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.

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-

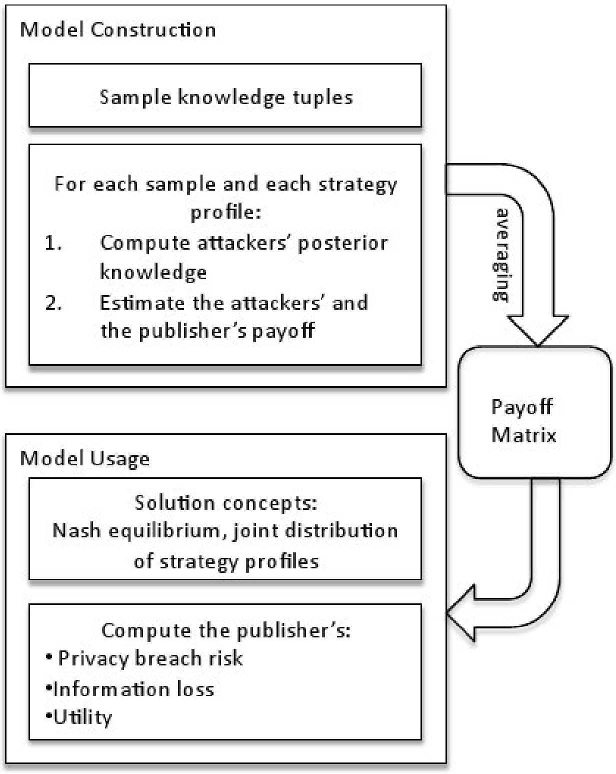

Figure 2: Overview of the strategic model of privacy attackers and database publisher.

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

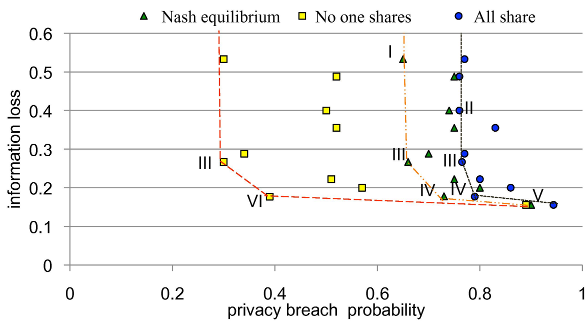

Figure 4: Expected privacy breach and information loss under various generalization actions and attacker behaviors.

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.

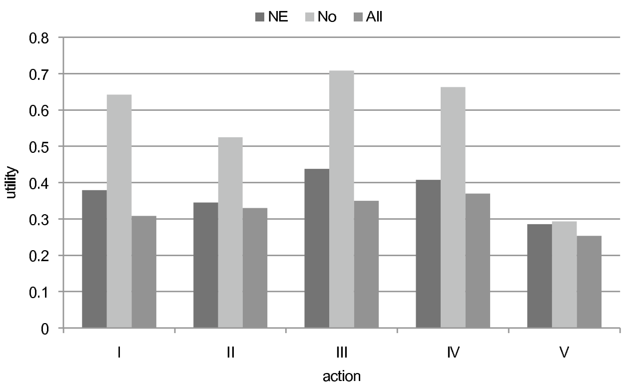

Figure 5: Database publisher's utility ($w = 0.25$) under different generalization actions and scenarios for attackers' behavior.

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.