| \leq 1$. If $|x_0| \leq 2$ and $|x_1| \leq 2$, then ξ is itself a τ-adic expansion. If $|x_0| \leq 2$ and $|x_1|=3$, then $\xi = x_0+(x_1\pm 4)\tau+x_2\tau^2+(x_3\pm 1)\tau^3-\tau^4$ is a τ-adic expansion of length $l \leq 5$.

If $|x_0|=3$, then from (6), we have

$$\begin{aligned}
&9(1-1/16)+4(x_2\pm 3/8)^2+(2-1/8)x_1^2\\
&+8(x_3-1/8x_1)^2 < \tfrac{2^4}{4(\sqrt{2}-1)^2},
\end{aligned} \tag{8}$$

which implies $|x_1| \leq 2$. If $x_1=0$, then $\xi=(x_0\pm 4)+(x_2\pm a_2)\tau^2+x_3\tau^3\pm\tau^4$ (if $|x_2\pm a_2| \leq 2$) or $\xi=(x_0\pm 4)+(x_2\pm a_2\pm 4)\tau^2+x_3\tau^3+(\pm 1\pm a_2)\tau^4\pm\tau^6$(if $|x_2\pm a_2|=3$) is a τ-adic expansion of length 5.

If $0<|x_1| \leq 2$ and $x_3=0$, then ξ is itself a τ-adic expansion or $\xi=(x_0\pm 4)+x_1\tau+(x_2\pm a_2\pm 4)\tau^2+(\pm 1\pm a_2)\tau^4\pm\tau^6$ is a τ-adic expansion of length $l \leq 5$.

If $0<|x_1| \leq 2$ and $|x_3|=1$, then from the equation (8) we obtain $|x_2| \leq 1$. Hence, $\xi=(x_0\pm 4)+x_1\tau+(x_2\pm a_2)\tau^2+x_3\tau^3\pm\tau^4$ is a τ-adic expansion of length $l \leq 5$.

If $|x_0|=4$, then from (6), we have $|x_2| < \frac{\sqrt{2^6/(\sqrt{2}-1)^2-15\times 16}}{8}+4/8 < 2$, and so $|x_2\pm a_2| \leq 2$. Hence, ξ is a τ-adic expansion of length $l \leq 4$.

d) If $q=3$, then $|a_2| \leq 3$. From (7), we have $|x_0| \leq 7$, $|x_1| \leq 4$, $|x_2| \leq 2$ and $|x_3| \leq 1$. If $|x_0| \leq 4$, then ξ is itself a τ-adic expansion of length 4. If $|x_0| \geq 5$ and $|a_2|=3$, then from (6) we have $3x_0^2+(6x_2\pm x_0)^2+9x_1^2+3(6x_3\pm x_1)^2 < 3^4/(\sqrt{3}-1)^2$, which implies $x_1=0$ or $x_3=0$. Thus, $\xi=x_0+x_1\tau+x_2\tau^2+x_3\tau^3=(x_0-9)+x_1\tau+(x_2-a_2)\tau^2+x_3\tau^3-\tau^4$ or $\xi=(x_0-9)+x_1\tau+(x_2-a_2\pm 9)\tau^2+x_3\tau^3+(1\pm a_2)\tau^4\pm\tau^6$ is a τ-adic expansion of length $l \leq 5$.

3. Suppose $a_2=0$. Then by Lemma 3, $|a_1|=1$ for $q=2,3$ and $|a_1|<\sqrt{q}$ for $q>3$. Since

$$\begin{aligned}
N(\xi)^2 &= 2x_0^2-a_1x_0x_1+a_1^2x_0x_2-(a_1^3+3a_1q)x_0x_3\\
&\quad+2qx_1^2-a_1qx_1x_2+a_1^2qx_1x_3+2q^2x_2^2\\
&\quad+2q^3x_3^2-a_1q^2x_2x_3\\
&=\tfrac{q^2(a_1^2+8q)}{a_1^2+4q}x_2^2+2q(x_1+\tfrac{a_1^2}{4}x_3-\tfrac{a_1}{4q}x_0-\tfrac{a_1}{4}x_2)^2\\
&\quad+\tfrac{q(16q^2-a_1^4)}{8}(x_3+\tfrac{-3a_1^3-12a_1q}{q(16q^2-a_1^4)}x_0-\tfrac{a_1}{4q+a_1^2}x_2)^2\\
&\quad+\tfrac{(a_1^2+8q)(a_1^2-q)}{q(a_1^2-4q)}x_0^2\\
&<\tfrac{q^4}{2(\sqrt{q}-1)^2}
\end{aligned}, \tag{9}$$

it follows that

$$\frac{(a_1^2-q)}{q(a_1^2-4q)}x_0^2+\frac{q^2}{a_1^2+4q}x_2^2 < \frac{q^4}{2(\sqrt{q}-1)^2(8q+a_1^2)} \tag{10}$$

and

$$\left\{\begin{array}{l}
|x_0| < \frac{\sqrt{q}q^2\sqrt{1+\frac{3q}{q-a_1^2}}}{\sqrt{2}(\sqrt{q}-1)\sqrt{8q+a_1^2}} < \frac{q^2\sqrt{1+\frac{3q}{2\sqrt{q}-1}}}{4(\sqrt{q}-1)}\\
|x_2| < \frac{q}{\sqrt{2}(\sqrt{q}-1)}\sqrt{1-\frac{4q}{8q+a_1^2}} < \frac{\sqrt{5}q}{3\sqrt{2}(\sqrt{q}-1)}
\end{array}\right. . \tag{11}$$

Similarly, we will get

$$\left\{\begin{array}{l}
|x_1| < \frac{q^2}{\sqrt{2}(q-\sqrt{q})}\sqrt{1-\frac{4q}{8q+a_1^2}} < \frac{\sqrt{5}q^2}{3\sqrt{2}(q-\sqrt{q})}\\
|x_3| < \frac{q}{\sqrt{2}(\sqrt{q}-1)\sqrt{8q+a_1^2}}\sqrt{1+\frac{3q}{q-a_1^2}} < \frac{q\sqrt{1+\frac{3q}{2\sqrt{q}-1}}}{4(q-\sqrt{q})}
\end{array}\right. . \tag{12}$$

If $q \geq 4$ then $|x_i| \leq q^2/2$ for $i=0,\ 1,\ 2,\ 3$. Hence, ξ is a τ-adic expansion of length at most 4.

Let $q=3$, then from (9), (10),(11) and (12), we have $|x_0| \leq 7$, $|x_1| \leq 3$, $|x_2| \leq 1$ and $|x_3| \leq 1$. Without loss of generality, suppose $a_1 = 1$ and $x_0 > 0$, then

$$\begin{aligned}\xi &= x_0 + x_1\tau + x_2\tau^2 + x_3\tau^3 (\text{ if } |x_0| \le 4)\\ &= (x_0 - 9) + (x_1 - 3)\tau + x_2\tau^2 + (x_3 - 1)\tau^3 - \tau^4\\ &\quad (\text{ if } |x_0| > 4 \text{ and } |x_1 - 3| \le 4)\\ &= (x_0 - 9) + (x_1 - 3 + 9)\tau + (x_2 + 3)\tau^2 + (x_3 - 1)\tau^3 + \tau^5\\ &\quad (\text{ if } |x_0| > 4 \text{ and } x_1 - 3 < -4)\end{aligned}$$

is a τ-adic expansion of length at most 5.

Similar discussion will show that ξ can also be represented as a τ-adic expansion of length at most 5 for $q = 2$.

4. Suppose $a_1 \ne 0$ and $a_2 \ne 0$. Then for $\xi = x_0 + x_1\tau + x_2\tau^2 + x_3\tau^3$, we have

$$\begin{aligned}&N(\xi)^2 =\\ &\quad 2q^3(x_3 - \tfrac{a_1}{4q}x_2 + \tfrac{a_1^2-2a_2}{4q^2}x_1 - \tfrac{a_1(a_1^2-3a_2+3q)}{4q^3}x_0)^2\\ &\quad + \tfrac{q(16q-a_1^2)}{8}(x_2 + \tfrac{a_1(a_1^2-2a_2-4q)}{q(16q-a_1^2)}x_1 + \tfrac{-a_1^4+3a_1^2a_2+qa_1^2-8qa_2}{q^2(16q-a_1^2)}x_0)^2\\ &\quad + \tfrac{-a_1^4+6a_1^2a_2-4qa_1^2-8a_2^2+32q^2}{16q-a_1^2}(x_1-\\ &\qquad\qquad \tfrac{a_1(16q^2-14qa_1^2-2a_1^4+13a_1^2a_2-20a_2^2+32qa_2)}{2q(-a_1^4+6a_1^2a_2-4qa_1^2-8a_2^2+32q^2)}x_0)^2\\ &\quad + \tfrac{qa_1^4-\frac{1}{4}a_1^2a_2^2-5qa_1^2a_2+7q^2a_1^2+a_2^3+2qa_2^2-4q^2a_2-8q^3}{q^2(a_1^2-2a_2-4q)}x_0^2\\ &= 2q^2(x_2 - \tfrac{a_1}{4}x_3 + \tfrac{a_1^2-2a_2}{4q^2}x_0 - \tfrac{a_1}{4q}x_1)^2\\ &\quad + \tfrac{q^2(16q-a_1^2)}{8}(x_3 + \tfrac{-3a_1^3+10a_1a_2-12a_1q}{q^2(16q-a_1^2)}x_0 + \tfrac{3a_1^2-8a_2}{q(16q-a_1^2)}x_1)^2\\ &\quad + \tfrac{(4a_2-8q-a_1^2)(a_1^4-3a_1^2a_2+3a_1^2q-2a_2q-4q^2)}{q^2(16q-a_1^2)}(x_0+\\ &\qquad\qquad \tfrac{a_1q(14a_1^2q+2a_1^4-32a_2q-13a_1^2a_2-16q^2+20a_2^2)}{2(4a_2-8q-a_1^2)(a_1^4-3a_1^2a_2+3qa_1^2-2qa_2-4q^2)}x_1)^2\\ &\quad + \tfrac{qa_1^4-\frac{1}{4}a_1^2a_2^2-5qa_1^2a_2+7q^2a_1^2+a_2^3+2qa_2^2-4q^2a_2-8q^3}{a_1^4-3a_1^2a_2+3a_1^2q-2a_2q-4q^2}x_1^2\\ &< \tfrac{q^4}{2(\sqrt{q}-1)^2}\end{aligned} \quad (13)$$

Let

$$\begin{cases} F = -qa_1^4 + \frac{1}{4}a_1^2a_2^2 + 5qa_1^2a_2 - 7q^2a_1^2 - a_2^3 - 2qa_2^2\\ \qquad +4q^2a_2 + 8q^3\\ G = -a_1^4 + 3a_1^2a_2 - 3a_1^2q + 2a_2q + 4q^2\\ H = -a_1^2 + 2a_2 + 4q\end{cases}.$$

Since $|a_1| \le 2\lfloor 2\sqrt{q}\rfloor$ and $2|a_1|\sqrt{q} - 2q < a_2 < a_1^2/4 + 2q$, we have $F \ge 1$, $G > 0$ and $H > 0$. Hence from (13) we get $|x_0| < \frac{\sqrt{2}q^3}{2(\sqrt{q}-1)}\sqrt{H/F}$ and $|x_1| < \frac{\sqrt{2}q^2}{2(\sqrt{q}-1)}\sqrt{G/F}$. Similarly, we will get $|x_2| < \frac{\sqrt{2}q^2}{2(q-\sqrt{q})}\sqrt{G/F}$ and $|x_3| < \frac{\sqrt{2}q^2}{2(q-\sqrt{q})}\sqrt{H/F}$. That is,

$$\begin{cases} |x_0| & < & \frac{\sqrt{2}q^3}{2(\sqrt{q}-1)}\sqrt{H/F}\\ |x_1| & < & \frac{\sqrt{2}q^2}{2(\sqrt{q}-1)}\sqrt{G/F}\\ |x_2| & < & \frac{\sqrt{2}q^2}{2(q-\sqrt{q})}\sqrt{G/F}\\ |x_3| & < & \frac{\sqrt{2}q^2}{2(q-\sqrt{q})}\sqrt{H/F}\end{cases}. \quad (14)$$

If $a_2 = 2q + a_1^2/4$ or $a_1 = \pm(2q + a_2)/(2\sqrt{q})$, then $F = 0$, which contradicts $F \ge 1$. While, it is very likely that H/F or G/F takes maximal values at $a_2 = 2q + a_1^2/4 - \theta$ or $-2q + 2|a_1|\sqrt{q}+\delta$, where $\theta = 1$ or $1/4$, $\delta = 1$ or $\lceil 2|a_1|\sqrt{q}\rceil - 2|a_1|\sqrt{q}$.

According to (10), if C is a curves with $a_2 = 2q + (a_1^2 - 1)/4$, then it may not be a hyperelliptic curve. Thus, we do not consider this case.

i) Let $a_2 = 2q + a_1^2/4 - 1$. Then, if $a_1 \ne \pm(4\sqrt{q} - 2)$, H/F and G/F are strictly increasing or decreasing with $a_1 > 0$ or $a_1 < 0$. Hence, if q is a square, then H/F and G/F reach their maximal values at $a_1 = \pm(4\sqrt{q} - 4)$, that is, $\frac{2(48q-56\sqrt{q}+128q^{3/2}-15)}{3(-160q+9+256q^2)}$ and $\frac{2(-1872q^2+8q^{3/2}+1152q^{5/2}+1353q-368\sqrt{q}-168)}{3(-160q+9+256q^2)}$, respectively. Thus, $H/F < \frac{3}{5}q^{-1/2}$ and $G/F < 3\sqrt{q}$. If q is non-square, without loss of generality, suppose $a_1 \ge 2$. Because both H/F and G/F are strictly increasing functions of a_1 except in a neighborhood of $a_1 = 4\sqrt{q} - 2$, they will reach their possible maximal values at $a_1 = 2(2\sqrt{q} - \varepsilon) - 2$ since $a_1 \le 2\lfloor 2\sqrt{q}\rfloor$ and a_1 is even, where $\varepsilon = 2\sqrt{q} - \lfloor 2\sqrt{q}\rfloor$. Replace a_1 and a_2 in H/F with $2(2\sqrt{q} - \varepsilon) - 2$ and $2q + a_1^2/4 - 1$, respectively. Then, since $\varepsilon = 2\sqrt{q} - \lfloor 2\sqrt{q}\rfloor > \frac{3}{2\lfloor 2\sqrt{q}\rfloor+1} > \frac{3}{5\sqrt{q}}$, we get

$$\begin{aligned}H/F &= \tfrac{-2(-4\sqrt{q}\varepsilon-4\sqrt{q}+\varepsilon^2+2\varepsilon+2)}{\varepsilon(4\varepsilon+\varepsilon^3+4\varepsilon^2-8\sqrt{q}\varepsilon^2-24\sqrt{q}\varepsilon-16\sqrt{q}+16q\varepsilon+32q)}\\ &\approx \tfrac{1}{4\sqrt{q}\varepsilon} < \tfrac{1}{4\sqrt{q}}\cdot\tfrac{5\sqrt{q}}{3} < \tfrac{5}{12}.\end{aligned}$$

Similarly, we have $G/F \approx \frac{9\sqrt{q}}{4\varepsilon} < \frac{9\sqrt{q}}{4}\cdot\frac{5\sqrt{q}}{3} < \frac{15q}{4}$.

ii) Let q be square and $a_2 = -2q + 2|a_1|\sqrt{q} + 1$. Without loss of generality, suppose $a_1 \ge 1$. Since $a_2 < a_1^2/4 + 2q$, it follows $a_1 \le 4\sqrt{q} - 3$. It is easy to show that H/F is strictly increasing for $1 \le a_1 \le 4\sqrt{q} - 3$. Hence, H/F will reach its maximal value at $a_1 = 4\sqrt{q} - 3$, and so, $H/F < \frac{4}{5}q^{-1/2}$. Similar discussion will induce $G/F < \frac{27}{5}q^{1/2}$.

iii) Let $a_2 = -2q + 2|a_1|\sqrt{q} + \delta$ with $\delta = \lceil 2|a_1|\sqrt{q}\rceil - 2|a_1|\sqrt{q}$ (q is non-square). Still suppose $a_1 \ge 1$. Then, $a_1 < 4\sqrt{q} - 2\sqrt{\delta}$. Let $a_1 = 4\sqrt{q} - 2 + \epsilon$, where $0 < \epsilon < 1$ such that a_1 is an integer, that is, $a_1 = \lceil 4\sqrt{q}-2\rceil$. Replace a_1 and a_2 in H/F with $4\sqrt{q}-2+\epsilon$ and $-2q+2a_1\sqrt{q}+\delta$, respectively. Then, since $\delta = \lceil 2a_1\sqrt{q}\rceil - 2a_1\sqrt{q} > \frac{1}{4\sqrt{q}}$, we have $H/F \approx -4q^{1/2}\frac{-8\sqrt{q}}{\delta 64q^{3/2}} = \frac{1}{2\delta\sqrt{q}} < 2$. Similar discussion will induce $G/F \approx \frac{9}{2}\sqrt{q}\delta^{-1} < 18q$.

From all the discussion above, we conclude that if $a_2 \ne 2q + (a_1^2 - 1)/4$, then

$$H/F < \begin{cases} \frac{4}{5}q^{-1/2} & \text{if } q \text{ is square}\\ 2 & \text{if } q \text{ is non-square}\end{cases},$$

and

$$G/F < \begin{cases} \frac{18}{5}q^{1/2} & \text{if } q \text{ is square}\\ 18q & \text{if } q \text{ is non-square.}\end{cases}.$$

Hence, if q is square, we have

$$\begin{cases} |x_0| & < & \frac{2}{\sqrt{5}}q^{9/4}\\ |x_1| & < & \frac{\sqrt{36}}{\sqrt{5}}q^{7/4}\\ |x_2| & < & \frac{\sqrt{36}}{\sqrt{5}}q^{5/4}\\ |x_3| & < & \frac{2}{\sqrt{5}}q^{3/4}\end{cases},$$

and if q is non-square, we have

$$\begin{cases} |x_0| & < & 2q^{5/2}\\ |x_1| & < & \sqrt{36}q^2\\ |x_2| & < & \sqrt{36}q^{3/2}\\ |x_3| & < & 2q\end{cases}.$$

In the following discussions, without loss of generality, we assure $a_1 > 0$ and $x_0 > 0$. And, for the worst case, we also assume that all x_i is near to its upper bound. Then, if q is a square no less than 49 and $a_2 > 0$(similar discussion for $a_2 < 0$), we have

$$\begin{aligned}\xi &= x_0 + x_1\tau + x_2\tau^2 + x_3\tau^3 + x_4\tau^4\\ &= (x_0 - d_0q^2) + (x_1 - d_0a_1q)\tau + (x_2 - d_0a_2)\tau^2\\ &\quad +(x_3 - d_0a_1)\tau^3 - d_0\tau^4\\ &= (x_0 - d_0q^2) + (x_1 - d_0a_1q + d_1q^2)\tau + (x_2 - d_0a_2\\ &\quad +d_1a_1q)\tau^2 + (x_3 - d_0a_1 + d_1a_2)\tau^3 + (-d_0 + d_1a_1)\tau^4\\ &\quad +d_1\tau^5\\ &= (x_0 - d_0q^2) + (x_1 - d_0a_1q + d_1q^2)\tau + (x_2 - d_0a_2\\ &\quad +d_1a_1q + d_2q^2)\tau^2 + (x_3 - d_0a_1 + d_1a_2 + d_2a_1q)\tau^3\\ &\quad +(-d_0 + d_1a_1 + d_0a_2)\tau^4 + (d_1 + d_2a_1)\tau^5 + d_2\tau^6,\end{aligned}$$

which implies that ξ is a τ-adic expansion of length at most 7, where d_0 is an integer close to $\frac{2}{\sqrt{5}}q^{1/4}$ such that $|x_0 - d_0q^2| \leq q^2/2$. $d_1 = 0, -1$ if $x_1 > 0$, or $d_1 = 1, 2$ if $x_1 < 0$. $d_2 = 0$ if $d_1 = 0, 1$, or $d_2 = 0, 1$ if $d_1 = -1$, or $d_2 = -1$ if $d_1 = 2$.

By almost the same discussions, we will show that ξ can be expressed as a τ-adic expansion of length at most 8 if q is a non-square no less than 37.

If q is a square smaller than 25 or a non-square smaller than 31, then ξ may go to a cyclic τ-adic expansion with the coefficients in R. But, we can easily show that if $a_2 \neq 2q + (a_1^2 - 1)/4$ and ξ does not go to a cyclic τ-adic expansion, then ξ will be a τ-adic expansions of length at most 8.

Our discussions and results above can be naturally generalized to the curves of genus $g \geq 3$, though it will be a bit more burdensome for high genus. In general, there exist fixed integers k_i and l_i non-related to q such that

$$|x_i| < \begin{cases} k_i q^{(5g-2i-1)/4} & \text{if } q \text{ is square} \\ l_i q^{(3g-i-1)/2} & \text{if } q \text{ is non-square} \end{cases} \tag{15}$$

hold for $i = 0, 1, \cdots, 2g-1$. And, every element $\xi \in \mathbb{Z}[\tau]$ can be represented as a τ-adic expansion of length at most $2g + 4$ as long as the related characteristic polynomial $P(T)$ will not lead to cyclic τ-adic expansions. □

5 Cyclic τ-adic expansions in $\mathbb{Z}[\tau]$

Let $q = 9$. Then, $a_1 = M_1 - 9 - 1 \leq 9, 6a_1 - 18 < a_2 \leq a_1^2/4 + 18$. If $a_1 = 9$, then $a_2 = 37$ or 38, and hence, $P(T) = T^4 + 9T^3 + 37T^2 + 81T + 81$ or $P(T) = T^4 + 9T^3 + 38T^2 + 81T + 81$. We have

$$\begin{aligned} 81 = & -a_2\tau^2 + (a_2 - 9)\tau^3 + ((89 - a_2)\tau^4 + (a_2 - 8)\tau^5 \\ & +8\tau^6 + \tau^7) \\ = & -a_2\tau^2 + (a_2 - 9)\tau^3 + -(a_2 - 8)\tau^4 \\ & -\tau((89 - a_2)\tau^4 + (a_2 - 8)\tau^5 + 8\tau^6 + \tau^7), \end{aligned}$$

which implies that 81 can only be expressed in a cyclic τ-adic expansion, that is, both $P(T) = T^4 + 9T^3 + 37T^2 + 81T + 81$ and $P(T) = T^4 + 9T^3 + 38T^2 + 81T + 81$ lead to cyclic τ-adic expansions.

If $a_1 = 8$, then $31 \leq a_2 \leq 33$. If $a_2 = 32, 33$, then ξ has a τ-adic expansion of length at most 8. If $a_2 = 31$, then we can only get a cyclic τ-adic expansion for $\xi = 81$. Thus, for the curves with the characteristic polynomial $P(T) = T^4 \pm 8T^3 + 31T^2 \pm 72T + 81$, we can not get a finite τ-adic expansion for 81 with the coefficients in $\{-\lceil q^2/2 \rceil + 1, \cdots, \lfloor q^2/2 \rfloor\}$. But if we add ± 41 to the coefficient set, then 81 will have a τ-adic expansion of length five.

Theorem 3 *Let C is a hyperelliptic curve of genus g over $\mathbb{F}_q$ and*

$$P(T) = T^{2g} + a_1T^{2g-1} + \cdots + a_gT^g + \cdots + a_1q^{g-1}T + q^g$$

be its characteristic polynomial with a root τ. If there exists $\xi \in \mathbb{Z}[\tau]$ such that ξ can only be expressed in a cyclic τ-adic expansion, then we call that $P(T)$ leads to cyclic τ-adic expansions.

Let $\tilde{C}$ be a quadratic twist of the hyperelliptic curve C and its characteristic polynomial $\tilde{P}(T)$ as

$$T^{2g} - a_1T^{2g-1} + \cdots + (-1)^g a_gT^g + \cdots - a_1q^{g-1}T + q^g$$

with a root of $\tilde{\tau}$. Then,

1) $P(T)$ leads to cyclic τ-adic expansions if and only if the following inequality (16) *holds.*

$$\sharp\mathbb{J}_C(\mathbb{F}_q) \leq \lfloor q^g/2 \rfloor \quad \text{or} \quad \sharp\mathbb{J}_{\tilde{C}}(\mathbb{F}_q) \leq \lfloor q^g/2 \rfloor. \tag{16}$$

2) There exists an element $\xi \in \mathbb{Z}[\tau]$ which has only a cyclic τ-adic expansion if and only if there exists an element $\tilde{\xi} \in \mathbb{Z}[\tilde{\tau}]$ which has only a cyclic $\tilde{\tau}$-adic expansion, that is, $P(T)$ leads to cyclic τ-adic expansions if and only if $\tilde{P}(T)$ leads to cyclic $\tilde{\tau}$-expansion.

Proof 1). Suppose ξ can only be expressed as a cyclic τ-adic expansion and

$$\begin{aligned} \xi = & \, x_0 + x_1\tau + x_2\tau^2 + x_3\tau^3 \\ = & \, r_0 \pm \tau(x_0 + x_1\tau + x_2\tau^2 + x_3\tau^3), \\ & x_0 > \lfloor q^2/2 \rfloor, |r_0| \leq \lfloor q^2/2 \rfloor. \end{aligned}$$

Let $x_0 - r_0 = dq^2$, then $x_0 = \pm(x_1 - da_1q)$, $x_1 = \pm(x_2 - da_2)$, $x_2 = \pm(x_3 - da_1)$ and $x_3 = \mp d$. and hence, $x_0 = -d - da_1 - da_2 - da_1q = d(q^2 - \sharp\mathbb{J}_C(\mathbb{F}_q))$ when $x_3 = -d$, or $x_0 = -d + da_1 - da_2 + da_1q = d(q^2 - \sharp\mathbb{J}_{\tilde{C}}(\mathbb{F}_q))$ when $x_3 = d$.

It follows $r_0 = -d\,\sharp\mathbb{J}_C(\mathbb{F}_q)$ and $d\,\sharp\mathbb{J}_C(\mathbb{F}_q) \leq \lfloor q^2/2 \rfloor$, or $r_0 = -d\,\sharp\mathbb{J}_{\tilde{C}}(\mathbb{F}_q)$ and $d\,\sharp\mathbb{J}_{\tilde{C}}(\mathbb{F}_q) \leq \lfloor q^2/2 \rfloor$. Hence

$$\sharp\mathbb{J}_C(\mathbb{F}_q) = |r_0|/d \leq \lfloor q^2/2 \rfloor$$

or

$$\sharp\mathbb{J}_{\tilde{C}}(\mathbb{F}_q) = |r_0|/d \leq \lfloor q^2/2 \rfloor.$$

Suppose ξ can be expressed as the following cyclic τ-adic expansion and

$$\begin{aligned} \xi = & \, x_0 + x_1\tau + x_2\tau^2 + x_3\tau^3 \\ = & \, r_0 + r_1\tau + \tau^2(x_0 + x_1\tau + x_2\tau^2 + x_3\tau^3), \\ & x_0 > \lfloor q^2/2 \rfloor, |r_i| \leq \lfloor q^2/2 \rfloor, i = 0, 1. \end{aligned}$$

Let $x_0 - r_0 = dq^2$ and $x_1 - da_1q - r_1 = eq^2$, then we have $x_0 = x_2 - da_2 - ea_1q$, $x_1 = x_3 - da_1 - ea_2$, $x_2 = -d - ea_1$ and $x_3 = -e$.

Thus, $x_0 = dq^2 + r_0 = -d - ea_1 - da_2 - ea_1q$ and $x_1 = da_1q + eq^2 + r_1 = -e - da_1 - ea_2$, which implies

$$r_0 + r_1 = -(d + e)\,\sharp\mathbb{J}_C(\mathbb{F}_q). \tag{17}$$

Hence, if $|d + e| \geq 2$, then

$$\sharp\mathbb{J}_C(\mathbb{F}_q) \leq (|r_0| + |r_1|)/|d + e| \leq \lfloor q^2/2 \rfloor.$$

If $d + e = 0$, then $r_0 = -d\,\sharp\mathbb{J}_{\tilde{C}}(\mathbb{F}_q)$, and so,

$$\sharp\mathbb{J}_{\tilde{C}}(\mathbb{F}_q) \leq |r_0|/d \leq \lfloor q^2/2 \rfloor/d \leq \lfloor q^2/2 \rfloor.$$

If $d + e = \pm 1$, then for $\tilde{\xi} = x_0 - x_1\tilde{\tau} + x_2\tilde{\tau}^2 - x_3\tilde{\tau}^3$, we have

$$\tilde{\xi} = r_0 - r_1\tilde{\tau} + \tilde{\tau}^2(x_0 - x_1\tilde{\tau} + x_2\tilde{\tau}^2 - x_3\tilde{\tau}^3).$$

It follows

$$r_0 - r_1 = (-2d + 1)\,\sharp\mathbb{J}_{\tilde{C}}(\mathbb{F}_q). \tag{18}$$

From the equations (17) and (18) we deduce that

$$\sharp\mathbb{J}_C(\mathbb{F}_q) \leq \lfloor q^2/2 \rfloor \quad \text{or} \quad \sharp\mathbb{J}_{\tilde{C}}(\mathbb{F}_q) \leq \lfloor q^2/2 \rfloor. \tag{19}$$

Similarly, we can easily show that Inequality (19) also holds if ξ has a longer period expression.

On the other hand, we suppose $\sharp\mathbb{J}_C(\mathbb{F}_q) \leq \lfloor q^2/2 \rfloor$(similar discussion for $\sharp\mathbb{J}_{\tilde{C}}(\mathbb{F}_q) \leq \lfloor q^2/2 \rfloor$), and let

$$\begin{cases} x_0 &= a_1q + a_2 + a_1 + 1 \\ x_1 &= a_2 + a_1 + 1 \\ x_2 &= a_1 + 1 \\ x_3 &= 1 \end{cases}$$

Then, we have

$$\begin{aligned} \xi &= x_0 + x_1\tau + x_2\tau^2 + x_3\tau^3 \\ &= \sharp\mathbb{J}_C(\mathbb{F}_q) + \tau(x_0 + x_1\tau + x_2\tau^2 + x_3\tau^3) \end{aligned}$$

is a cyclic τ-adic expansion.

2). Suppose

$$\begin{aligned} \xi &= x_0 + x_1\tau + x_2\tau^2 + x_3\tau^3 \\ &= r_0 + \tau(x_0 + x_1\tau + x_2\tau^2 + x_3\tau^3) \\ &\text{(with } x_0 > \lfloor q^2/2 \rfloor, |r_0| \leq \lfloor q^2/2 \rfloor) \end{aligned}$$

is a cyclic τ-adic expansion, then

$$\begin{aligned} \tilde{\xi} &= x_0 - x_1\tilde{\tau} + x_2\tilde{\tau}^2 - x_3\tilde{\tau}^3 \\ &= r_0 - \tilde{\tau}(x_0 - x_1\tilde{\tau} + x_2\tilde{\tau}^2 - x_3\tilde{\tau}^3) \end{aligned}$$

is is a cyclic $\tilde{\tau}$-adic expansion.

Similar discussions will show that Theorem 3 still holds for the hyperelliptic curve of genus $g > 2$. □

For example, The curves with $P(T) = T^4 \pm 9T^3 + 38T^2 \pm 81T + 81$, $P(T) = T^4 - 5T^3 + 15T^2 - 25T + 25$ or $P(T) = T^6 - 7T^5 + 21T^4 - 49T^3 + 147T^2 - 343T + 343$ will lead to cyclic τ-adic expansions. For $q = g = 2$, only the non-supersingular curves with $P(T) = T^4 \pm 2T^3 + 2T^2 \pm 4T + 4$ lead to cyclic τ-adic expansions. For $q = 3$ and $g = 2$, only the non-supersingular curves with $P(T) = T^4 \pm 2T^3 + 2T^2 \pm 6T + 9$ or $T^4 \pm T^3 - 2T^2 \pm 3T + 9$ or $T^4 \pm 3T^3 + 5T^2 \pm 9T + 9$ will lead to cyclic τ-adic expansions.

Based on Hasse-Weil Theorem, that is, $(\sqrt{q} - 1)^{2g} \leq \sharp\mathbb{J}_C(\mathbb{F}_q) \leq (\sqrt{q}+1)^{2g}$, if q is an integer such that $(\sqrt{q}-1)^{2g} \geq q^g/2$, then, the corresponding hyperelliptic curves will not lead to cyclic expansion. For $g = 2$, 3 and 4, we have $q \geq 37$, 83 and 139, respectively.

If the curve C with $P(T)$ leads to cyclic expansion, then by Theorem 3, $P(1) \leq \lfloor q^g/2 \rfloor$ or $P(-1) \leq \lfloor q^g/2 \rfloor$, hence we can add $\pm(P(1) - q^g)$ or $\pm(P(-1) - q^g)$ to the coefficient set to make the expansion finite. For example, if the coefficient set is $\{0, \pm 1, \cdots, \pm 39, \pm 40\} \bigcup \{\pm 51\}$, then for the hyperelliptic curves with $P(T) = T^4 \pm 9T^3 + 38T^2 \pm 81T + 81$, all τ-adic expansions are finite.

6 Optimizing the length of the τ-expansion

Let $\beta = b_0 + b_1\tau + \cdots + b_{2g-1}\tau^{2g-1} \in \mathbb{Z}[\tau]$, then since $P(T)$ is irreducible, $P(T)$ and $B(T) = b_0 + b_1T + \cdots + b_{2g-1}T^{2g-1}$ are coprime. Hence, there exist polynomials $U(T), V(T) \in \mathbb{Z}[T]$ such that

$$U(T)B(T) + V(T)P(T) = 1.$$

Replace T with τ, we have $\beta^{-1} = U(\tau) \in \mathbb{Z}[\tau]$. That is, we can use the extended Euclidean algorithm to compute the inverse of any element of $\mathbb{Z}[\tau]$.

For any divisor $D \in \mathbb{J}(\mathbb{F}_{q^n})$, we have $\phi^n(D) = D$. Furthermore, by Lemma 1, we have

$$\sharp\mathbb{J}(\mathbb{F}_{q^n}) = \prod_{i=1}^{2g}(1 - \tau_i^n) = \prod_{i=1}^{2g}(1 - \tau_i)\prod_{i=1}^{2g}(\sum_{j=0}^{n-1}\tau_i^j).$$

If for some $D \in \mathbb{J}(\mathbb{F}_{q^n})$, $\phi(D) = D$, that is, $(1 - \tau)D =< 1, 0 >$, then $\prod_{i=1}^{2g}(1 - \tau_i)D =< 1, 0 >$, that is, $\sharp\mathbb{J}(\mathbb{F}_q)D =< 1, 0 >$. For practical cryptosystems, the divisor D should be chosen to be of large order. Hence, $\sharp\mathbb{J}(\mathbb{F}_q)D =< 1, 0 >$ generally does not hold. Thus, for a large multiplier m, we can obtain the divisor multiplication mD by computing $\bar{m}D$ with $\bar{m} = m \bmod (\tau^n - 1)$ or $m \bmod (\frac{\tau^n-1}{\tau-1})$.

Similar to Theorem 3.1 in (8), we have the following Lemma 5.

Lemma 5 *For any positive integers n and m, there exists $\bar{m} \in \mathbb{Z}[\tau]$, such that*

1. $\bar{m} = m \bmod (\tau^n - 1)$,

2. $N(\bar{m}) \leq \frac{\sqrt{g}(q^g-1)(\sqrt{q}^n+1)}{2(\sqrt{q}-1)}$.

We can prove this lemma by letting $s = m/(\tau^n - 1) = \sum\limits_{i=0}^{2g-1} s_i\tau^i \in \mathbb{Q}[\tau]$, $r = \sum\limits_{i=0}^{2g-1} \lfloor s_i + 1/2 \rfloor\tau^i$ and $\bar{m} = m - r(\tau^n - 1)$.

By letting $\alpha = m(\tau - 1)/(\tau^n - 1) = \sum\limits_{i=0}^{2g-1} \alpha_i\tau^i \in \mathbb{Q}[\tau]$, $\gamma = \sum\limits_{i=0}^{2g-1} \lfloor \alpha_i + 1/2 \rfloor\tau^i \in \mathbb{Z}[\tau]$ and $\bar{m} = m - \gamma(\tau^n - 1)/(\tau - 1)$, we can prove the following Lemma 6.

Lemma 6 *For any positive integers n and m, there exists $\bar{m} \in \mathbb{Z}[\tau]$, such that*

1. $\bar{m} = m \bmod ((\tau^n - 1)/(\tau - 1))$,

2. $N(\bar{m}) \leq \frac{\sqrt{g}(q^g-1)(\sqrt{q}^n-1)}{2(\sqrt{q}-1)^2}$.

Thus, by Theorem 1, Theorem 2, Lemma 5 and Lemma 6, we obtain the following Theorem 4.

Theorem 4 *Let C be a hyperelliptic curve of genus g over $\mathbb{F}_q$, and let τ be a root of C's irreducible characteristic polynomial $P(T)$. If C does not lead to cyclic τ-adic expansions, then for a large positive integer m, m is congruent, modulo $\tau^n - 1$ or $(\tau^n - 1)/(\tau - 1)$, to a τ-adic expansion of length at most $n + 4g + 5$ or $n + 4g + 4$, respectively. If C leads to cyclic τ-adic expansions, then this conclusion still holds if $P(-1) - q^g$ or $P(1) - q^g$ is added to the coefficient set.*

7 An efficient scalar multiplication algorithm

According to the above discussion, we can obtain an efficient scalar multiplication algorithm. This algorithm is composed of the four steps: pre-computing, reducing the multiplier, converting the reduced multiplier into Frobenius expansion and computing scalar multiplications.

Let $a_0 = 1$, $P[i] = a_iq^{g-i}$ and $P[g + i] = a_{g-i}$ for $i = 1, \ldots, g$.

Algorithm 1 Compute scalar multiplication by τ-adic Expansion

```
Input: a large positive integer multiplier $m$ and a divisor $D \in \mathbb{J}(\mathbb{F}_{q^n})$.
Output: $mD$.
I) Pre-computing :
    1. For $0 < r \leq \lfloor q^g/2 \rfloor$, compute $rD$ by (signed) binary
       method and store it as $D_r$.
    2. For $-\lceil q^g/2 \rceil + 1 \leq r < 0$, set

           $rD := < x((-r)D), -y((-r)D) - h(u) >$
    and store it as $D_r$, where $x((-r)D)$ and $y((-r)D)$ denote
    the first polynomial and the second polynomial of the divisor
    $(-r)D$, respectively.
II) Computing $m$ mod $(\tau^n - 1)$ :
  1. Find integers $s[i]$ such that
         $s = \sum_{i=0}^{2g-1} s[i]\tau^i = \tau^n - 1$ :
     1) Initialize $s[i] := -P[i]$ for $0 \le i \le 2g-1$.
     2) For $k$ from 1 to $n - 2g$, do
        (a) $\Delta := s[2g-1]$.
        (b) For $i$ from 1 to $2g-1$, set
            $s[i] := s[i-1] - P[i]\Delta$, and $s[0] := -P[0]\Delta$.
     3) Set $s[0] := s[0] - 1$.
     4) Set $s := \sum_{i=0}^{2g-1} s[i]\tau^i$.
  2. Applying Extended Euclidean Algorithm, there exist $t$, $u \in \mathbb{Z}[\tau]$
     with $\deg_\tau t \le 2q - 1$, such that
         $t \cdot s + u \cdot P(\tau) = 1.$
  3. For $t := \sum_{i=0}^{2g-1} t_i\tau^i$, set
         $\alpha := \sum_{i=0}^{2g-1} \lfloor m \cdot t_i + 1/2 \rfloor \tau^i.$
  4. Set $\bar{m} := m - s\alpha \bmod P(\tau)$.
III) Supposing $\bar{m} = \sum_{i=0}^{2g-1} \bar{m}_i\tau^i$ and converting $\bar{m}$ into a $\tau$-adic expansion:
  1. $j := 0$; $k := 0$.
  2. If $\bar{m} \neq 0$, then do
     (a) Select
         $r_j \in \{-\lceil q^g/2 \rceil + 1, \ldots, -1, 0, 1, \ldots, \lfloor q^g/2 \rfloor\}$
         such that $q^g | (\bar{m}_0 - r_j)$.
     (b) Set $d := (\bar{m}_0 - r_j)/q^g$.
     (c) Set $\bar{m} := \sum_{i=0}^{2g-2} (\bar{m}_{i+1} - P[i+1]d)\tau^i - d\tau^{2g-1}$.
     (d) Set $j := j + 1$ and $k := k + 1$, and go back.
IV) Computing $\bar{m}D$ :
  1. Initialize $B := D_{r_{k-1}}$.
  2. For $i$ from $k - 2$ downto 0 do
     (a) Set $B := \phi(B)$.
     (b) Set $B := B + D_{r_i}$.
V) Output $B$ as $mD$.
```

Since the multiplier m can also be reduced by modulo $(\tau^n - 1)/(\tau - 1)$, Steps 1 in Step II) can be replaced by the following steps:

```
Step II*) Computing $m$ mod $(\tau^n - 1)/(\tau - 1)$:
  1. Find integers $s[i]$ such that
     $s = \sum_{i=0}^{2g-1} s[i]\tau^i = (\tau^n - 1)/(\tau - 1)$:
     1) Initialize $0 \le i \le 2g - 1$, set $s[i] := 1 - P[i]$ and
        $t[i] := -P[i]$;
     2) For $k$ from 1 to $n - 2g - 1$ do
        (a) Set $\Delta := t[2g-1]$ and $t[0] := -P[0]\Delta$;
        (b) For $i$ from 1 to $2g - 1$, set $t[i] := t[i-1] - P[i]\Delta$
            and $s[i] := s[i] + t[i]$;
     3) Set $s := \sum_{i=0}^{2g-1} s[i]\tau^i$;
```

Note that if $P(1) \le \lfloor q^g/2 \rfloor$ or $P(-1) \le \lfloor q^g/2 \rfloor$, then add $P(1) - q^g$ or $P(-1) - q^g$ to the coefficient set. Take $P(1) < q^g/2$ for example, we only make some minor modifications in Algorithm 1 as follows:

First, add the computation of $\pm(q^g - P(1))D$ in the precomputation step; Second, change the step (a)-(b) in Step III) as follows:

(a*) If $|\bar{m}_0| \le \lfloor q^g/2 \rfloor$ or $\bar{m}_0 = P(1) - q^g$, then set $r_j := \bar{m}_0$, otherwise, go to the next step:

(b*) Select
$r_j \in \{P(1) - q^g, -\lceil q^g/2 \rceil + 1, \cdots, -1, 0, 1, \cdots, \lfloor q^g/2 \rfloor\}$ such that $q^g | (\bar{m}_0 - r_j)$.

We implement Step II)-III) in Algorithm 1 in Maple for five hyperelliptic curves and get the Table 1. Table 1 lists five hyperelliptic curves and the bits of the orders of their corresponding Jacobian groups, the average lengths and densities of the τ-adic expansions of the multipliers approximate to the Jacobian group orders, and the average lengths (1)-(2) and densities (1)-(2) of the τ-adic expansions of the multipliers after reduced by modulo $(\tau^n - 1)/(\tau - 1)$ or $(\tau^n - 1)$, respectively. The density means the ratio of the number of the non-zero coefficients to the length in a τ-adic expansion.

The corresponding characteristic polynomials of the five hyperelliptic curves in Table 1 are $T^4 + 2T^3 + 3T^2 + 4T + 4$, $T^4 - 2T^3 + 2T^2 - 6T + 9$, $T^6 + 2T^4 - 2T^3 + 4T^2 + 8$, $T^4 - 4T^3 + 11T^2 - 20T + 25$, and $T^6 + 2T^5 + 4T^4 + 14T^3 + 20T^2 + 50T + 125$, respectively.

Table 1 shows that, when the multipliers are reduced by modulo $(\tau^n - 1)/(\tau - 1)$ or $\tau^n - 1$, the average lengths of the τ-adic expansions are between $n - 2$ and $n + g$, or between $n + 1$ and $n + g + 1$, respectively. It also shows that, if the multipliers are not reduced, then the average length of τ-adic expansions is about q^g times of the extension degree of the field. While their average densities are almost the same whether the multipliers are reduced or not.

Suppose the multiplier $m \sim q^{gn}$(near to the Jacobian order). Then, to compute mD, the binary method needs on average $\frac{ng}{3} \log_2 q$ divisor additions and $ng \log_2 q$ divisor doublings. While according to our experiments, Algorithm 1 needs on average $n + \frac{g}{2}$ divisor additions and $g \log_2 q - 1$ divisor doublings, plus about $n + \frac{g}{2}$ divisor evaluations of Frobenius endomorphism. If we implement Algorithm 1 in some *normal basis*, then the Frobenius evaluation cost can be considered free. Hence, according to Theorem 14 in (11), Algorithm 1 will cost about 55% less than the signed binary method for the curves listed in Table 1. It follows that our algorithm will greatly speed up the implementation of hyperelliptic curve cryptosystems since the divisor scalar multiplication is the most time-consuming operation.

8 Conclusion

In this paper, we have applied Frobenius endomorphism and Euclidean length to reduce the multipliers in divisor scalar multiplications by modulo $\tau^n - 1$ or $(\tau^n - 1)/(\tau - 1)$, and show that the upper bound of the lengths of the reduced multipliers' τ-adic expansions is $n + 4g + 5$. In addition, our experiment results show that the lengths of the multipliers' τ-adic expansions

Hyperelliptic curves	q	n	bits of orders	average length	average density	reduced average length(1)	reduced average density(1)	reduced average length(2)	reduced average density(2)
$v^2 + (u^2 + u + 1)v = u^5 + u^4 + u^3 + u$	2	61	122	242.125	0.749	62.000	0.761	62.833	0.756
		67	134	264.250	0.738	64.667	0.732	68.000	0.733
		73	146	291.000	0.758	70.500	0.756	73.000	0.752
		89	178	355.000	0.736	87.833	0.784	87.500	0.760
		113	226	451.000	0.741	110.667	0.768	112.333	0.712
$v^2 = u^5 + u^4 - u^3 + u^2 - u + 2$	3	61	194	244.000	0.832	61.833	0.865	62.000	0.869
		67	213	267.500	0.828	67.667	0.872	68.167	0.905
		97	308	386.833	0.844	97.667	0.887	99.500	0.901
		103	327	412.333	0.826	104.500	0.890	105.333	0.889
		113	359	452.000	0.824	112.667	0.873	114.667	0.903
$v^2 + v = u^7 + u^6 + u^5$	2	29	87	168.400	0.828	30.333	0.890	29.667	0.832
		37	111	210.000	0.861	37.667	0.858	38.333	0.878
		43	129	254.000	0.867	42.833	0.883	45.500	0.865
		59	177	352.000	0.845	60.500	0.859	59.833	0.847
		67	201	398.000	0.865	67.167	0.868	70.000	0.877
$v^2 = u^5 + u^4 + 2u^3 + u^2 + u + 2$	5	61	284	246.000	0.906	62.167	0.949	64.000	0.958
		67	312	270.000	0.916	68.667	0.973	70.500	0.962
		71	330	285.600	0.936	72.167	0.972	74.167	0.951
		79	367	318.000	0.936	81.167	0.955	81.667	0.943
		83	386	334.000	0.930	84.500	0.955	85.167	0.967
$v^2 = u^7 + u^5 + u^3 + u - 1$	5	29	203	174.000	0.986	31.667	0.984	33.500	0.985
		31	216	186.000	0.985	33.667	0.980	34.833	0.990
		37	258	222.000	0.998	40.833	0.988	40.333	0.979
		43	300	258.000	0.979	44.167	0.985	47.333	0.986
		53	370	318.000	0.991	55.167	0.988	57.667	0.991

Figure 1: Average Lengths and Densities of τ-adic Expansions

are actually between $n - 2$ and $n + g + 1$. While Günther et al(8) did experimentally show that the two hyperelliptic curves $v^2 + uv = u^5 + \alpha u^2 + 1$ ($\alpha = 0, 1$) have some τ-adic expansions of length about $n + \frac{4}{3}$ (which are near to our experimental result, but they did not give a theoretical proof.

In practical hyperelliptic curve cryptosystems, since the parameters q, g, n and the basic divisor D are relatively fixed, we can pre-compute $\phi^i(r_j D)$ for $1 \leq i \leq n + 4g + 4$ and $-\lfloor q^g/2 \rfloor + 1 \leq j \leq \lfloor q^g/2 \rfloor$ and then store the results as a table. If we employ this table, then our algorithm only needs at most $n + 4g + 4$ divisor additions, which is approximately one third computation expense that the binary method does. In addition, based on the Proposition 3.4 in (12), the elliptic curve rational point group $E_C(\mathbb{F}_{q^n})$ is isomorphic to the Jacobian group $\mathbb{J}_C(\mathbb{F}_{q^n})$ under their group law definitions when the curve C's genus $g = 1$, hence, our Algorithm 1 is also applicable to the scalar multiplication computations on $E_C(\mathbb{F}_{q^n})$.

Acknowledgement

This research is supported by the National Science Foundation of China (No.60763009), the science and technology Key Project of the Ministry of Education of China(No.207089), and Zhejiang Natural Science Foundation of Outstanding Youth Team Project(No.R1090138).

References

[1] N. Koblitz(1989). A Family of Jacobians suitable for discrete log cryptosystems, Advances in Cryptology-Crypto'88, LNCS 403, Springer-Verlag, pp.94-99.

[2] N. Koblitz(1989). Hyperelliptic cryptosystems, Journal of Cryptology, No.1, pp.139-150.

[3] J. Pollard(1978). Monte Carlo methods for index computation mod p, Mathematics of Computation, No.32, pp. 918ÍC924.

[4] S. C. Pohlig, M. E. Hellman(1978). An improved algorithm for computing logarithms over $GF(p)$ and its cryptographic significance. IEEE Trans. Information Theory, IT-24(1), pp.106ÍC110.

[5] A. Weil(1949). Number of solutions of equations in finite fields. Bull. AMS. 55, pp.497-508.

[6] V. Müller(1998). Fast multiplication on Elliptic Curve over Small fields of Characteristic Two. Journal of Cryptology. 11, pp. 219-234.

[7] N. P. Smart(1999). Elliptic Curve Cryptosystems over Small Fields of Odd Characteristic. Journal of Cryptology. 12, pp.141-151.

[8] Ch. Günther, T. Lange and A.Stein(2001). Speeding Up the Arithmetic on Koblotz Curves of Genus Two. LNCS 2012,pp. 106-117.

[9] K. Matsuo, J.Chao, S.Tsujii(2002). An improved baby step giant step algorithm for point counting of hyperelliptic curves over finite fields. LNCS. 2369, pp. 461-474.

[10] D. Maisner, E. Nart(2002). Abelian surfaces over finite fields as Jacobians. Experimental Mathematics. 11, pp. 321-337.

[11] A. Enge(2001). The Extended Euclidean Algorithm on Polynomials. and the Com-putational Efficiency of Hyperelliptic Cryptosystems,Design, Codes and Cryptography. 23(1), pp.53-74.

[12] J. H. Silverman(1986). The Arithmetic of Elliptic Curves, Spriger-Verlag, pp.66-67.

On the Security of Two Group Signature Schemes with Forward Security

Kitae Kim
Department of Mathematics, and
ISRL, Graduate School of IT&T, Inha University, Republic of Korea
ktkim@inha.ac.kr

Ikkwon Yie
Department of Mathematics, Inha University, Republic of Korea
ikyie@inha.ac.kr

Daehun Nyang
ISRL, Graduate School of IT&T, Inha University, Republic of Korea
nyang@inha.ac.kr

Keywords: digital signature, group signature, forward security, cryptanalysis

Received: September 22, 2008

A group signature scheme allows a group member of a group to sign messages on behalf of the group anonymously. In case of dispute, a special entity of the group, group manager, can reveal the signer of a valid group signature. In 2005, Zhang et al. proposed a new group signature with forward security based on their earlier scheme in ICICS 2003. Recently, Zhou et al. proposed a dynamic group signature scheme with forward security at GCC 2007, In the year of 2008, Zhang and Geung pointed out the scheme is insecure and suggested an improvement. In this paper, we analyze a security analysis of Zhang et al.'s group signature scheme and Zhou et al.'s group signature scheme. We also discuss why the improved Zhou et al.'s scheme by Zhang et al. is still insecure.

Povzetek: Analizirane so varnosti skupinskega podpisovanja dveh modelov: Zhou in Zhang.

1 Introduction

Following the first work by Chaum and van Heyst (10) in the year of 1991, many group signature schemes have been proposed and analyzed. In such a scheme, individual members of a group is allowed to sign messages on behalf of the group anonymously. Moreover, group signature schemes allow the group manager to reveal a signer's identity in case of dispute. Unforgeability, anonymity and traceability were noted as basic security requirements for group signature schemes by Chaum and van Heyst (10). Later, more security requirements such as unlinkability, collision-resistance, exculpability, and framing have been introduced. Informally, a secure group signature scheme must satisfy the following properties :

Unforgeability : Without knowledge of the secret key(s), no one can generate a valid group signatures. In other words, only the group members can sign messages on behalf of the group.

Anonymity : Anybody except the group manager has no information of the member's secret keys. Particularly, given a valid group signature, no one except the group manager can identify the signer.

Unlinkability : Even though seeing a list of signatures, anyone except the group manager can not relate two signatures together as being produced by the same member.

Traceability : It is not possible to produce signatures which can not be traced to one of the group that has produced the signature. That is, for given a group signature, the group manager is always able to determine who is the signer of the signature.

Exculpability : Neither member of the group nor the group manager can produce signatures on behalf of other group members. Sometimes, the requirement is used in a weaker form that group members except the group manager can not produce a valid signature that traced to other member of the group.

Coalition Resistance : Even though a set of group members collude together, it is not possible to generate signatures that cannot be traced to any of them. A weaker form that a set of members cannot produce a signature that is traced to other member than the set is sometimes called Framing.

In 2003, Zhang et al. (15) proposed a group signature scheme with forward security. However, G. Wang showed that Zhang et al.'s scheme is insecure by presenting several attacks (14). After that, Zhang et al. suggested a new group signature scheme in (16) and claimed that their scheme satisfies all the above requirements, and, in addition to, forward security.

Recently, Zhou et al. (17) proposed a dynamic group signature scheme with forward security. But, Zhang and Geng showed that the Zhou et al.'s scheme is universally forgeable and suggested an improvement in (18). The authors claimed that their improvement can be proved to be secure without presenting the details of the proof.

In this paper, we analyze the Zhang et al.'s new group signature scheme in (16) and the improvement of Zhou et al's group signature scheme (as well as the original Zhou et al.'s scheme). More precisely, we point out the open algorithm of new Zhang et al.'s scheme does not properly operate, and show their scheme is not secure even if the algorithm can be improved. In addition, we show that the Zhou et al.'s scheme and its improvement are insecure.

2 Zhang et al.'s Group Signature Scheme

Before presenting Zhang et al.'s group signature scheme, we briefly review some notions used in the scheme.

For a positive integer n, the *Euler phi function* (or *Euler totient function*) $\phi(n)$ is defined to be the number of positive integers less then n which are relatively prime to n. If a positive integer n is a composite of two primes, say $n = pq$, then $\phi(n) = (p-1)(q-1)$.

As RSA-like schemes, their group signature scheme is constructed on the group $\mathbb{Z}_n^\times$, where $\mathbb{Z}_n^\times = \{k : 0 < k < n \text{ and } \gcd(k,n) = 1\}$. Due to the security, n is usually chosen to be a product of two strong primes of the same size. A prime p is called *strong prime* if $(p-1)/2$ is also prime. To summarize, n is chosen to be $n = p_1 p_2$ such that p_1 and p_2 are large strong primes of the same size.

Additionally, two cryptographic primitives, a hash function and a signature of knowledge, are used in Zhang et al.'s group signature scheme. More precisely, it employs a coalition resistant hash function $h(\cdot)$, and a signature of knowledge SPK on the discrete logarithm : Given g and $y = g^\gamma$ for some γ, $SPK\{\gamma : y = g^\gamma\}()$ is a (non-interactive) proof of knowledge of γ

2.1 The Scheme

We briefly describe Zhang et al.'s new group signature scheme (16).

Setup. The group manager (GM) randomly chooses two strong primes p_1, p_2. Let $n = p_1 p_2$ and $G =< g >$ be a cyclic subgroup of $\mathbb{Z}_n^\times$. GM chooses an integer x as his secret key, and computes the public key $y = g^x \bmod n$. GM selects a random integer e such that $\gcd(e, \phi(n)) = 1$ and computes d such that $de = 1 \bmod \phi(n)$. The expected system life-time is divided into T intervals which are publicly known. Finally, GM publishes the public key $(y, n, g, e, h(\cdot), ID_{GM}, T)$, where $ID_{GM} \in \mathbb{Z}_n^\times$ is the identity of the group manager and $(c, s) = SPK\{\gamma : y = g^\gamma\}()$.

Join. If a user, say Bob, wants to join to the group, Bob executes an interactive protocol with GM as follows :

1. Bob chooses a random $k \in \mathbb{Z}_n^\times$ as his private key, and computes his identity $ID_B = g^k \bmod n$. Then he generates $(c, s) = SPK\{\gamma : ID_B = g^k\}()$. Finally, Bob keeps k privately and sends $(ID_B, (c, s))$ to the group manager.
2. Upon receiving $(ID_B, (c, s))$, GM verifies the signature of knowledge (c, s). If the verification holds, GM choose a random $\alpha \in \mathbb{Z}_n^\times$ and computes a triple (r_B, s_B, W_{B_0}) from
$$\begin{aligned} r_B &= g^\alpha \bmod n, \\ s_B &= \alpha + r_B x, \\ w_{B_0} &= (ID_{GM} r_B ID_B)^{-d^T} \bmod n. \end{aligned}$$
Then GM sends Bob (s_B, r_B, w_{B_0}) via a private channel, and stores $(ID_B, (c, s))$ together with (r_B, s_B, w_{B_0}) in his local database.
3. After Bob receives (s_B, r_B, w_{B_0}), he verifies
$$g^{s_B} \overset{?}{\equiv} r_B y^{r_B} \bmod n \quad (1)$$
$$ID_{GM} ID_B r_B \overset{?}{\equiv} w_{B_0}^{-e^T} \bmod n \quad (2)$$
If the equations (1) and (2) hold, Bob stores (s_B, r_B, w_{B_0}) as his initial membership certificate.

Evolve. Assume that Bob has the group membership certificate (s_B, r_B, w_{B_j}) at time period j. Then at time period $j+1$, he updates his group membership certificate as $(s_B, r_B, w_{B_{j+1}})$ by computing
$$w_{B_{j+1}} = \left(w_{B_j}\right)^e \bmod n,$$
where $w_{B_j} = (r_B ID_{GM} ID_B)^{-d^{T-j}} \bmod n$.

Sign. Suppose that Bob has the group membership certificate (s_B, r_B, w_{B_j}) at time period j. To sign a message m, Bob chooses random numbers $q_1, q_2, q_3 \in \mathbb{Z}_n^\times$, and computes
$$\begin{aligned} z_1 &= g^{q_1} y^{q_2} q_3^{e^{T-j}} \bmod n, \\ u &= h(z_1, m), \\ r_2 &= q_3 w_{B_j}^u \bmod n, \\ r_1 &= q_1 + (s_B + k)u, \\ r_3 &= q_2 - r_B u. \end{aligned}$$
The resulting group signature on m is $\sigma = (u, r_1, r_2, r_3, m, j)$.

Verify. Given $\sigma = (u, r_1, r_2, r_3, m, j)$, a verifier computes

$$z_1' = ID_{GM}^u g^{r_1} r_2^{e^{T-j}} y^{r_3} \bmod n,$$

and then checks $u' \overset{?}{=} h(z_1', m)$. If so, the verifier accepts the signature as a valid group signature from a legal group member.

Open. In case of a dispute, GM can open signature to reveal the actual identity of the signer. If $\sigma = (u, r_1, r_2, r_3, m, j)$ is a valid signature, GM operates as follows to find the signer's identity :

1. Computes $\eta = u^{-1} \bmod \phi(n)$.
2. Compute

$$z_1' \quad = \quad ID_{GM}^u g^{r_1} r_2^{e^{T-j}} y^{r_3} \bmod n.$$

3. Find w_B using (ID_B, r_B, w_{B_0}) in his local database satisfying

$$r_2/w_B^\eta \overset{?}{\equiv} (z'/g^{r_1}y^{r_3})^{d^{T-j}} \bmod n.$$

Revoke. Suppose GM wants to revoke Bob's membership certificate at time period j. Then GM performs as follows :

1. Compute $R_j = w_B(r_B ID_B)^{d^{T-j}} \bmod n$.
2. Publish (R_j, j) in the certificate revocation list (CRL).

Given a valid signature $\sigma = (u, r_1, r_2, r_3, m, j)$, a verifier can identify whether σ is produce by a revoked member. For this sake, he performs as follows :

1. Compute

$$z_1' \quad = \quad ID_{GM}^u g^{r_1} r_2^{e^{T-j}} y^{r_3} \bmod n \qquad (3)$$

2. Check

$$z_1' \left(r_2^{-1} R_j^u\right)^{e^{T-j}} \quad \overset{?}{\equiv} \quad g^{r_1} y^{r_3} \bmod n \qquad (4)$$

If the signature satisfies the equation of (3) and (4) then the verifier concludes that the signature is revoked.

2.2 Security Analysis of Zhang et al.'s Scheme

In (16), Zhang et al. analyzed the security of their scheme, and concluded that their scheme satisfies the security requirements of group signature schemes. However, we find the open algorithm is incorrectly designed. Moreover, their scheme does not satisfy the unforgeability even if one can improve the open algorithm to work correctly.

2.2.1 Incorrectness of the open algorithm

Suppose that $\sigma = (u, r_1, r_2, r_3, m, j)$ is a valid group signature signed by Bob with valid certificate (s_B, r_B, w_{B_j}). Then since

$$\begin{aligned} z_1 &= g^{q_1} y^{q_2} q_3^{e^{T-j}} \bmod n \\ &= ID_{GM}^u g^{r_1} r_2^{e^{T-j}} y^{r_3} \bmod n, \end{aligned}$$

we have $\frac{z_1}{g^{r_1}y^{r_3}} \equiv ID_{GM}^u r_2^{e^{T-j}} \pmod n$, and so

$$\left(\frac{z_1}{g^{r_1}y^{r_3}}\right)^{d^{T-j}} \equiv ID_{GM}^{ud^{T-j}} r_2 \bmod n \qquad (5)$$

$$= ID_{GM}^{ud^{T-j}} q_3 w_{B_j}^u \bmod n. \qquad (6)$$

On the other hand,

$$\frac{r_2}{w_{B_j}^\eta} \equiv r_2 w_{B_j}^{-u^{-1}} \bmod n \qquad (7)$$

$$\equiv q_3 w_{B_j}^u w_{B_j}^{-u^{-1}} \bmod n. \qquad (8)$$

We can easily see the quantities (6) and (8) are not the same : If these are equal then $ID_{GM}^{ud^{T-j}} \equiv w_{B_j}^{-u^{-1}} \pmod n$. Powering ue^{T-j} on both sides we have $ID_{GM}^{u^2} \equiv ID_{GM} ID_B r_B \pmod n$, which leads to a contradiction.

Remark 2.1. Before the invention of the above scheme, Zhang et al. already proposed a group signature scheme (15) entitled with "A novel group signature scheme with forward security" in ICICS 2003. At the same time, Wang suggested several attacks against the scheme (14). Lately, Zhang et al. proposed a new group signature scheme described above. Considering Zhang et al.'s early scheme, the following modification is seemed to be natural :

Given a valid group signature $\sigma = (u, r_1, r_2, r_3, m, j)$, the group manager does the following :

1. Compute $\eta = u^{-1} \bmod \phi(n)$.
2. Compute $z_1' = ID_{GM}^u g^{r_1} r_2^{e^{T-j}} y^{r_3} \bmod n$.
3. Find $(s_B, r_B, w_{B_j}, ID_B)$ in his local database satisfying

$$\left(\frac{r_2^\eta}{w_{B_j}}\right)^{e^{T-j}} \overset{?}{\equiv} \left(\frac{z_1}{g^{r_1}y^{r_3}}\right)^\eta ID_B r_B \pmod n.$$

However, this modification is not a correct improvement. Indeed, for a valid signature $\sigma = (u, r_1, r_2, r_3, m, j)$, every (w_{B_j}, ID_B, r_B) (not necessarily membership certificate of actual signer) satisfies the equation of 3.

2.2.2 Forgery attack

The above subsection illustrates that the open algorithm of Zhang et al.'s group signature scheme does not correctly work. Of course, there might be an improvement of the open algorithm while we couldn't find such one. However, we can break the scheme even if the algorithm can be modified to operate correctly. We remark that the attack in subsection is much similar to Wang's attack (14). Now, we describe our attack which can be mounted by anyone, not necessarily a group member.

Suppose that a group member Bob with certificate (r_B, s_B, w_{B_j}) was revoked by GM at time period j. Then the CRL should contain (R_j, j) where $R_j = w_{B_j}(r_B ID_B)^{d^{T-j}}$. Now an attacker Oscar (not a group member, outsider) can sign on any message M chosen by himself as follows :

1. Choose $q_1, q_2, q_3, \alpha, \beta \in \mathbb{Z}_n^\times$.
2. Compute

$$\begin{aligned} z_1 &= g^{q_1} y^{q_2} q_3^{e^{T-j}} \bmod n, \\ u &= h(z_1, M), \\ r_2 &= R_j^u g^{-\alpha} y^{\beta} q_3 \bmod n \\ r_1 &= q_1 + \alpha e^{T-j} \bmod n, \\ r_3 &= q_2 - \beta e^{T-j} \bmod n. \end{aligned}$$

In order to show that the tuple (u, r_1, r_2, r_3, M, j) is a valid group signature, it is enough to show that $z_1' = z_1$, where

$$\begin{aligned} z_1 &= g^{q_1} y^{q_2} q_3^{e^{T-j}} \bmod n, \\ z_1' &= ID_{GM}^u g^{r_1} r_2^{e^{T-j}} y^{r_3} \bmod n. \end{aligned}$$

We first note that

$$\begin{aligned} R_j &= w_{B_j}(r_B ID_B)^{d^{T-j}} \bmod n \\ &= (r_B ID_B ID_{GM})^{-d^{T-j}} (r_B ID_B)^{d^{T-j}} \bmod n \\ &= ID_{GM}^{-d^{T-j}} \bmod n. \end{aligned}$$

Then

$$\begin{aligned} z_1' &= ID_{GM}^u g^{r_1} r_2^{e^{T-j}} y^{r_3} \bmod n \\ &= ID_{GM}^u g^{r_1} (R_j^u g^{-\alpha} y^{\beta} q_3)^{e^{T-j}} y^{r_3} \bmod n \\ &= ID_{GM}^u g^{q_1 + \alpha e^{T-j}} R_j^{ue^{T-j}} g^{-\alpha e^{T-j}} \\ &\quad \cdot y^{\beta e^{T-j}} q_3^{e^{T-j}} y^{q_2 - \beta e^{T-j}} \bmod n \\ &= ID_{GM}^u ID_{GM}{}^{-d^{T-j} ue^{T-j}} \\ &\quad \cdot g^{q_1} y^{q_2} q_3^{e^{T-j}} \bmod n \\ &= g^{q_1} y^{q_2} q_3^{e^{T-j}} \bmod n = z_1 \end{aligned}$$

Thus, once the GM releases a revocation token (R_j, j) for a group member at time period j, everyone can generate valid group signatures during the same time period on any message. Since $R_i = R_j^{e^{i-j}}$, one can compute R_i for all $i > j$ from the token R_j and then mount the above attack. Therefore, one can generate valid signatures for any time period i where $i \geq j$.

3 Zhou et al.'s Group Signature Scheme

At GCC 2007, Zhou et al. proposed a dynamic group signature with forward security (17). Later, Zhang and Geung (18) showed the scheme is insecure by presenting a universal forgery attack, and proposed an improvement. However, we find that the improvement as well as the original scheme is insecure.

3.1 Brief Review of Zhou et al.'s Scheme

We first briefly describe the Zhou et al.'s group signature scheme in (17) as follows:

Setup. Let F_q be a finite field, $E : y^2 = x^3 + ax + b$ be an elliptic curve over the field , where q is a prime of n bits and $4a^3 + 27b^2 \bmod q \neq 0$, and $P \in E(F_q)$ be a (base) point whose order is a large prime number l. Let $\#E(F_q)$ and ψ denote the order of the elliptic curve and a function which makes the conversion from a point $P = (x, y) \in E(F_q)$ to x, respectively. We use $(P)_x$ instead of $\psi(P)$. Now, the group manager GM chooses a random $k_{GM} \in \mathbb{Z}_l^\times$ and then computes $K_{GM} = k_{GM}P$ as its public key. The GM's secret key is k_{GM}, and the group public key is K_{GM}. We assume that each user B has its identity ID_B which is an element of $E(F_q)$.

Join. When a user B wants to join the group, GM and B perform the following protocol :

1. B chooses a random $k_B \in \mathbb{Z}_l^\times$ as private key, and computes $K_B = k_B P$ as private key. Then B sends (K_B, ID_B) to the group manager.
2. Upon receiving (K_B, ID_B), GM chooses a random $u_B \in \mathbb{Z}_l^\times$, and computes $ID_B' = h(u_B || (ID_B)_x)P$. Then he sends ID_B' to the GM.
3. GM selects a random $v \in \mathbb{Z}_l^\times$, and computes $V_B = vP$ and

$$s_B = k_{GM} h((ID_B')_x || (V_B)_x) + v \bmod l.$$

GM sends (V_B, s_B) to the user B, and stores (ID_B, ID_B', u_B) in his local data base.

Finally, B gets its membership certificate (K_B, ID_B', V_B, s_B) and becomes a member of the group.

Sign. To sign a message $m \in \mathbb{Z}_l^\times$, a member B chooses a random $r \in \mathbb{Z}_l^\times$, and computes

$$\begin{aligned} R &= rP, \\ s &= (k_B - m(R)_x) r^{-1} \bmod l, \\ I &= ID_B' + ID_B + k_B K_{GM}. \end{aligned}$$

Then $\sigma = (m, s, R, I, K_B)$ is a group signature on the message m.

Verify. A verifier accepts a signature $\sigma = (M, \sigma_1, \sigma_2, \sigma_3, \sigma_4)$ if the following equations are satisfied

$$\begin{aligned} s_B P &\overset{?}{=} h((ID'_B)_x || (V_B)_x) K_{GM} + V_B, \\ \sigma_4 &\overset{?}{=} \sigma_1 \sigma_2 + m(\sigma_2)_x P. \end{aligned}$$

Note that the verifier must be able to search (s_B, V_B) corresponding to $\sigma_4 = K_B$. That is, this algorithm assumes that tuples (s_B, V_B, K_B) of all group members are public.

Open. Omitted (see (17))

3.2 A Comment on Zhou et al.'s Scheme and on its improvement by Zhang et al.

Zhang et al. (18) pointed out that Zhou et al.'s scheme is forgeable, and presented an improvement by including another hash function $H(\cdot) : \{0,1\}^* \times E(F_q) \to \mathbb{Z}_l^\times$ and revising the Sign and Verify algorithms. In particular, the revised Sign procedure is as follows :

To sign message m, a member B randomly selects $r \in \mathbb{Z}_l^\times$ to compute

$$\begin{aligned} R &= rP, \\ s &= (k_B - H(m,R)(R)_x) r^{-1} \bmod l, \\ I &= ID'_B + ID_B + k_B K_{GM}. \end{aligned}$$

Then the group signature on m is $\sigma = (m, s, R, I, K_B)$.

Though Zhang et al. claimed that their improved scheme is proved to be secure without detailed proofs, this revision as well as the Zhou et al.'s scheme is obviously linkable since two same pieces of information I and K_B are included in all group signatures generated by the same group member.

To avoid the linkability property, the deterministic information depending on the actual signer should be randomized. In this case, however, the group manager cannot trace the actual signer because the information I is used by the group manager in opening process. In other words, the Open algorithm cannot properly operate. Even worse, since K_B is critical value for signatures generated by B to be verified, no one can verify the signatures if the information is randomized. As a result, we conclude that the Zhou et al.'s scheme as well as Zhang et al.'s improvement cannot be repaired.

4 Conclusion

Zhang et al.'s new group signature scheme described in section 2 is based on their earlier version in (15). The earlier scheme was analyzed by Wang (14) and Cao (9), but no attack against the later scheme was announced. In this paper, we firstly presented security analysis of Zhang et al.'s new group signature scheme. By our analysis, the open algorithm of their scheme is incorrectly designed. Moreover, the scheme is not secure even though the open algorithm can be improved. Finally, we analyze Zhou et al.'s group signature scheme (17) and an improved scheme (18). The Zhou et al.'s scheme and the improved scheme are always linkable because each signature in the schemes includes deterministic values corresponding to a group member.

Acknowledgement

The authors would like to thank anonymous reviewers for their valuable comments. This work was supported by the IT R&D program of MIC/IITA,[2008-F-036- 01,Development of Anonymity-based u-Knowledge Security Technology].

References

[1] G. Ateniese, J. Camenisch, M. Joye, G. Tsudik (2000), *A practical and provably secure coalition-resistant group signature scheme*, CRYPTO'00, LNCS 1880, Springer-Verlag, pp. 255-270.

[2] J.H. An, Y. Dodis, T. Rabin (2002), *On the security of joint signature and encryption*, EUROCRYPT'02, LNCS 2332, Springer-Verlag, pp. 83-107.

[3] G. Ateniese, B. de Medeiros (2003), *Efficient group signatures without trapdoors*, ASIACRYPT'03, LNCS 2894, Springer-Verlag, pp. 246-268.

[4] M. Bellare, D. Micciancio, B. Warinschi (2003), *Foundations of group signatures : Formal definitions, simplified requirements and a construction based on general assumptions*, EUROCRYPT'03, LNCS 2656, Springer-Verlag, pp. 614-629.

[5] D. Boneh, X. Boyen, H. Shacham (2004), *Short group signatures*, CRYPTO'04, LNCS 3152, Springer-Verlag, pp. 45-55.

[6] X. Boyen, B. Waters (2006), *Compact group signatures without random oracles*, EUROCRYPT'06, LNCS 4004, Springer-Verlag, pp. 427-444.

[7] J. Camenisch and A. Lysyanskaya (2002), *Dynamic accumulators and application to efficient revocation of anonymous credentials*, CRYPTO'02, LNCS 2442, Springer-Verlag, pp. 61-76.

[8] J. Camenisch and M. Stadler (1997), *Efficent group signature schemes for large groups*, CRYPTO'97, LNCS 1294, Springer-Verlag, pp. 410-424.

[9] Z. Cao (2005), *Untraceability of Two Group signature Schemes*, Cryptology ePrint archive, http://eprint.iacr.org/2005/055.

[10] D. Chaum, E.V. Heyst (1992), *Group signatures*, EUROCRYPT'91, LNCS 547, Springer-Verlag, pp. 257-265.

[11] H. Park, S. Lim, I. Yie, K. Kim, J. Song (2009), *Strong unforgeability in group signature schemes*, to apper in Elsevier.

[12] D.X. Song (2001), *Practical forward secure group signature schemes*, Proc. of the 8th ACM CCS 2001, ACM press, pp. 225-234.

[13] G. Tsudik and S. Xu (2003), *Accumulating composites and improved group signing*, ASIACRYPT 2003, LNCS 2894, Springer-Verlag, pp. 269-286.

[14] G. Wang (2003), *On the security of a Group Signature with Forward Security*, ICISC 2003, LNCS 2971, Springer-Verlag, pp. 27-39.

[15] J. Zhang, Q. Wu and Y. Wang (2003), *A Novel Efficient Group Signature With Forward Security*, ICICS 2003, LNCS 2836, Springer-Verlag, pp. 292-300.

[16] J. Zhang, Q. Wu and Y. Wang (2005), *A New Efficient Group Signature With Forward Security*, Informatica, Vol. 29, No. 3, Slovenian Society Informatika, pp. 321-325.

[17] X. Zhou, X. Yang, P. Wei and Y. Hu (2007), *Dynamic group signature with forward security and its application*, Proc. of the 6th International Conference on Grid and Cooperative Computing (GCC 2007), IEEE Compupter Society, pp. 473-480.

[18] J. Zhang and Q. Geng (2008), *On the Security of Group Signature Scheme and Designated Verifier Signature Scheme*, Proc. of International Conference on Networking, Architecture, and Storage, IEEE Computer Society, pp. 351-358.

A Web-Mining Approach to Disambiguate Biomedical Acronym Expansions

Mathieu Roche
LIRMM - UMR 5506, CNRS
Univ. Montpellier 2,
34392 Montpellier Cedex 5 - France

Violaine Prince
LIRMM - UMR 5506, CNRS
Univ. Montpellier 2,
34392 Montpellier Cedex 5 - France

Keywords: web-mining, text-mining, natural language processing, BioNLP, named entities recognition, acronym, quality measures

Received: September 25, 2008

Named Entities Recognition (NER) has become one of the major issues in Information Retrieval (IR), knowledge extraction, and document classification. This paper addresses a particular case of NER, acronym expansion (or definition) when this expansion does not exist in the document using the acronym. Since acronyms may obviously expand into several distinct sets of words, this paper provides nine quality measures of the relevant definition prediction based on mutual information (MI), cubic MI (MI3), and Dice's coefficient. A combinaison of these statistical measures with the cosine approach is proposed. Experiments have been run on biomedical domain where acronyms are numerous. The results on our biomedical corpus showed that the proposed measures were accurate devices to predict relevant definitions.

Povzetek: Predstavljene so metode spletnega preiskovanja dvoumnih akronimov v domeni biomedicinskih baz.

1 Introduction

Named Entities Recognition (NER) has become one of the major issues in Natural Language Processing (NLP). The state-of-the-art literature in NER mostly focuses on proper names, temporal information, specific expressions in some technical or scientific fields for domain ontologies building, and so forth. A lot of work has been done on the subject, among which on acronyms, seen as particular named entities. Acronyms are very widely used in every type of text, and therefore have to be considered as a research issue as linguistic objects and as named entities.

An **acronym** is composed from the first letters of a set of words, written in uppercase style. This set of words is generally frequently addressed, which explains the need for a shortcut. It is also a specific multiword expression, such as 'Named Entities Recognition', abbreviated into NER, sometimes completely domain dependent (as NER or NLP are). In some cases, acronyms become proper names referring to countries or companies (like USA or IBM). However, most of the time, acronyms are domain or period dependent. They are contracted forms of multiword expressions where words might belong to the common language. As contracted forms, they might be highly ambiguous since they are created out of words first letters. For instance, NER, the acronym we use for `Named Entities Recognition` might also represent `Nippon Electrical Resources` or `Natural Environment Restoration`. An **expansion** (called **definition** too) is the set of words that defines the acronym.

In all cases, an acronym behaves like a named entity. However, the intrinsic ambiguity in most acronyms enhances the difficulty of finding which exact entity is referred by this artificial name. Literature has been addressing acronym building and expansion (see section 'related work') when the acronym definition is given in the text. However, choosing the right expansion for a given acronym in a given document, if no previous definition has been provided in the text, is an issue definitely belonging to NER, and not yet exhaustively tackled. The difficulty in acronym disambiguation is to automatically choose, as an expansion, the most appropriate set of words. This article tries to deal with this issue by offering a **quality measure** for each candidate expansion. In this context, let us name a a given acronym. For every a which expansion is lacking in a document d, we consider a list of n possible expansions for a: $a^1...a^n$. For instance, if NER is the acronym at stake, we could have NER^1 = `Named Entities Recognition`, NER^2 =`Nippon Electrical Resources`, and NER^3 =`Natural Environment Restoration`. Some web resources exist for providing acronym definitions as http://www.sigles.net/

or specialized biomedicine resources given by http://www.nactem.ac.uk/software/acromine/. In the experiments of this paper we have focused on biomedical data (18) because this domain uses a lot of polysemic acronyms.

The aim of our approach is to determine k ($k \in [1, n]$) such that a^k is the relevant expansion of a in the document d. To make such a choice, we provide different quality measures which relie on Web resources.

The presentation is structured as following: section 2 discusses the output of the related literature, section 3 focuses on the quality measure $AcroDef$, where context and web resources are essential characteristics to be taken into account. The section 4 extends the Turney's measures that we call $IADef$ measures. Section 5 gives an example of the nine quality measures based on $AcroDef$ and $IADef$ measures. Section 7 describes some experiments about $AcroDef$ and $IADef$ measures on biomedical domain. Finally conclusion and future work are suggested in section 8.

2 Related work

Among the several existing methods for acronyms and acronyms expansion extraction in the literature, we present here some significant works. First, acronyms detection within texts is an issue by itself. It involves recognizing a character chain as an acronym and not as an unknown or misspelled word. Most acronyms detecting methods rely on using specific linguistic markers.

Yates' method (28) involves the following steps: First, separating sentences by segments using specific markers (brackets, points) as frontiers.

For instance, the sentence:

```
The NER (Named Entity Recognition) system is
                presented.
```

will become

```
The NER | Named Entity Recognition | system is
                presented |
```

The second step compares each word of each segment with the preceding and following segments. In our example, the following comparisons are performed:

- `The` with `Named Entity Recognition`
- `NER` with `Named Entity Recognition`
- `Named` with `The NER`
- `Entity` with `The NER`
- and so forth...

Then the couples acronym/expansion are tested. The candidates acronym/definition are accepted if the acronym characters correspond to the first letters of the potential definitions words. In our example, the pair 'NER/Named Entity Recognition' is a good acronym/expansion candidate. The last step uses specific heuristics to select the relevant candidates. These heuristics rely on the fact that acronyms length is smaller than their expansion length, that they appear in upper case, and that long expansions of acronyms tend to use 'stop-words' such as determiners, prepositions, suffixes and so forth. In our example, the pair 'NER/Named Entity Recognition' is valid according to these heuristics.

Other works (2; 10) use similar methods based on the presence of markers associated to linguistic and/or statistical heuristics. For example, some recent works as (15) use statistical measurements from terminology extraction field. Okazaki and Ananiadou apply the C-value measure (7; 14) initially used to extract terminology. This one favors a candidate term that not appears often in a longer term. For instance, in a specialized corpus (Ophthalmology), the authors found the irrelevant term 'soft contact' while the frequent and longer term 'soft contact lens' is relevant. The advantage of the measure proposed by (15) is the independence of the characters alignment (actually, a lot of acronyms/definitions are relevant while the letters are in a different order as 'AW / water activity').

Other approaches based on supervised learning methods consist in selecting relevant expansions. In (27), the authors use the SVM approach (Support Vector Machine) with features based on acronyms/expansions informations (length, presence of special characters, context, etc). The work of (24) presents a comparative study of the main approaches (supervised learning methods, rules-based approaches) by combining domain-knowledge.

Our method is closer than Word Sense Disambiguation (WSD) approaches summarized in (13). A part of these WSD approaches uses machine-learning techniques to learn a classifier from labeled training sets (22; 9). In our case, we consider our method like unsupervised. But our system based on statistical measures and web-mining techniques differs with "bag of words" approaches described in (13). Note that our method will be combined with approaches of the literature to disambiguate definitions of biomedical domain (see section 6).

Larkey *et al.*'s method (10) uses a search engine to enhance an initial corpus of Web pages useful for acronym detection. To do so, starting from a list of given acronyms, queries are built and submitted to the AltaVista search engine.[1] Queries results are Web pages which URLs are explored, and eventually added to the corpus. Our method shares with (10) the usage of the Web. However, we do not look for existing expansions in text since we try to determine a possible expansion that would be lacking in the text where the acronym is detected. From that point of view, we are closer to works like Turney's (25), which are not

[1] http://www.altavista.com/

specifically about acronyms but which use the Web to define a ranking function. The algorithm PMI-IR (Pointwise Mutual Information and Information Retrieval) described in (25) queries the Web via the AltaVista search engine to determine appropriate synonyms to a given query. For a given word, noted $word$, PMI-IR chooses a synonym among a given list. These selected terms, noted $choice_i$, $i \in [1, n]$, correspond to the TOEFL questions. The aim is to compute the $choice_i$ synonym that gives the better score. To obtain scores, PMI-IR uses several measures based on the proportion of documents where both terms are present. Turney's formula is given below (1): It is one of the basic measures used in (25). It is inspired from Mutual Information described in (3).

$$score(\ choice_i\ ) = \frac{nb(\ word\ NEAR\ choice_i\ )}{nb(\ choice_i\ )} \qquad (1)$$

- $nb(x)$ computes the number of documents containing the word x,
- $NEAR$ (used in the 'advanced research' field of AltaVista) is an operator that precises if two words are present in a 10 words wide window.

With this formula (1), the proportion of documents containing both $word$ and $choice_i$ (within a 10 words window) is calculated, and compared with the number of documents containing the word $choice_i$. The higher this proportion is, the more $word$ and $choice_i$ are seen as synonyms. More sophisticated formulas have also been applied: They take into account the existence of negation in the 10 words windows. For instance, the words 'big' and 'small' are not synonyms if, in a given window, a negation associated to one of these two words has been detected, which is likely to happen, since they are antonyms (opposite meanings).

To enhance relevance to the document, our $AcroDef$ approach described in section 3 calculates the dependency between the words composing the possible expansions in order to rank them. In that sense, it is close to Daille's approach (4) which uses statistical measures to rank terms. Also, as defended in next section, we use other quality measures and attempt to relate as much as possible to the context, in order to significatively enhance basic measures.

3 Defining the $AcroDef$ measure

Several quality measures in the literature (8) are based on ranking function. They are brought out of various fields: Association rules detection, terminology extraction, and so forth.

To determine the expansion of an acronym starting from a list of co-occurrences of set of words, our aim is to provide a relevance ranking of this set using statistical measures. The most appropriate definition has to be placed at the top of the list by the $AcroDef$ (section 3) and $IADef$ (section 4) measures described in the following sections.

3.1 Basic $AcroDef$ measure based on Dice's coefficient

In this paper, the $AcroDef$ measure based on the Dice's coefficient is described. Other statistical measures like Mutual Information (MI) (3) and Cubic MI (26; 5) can be used. They are presented in the subsection 3.2.

Dice's Coefficient and Mutual Information are simple and effective because they use weak knowledge. Actually, they are based on a number of examples (in our case, the number of pages provided by a search engine and queries with the words of expansions) without the need to determine the counter-examples. Indeed, the counter-examples (used by a lot of quality measures (8)) are often more difficult to find in an unsupervised context based on statistical data from the Web.

The Dice's coefficient (21) used by our basic $AcroDef$ measure computes a sort of relationship between the words composing what is called a **co-occurrence**. This measure is defined by the following formula:

$$D(x,y) \quad = \quad \frac{2 \times P(x,y)}{P(x) + P(y)} \qquad (2)$$

For instance, with the acronym 'IR', x might represent the word 'Information' and y the word 'Retrieval'. It might also be a pair such as 'International' and 'Relations'.

Formula (2) leads directly to formula (3).[2]

$$Dice(x,y) \quad = \quad \frac{2 \times nb(x,y)}{nb(x) + nb(y)} \qquad (3)$$

Petrovic *et al.* (17) present an extension of the original Dice formula to three elements. In a natural way, we could extend this approach to n elements as follows:

$$Dice(x_1, ..., x_n) \quad = \quad \frac{n \times nb(x_1, ..., x_n)}{nb(x_1) + ... + nb(x_n)} \qquad (4)$$

Since our work, like many others, relies on Web resources, the nb function used in the preceding measures represents the number of pages provided by the search engine Exalead (http://www.exalead.fr/). The choice of Exalead has been determined by the fact that this search engine uses the NEAR function like the Turney's approach (formula (1)). This function will be used in other quality measures (i.e. $IADef$ measures) described in section 4.

Starting from the n extended Dice's formula (4), and using statistics provided by search engines we propose the basic $AcroDef$ measure (formula (5)).

[2] by writing $P(x) = \frac{nb(x)}{nb\_total}$, $P(y) = \frac{nb(y)}{nb\_total}$, $P(x,y) = \frac{nb(x,y)}{nb\_total}$

$$AcroDef_{Dice}(a^j) = \frac{\left|\{a_i^j | a_i^j \notin M_{stop}\}_{i\in[1,n]}\right| \times nb(\bigcap_{i=1}^{n} a_i^j)}{\sum_{i=1}^{n} nb(a_i^j | a_i^j \notin M_{stop})} \quad \text{where } n \geq 2 \tag{5}$$

- $\bigcap_{i=1}^{n} a_i^j$ represents the set of words a_i^j ($i \in [1, n]$) seen as a string (using *brackets* with Exalead and illustrated as follows: "$a_1^j...a_n^j$"). Then an important point of this formula is that the order of the words a_i^j is taken into account to calculate their dependency.
- M_{stop} is a set of stop-words (prepositions, determiners, etc). Then the pages containing only these words are not taken into account.
- $|.|$ represents the number of words of the set.

We used the acronym 'IR' as a basic example. With $a =$`IR`, two definitions are available:

a^1: `Information Retrieval`
and a^2: `International Relations`

Let us precise that the resulting pages numbers with both definitions are:

- $a_1^1 \cap a_2^1 =$ `Information`$\cap$`Retrieval`: $366,508$ resulting pages
- $a_1^2 \cap a_2^2 =$ `International` $\cap$ `Relations`: $1,021,054$ resulting pages

The obtained values with the $AcroDef$ formula (5) are:

$AcroDef_{Dice}(\texttt{IR}^1) = \frac{2\times nb(\texttt{Information } \cap \texttt{Retrieval})}{nb(\texttt{Information})+nb(\texttt{Retrieval})} = \frac{2\times 366508}{513072210+3202458} = 0.0014$

$AcroDef_{Dice}(\texttt{IR}^2) = \frac{2\times nb(\texttt{International } \cap \texttt{Relations})}{nb(\texttt{International})+nb(\texttt{Relations})} = \frac{2\times 1021054}{234463128+47716188} = 0.0072$

Practically, the first result comes back to submitting the three following queries to Exalead: `"Information Retrieval"` (`Information` $\cap$ `Retrieval`), `Information` and `Retrieval`.

In languages, many noun phrases contain stop-words such as determiners or prepositions, and thus, several acronym expansions will be composed of such elements. So, when the definition of an acronym contains a stop-word, it is neglected in the formula denominator. In English, stop-words are scarce, but sometimes appear in the acronym: Part-Of-Speech in often referred to as POS in computational and general linguistics. It designates the grammatical/lexical category to which the word belongs (verb, noun, etc). The preposition 'of' has given its first letter to the acronym, probably because it simplifies the acronym pronunciation.

3.2 Basic $AcroDef$ measure based on mutual information (MI and MI3)

We can use other statistical measures to calculate the dependancy between the words x and y: Mutual Information (MI) – formula (6) – and Cubic MI – formula (7). These measures are described in (20).

$$MI(x, y) = \frac{nb(x, y)}{nb(x) \times nb(y)} \tag{6}$$

$$MI3(x, y) = \frac{nb(x, y)^3}{nb(x) \times nb(y)} \tag{7}$$

Let us note that MI tends to extract rare and specific cooccurrences according to (23). Vivaldi *et al.* have estimated that the Cubic MI (MI3) was the best behaving measure (26). Then MI3 is used in several works related to terminology (26) and complex named entities extraction in texts (5).

Then we can use these formulas ((6) and (7)) in order to define other $AcroDef$ measures, respectively based on MI and Cubic MI. $AcroDef_{MI}$ and $AcroDef_{MI3}$ are given as follows:

$$AcroDef_{MI}(a^j) = \frac{nb(\bigcap_{i=1}^{n} a_i^j)}{\prod_{i=1}^{n} nb(a_i^j | a_i^j \notin M_{stop})} \quad \text{where } n \geq 2 \tag{8}$$

$$AcroDef_{MI3}(a^j) = \frac{nb(\bigcap_{i=1}^{n} a_i^j)^3}{\prod_{i=1}^{n} nb(a_i^j | a_i^j \notin M_{stop})} \quad \text{where } n \geq 2 \tag{9}$$

These measures enable to provide different experiment comparisons in section 7.

These basic formulas ((5), (8), (9)) do not take the context into account. This is a severe liability. Therefore, next subsection details a measure that relies on context to define a more relevant expansion choice for a given acronym.

3.3 Contextual $AcroDef$

In this paper, context is defined as a set of significant words present in the page where the acronym to expand is found. Of course, other definitions of the context notions have to be considered as extensions to this preliminary approach. However, even in this restricted point of view, several operational expressions of the context could be used:

- The n most frequent words (excepting stop words);
- The n most frequent proper names;
- The n most rare words;
- POS tags (1) or terminological information present in the surroundings of the considered item.

A combination of these expressions could also be envisaged. The experiments presented in this article (section 7) use a context represented by the most frequent words, and give satisfying results. In a sequel work, we plan to define the context with a richer set of information, namely, linguistic knowledge (lexical, syntactic, semantic) as the WSD (Word Sense Disambiguation) approaches (13) do.

Adding contextual information to $AcroDef$ (formula (5)) leads to formula (10). The principle underlying this formula is to apply statistical measures on a set of words of a given domain. So, the goal is not to count the dependency between the words of an acronym definition and those of the context, but to restrict the searching space. This restriction is a requirement for the word dependency computation (and not otherwise). The formula is written as follows:

$$AcroDef_{Dice}(a^j) = \frac{\left|\{a_i^j \text{ AND } C | a_i^j \notin M_{stop}\}_{i\in[1,n]}\right| \times nb(\bigcap_{i=1}^n a_i^j \text{ AND } C)}{\sum_{i=1}^n nb(a_i^j \text{ AND } C | a_i^j \notin M_{stop})}$$

$$\text{where } n \geq 2 \qquad (10)$$

In this formula, a_i^j AND C represents the pages containing the word a_i^j with all the words of the context C. For this we use the *AND* operator of Exalead. Our experiments presented in (20) show that the use of a context improves the results. If we consider our example a =`IR` with its two possible expansions (`Information Retrieval` and `International Relations`), the favored definition with $AcroDef$ is still `International Relations` with the 0.0072 value against the 0.0014 value for `Information Retrieval`. If we take as a context the following $C = \{\texttt{corpus}\}$ then we have:

$$AcroDef_{Dice}(\texttt{IR}^1) = \frac{2\times nb(\texttt{Information } \cap \texttt{Retrieval} \text{ AND } \texttt{corpus})}{nb(\texttt{Information} \text{ AND } \texttt{corpus}) + nb(\texttt{Retrieval} \text{ AND } \texttt{corpus})}$$
$$= \frac{2\times 19270}{2079155+55253} = 0.0181$$

$$AcroDef_{Dice}(\texttt{IR}^2) = \frac{2\times nb(\texttt{International } \cap \texttt{Relations} \text{ AND } \texttt{corpus})}{nb(\texttt{International} \text{ AND } \texttt{corpus}) + nb(\texttt{Relations} \text{ AND } \texttt{corpus})}$$
$$= \frac{2\times 5075}{1020428+281055} = 0.0078$$

In this example the relevant expansion choosen (i.e. having the best score) is the first definition (i.e. `Information Retrieval`).

We can add the context C in the basic measures based on MI and MI3 measures (formulas (8) and (9)) presented in section 3.2. $AcroDef_{MI}$ and $AcroDef_{MI3}$ using the context are given as follows:

$$AcroDef_{MI}(a^j) = \frac{nb(\bigcap_{i=1}^n a_i^j \text{ AND } C)}{\prod_{i=1}^n nb(a_i^j \text{ AND } C | a_i^j \notin M_{stop})}$$
$$\text{where } n \geq 2 \qquad (11)$$

$$AcroDef_{MI3}(a^j) = \frac{nb(\bigcap_{i=1}^n a_i^j \text{ AND } C)^3}{\prod_{i=1}^n nb(a_i^j \text{ AND } C | a_i^j \notin M_{stop})}$$
$$\text{where } n \geq 2 \qquad (12)$$

These different measures are language independent. They are tested in section 7, dedicated to experimentating $AcroDef$ on 'real' biomedical data.

4 $IADef$ measure: an expansion of Turney's measure

In the previous sections we presented the $AcroDef$ measures, that compute dependency between words forming the expansions. Such measures help choosing the relevant definitions. This approach is close to the work based on terminology extraction techniques (ranking of extracted terms) (4).

The $IADef$ (**I**ndependency between **A**cronyms and **Def**initions) measures presented in this section are closer to Turney's method described in section 2. $IADef$ computes the dependency between acronyms and definitions.

4.1 Basic Turney's measure for the acronym disambiguisation

P. Turney (25) has provided a formula (13) calculating the dependency between an acronym a and a candidate definition $\bigcap_{i=1}^n a_i^j$ (using *brackets* with Exalead). This formula is based on the standard measure of Mutual Information (MI).[3]

$$IADef_{MI}^{And}(\ a^j\ ) = \frac{nb(\ a \text{ AND } \bigcap_{i=1}^n a_i^j\ )}{nb(\ \bigcap_{i=1}^n a_i^j\ )} \qquad (13)$$

For instance, $nb(\ a \text{ AND } \bigcap_{i=1}^n a_i^j\ )$ with a = `IR` and $\bigcap_{i=1}^2 a_i^j$ = `Information` $\cap$ `Retrieval` calculates the number of pages returned by the query `IR AND "Information Retrieval"`. Thus, we compute the number of times where the terms IR and 'Information Retrieval' are present in the same page.

To be more precise in the calculation of the dependency between both words a (e.g. 'IR') and $\bigcap_{i=1}^n a_i^j$ (e.g. 'Information Retrieval'), we can compute the number of pages where the words are in a same window using the NEAR operator of Exalead. Actually, this operator requires that both words are within 16 words of each other.[4] The formula (14) calculates this dependency:

[3] In this formula, the constant $\frac{1}{nb(a)}$ is not taken into account because it does not change the order of expansions given by the statistical measure.

[4] Informations about the use of the NEAR operator of Exalead : http://www.searchengineshowdown.com/blog/exalead/ or http://moritzlegalinformation.blogspot.com/ 2006_06_01_archive.html

$$IADef_{MI}^{Near}(\, a^j \,) = \frac{nb(\, a \text{ NEAR } \bigcap_{i=1}^{n} a_i^j \,)}{nb(\, \bigcap_{i=1}^{n} a_i^j \,)} \tag{14}$$

4.2 Turney's measure based on different statistical measures

Turney's Measure can be extended using other statistical criteria that have been described in section 3: Cubic Mutual Information and Dice's coefficient.

Cubic Mutual Information gives a greater weight in the score of the formula's numerator that calculates the dependency between terms (acronym and definition). Formulas (15) and (16) describe such measures. They use the functions AND (formula (15)) and NEAR (formula (16)) of the search engine Exalead.

$$IADef_{MI3}^{And}(\, a^j \,) = \frac{nb(\, a \text{ AND } \bigcap_{i=1}^{n} a_i^j \,)^3}{nb(\, \bigcap_{i=1}^{n} a_i^j \,)} \tag{15}$$

$$IADef_{MI3}^{Near}(\, a^j \,) = \frac{nb(\, a \text{ NEAR } \bigcap_{i=1}^{n} a_i^j \,)^3}{nb(\, \bigcap_{i=1}^{n} a_i^j \,)} \tag{16}$$

In addition to conventional measures such as MI and MI3, we propose to use the Dice's coefficient applied to the $IADef$ measure (formulas (17) and (18)).

$$IADef_{Dice}^{And}(\, a^j \,) = \frac{2 \times nb(\, a \text{ AND } \bigcap_{i=1}^{n} a_i^j \,)}{nb(\, a \,) + nb(\, \bigcap_{i=1}^{n} a_i^j \,)} \tag{17}$$

$$IADef_{Dice}^{Near}(\, a^j \,) = \frac{2 \times nb(\, a \text{ NEAR } \bigcap_{i=1}^{n} a_i^j \,)}{nb(\, a \,) + nb(\, \bigcap_{i=1}^{n} a_i^j \,)} \tag{18}$$

4.3 Contextual $IADef$

Like $AcroDef$ measures, we can take into account a context C (see section 3.3) with these new measures described in section 4.

Then we add a context C (using the 'AND' operator) to the queries of the formulas (13), (14), (15), (16), (17), and (18). This context enables to enhance the original measure of P. Turney (25).

5 Applying those measures: a few examples

This section provides examples of the nine quality measurements, applied to the acronym 'IR'. Actually, with these measures, we calculate the obtained score with the possible expansion 'Information Retrieval':

- $AcroDef_{Dice}$ – formula (10): $\frac{2 \times nb(\mathtt{Information} \cap \mathtt{Retrieval})}{nb(\mathtt{Information}) + nb(\mathtt{Retrieval})}$
- $AcroDef_{MI}$ – formula (11): $\frac{nb(\mathtt{Information} \cap \mathtt{Retrieval})}{nb(\mathtt{Information}) \times nb(\mathtt{Retrieval})}$
- $AcroDef_{MI3}$ – formula (12): $\frac{nb(\mathtt{Information} \cap \mathtt{Retrieval})^3}{nb(\mathtt{Information}) \times nb(\mathtt{Retrieval})}$
- $IADef_{Dice}^{And}$ – formula (17): $\frac{2 \times nb(\mathtt{IR}\ \text{AND}\ (\mathtt{Information} \cap \mathtt{Retrieval}))}{nb(\mathtt{IR}) + nb(\mathtt{Information} \cap \mathtt{Retrieval})}$
- $IADef_{Dice}^{Near}$ – formula (18): $\frac{2 \times nb(\mathtt{IR}\ \text{NEAR}\ (\mathtt{Information} \cap \mathtt{Retrieval}))}{nb(\mathtt{IR}) + nb(\mathtt{Information} \cap \mathtt{Retrieval})}$
- $IADef_{MI}^{And}$ – formula (13): $\frac{nb(\mathtt{IR}\ \text{AND}\ (\mathtt{Information} \cap \mathtt{Retrieval}))}{nb(\mathtt{Information} \cap \mathtt{Retrieval})}$
- $IADef_{MI}^{Near}$ – formula (14): $\frac{nb(\mathtt{IR}\ \text{NEAR}\ (\mathtt{Information} \cap \mathtt{Retrieval}))}{nb(\mathtt{Information} \cap \mathtt{Retrieval})}$
- $IADef_{MI3}^{And}$ – formula (15): $\frac{nb(\mathtt{IR}\ \text{AND}\ (\mathtt{Information} \cap \mathtt{Retrieval}))^3}{nb(\mathtt{Information} \cap \mathtt{Retrieval})}$
- $IADef_{MI3}^{Near}$ – formula (16): $\frac{nb(\mathtt{IR}\ \text{NEAR}\ (\mathtt{Information} \cap \mathtt{Retrieval}))^3}{nb(\mathtt{Information} \cap \mathtt{Retrieval})}$

Of course, we add a context C with these basic measures. The section 7 gives the results of these nine quality measures.

6 A hybrid approach

The context used by the $AcroDef$ and $IADef$ measures is very small (often less than three words). Therefore, results are less attractive than with methods using a large context based on "bags of words" representations.

The work presented in this section proposes a hybrid method relying on a vector representation and $AcroDef/IADef$ measures, in order to improve results of the precision (see section 7.4). This hybrid measure is called $IAcos$.

6.1 A vector space model to disambiguate biomedical definitions

Expanding ambiguous biomedical abbreviations is an asset. Thus, several Word Sense Disambiguation (WSD) techniques use a Vector Space Model (16; 22) to represent various possibilities. The hybrid method represents the context of an abbreviation to disambiguate, by a vector which elements are the occurrences of its close words. With such a representation, several machine learning techniques can be applied (13), particularly in the biomedical domain: SVM (22; 9), Naive Bayes (22; 9), Decision trees (9), and so forth. These techniques can use a richer representation based on linguistic features like Part-of-Speech tags, bigrams (two consecutive words that occur together) (9), or semantic knowledge like MeSH (22), UMLS (12). In this paper, domain-knowledge is not addressed, since our approach is not specific to the sole biomedical field; It can be adapted to other domains and languages (20).

Here, an unsupervised approach is applied, a technique rather seldom developed in the biomedical disambiguisation literature. Among the few who have investigated such a process, one of the most representative is the work of (16), which consists in building contexts (bag of words) in order to predict the relevant meaning of an acronym. This context is provided by three types of corpora (i.e. Unrestricted Web, Medline abstract, Mayo Clinic). For each definition, the process developed by (16) allows to generate a context vector of lexical items and their frequency (using a window of ±20 words). The last step of the process is based on the computation of the vectors closeness. The largest cosine is selected in order to choose the adapted definition (meaning) of an acronym in a given context. Our hybrid approach, detailed in sections 6.2 and 7.4 uses this principle associated with the $IADef$ and $AcroDef$ measures.

6.2 Our $IAcos$ method

In the first step, $IAcos$ consists in building a context (1) for the candidate definitions based on a web corpus and (2) for the document where the acronym must be defined. Like (16)'s approach, our method consists in selecting the definition having the best cosine value.

In the second step, only the definitions which are in the first positions with the $AcroDef$ or $IADef$ measures are selected. Selection aims at improving the quality of the relevant expansions returned by the system. This technique takes into account both informations returned by the cosine and web-mining methods ($AcroDef$ and $IADef$). The formula (19) gives the $IAcos$ measure.

$$\begin{aligned} IAcos_i = \max_j \{ & cos(d, \text{context}(a^j)) \\ & / \ a^j \text{ is in the } i \text{ first definitions} \\ & \text{returned by } IADef \text{ or } AcroDef \} \end{aligned} \tag{19}$$

In this formula (19):

- d represents the vector of the document where the acronym have to be defined.
- context(a^j) is the context of candidate-definition a^j.

The method to build the context and the experimental protocol are detailed in section 7.4.

7 Experiments

7.1 Experimental protocol

In our experiments, we have focused on a classification of biological data definitions, provided by the Acromine application.[5] For any given acronym in this area, Acromine provides a list of its possible expansions. 102 pairs acronym/definitions have been randomly extracted from Acromine, which provided, for each tested item, from 4 to 6 possible definitions. The acronyms we study can be either two, three or four character strings. For instance, JA, PKD, and ABCD are possible acronyms, and for the latter, its definitions are described in the table 1. As one can see, it might range from medicine to biochemistry, dentistry, etc.

polycystic kidney disease
protein kinase D
proliferative kidney disease
paroxysmal kinesigenic dyskinesia
pyruvate kinase deficiency

Table 1: Extract of some definitions of the PKD acronym in biomedicine.

For each of these pairs, articles abstracts have been extracted from the specialized bibliographical data base Medline,[6] containing acronyms and their expansions. This base contains 204 documents (two documents per couple acronym/expansion, manually extracted). The goal of this experiment is to determine whether, for each document, the definition could be correctly predicted by classifying the candidate definitions with our quality measures. The distribution of the 204 documents according of the number of plausible candidate expansions for acronyms is given in the table 2. This table shows we need $12 \times 6 + 120 \times 5 + 72 \times 4 = 960$ expansions to test.

This experiment has needed the run of **7340 queries**:[7]

- Calculation of the 6 $IADef$ measures: $IADef$ measures require 2×960 queries for the numerator (with the AND and NEAR operators) and 2×960 for the denominator (for Dice measure): 3840 queries.
- Calculation of the 3 $AcroDef$ measures: $AcroDef$ requires 960 queries for the numerator and 2540 for the denominator (the number of queries for the denominator depends of the number of words of each expansion): 3500 queries.

Nb of documents	Nb of possible expansions per document
12	6
120	5
72	4

Table 2: Number of possible acronym definitions for the 204 documents.

[5] http://www.nactem.ac.uk/software/acromine/

[6] http://www.ncbi.nlm.nih.gov/PubMed/

[7] Experiments conduced in august 2009.

7.2 Results of $AcroDef$ and $IADef$ measures

Table 3 presents the results of these experiments. For each of the $AcroDef$ and $IADef$ measures:

- The first column value is the number of times where the correct definition has been given, as a first item,
- the second column value corresponds to the number of times it has been predicted among the two first definitions (ranks 1 and 2 according to the measure classification),
- and the third value corresponds to the number of times it appears among the first three.

Ranks	1	1 or 2	1, 2, or 3
$AcroDef_{Dice}$	73 (35.8%)	127 (62.3%)	161 (78.9%)
$AcroDef_{MI}$	62 (30.4%)	111 (54.4%)	149 (73.0%)
$AcroDef_{MI3}$	72 (35.3%)	118 (57.8%)	165 (80.9%)
$IADef_{Dice}^{And}$	111 (**54.4%**)	150 (**73.5%**)	174 (**85.3%**)
$IADef_{Dice}^{Near}$	104 (51.0%)	142 (69.6%)	174 (85.3%)
$IADef_{MI}^{And}$	94 (46.1%)	139 (68.1%)	169 (82.8%)
$IADef_{MI}^{Near}$	90 (44.1%)	137 (67.1%)	170 (83.3%)
$IADef_{MI3}^{And}$	104 (51.0%)	145 (71.1%)	174 (85.3%)
$IADef_{MI3}^{Near}$	102 (50.0%)	146 (71.6%)	170 (83.3%)

Table 3: Number of correct definitions based onthe expansions ranks provided by the statistical measure (Medline Abstracts)

Measure	Sum
$AcroDef_{Dice}$	470
$AcroDef_{MI}$	516
$AcroDef_{MI3}$	481
$IADef_{Dice}^{And}$	**389**
$IADef_{Dice}^{Near}$	403
$IADef_{MI}^{And}$	422
$IADef_{MI}^{Near}$	424
$IADef_{MI3}^{And}$	401
$IADef_{MI3}^{Near}$	405

Table 4: Sums of the Ranks of Relevant Definitions.

Experiments have been led with a one-word context only, i.e., the most frequent word in each document. Working on a specialized domain, queries with more than one word have null pages results with a general search engine such as Exalead.

Table 3 shows some important facts, that might provide answers to the following questions and meet some of the assigned goals:

- **Which are the best quality measures?**

Table 3 shows that $IADef$ measures give better results than $AcroDef$ measures. It seems that the calculation of the acronyms and expansions dependency is more relevant than the dependency between the expansions words. Another important conclusion is that the *$IADef$ measure based on Dice's coefficient gives the best result*. This one is best than the result obtained with the original Turney's measure (quality measures based on MI). Note that MI3 provides good results too (close to Dice's coefficient).

Table 3 shows that the performance of AND and NEAR operators is very close. This result differs from the study presented in (25). It can be explained by the specificity of acronyms usage. Indeed, acronyms and their expansions are often very close, in the same sentence, in the documents returned by the search engine. Thus, there are only little differences in the documents returned by AND and NEAR operators.

In order to determine more precisely the quality of these measures, we have computed the sum of the ranks of relevant definitions. The best measure is the one that has the smallest sum. This method, while evaluating rank functions, is equivalent to approaches based on ROC (Receiver Operating Characteristics) curves and to the calculus of surfaces under them (6; 19). Therefore, Table 4 confirms that $IADef_{Dice}$ behaves as the best measure in specialized documents belonging to biomedicine. Also, every $IADef$ measures has a better rank (smaller sum) than the best $AcroDef$ measure: The 'worst' $IADef$ result, 424, is above $AcroDef_{Dice}$, the best one among $AcroDef$ results, with 470. Note that Dice's coefficient enhances both measures results.

- **Significance of results:**
  $AcroDef_{Dice}^{And}$ hits the good definition on rank 1 in 54.4% of the cases. This is significantly better than a random prediction, which scores 22%. We calculated this random prediction as such: 1 chance over 4 to put the relevant definition as the first one in 72 cases, 1 over 5 in 120 cases, and 1 over 6 in 12 cases, which are the number of documents with respectively 4, 5, and 6 possible definitions (in Table 2).

- **Restricting the definition space:**
  The high predictive values for the first three definitions ranked by $IADef$ measures restrict the search space. It is useless to go down further in the list, and in the 204 documents where more than 4 definitions occur, it would be efficient to restrict to the first three chosen by our measures, and give the user the opportunity of choosing the best one. Further, they might be close definitions as we will show it in a deeper study of the data content.

7.3 Data properties

The retrieved definitions has led us to formulate some comments. Among the difficulties encountered in NLP research in the biomedical domain, the fact that several terms could address the same or very similar concepts is a very classical issue. For instance, when we retrieved the acronym ZO we had the following definitions: `zonula occludens`, `zona occludens`, `zonulae occludentes`. As one can see, these are either flexions of the same term (plural vs singular) or very close terms (*zonula* meaning 'small zone' vs *zona*). Variations are explained by linguistic functions or properties. Therefore, quite a fair amount of prediction errors could be caused by linguistic variations on the same basic lexical item.

On the other hand, some equivalent definitions cannot be fathomed without the help of a domain expert. If *terminal* and *termini* could be seen as Latin flexions in the following example: `carboxy terminal`, `carboxy termini`, or in the pair `COOH-terminal`, `COOH-termini`, or in `CO2H-terminal`, `CO2H termini`, the idea that *COOH*, *CO2H* and *carboxy* are equivalent forms (which makes all these pairs totally equivalent to each other) is not automatically deductible and needs expertise.

7.4 Evaluation of $IAcos$

The cosine measure has been applied, as a similarity metric in the document vector space (all the words except stopwords). The vector components are figures, representing word frequencies in the documents where the acronym has to be defined. The context of the candidate-definition is based on its close words (window of 20 to 30 words). This context is extracted from the first 10 pages returned by the Exalead search engine (this kind of context gives approximatively the same amount of words as provided by the documents). We use Exalead because specialized search engines were used to build test corpus. Then the cosine between document vector and the context of candidate definitions vector, is calculated, to predict the relevant expansion. The results presented in Table 5 show good results given by this method, i.e., 146 relevant definitions are predicted on the 204 documents (71.5% relevant definitions are ranked at the first position). The average value of the correctly predicted definitions cosine is 0.51.

Then, when shifting to the hybrid approach to improve accuracy, we calculate the $IAcos$ measure, based on the cosine and $IADef$ measures (see section 6). We select $IADef_{Dice}^{And}$ because it offers the best results in the experiments presented in section 7.2. The $IAcos$ measure consists in selecting the definitions that have the best cosine, and that are ranked at the first positions by applying $IADef$ measures.

The results are presented in Table 5 according the precision (formula (20)) and recall (formula (21)).

$$P = \frac{\text{Number of returned relevant definitions}}{\text{Number of returned definitions}} \quad (20)$$

$$R = \frac{\text{Number of returned relevant definitions}}{\text{Number of relevant definitions}} \quad (21)$$

Measures	Rate P	P (%)	Rate R	R (%)
$cosine$	146/204	71.5	146/204	**71.5**
$IADef_{Dice}^{And}$	111/204	54.4	111/204	54.4
$IAcos_1$	85/99	**85.8**	85/204	41.7
$IAcos_2$	113/142	79.5	113/204	55.4
$IAcos_3$	125/142	75.3	125/204	61.3

Table 5: Precision and Recall of the $cosine$, $IADef$, and $IAcos$ approaches ($IAcos_i$ where i represents the number of i first definitions taken into account by $IADef$).

Table 5 shows that we obtain either a best precision ($IAcos$) or a best recall ($cosine$), but not both with the same measure. This means $IAcos$ selects fewer definitions but these are more relevant. Depending on the task, the expert might want to retrieve an expansion requiring either high precision or high recall, we can use the appropriate method, i.e, $cosine$ or $IAcos$. Note that all $IAcos$ variants are on the Pareto front (11), so they are relevant. Only the $IADef$ measure used alone is dominated (see Figure 1), this is the reason why we have proposed to combine it with $cosine$ technique based on a largest context.

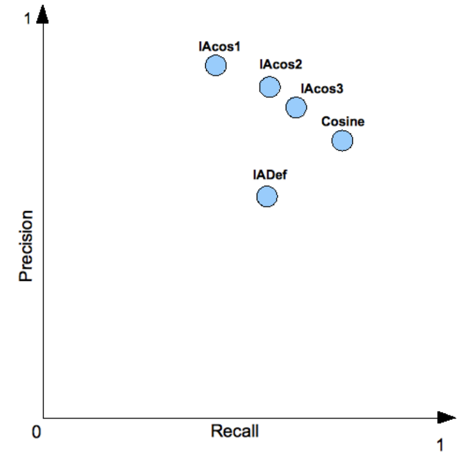


Figure 1: Pareto front of the $cosine$, $IADef$, and $IAcos$ approaches.

8 Conclusion and future work

Acronyms are widely used words that act as proper names for organizations or associations, or as shortcuts in denominating very frequent concepts or notions. As such, they are representative of the named entities issue currently tackled by the text mining scientific community. Acronyms recognition is one part of the issue, but ambiguous acronyms expansion, especially when the acronym definition is not present in the considered document, is another. This paper offers a set of quality measures to determine the choice of the best expansion for an acronym not defined in the

Web page that uses it, the $AcroDef$ and $IADef$ measures. The method uses statistics computed on Web pages to determine the appropriate definition. Measures are deeply context-based and rely on the assumption that the most frequent words in the page are related semantically or lexically to the acronym expansion. An evaluation on specialized corpora extracted from biomedical databases showed that measures still significantly operated, although contexts were much similar, and expansions very close to each other, reducing the measures ability to discriminate. However, within a context of one word (the only one with which search engines were able to retrieve pages for specific domains), the relevant definitions appeared in the first three elected by the $IADef$ measure based on Dice's coefficient with a probability of 85%. The hybrid approach presented in this paper, i.e. $IAcos$, combines a vector representation of the context (a very rich context) and $IADef$ mesure. This method improves the basic measure results precision.

$IADef$ errors are explained by the fact that they originate from too general words within contexts. If the most frequent words in the page are highly polysemous, too widely used, or vague, this has an impact on the best expansion choice, since the semantic constraint is looser. If the corpus in which acronyms have to be expanded belongs to a given domain, an interesting perspective would be to use as heuristics domain-based features (proper names, terms), or even better, a domain ontology. The experiments conducted on the biomedical corpus has clearly aimed at this direction.

Every method has its limitations and needs to be enhanced. Our approach has difficulties in building a context when the Web page in which the acronym has been found only contains a short text (a few lines for instance). Context extraction relies on words frequency as a cornerstone for thematic detection. If words are few, frequency becomes meaningless. An interesting perspective would be to represent documents as semantic vectors defined to get a thematic information on the text. These vectors project the document on a Roget-based ontology and thus do not need quantities of words to sketch a thematic environment for the acronym. That complementary information, associated with $AcroDef$ and $IADef$, would help predicting acronym definitions in the case of short texts. This work is currently undergoing as a sequel to the acronym expansion issue that we have been dealing with for a couple of years.

References

[1] E. Brill. Some advances in transformation-based part of speech tagging. In *AAAI, Vol. 1*, pages 722–727, 1994.

[2] J. Chang, H. Schtze, and R. Altman. Creating an online dictionary of abbreviations from medline. *Journal of the American Medical Informatics Association*, 9:612–620, 2002.

[3] K.W. Church and P. Hanks. Word association norms, mutual information, and lexicography. In *Computational Linguistics*, volume 16, pages 22–29, 1990.

[4] B. Daille. Study and Implementation of Combined Techniques for Automatic Extraction of Terminology. In *The Balancing Act: Combining Symbolic and Statistical Approaches to Language, MIT Press*, pages 49–66, 1996.

[5] D. Downey, M. Broadhead, and O. Etzioni. Locating complex named entities in web text. In *Proceedings of IJCAI'07*, pages 2733–2739, 2007.

[6] C. Ferri, P. Flach, and J. Hernandez-Orallo. Learning decision trees using the area under the ROC curve. In *Proceedings of 9th International Conference on Machine Learning, ICML'02*, pages 139–146, 2002.

[7] K. Frantzi, S. Ananiadou, and H. Mima. Automatic recognition of multi-word terms: the C-value/NC-value method. *International Journal on Digital Libraries*, 3(2):115–130, 2000.

[8] F. Guillet and H.J. Hamilton. *Quality Measures in Data Mining*. Springer Verlag, 2007.

[9] M. Joshi, S. Pakhomov, T. Pedersen, and C. G. Chute. A comparative study of supervised learning as applied to acronym expansion in clinical reports. In *Proceedings of the Annual Symposium of the American Medical Informatics Association*, pages 399–403, 2006.

[10] L.S. Larkey, P. Ogilvie, M.A. Price, and B. Tamilio. Acrophile: An automated acronym extractor and server. In *Proceedings of the Fifth ACM International Conference on Digital Libraries*, pages 205–214, 2000.

[11] H. A. Leiva, S. C. Esquivel, and R. H. Gallard. Multiplicity and local search in evolutionary algorithms to build the pareto front. In *SCCC*, pages 7–13, 2000.

[12] H. Liu, A.R. Aronson, and C. Friedman. A study of abbreviations in medline abstracts. In *Proceedings of the Annual Symposium of the American Medical Informatics Association*, pages 464–468, 2002.

[13] R. Navigli. Word sense disambiguation: A survey. *ACM Comput. Surv.*, 41(2), 2009.

[14] G. Nenadic, I. Spasic, and S. Ananiadou. Terminology-Driven Mining of Biomedical Literature. *Bioinformatics*, 19(8):938–943, 2003.

[15] N. Okazaki and S. Ananiadou. Building an abbreviation dictionary using a term recognition approach. *22*, Bioinformatics(24):3089–3095, 2006.

[16] S. Pakhomov, T. Pedersen, and C. G. Chute. Abbreviation and acronym disambiguation in clinical discourse. In *Proceedings of the Annual Symposium of*

the American Medical Informatics Association, pages 589–593, 2005.

[17] S. Petrovic, J. Snajder, B. Dalbelo-Basic, and M. Kolar. Comparison of collocation extraction measures for document indexing. In *Proc of Information Technology Interfaces (ITI)*, pages 451– 456, 2006.

[18] V. Prince and M. Roche, editors. *Information Retrieval in Biomedicine: Natural Language Processsing for Knowledge Integration*. Medical Information Science Reference, IGI Gobal, 460 pages, 2009.

[19] M. Roche and Y. Kodratoff. Pruning Terminology Extracted from a Specialized Corpus for CV Ontology Acquisition. In *Proceedings of onToContent Workshop - OTM'06, Springer Verlag, LNCS*, pages 1107–1116, 2006.

[20] M. Roche and V. Prince. Managing the Acronym/Expansion Identification Process for Text-Mining Applications. *International Journal of Software and Informatics*, 2(2):163–179, 2008.

[21] F. Smadja, K. R. McKeown, and V. Hatzivassiloglou. Translating collocations for bilingual lexicons: A statistical approach. *Computational Linguistics*, 22(1):1–38, 1996.

[22] M. Stevenson, Y. Guo, A. Alamri, and R. Gaizauskas. Disambiguation of biomedical abbreviations. In *Proceedings of the BioNLP 2009 Workshop*, pages 71–79, Boulder, Colorado, June 2009. Association for Computational Linguistics.

[23] A. Thanopoulos, N. Fakotakis, and G. Kokkianakis. Comparative Evaluation of Collocation Extraction Metrics. In *Proceedings of LREC'02*, pages 620–625, 2002.

[24] M. Torii, Z.Z. Hu, M. Song, C.H. Wu, and H. Liu. A comparison study on algorithms of detecting long forms for short forms in biomedical text. *BMC Bioinformatics*, 2007.

[25] P.D. Turney. Mining the Web for synonyms: PMI–IR versus LSA on TOEFL. *Proceedings of the 12th European Conference on Machine Learning (ECML), LNCS*, 2167:491–502, 2001.

[26] J. Vivaldi, L. Màrquez, and H. Rodríguez. Improving term extraction by system combination using boosting. In *Proceedings of the 12th European Conference on Machine Learning (ECML)*, pages 515–526, 2001.

[27] J. Xu and Y. Huang. Using svm to extract acronyms from text. *Soft Comput.*, 11(4):369–373, 2007.

[28] S. Yeates. Automatic extraction of acronyms from text. In *New Zealand Computer Science Research Students' Conference*, pages 117–124, 1999.

A Risk Management System to Oppose Cyber Bullying in High School: Warning System with Leaflets and Emergency Staffs

Hirohiko Yasuda
Shimonoseki Technical High School
Tomitou-cyo, 4-cyome, 1-1, Shimonoseki City, postcode 759-6613, Japan
E-mail: yasuda.hirohiko@ysn21.jp, konippon@pluto.dti.ne.jp
URL:http/www.sekko-t.ysn21.jp

Keywords: prevention of cyber bullying, school violence, cognitonics, school education, mobile phone, information society, internet, parents education, cyber crime

Received: February 11, 2010

The aim of this study is preventing cyber-bullying with a risk management system. This paper examines the importance of the guardians role in preventing cyber bullying and considers how the schools can support guardians and students. School organizes Information System to support students and guardians. School supports them, without the means of class, to gain extensive knowledge on an information society from the view point of Web Science.

Povzetek: Predstavljen je sistem za preprečevanje spletnega nasilništva.

1 Introduction

Information communication technology is changing society rapidly. It has brought not only positive but also negative aspects to information society. One of the most serious threats is cyber-bullying. There is no solid solution or countermeasure to stop and prevent cyber-bullying. There are two reasons why school can not stop cyber-bullying. The first reason is that technology is changing rapidly and knowledge does not last long. The second reason is that High School students do not like any kind of moral education. We have been practicing Information Education in interdisciplinary approach since 2002 [2]. We have used two methods. One is using leaflets which mainly deal with case studies of computer science and social science. The other is organizing Information System of school to support students and guardians.

In 2006 Tim Berners-Lee and Wendy Hall, Nigel Shadbolt, Daniel J. Weitzner advocated Web Science [1] and founded Web Science Research Initiative (WSRI) [3]. Web Science focuses on understanding, designing, and applications that make up the World Wide Web [3].

In 2006, Vladimir Fomichov and Olga Fomichova advocated an interdisciplinary approach – Cognitonics [4][5][6]. This approach aims at describing distortions in the development of the personality and national cultures which are caused by the stormy development of information technology. It is seeking systematic solutions to compensate the negative implications.

The method of this paper, Web Science method, and Cognitonics approach have similar aims. This paper shows that Web Science method is useful not only for High School education but also for guardians' education. This paper shows the guardians' difficulties when they give instructions to their children.

2 Background

On January 31, 2009, Ministry of Education, Culture, Sports, Science and Technology (MEXT) notified all Prefectural Boards of Education in Japan that mobile phones are now banned in principle in Elementary and Junior High Schools, and no use of mobile phones allowed in High Schools. [7] High school students most often use a Profile home page service, which is a free home page service for an individual person on the mobile phone internet (Figure 1).

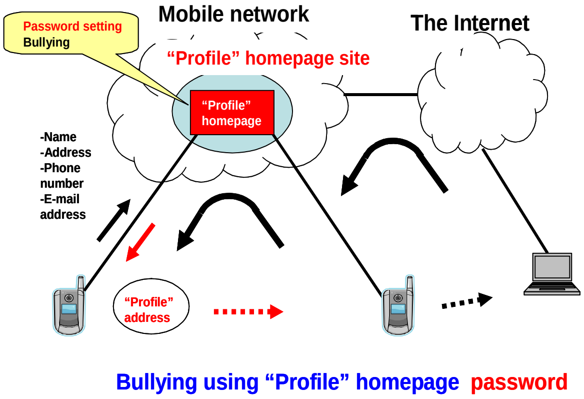


Figure 1: Profile homepage

Profile home pages are a breeding ground for cyber-bullying now. They often use a social networking service (SNS) too. "MOBAGE Town" is a free game site on the mobile phone internet. 50 percent of Junior High and High School students use it (April 2008). It has the same function as a dating site. Most of Junior High and High School bullying are related to profile home pages and SNS. They provide the students with opportunities for contact with harmful information and adults with inappropriate intentions.

3 Problems of mobile phones

3.1 The crux of the problems with mobile phones

The most remarkable characteristics of mobile phones are their mobility and high performance. Children can use mobile phones privately and are able to accomplish many extraordinary things without adults knowing.

The biggest reason that cyber-bullying and illegal acts are continuing is the misunderstanding that the net is anonymous. Students do not understand what strong weapons mobile phones and the Internet are. They can not realize how many people read their messages and pictures.

Once the message has been sent on the mobile phone internet and the Internet, it is hard to delete it because of the constitution, “secret of communication”, “freedom of expression”, “freedom of speech”, and the provider law. Even though the sender is detected and the original information or data defaming others is deleted, a defaming message, or illegal personal information or pictures, which seriously invade people’s human rights, have been already copied repeatedly through the bulletin board system and peer-to-peer file exchanging software. Actually it is impossible to delete them all.

3.2 The limits of school guidance

It is essential for guardians to instruct their children in use of mobile phones and the Internet to prevent cyber-bullying.

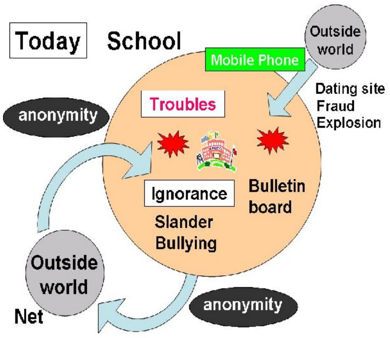


Figure 2: School and mobile phones.

Students: How many hours do you use your mobile phones?
Questionnaire: S High School, June 2008
# responses 447(m: 439, f: 8), # students 461

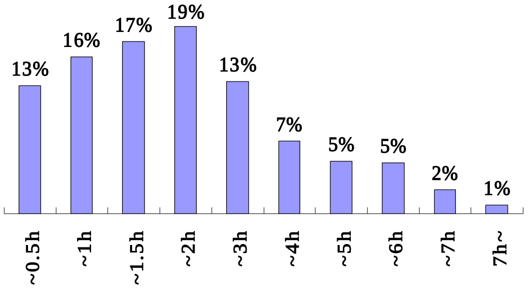


Figure 3: Excessive use of mobile phones.

Schools do not even have the authority to ask providers and administrators of the sites to delete the illegal defamations on the net, because of the laws and regulations (Figure 2).

The Biggest reason guardians can not instruct

Guardians have the lack of knowledge on technology and Information Society. It is the biggest reason that they can not instruct on the safe use of mobile phones. Guardians do not understand the increased importance of mobile phones in a student’s life. They can hardly keep up with teenagers’ use of mobile phone. 20 percent of students use mobile phones for more than 3 hours a day. (Figure 3)

Guardians do not understand how students use the mobile phones and the Internet. Only 47 percent of guardians know the actual use by students [8] (S High School, 2007).

A fixed sum system of paying for mobile phones makes it difficult for guardians to know the exact use by students. They are confused about the student’s way of thinking about mobile phones and the anxiety over the threat of troubles caused by it.

Guardians should have responsibility of minors’ use of mobile phones and the Internet. School have to make their position clear, inside and out-side of school, that guardians should make themselves responsible for the use of mobile phones and the Internet by students.

3.3 Guardians’ difficulties

Guardians feel strongly that they should give some instruction to their children themselves. But they think it very hard to do, actually. According to our survey such instructions are the following [9].

How to cope with the troubles when their children are victims or perpetrators of cyber crime (24%), the risks of dating sites on mobile phone internet (22%), the risk of drug site (22%), defaming others by mobile phones (20%), the download of illegal images (child pornography) (20%).

4 Purpose of the study

School help students to learn information morality through case studies and understand how our information society will be changing.

Students learn how information technology influences people, their daily life, the market and the law. They learn to recognize the information society as a system composed of many social factors. Then they will be able to adapt themselves to deal with unknown situations more effectively in the future.

School help students to understand the intention of the information sender on the Internet by asking them why these cases or events have happened. We can prevent students from getting involved in troubles as a victim or a perpetrator, resulting from a lack of knowledge of the Internet and information technology.

School help students and parents to obtain the ability to adapt themselves to the information society, which is rapidly changing. School help students to obtain the ability and aptitude to make judgments using their own initiative without being confused by transient incidents or social trends.

4.1 Merits of a Web Science approach in school: Moving from a ban to understanding information society

Students neglect school instructions to keep safe on the Net. They do not like any kind of moral education and never listen to serious instructions of principles. A merit of a Web Science approach is that it does not depend on compulsion of morality or prohibition. Learning the events or cases, students understand why they happened, or think how technology related to them. Then they come to recognize the changes of society as a process of social development, they gain an insight into information society, and adapt themselves to the changes of a society.

An insight into a technological society is useful to decide their courses in the future. It helps students to choose an occupation that they will not loose in 10 or 20 years hence.

5 Target and development of a Web Science Method

The aims of this method are as follows.

(1) To put an emphasis on timely instruction. (2) To reduce the frequency of moral education. (3) To keep a way to always provide information to guardians. (4) To utilize and maximize the instructions of guardians. (5) Guidance to be short. (6) To require no prior knowledge on Web Science. (7) To be used repeatedly. (8) To be a Web Science approach.

6 Method

6.1 Information system

School provides Information System for complete prevention and understanding of information society (Figure 4). Information and warning of the Net are given to students and guardians with leaflets at every opportunity, morning homeroom class, classes, long homeroom class, school meetings, PTA committee, PTA general meetings.

Emergency staff prevents repeat suffering and secondary victims. Emergency staff solves cyber crime at the early stage, keeping victim's privacy.

6.2 Guidance with leflets in morning home room class

We have published and produced a regular series of leaflets," The SEKIKO Good Net News for Family "[10].

Every homeroom teacher hands out leaflets to students during their morning homeroom sessions and comments on a story just for one minute. Sometimes, the teacher will caution the students not to bully with mobile phones or the Internet. After students read each leaflet, they hand it over to their parents. We help parents to understand new technologies and what students are doing at school or in their daily lives. Every year, we edit all of the distributed leaflets for that year and write an Annual Textbook of Information-related Education.

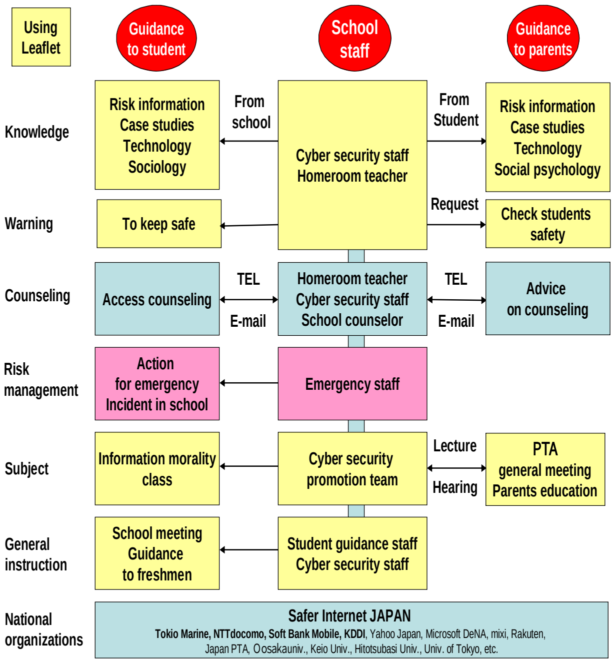


Figure 4: Information System.

6.3 Guidance given to new freshmen and guardians prior to starting school

The Annual Textbook will be given to all new freshmen and their parents of next year when they attend a briefing. They are promoted to understand the danger of the Net with the Textbook and to talk about that before entering High School. Freshmen have less knowledge on mobile phones and the Internet. They are a high-risk group, who can easily become involved in trouble. This intensive instruction targets the high-risk group.

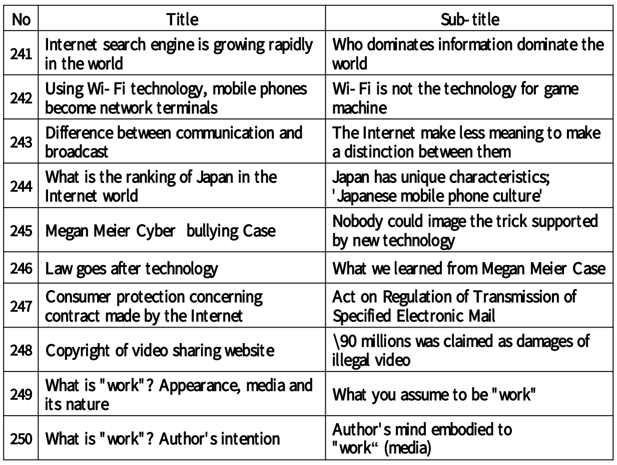
6.4 An immediate and intensive instruction

An immediate and intensive program of guidance is given to freshmen in their first three months. Freshmen are taught a program about characteristics and dangers of the Net again and again. The program prevents freshmen becoming either a victim or perpetrator.

6.5 Contents of leaflet

At least one quiz is made to present a problem and let students think why the event happened and what the point of the problem is. We have issued 240 leaflets since 2002 (Table 1).

Contents cover the following. (1) Events in school (2) Risk information (3) About cyber-bullying (4) Request of guardians' instruction (5) Explanation on information technology (6) Impact given to industry by information technology (7) Information-oriented society (8) The future built by technology



No	Title	Sub-title
241	Internet search engine is growing rapidly in the world	Who dominates information dominate the world
242	Using Wi-Fi technology, mobile phones become network terminals	Wi-Fi is not the technology for game machine
243	Difference between communication and broadcast	The Internet make less meaning to make a distinction between them
244	What is the ranking of Japan in the Internet world	Japan has unique characteristics; 'Japanese mobile phone culture'
245	Megan Meier Cyber bullying Case	Nobody could image the trick supported by new technology
246	Law goes after technology	What we learned from Megan Meier Case
247	Consumer protection concerning contract made by the Internet	Act on Regulation of Transmission of Specified Electronic Mail
248	Copyright of video sharing website	\90 millions was claimed as damages of illegal video
249	What is "work"? Appearance, media and its nature	What you assume to be "work"
250	What is "work"? Author's intention	Author's mind embodied to "work" (media)

Table 1: Example titles of leaflets.

7 Results and conclusion

The methods using leaflets are very effective. School has practiced these methods for three years, and could get rid of all trouble caused by mobile phones and the Internet in school. Leaflets are so flexible that the school can use it for any purpose, in combination with various methods. This practice proved that these combinations have been very effective. We can make a leaflet quickly, take it anywhere and teach immediately. The merits of flexibility, which class lessons do not have, have extended the target from students to guardians and increased the chances of timely intervention.

All results show that there is no difference between the basic concept of prevention against cyber-bullying and that of traditional bullying. The different point is that significant knowledge of the power of technology is necessary to prevent cyber-bullying.

7.1 Effect of timely publishing

As soon as a serious event or a dangerous case takes place, school stopped students having to face the same risks as these cases (Picture 1). Repeated warning to the freshmen year and reduced exposure of personal information on the Net, which is often sent with mobile phone cameras from school.

With a leaflet school gives information to guardians at anytime (Picture 2). As soon as a profile homepage caused a suicide, school informed both guardians and students details of an event, and gave instruction and guidance to check the students' use of mobile phones is safe. (Figure 5)

Picture 1: A lecture to students.

Picture 2: A lecture to parents.

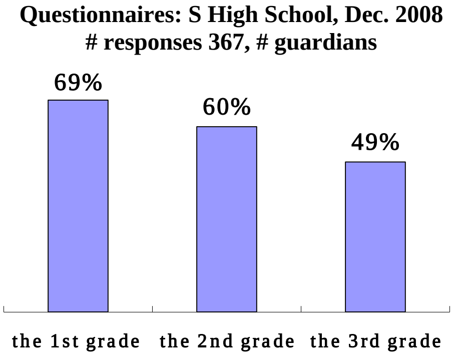


Figure 5: Rate of guardian who guided children according to warning of school.

7.2 Campaign for communication with guardians

73 percent of guardians think that school guidance with leaflets is useful (see Figure 6, Figure 7).

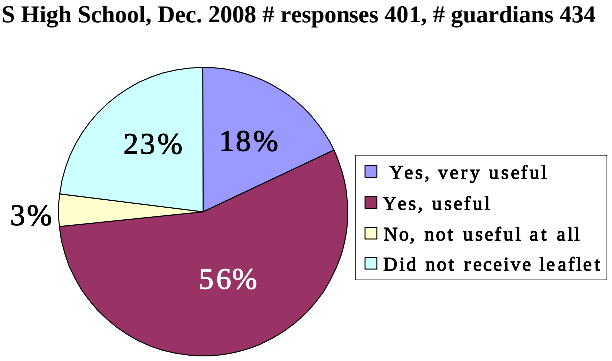


Figure 6: Guardians: Are leaflets useful?

As a result, leaflets give an opportunity to talk about the use of mobile phones and the Internet at home. In response to the schools request for safety checks, more guardians discussed safety and use of mobile phones with students.

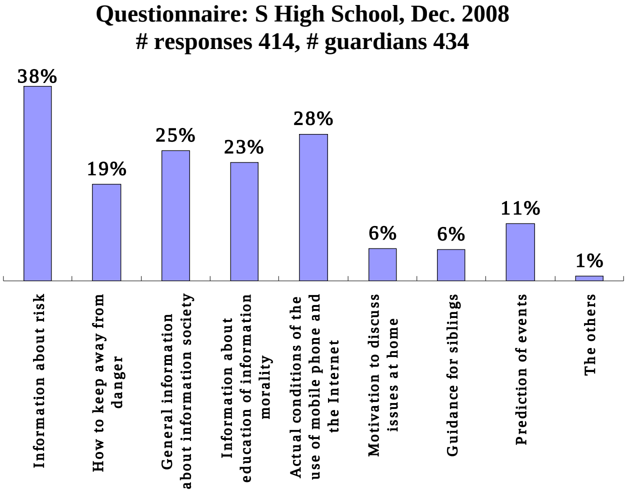


Figure 7: Guardians: What content of leaflets is useful?

7.3 Monitoring the effects of our guidance

According to the survey, 41 percent of students don't keep to the school regulations about using their mobile phone in school [11]. Also, only 45 percent of parents agreed to students using mobile phones at school [12]. Therefore, mobile phones remain banned at school.

8 Issues and solutions

Actually, there is the possibility that cyber-bullying goes underground, continues secretly and is becoming more serious. School keeps sufficient communication with guardians. School has established a system that guardians inform the school if they suspect any symptoms of cyber-bullying at home. 25 percent of students do not read the leaflet [13]. We take a countermeasure that the homeroom teacher urges students to read leaflet and give short comment.

A leaflet can not substitute a class. It is important that essential instruction should be done in a class and a whole school meeting too, especially a whole school meeting has been very effective. A whole school meeting can make students understand the matter is a serious problem.

25 percent of guardians do not receive the leaflets at all. Students do not pass them on to parents [14]. Students don't want to pass the warning information to ban the use of mobile phones and the Internet. We take a countermeasure that school send a leaflet to guardians through their e-mail.

References

[1] J. Hendler, N. Shadbold, W. Hall, T. Berners-Lee, D. Weitzner, *Web science: An interdisciplinary approach to understanding the World Wide Web.* New York: Communications of the ACM, Volume 51, Issue 7 (July 2008)

[2] Hirohiko Yasuda, *SEKIKO good-net NEWS.* Shimonoseki Technical High School, Shimonoseki, Japan, 2002 (http://sweb.nctd.go.jp/g_support/-others/k_ys_s01.pdf)

[3] Web Science Research Initiative (http://webscience.org/)

[4] V.A. Fomichov, O.S. Fomichova. *Cognitonics as a New Science and Its Significance for Informatics and Information Society*. Special Issue on Developing Creativity and Broad Mental Outlook in the Information Society, Informatica. An Intern. Journal of Computing and Informatics, Slovenia, 2006, Vol. 30, No. 4, pp. 387-398.

[5] O.S. Fomichova, V.A. Fomichov. *Cognitonics as a New Science and Its Social Significance In the Age of Computers and Globalization*. IIAS-Transactions on Systems Research and Cybernetics. Vol. VII, No. 2. Intern. Journal of the International Institute for Advanced Studies in Systems Research and Cybernetics. Published by The International Institute for Advanced Studies in Systems Research and Cybernetics (IIAS), Tecumseh, Ontario, Canada, 2007, pp. 13-21.

[6] Olga Fomichova and Vladimir Fomichov. *Cognitonics as an Answer to the Challenge of Time*. Proceedings of the First International Workshop on Cognitonics in conjunction with the International Multiconference Information Society 2009, Slovenia, Ljubljana, 12 – 16 October 2009, available online at http://is.ijs.si/-zborniki.asp?lang=eng

[7] MEXT, *Survey on school bans on mobile phones*. Tokyo, 2008 (http://www.mext.go.jp/b_menu/-houdou/21/01/1234723.htm) (http://www.mext.go.jp/b_menu/houdou/21/01/__icsFiles/afieldfile/2009/02/02/1234723_1_1.pdf)

[8] Shimonoseki Technical High School, *Survey on guardians' opinions about use of mobile phones*, Shimonoseki, Japan, July 2007, Questionnaire: # of responding 360, # of guardians 479

[9] Shimonoseki Technical High School, *Annual survey on guardians about Information Ethics*, Shimonoseki, Japan, January 2010, Questionnaire: # of responding 437, # of guardians 476

[10] Hirohiko Yasuda, *SEKIKO good-net NEWS for family.* Shimonoseki Technical High School, Shimonoseki, Japan, 2007 (http://www.jasrac.or.jp/seminar/5th.html) (http://www.jasrac.or.jp/seminar/pdf/5th_pdf010.pdf)

[11] Shimonoseki Technical High School, *Survey on students' use of mobile phones*, Shimonoseki, Japan, June 2007, Questionnaire: # of responding 300, # of students (1^{st},2^{nd} grade) 309

[12] Shimonoseki Technical High School, *Survey on guardians 'opinions about use of mobile phones*, Shimonoseki, Japan, July 2007, Questionnaire: # of responding 378, # of guardians 479

[13] Shimonoseki Technical High School, *Survey on Information Morality of students*, Shimonoseki, Japan, Dec. 2005, Questionnaire: # of responding 524, # of students 560

[14] Shimonoseki Technical High School, *Survey on guardians 'opinions about use of mobile phones*, Shimonoseki, Japan, July 2007, Questionnaire: # of responding 374, # of guardians 479

Text Mining for Discovering Implicit Relationships in Biomedical Literature

Ingrid Petrič
University of Nova Gorica, Vipavska 13, 5000 Nova Gorica, Slovenia
ingrid.petric@ung.si, http://slais.ijs.si/theses/2009-10-30-Petric.pdf

Thesis Summary

Keywords: text mining, knowledge discovery, hypotheses generation

Received: April 15, 2009

This article presents an innovative methodology for knowledge discovery in text databases that can improve the existing methods of exploring implicit relationships across different domains of expertise by providing a more intuitive computer aided search of unexplored links in literature. The literature mining method, called RaJoLink is based on rare pieces of information in a given domain. When these relations are interesting from a medical point of view and can be verified by medical experts, they represent new pieces of knowledge and can contribute to better understanding of diseases.

Povzetek: Članek opisuje inovativno metodologijo odkrivanja znanja iz tekstovnih baz podatkov.

1 Introduction

Automated knowledge discovery based on text data sets in the field of biomedicine is an intriguing challenge as it requires intensive collaboration with domain experts during the processes of both domain-specific text analysis and evaluation. Hence an interactive approach is recommended when text mining and decision support are combined. Also, it is beneficial to apply improved methods of literature mining, searching indirect connections and bisociative knowledge discovery from extensive text databases such as MEDLINE. Namely, information that is related across different contexts is difficult to identify with conventional associative approaches. The context-crossing associations, however, are the ones often needed for innovative discoveries. Such associations are called bisociations. The major aim here is to unravel the still hidden relations between the researched phenomena and their potential causes.

We developed a literature mining method called RaJoLink [1-3, 6] that uncovers hidden relations from large sets of scientific articles in a given domain. The method searches for logically connected pieces of literature on rare terms identified in literature on a given phenomenon under investigation (e.g., disease). This way it supports human expert in the process of generating and testing hypotheses in the domain under study. RaJoLink supports biomedical experts in both open and closed discovery process. In the open knowledge discovery process, hypotheses have to be generated, while in the closed knowledge discovery process, given hypotheses are tested. By identifying relations between biomedical concepts in disjoint sets of articles, the method implements the Swanson's [4] ABC model approach. However, the RaJoLink method analyses such relations in a new way and expands the Swanson's ABC model by suggesting hypotheses in advance, as a result of the open knowledge discovery process. The main novelty is a semi-automated suggestion of candidates for agents *A* that might be logically connected with a given phenomenon *C* under investigation. The choice of candidates for *A* is based on rare terms identified in the literature on the topic *C*. As rare terms are not part of the typical range of information, which describe the phenomenon under investigation, such information might be considered as unusual observations about the phenomenon *C*. If literatures on these rare terms have an interesting term in common, this joint term is declared as a candidate for *A*. Linking terms *B* between literature on *A* and literature on *C* are then searched for in the closed discovery process to provide supportive evidence for uncovered connections.

2 RaJoLink

The method is named RaJoLink after its key procedural elements, which are: rare terms, joint terms and linking terms. Consequently, the entire RaJoLink's approach consists of three principal steps, Ra, Jo and Link. In step Ra, a specified number (set by user as a parameter value) of interesting rare terms in literature about the phenomenon *C* under investigation are identified. In step Jo, all available articles about the selected rare terms are inspected and interesting joint terms that appear in the intersection of the literatures about rare terms are identified. One of them is selected as the candidate for *A*. In step Link, linking terms *B*, which bridge literature about *A* and literature about *C*, are searched for. Relations between *A* and *C* are established via *AB* and *BC* relations. Evaluation of pairs (*AB*, B*C*) as support for potential hypotheses about the relation between *A* and *C* is carried out by the domain expert.

We have applied the RaJoLink method to the scientific literature on autism and have used MEDLINE as a source of data. Autism was selected as the problem domain due to its complexity, insufficient and partial knowledge about its various causes, and because of the strong focus of current medical research towards early diagnosis of this disorder. In the autism domain we discovered a relation between autism and calcineurin and between autism and transcription factor NF-kappaB, which have been evaluated by a medical expert as relevant for better understanding of autism [3]. To assess the usefulness of RaJoLink in general, we evaluated the potential of our method also in the migraine-magnesium experiment, which represents a gold standard for the literature-based discovery [5]. For all these purposes we also developed a software tool, which implements the RaJoLink method and provides decision support to experts in the process of generating and testing of the scientific hypotheses in biomedical domains.

The unique contribution of the RaJoLink method and a fundamental difference from the previously proposed models of the literature-based knowledge discovery approach lies in the rarity principle that we apply to the open literature-based discovery. In fact, we use rare terms identified in the literature *C* to guide the search for new hypotheses. To this end, we have applied the rarity principle together with the notion of bisociation. In fact, the context-crossing connections, called bisociations, are often needed for creative, innovative discoveries. Bisociative relationships can only be discovered on the basis of a sufficiently large and diverse underlying corpus of information. In our case this corpus are MEDLINE papers. The larger the corpus is, the more likely it is to contain bisociative relationships. The RaJoLink approach has the potential for bisociative relation discovery as it allows switching between contexts (papers from different areas) by exploring rare terms in the intersection between contexts.

Besides this, we contributed an innovative approach also to the closed discovery process. The closed discovery process in RaJoLink is based on the outliers' detection in the content similarity graphs. In fact, having two disjoint literatures *A* and *C*, we automatically search for linking terms that are mentioned in both, the literature *A* as well as in the literature *C*. Pairs of documents with the same linking terms are subject to closer inspection in order to find out whether by putting statements about a linking term in these two articles together supports the hypothesis about a meaningful relation between previously disjoint literatures. In this manner, our search for linking terms is done in a semi-automated way that reduces manual work and efficiently points to meaningful relations between the domains *A* and *C*. In the closed discovery process we presented important connections between autism and calcineurin literature, as well as between autism and NF-kappaB literature. We discovered such connections by analysing outliers in the published evidence of some autism findings on one hand that coincide with specific calcineurin and NF-kappaB observations on the other hand.

3 Conclusion

The knowledge gathered by various specialised sciences throughout the digital era has resulted in large volumes of data and complex data interrelationships. To support biomedical experts in their knowledge discovery process, we have developed a literature mining method called RaJoLink. One of the main advantages of RaJoLink lays in the support of knowledge discovery by the innovative use of rare terms from the problem domain literature to guide the generation of new hypotheses. Accordingly, the crucial step of the method consists of selecting rare terms from the problem domain literature. The intuition behind this research idea was that the rarer a term is in the domain literature, the higher is the probability to encounter observations that represent something unexpected that may lead to creative discovery of new knowledge. This way we managed to employ rarity as a principle and means to find new interesting pieces of knowledge that were previously available in the dispersed literature and could be linked together. The results of the experimental case studies in autism [1-3, 6] and migraine [5] domain showed that the RaJoLink method can enhance the state-of-the-art methods for the literature-based knowledge discovery.

References

[1] Petrič, I.; Urbančič, T.; Cestnik, B. (2007). Discovering hidden knowledge from biomedical literature. *Informatica* 31(1), pp. 15-20.

[2] Petrič, I.; Urbančič, T.; Cestnik, B. (2006). Literature mining: potential for gaining hidden knowledge from biomedical articles. *Proceedings of the 9th International multi-conference Information Society IS-2006*, Ljubljana, pp. 52-55.

[3] Petrič, I.; Urbančič, T.; Cestnik, B.; Macedoni-Lukšič, M. (2009). Literature mining method RaJoLink for uncovering relations between biomedical concepts. *Journal of Biomedical Informatics* 42(2), pp. 219-227.

[4] Swanson, D. R. (1986). Undiscovered public knowledge. *Library Quarterly* 56(2), pp. 103-118.

[5] Urbančič, T.; Petrič, I.; Cestnik, B. (2009). A method for finding seeds of future discoveries in nowadays literature. *Foundations of intelligent systems.* Springer, Berlin, pp. 129-138.

[6] Urbančič, T.; Petrič, I.; Cestnik, B.; Macedoni-Lukšič, M. (2007). Literature mining: towards better understanding of autism. *Proceedings of the 11th Conference on Artificial Intelligence in Medicine in Europe*, Amsterdam, pp. 217-226.

Parsing with Intraclausal Coordination and Clause Detection

Domen Marinčič
Institut "Jožef Stefan", Jamova cesta 39, 1000 Ljubljana
E-mail: domen.marincic@ijs.si

Thesis Summary

Keywords: clause and intraclausal coordination detection, parsing, machine learning

Received: January 26, 2010

This paper presents the work on syntactic analysis of Slovene text. A new algorithm for parsing using intraclausal coordination and clause detection is described. The experiments show that the algorithm achieves a significant decrease in the number of parsing errors.

Povzetek: Članek opisuje nov algoritem za skladenjsko razčlenjevanje z iskanjem naštevanj in stavkov.

1 Introduction

Syntactic analysis, i.e., parsing of text is used during various tasks, e.g., machine translation, question answering, etc. The structure of a sentence is represented with a tree. Parsing long sentences is a difficult task. The motivation was to analyze sub-units of the sentence independently, which could improve the overall parsing accuracy. We developed a new parsing algorithm that includes intraclausal coordination and clause detection.

Parsing using clause detection was first tried by Abney (1), whose algorithm delimits non-embedded clauses before the complete parse is made. In (2), there is a short description of a rule-based parser where clause identification is included in the parsing process. A detailed description of our new algorithm can be found in (3).

To our knowledge, the algorithm is the first to use intraclausal coordination detection in cojunction with clause detection before parsing. The most important contribution is the decrease in the number of parsing errors by 7.1% and 6.4% for Slovene, compared to the Malt (4) and MSTP (5) baseline parsers, respectively.

2 The algorithm

The first phase is a loop for intraclausal coordination and clause detection. It begins by splitting the sentence into segments. Punctuation tokens and conjunctions are delimiters between the segments (the vertical line in Fig. 1). Then, the intraclausal coordinations are detected and reduced into the meta tokens. In the example in Fig. 1, one intraclausal coordination is found. In the next step, the sentence is split into segments again. At the end, clause detection and reduction is performed. The loop iterates until no more units can be retrieved or only one segment remains. Detection in the example sentence in Fig. 1 finishes in the step b).

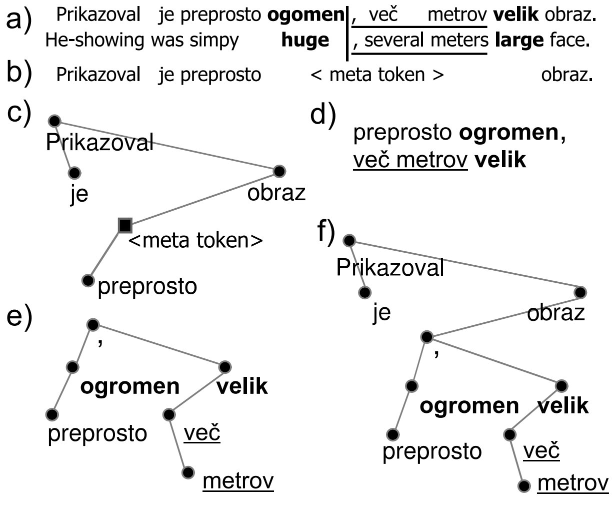

a) Prikazoval je preprosto ogomen , več metrov velik obraz.
He-showing was simpy huge , several meters large face.
b) Prikazoval je preprosto < meta token > obraz.
c)
Prikazoval
je
obraz
<meta token>
preprosto
d) preprosto ogromen, več metrov velik
e)
,
ogromen
velik
preprosto
več
metrov
f)
Prikazoval
je
obraz
,
ogromen
velik
preprosto
več
metrov

Figure 1: An example how a sentence is processed.

Detection of intraclausal coordinations and clauses is made in two steps: (i) candidate search and (ii) candidate classification using the AdaboostM1 algorithm. The candidates for intraclausal coordinations are searched for with the heuristic rule, stating that all the head words (in bold in Fig. 1a) must have the same part-of-speech and case. The candidates are then machinely classified using the features (presence of an adverb, noun/adjective matching with the head word) from the text between the head words, underlined in Fig. 1a. For the clause candidates, all the verb segments are taken. The following features present in the segment are used for machine classification: conjunctions, pronouns, punctuation tokens, auxiliary verbs, possible crossing intraclausal coordinations. The positively classified candidates are reduced.

The second phase builds the parse tree. It begins by parsing the sequence remaining after the first phase by the base parsers into the initial sentence tree, Fig. 1c. Certain errors in the initial tree are corrected by a newly developed rule-based parser. Then, the meta tokens are processed in a loop containing three steps: (i) the tokens of the meta-token subtree are joined with the unit that corresponds with the meta token, Fig. 1d; (ii) the new sequence is parsed, Fig. 1e; (iii) the subtree is merged with the sentence tree, Fig. 1f.

3 Evaluation

The experiments for estimating parsing accuracy are presented (Table 1). The part of the SDT corpus (6) from the Orwell's novel "1984" was used as the train and test set. Each experiment was carried out either with the MSTP parser or the Malt parser in the role of the base parsers. As the accuracy measure, the quotient between the nodes (punctuation excluded) assigned the correct parent and all the nodes in the tree was used. The accuracies of the plain MSTP and Malt parsers represent the baseline results.

Various versions of the new algorithm were compared: (i) the baseline parsers without detection; (ii) detection without classification and the rule-based parser; (iii) the classifiers turned on, no rule-based parser; (iv) the full version, achieving the 6.4% and 7.1% relative decrease of error compared to the baseline results.

Parsing algorithm	Malt	MSTP
Baseline	73,28 %	80,24 %
No classif., no rule-based p.	*74,63 %	*81,05 %
No rule-based parser	*74,83 %	*81,34 %
Full detection	*75,19 %	*81,51 %

Table 1: Parsing accuracy. The results marked with * are statistically significantly different from the baseline results.

4 Conclusion

The experiments show that by dividing complex sentences into smaller, more easily manageable units parsing accuracy can be increased. This was made possible by encoding background knowledge about the structure of clauses and intraclausal coordinations into the heuristic rules and classifiers used at the detection phase. Such knowledge apparently cannot be mined from the data by language-independent parsers. The most important idea for the future work seems to be the following: encode the information about the intraclausal coordination and clause structure as additional features of the words to enable a parser to combine this information with other knowledge about the parsed text more smoothly.

References

[1] Abney, S. P. (1990) In Proc. of the 6th New OED Conference : pp. 1–9.

[2] Holán, T. and Žabokrtský, Z. (2006) In Proc. of the TSD conference : pp. 95–102.

[3] Marinčič, D. Automatic text parsing with intraclausal coordination and clause detection, PhD thesis , Jozef Stefan International Postgraduate School (2008).

[4] Nivre, J. (2006) Inductive Dependency Parsing, Springer, The Netherlands.

[5] McDonald, R., Pereira, F., Ribarov, K., and Hajié, J. (2005) In Proc. of the HLT-EMNLP conference : pp. 523–530.

[6] Džeroski, S., Erjavec, T., Ledinek, N., Pajas, P., Žabokrtský, Z., and Žele, A. (2006) In Proc. of the LREC conference : pp. 1388–1391.

JOŽEF STEFAN INSTITUTE

Jožef Stefan (1835-1893) was one of the most prominent physicists of the 19th century. Born to Slovene parents, he obtained his Ph.D. at Vienna University, where he was later Director of the Physics Institute, Vice-President of the Vienna Academy of Sciences and a member of several scientific institutions in Europe. Stefan explored many areas in hydrodynamics, optics, acoustics, electricity, magnetism and the kinetic theory of gases. Among other things, he originated the law that the total radiation from a black body is proportional to the 4th power of its absolute temperature, known as the Stefan–Boltzmann law.

The Jožef Stefan Institute (JSI) is the leading independent scientific research institution in Slovenia, covering a broad spectrum of fundamental and applied research in the fields of physics, chemistry and biochemistry, electronics and information science, nuclear science technology, energy research and environmental science.

The Jožef Stefan Institute (JSI) is a research organisation for pure and applied research in the natural sciences and technology. Both are closely interconnected in research departments composed of different task teams. Emphasis in basic research is given to the development and education of young scientists, while applied research and development serve for the transfer of advanced knowledge, contributing to the development of the national economy and society in general.

At present the Institute, with a total of about 800 staff, has 600 researchers, about 250 of whom are postgraduates, nearly 400 of whom have doctorates (Ph.D.), and around 200 of whom have permanent professorships or temporary teaching assignments at the Universities.

In view of its activities and status, the JSI plays the role of a national institute, complementing the role of the universities and bridging the gap between basic science and applications.

Research at the JSI includes the following major fields: physics; chemistry; electronics, informatics and computer sciences; biochemistry; ecology; reactor technology; applied mathematics. Most of the activities are more or less closely connected to information sciences, in particular computer sciences, artificial intelligence, language and speech technologies, computer-aided design, computer architectures, biocybernetics and robotics, computer automation and control, professional electronics, digital communications and networks, and applied mathematics.

The Institute is located in Ljubljana, the capital of the independent state of **Slove**nia (or S♡nia). The capital today is considered a crossroad between East, West and Mediterranean Europe, offering excellent productive capabilities and solid business opportunities, with strong international connections. Ljubljana is connected to important centers such as Prague, Budapest, Vienna, Zagreb, Milan, Rome, Monaco, Nice, Bern and Munich, all within a radius of 600 km.

From the Jožef Stefan Institute, the Technology park “Ljubljana” has been proposed as part of the national strategy for technological development to foster synergies between research and industry, to promote joint ventures between university bodies, research institutes and innovative industry, to act as an incubator for high-tech initiatives and to accelerate the development cycle of innovative products.

Part of the Institute was reorganized into several high-tech units supported by and connected within the Technology park at the Jožef Stefan Institute, established as the beginning of a regional Technology park "Ljubljana". The project was developed at a particularly historical moment, characterized by the process of state reorganisation, privatisation and private initiative. The national Technology Park is a shareholding company hosting an independent venture-capital institution.

The promoters and operational entities of the project are the Republic of Slovenia, Ministry of Higher Education, Science and Technology and the Jožef Stefan Institute. The framework of the operation also includes the University of Ljubljana, the National Institute of Chemistry, the Institute for Electronics and Vacuum Technology and the Institute for Materials and Construction Research among others. In addition, the project is supported by the Ministry of the Economy, the National Chamber of Economy and the City of Ljubljana.

Jožef Stefan Institute
Jamova 39, 1000 Ljubljana, Slovenia
Tel.:+386 1 4773 900, Fax.:+386 1 251 93 85
WWW: http://www.ijs.si
E-mail: matjaz.gams@ijs.si
Public relations: Polona Strnad

INFORMATICA
AN INTERNATIONAL JOURNAL OF COMPUTING AND INFORMATICS
INVITATION, COOPERATION

Submissions and Refereeing

Please submit an email with the manuscript to one of the editors from the Editorial Board or to the Managing Editor. At least two referees outside the author's country will examine it, and they are invited to make as many remarks as possible from typing errors to global philosophical disagreements. The chosen editor will send the author the obtained reviews. If the paper is accepted, the editor will also send an email to the managing editor. The executive board will inform the author that the paper has been accepted, and the author will send the paper to the managing editor. The paper will be published within one year of receipt of email with the text in Informatica MS Word format or Informatica LaTeX format and figures in .eps format. Style and examples of papers can be obtained from http://www.informatica.si. Opinions, news, calls for conferences, calls for papers, etc. should be sent directly to the managing editor.

QUESTIONNAIRE

□ Send Informatica free of charge

□ Yes, we subscribe

Please, complete the order form and send it to Dr. Drago Torkar, Informatica, Institut Jožef Stefan, Jamova 39, 1000 Ljubljana, Slovenia. E-mail: drago.torkar@ijs.si

Since 1977, Informatica has been a major Slovenian scientific journal of computing and informatics, including telecommunications, automation and other related areas. In its 16th year (more than sixteen years ago) it became truly international, although it still remains connected to Central Europe. The basic aim of Informatica is to impose intellectual values (science, engineering) in a distributed organisation.

Informatica is a journal primarily covering the European computer science and informatics community - scientific and educational as well as technical, commercial and industrial. Its basic aim is to enhance communications between different European structures on the basis of equal rights and international refereeing. It publishes scientific papers accepted by at least two referees outside the author's country. In addition, it contains information about conferences, opinions, critical examinations of existing publications and news. Finally, major practical achievements and innovations in the computer and information industry are presented through commercial publications as well as through independent evaluations.

Editing and refereeing are distributed. Each editor can conduct the refereeing process by appointing two new referees or referees from the Board of Referees or Editorial Board. Referees should not be from the author's country. If new referees are appointed, their names will appear in the Refereeing Board.

Informatica is free of charge for major scientific, educational and governmental institutions. Others should subscribe (see the last page of Informatica).

ORDER FORM – INFORMATICA

Name: ....

Title and Profession (optional): ....

....

Home Address and Telephone (optional): ....

....

Office Address and Telephone (optional): ....

....

E-mail Address (optional): ....

Signature and Date: ....

Informatica WWW:

http://www.informatica.si/

Referees from 2008 on:

Informatica

An International Journal of Computing and Informatics

Web edition of Informatica may be accessed at: http://www.informatica.si.

Subscription Information Informatica (ISSN 0350-5596) is published four times a year in Spring, Summer, Autumn, and Winter (4 issues per year) by the Slovene Society Informatika, Vožarski pot 12, 1000 Ljubljana, Slovenia.
The subscription rate for 2010 (Volume 34) is
– 60 EUR for institutions,
– 30 EUR for individuals, and
– 15 EUR for students
Claims for missing issues will be honored free of charge within six months after the publication date of the issue.

Typesetting: Borut Žnidar.
Printing: Dikplast Kregar Ivan s.p., Kotna ulica 5, 3000 Celje.

Orders may be placed by email (drago.torkar@ijs.si), telephone (+386 1 477 3900) or fax (+386 1 251 93 85). The payment should be made to our bank account no.: 02083-0013014662 at NLB d.d., 1520 Ljubljana, Trg republike 2, Slovenija, IBAN no.: SI56020830013014662, SWIFT Code: LJBASI2X.

Informatica is published by Slovene Society Informatika (president Niko Schlamberger) in cooperation with the following societies (and contact persons):
Robotics Society of Slovenia (Jadran Lenarčič)
Slovene Society for Pattern Recognition (Franjo Pernuš)
Slovenian Artificial Intelligence Society; Cognitive Science Society (Matjaž Gams)
Slovenian Society of Mathematicians, Physicists and Astronomers (Bojan Mohar)
Automatic Control Society of Slovenia (Borut Zupančič)
Slovenian Association of Technical and Natural Sciences / Engineering Academy of Slovenia (Igor Grabec)
ACM Slovenia (Dunja Mladenič)

Informatica is surveyed by: ACM Digital Library, Citeseer, COBISS, Compendex, Computer & Information Systems Abstracts, Computer Database, Computer Science Index, Current Mathematical Publications, DBLP Computer Science Bibliography, Directory of Open Access Journals, InfoTrac OneFile, Inspec, Linguistic and Language Behaviour Abstracts, Mathematical Reviews, MatSciNet, MatSci on SilverPlatter, Scopus, Zentralblatt Math

The issuing of the Informatica journal is financially supported by the Ministry of Higher Education, Science and Technology, Trg OF 13, 1000 Ljubljana, Slovenia.

Volume 34 Number 2 June 2010 ISSN 0350-5596

Informatica

An International Journal of Computing and Informatics

End of Special Issue / Start of normal papers

Informatica **34** (2010) Number 2, pp. 127–265