

Volume 36 Number 1 March 2012

ISSN 0350-5596

# *Informatica*

**An International Journal of Computing  
and Informatics**

Special Issue:

**IPTV and Multimedia Services**

Guest Editor:

**E. Mikóczy**

**I. Vidal**

**D. Kanellopoulos**



1977

## EDITORIAL BOARDS, PUBLISHING COUNCIL

Informatica is a journal primarily covering intelligent systems in the European computer science, informatics and cognitive 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.

### Executive Editor – Editor in Chief

Anton P. Železnikar  
Volaričeva 8, Ljubljana, Slovenia  
s51em@lea.hamradio.si  
<http://lea.hamradio.si/~s51em/>

### Executive Associate Editor - Managing Editor

Matjaž Gams, Jožef Stefan Institute  
Jamova 39, 1000 Ljubljana, Slovenia  
Phone: +386 1 4773 900, Fax: +386 1 251 93 85  
matjaz.gams@ijs.si  
<http://dis.ijs.si/mezi/matjaz.html>

### Executive Associate Editor - Deputy Managing Editor

Mitja Luštrek, Jožef Stefan Institute  
mitja.lustrek@ijs.si

### Executive Associate Editor - Technical Editor

Drago Torkar, Jožef Stefan Institute  
Jamova 39, 1000 Ljubljana, Slovenia  
Phone: +386 1 4773 900, Fax: +386 1 251 93 85  
drago.torkar@ijs.si

### Editorial Board

Juan Carlos Augusto (Argentina)  
Costin Badica (Romania)  
Vladimir Batagelj (Slovenia)  
Francesco Bergadano (Italy)  
Marco Botta (Italy)  
Pavel Brazdil (Portugal)  
Andrej Brodnik (Slovenia)  
Ivan Bruha (Canada)  
Wray Buntine (Finland)  
Zhihua Cui (China)  
Ondrej Drbohlav (Czech Republic)  
Hubert L. Dreyfus (USA)  
Jozo Dujmović (USA)  
Johann Eder (Austria)  
Ling Feng (China)  
Vladimir A. Fomichov (Russia)  
Maria Ganzha (Poland)  
Marjan Gušev (Macedonia)  
N. Jaisankar (India)  
Dimitris Kanellopoulos (Greece)  
Samee Ullah Khan (USA)  
Hiroaki Kitano (Japan)  
Igor Kononenko (Slovenia)  
Miroslav Kubat (USA)  
Ante Lauc (Croatia)  
Jadran Lenarčič (Slovenia)  
Shiguo Lian (China)  
Huan Liu (USA)  
Suzana Loskovska (Macedonia)  
Ramon L. de Mantras (Spain)  
Angelo Montanari (Italy)  
Pavol Návrat (Slovakia)  
Jerzy R. Nawrocki (Poland)  
Nadia Nedjah (Brasil)  
Franc Novak (Slovenia)  
Marcin Paprzycki (USA/Poland)  
Ivana Podnar Žarko (Croatia)  
Karl H. Pribram (USA)  
Luc De Raedt (Belgium)  
Shahram Rahimi (USA)  
Dejan Raković (Serbia)  
Jean Ramaekers (Belgium)  
Wilhelm Rossak (Germany)  
Ivan Rozman (Slovenia)  
Sugata Sanyal (India)  
Walter Schempp (Germany)  
Johannes Schwinn (Germany)  
Zhongzhi Shi (China)  
Oliviero Stock (Italy)  
Robert Trappl (Austria)  
Terry Winograd (USA)  
Stefan Wrobel (Germany)  
Konrad Wrona (France)  
Xindong Wu (USA)

## Editorial: Special Issue on “IPTV and Multimedia Services”

### 1 Introduction

Nowadays, there is a growing interest in delivery of TV services via IP networks, known as IPTV (Internet Protocol Television). Operators and vendors are currently working on IPTV standardization efforts (e.g., ATIS/IIF, ITU-T IPTV-GSI, ETSI TISPAN) to bear wider availability and interoperability of IPTV as a secure, reliable, managed multimedia service. Although, technologies for packet video have been explored for some time, there are many remaining issues in the design, development and deployment of commercially viable IPTV services. These issues include mainly the standardization of architectural elements, content protection and service aspects including scalability, portability, interoperability, performance and accounting.

This special issue aims to bring together research work in the area of IPTV and multimedia services, investigate the novel solutions and discuss the future trends in this field. This special issue of the *Informatica Journal* invited authors to submit their original work that communicates current research on IPTV and multimedia services regarding both the novel solutions and future trends in the field.

In this special issue, we have five papers, which can demonstrate advanced works in the field including: IPTV evolution towards next-generation networks (NGN) and hybrid scenarios, IPTV services personalisation, implementation of the IPTV Media Function, and secure key exchange scheme for IPTV Broadcasting.

### 2 The papers in this special issue

The future evolution of IPTV architecture and services will depend on the acceptance of the NGN based IPTV concept by operators and vendors. In the first paper, entitled: “IPTV evolution towards NGN and hybrid scenarios” *E. Mikóczy, I. Vidal and D. Kanellopoulos* survey actual NGN-based IPTV standards and the development of several new technologies that can have an impact on content services in the next years.

Generally, the advances in IPTV technology enable a new user-centric and interactive TV model, in which context-awareness is promising in making the user’s interaction with the TV dynamic and transparent. In the second paper, *S. Song, H. Moustafa and H. Afifi* present a solution for IPTV services personalization by introducing context-awareness on top of the IPTV architecture to gather different information of the user and his/her environment. The proposed solution allows each user to be distinguished to the system in a unique manner. The authors implemented the proposed solution on top of an IPTV platform considering the NGN IPTV architecture as a proof of concept and as a means to evaluate the performance.

From another perspective, recommendation services for IPTV should provide users with referrals of items they will appreciate based upon their personal preferences. In the third paper, *A. Elmisery and D. Botvich* introduce a framework for private recommender service based on *Enhanced Middleware for Collaborative Privacy* (EMCP). EMCP executes a two-stage concealment process that gives the user a complete control on the privacy level of his profile. The authors utilize a trust mechanism to augment recommendation’s accuracy and privacy. Trust heuristic spot users whom are trustworthy with respect to the user requesting recommendation (target-user). Later, the neighbourhood formation is calculated using proximity metrics based on these trustworthy users. Finally, users submit their profiles in an obfuscated form without revealing any information about their data, and the computation of recommendations proceeds over the obfuscated data using secure multi-party computation protocol. The authors expand the obfuscation scope from single obfuscation level for all users to arbitrary obfuscation levels based on trustworthiness between users. In particular, they correlate the obfuscation level with different trust levels, so the more trusted a target user is the less obfuscation copy of users’ profile he can access. The authors also provide an IPTV network scenario and experimentation results. Their results and analysis show that their two-stage concealment process not only protects the users’ privacy, but also can maintain the recommendations accuracy.

Multimedia in IPTV is handled by a separate unit, the *Media Function* (MF), which is made up of *Media Control* and *Media Delivery Functions* (MCF & MDF). According to the different specifications of an IP Multimedia Subsystem (IMS)-based IPTV architecture, the *User Equipment* (UE) is expected to use the Real Time Streaming Protocol (RTSP) protocol as a media control protocol to interact with the MCF, and gets the delivery of media from the MDF using the Real-time Transport Protocol (RTP) protocol. This also means that the streaming session can be initiated from the media controller on behalf of the user but the delivery of media is sent to the UE from the media server. Due to the lack of free and open source Media Servers and on the contrary, the availability of free and open source streaming servers; the ideal choice for the delivery of IPTV services by the research community is *Streaming Servers*. Nevertheless, because of denial of service attack and other issues, most streaming servers do not allow a different location for the session setup request and the delivery of media of the streaming session. Speaking more precisely, most streaming servers are not designed to be controlled by some other entity than the RTSP client that consumes the media. This makes it difficult to have a separate media control unit for IPTV service in

IMS, if one wants to use a streaming server as an MDF unit. Consequently, it is better to find a work around so as to use streaming servers to develop and test IPTV services in IMS environments than waiting for streaming servers to work in this manner.

For this purpose, in the fourth paper entitled: “An RTSP Proxy for implementing the IPTV Media Function using a streaming server”, *Z. Shibeshi, A. Terzoli and K. Bradshaw* propose another component (an RTSP proxy and relay unit) to be part of the IPTV Media Function (MF) and mediate between the MFC and MDF. The proposed unit properly relays media control commands from the UE and MFC to the MDF and the RTP packets from the MDF to the UE. This unit also helps one to implement other streaming functionalities that are required for IPTV service delivery and which are not implemented in the current open source streaming servers. Additional services can also be easily implemented with the help of this unit. This facilitates the development of an IPTV service using the readily available open source streaming servers. The authors show how this RTSP proxy unit can be integrated into the Media Function of the IPTV architecture to ease the media delivery process of IMS based IPTV service.

In the last paper, *Ravi Singh Pippal, Jaidhar C. D. and Shashikala Tapaswi* present a secure mutual authentication and key exchange scheme between set-top box (STB) and smart card for IPTV broadcasting. The proposed scheme provides dynamic session key agreement and mutual authentication. Security analysis proves that the proposed scheme is strong against subscriber and STB impersonation attacks, replay attack, stolen verifier attack, smart card loss attack, man-in-the-middle attack and attack on perfect forward secrecy, which are considered as common threats in IPTV environment.

The list of the papers follows:

- E. Mikóczy, I. Vidal and D. Kanellopoulos. “IPTV evolution towards NGN and hybrid scenarios”.
- S. Song, H. Moustafa and H. Afifi. “IPTV services personalization using context-awareness”.
- A. Elmisery and D. Botvich. “Privacy aware recommender service using multi-agent middleware—an IPTV network scenario”.
- Z. Shibeshi, A. Terzoli and K. Bradshaw. “An RTSP Proxy for implementing the IPTV Media Function using a streaming server”.
- Ravi Singh Pippal, Jaidhar C.D. and Shashikala Tapaswi. “Secure key exchange scheme for IPTV Broadcasting”.

## Acknowledgments

The guest editors wish to thank Prof. Anton P. Zeleznikar (Editor-in-Chief of the Informatica Journal) and Prof. Matjaz Gams (Managing Editor) for providing the opportunity to edit this special issue on “IPTV and Multimedia Services”. We would also like to thank the authors for submitting their works as well as the referees who have critically evaluated the papers within the short stipulated time. Finally, we hope the reader will share our joy and find this special issue very useful.

*E. Mikóczy, I. Vidal, D. Kanellopoulos*  
Guest Editors

# IPTV Evolution Towards NGN and Hybrid Scenarios

Eugen Mikóczy  
Slovak University of Technology  
Bratislava, Slovakia  
E-mail: eugen.mikoczy@stuba.sk

Iván Vidal  
Department of Telematic Engineering,  
University Carlos III of Madrid,  
Madrid, Spain  
E-mail: ivaldal@it.uc3m.es

Dimitris Kanellopoulos  
Educational Software Development Laboratory (ESDLab)  
Department of Mathematics,  
University of Patras, Greece  
E-mail: d\_kan2006@yahoo.gr

**Keywords:** IPTV, multimedia system, NGN, content delivery network

**Received:** November 1, 2011

*Internet Protocol-based television (IPTV) concerns video entertainment and represents a solution for interactive television-like services over IP-based networks. Operators and vendors are currently working on IPTV standardization efforts (e.g., ATIS/HIF, ITU-T IPTV-GSI, ETSI TISPAN) to bear wider availability and interoperability of IPTV as a secure, reliable, managed multimedia service. In this paper, we present a generic IPTV architecture, and explain standardization efforts such as TISPAN, OIPF, ITU-T and ATIS specifications for the next generation IPTV. We review new approaches in multimedia services and media delivery, and present open issues in IPTV networks. Finally, we give research directions for future work.*

*Povzetek: Članek naredi pregled televizije IP, njene arhitekture in standardov.*

## 1 Introduction

Television in IPTV is distributed to subscribers using a broadband connection over the IP. IPTV is expected to grow rapidly as broadband is now available to more than 500 million households worldwide [1]. Currently, most of telecommunication companies (telcos) and cable operators are actively providing IPTV-based services. Total IPTV subscriber numbers have now reached 48.2 million at the end of the first quarter 2011 [1] with growth over 34% in the last 12 months (till Q1 2011). Telcos acquire and manage video content for distribution, and ultimately deliver this content to the end-user using various transmission techniques (i.e., unicast, multicast, or broadcast). As far as the content in IPTV is concerned, there are three different categories:

- *Linear content.* In this case, viewers adapt to the usual grid-style programming schedule that now exists on cable, satellite, and over-the-air TV. This is a conventional multichannel pay-TV service.
- *On-demand content.* Programming is streamed “on-demand” to the viewer from a video server system to a set-top box (STB) device. Accessible

cable-based VOD services provide many programming choices such as movies at the request of individual subscribers.

- *Exclusive content.* It is programming that is unique to an individual service provider (SP). Exclusive content can provide consumers with a reason to choose one SP’s service over another SP’s (e.g., movie blockbusters, sport events).

The distinctive feature of IPTV service is that interactivity can be easily provided, as the end-user can compile the program according to his/her individual preferences. Actually, available IPTV services include: (1) targeted advertising (e.g., banner advertising in an electronic programming guide); (2) (EPG) or sponsored advertising for on-demand content; (3) in-program electronic messaging; (4) personal TV channels; (5) sharing of photos, movies, and interests; (6) walled garden portals – weather, sports, recipes, etc.; (7) EPG-based electronic messaging and social networking; (8) home security and management services; (9) whole home DVR; (10)

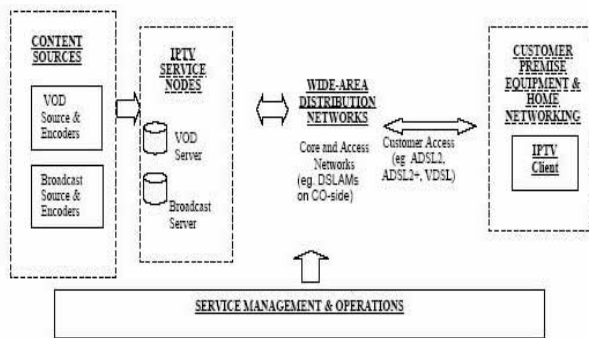


Figure 1: IPTV network architecture [Source: <http://www.networkdictionary.com/networking/IPTV.php>]

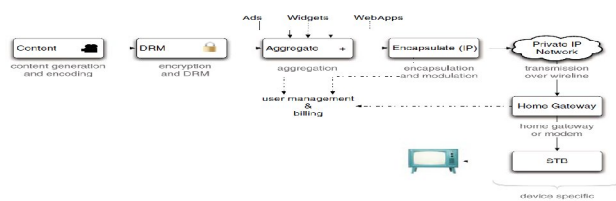


Figure 2: IPTV developments [23].

network-based time shifting archive; (11) sports participation and gaming; (12) integration with Voice over IP telephony for TV display of call information and call routing, as well as caller ID and blocking, displayed on the screen and call forwarding; (13) PVR programming via mobile phone; (14) accessing Internet services via TV; (15) voting.

The rest of the paper is organized as follows. Section 2 presents a generic IPTV architecture, while section 3 describes standardization efforts such as TISPAN, OIPF, ITU-T and ATIS specifications for the next generation IPTV. Section 4 reviews new approaches in multimedia services and media delivery. Section 5 presents open issues in IPTV networks, while section 6 concludes the paper.

## 2 IPTV Architecture

The architecture of an IPTV network is shown in Figure 1, while current IPTV developments are depicted in Figure 2. In an IPTV network, telcos acquire and manage video content for distribution, and deliver this content to the end-user over broadband IP infrastructure. This is a complex process with precise business and technical requirements.

### 2.1 Content acquisition and management

Not all-video content is produced equal. However, compression efficiency and visual quality should be maximized with existing content types dictating the outcome. For example, a static news program dictates that may require less bandwidth than a fast moving sports program. From another viewpoint, a complete IPTV solution requires effective content management

that refers to programming and advertising. Content management encapsulates a number of critical systems, which can be categorized into three distinct aspects [2]:

- *Reception and encoding*
- *Rights management*
- *Back office billing, provisioning, activation, and monitoring.*

*Reception and encoding:* A usual IPTV deployment employs advanced compression technologies to deliver content most efficiently. The encoding process uses sophisticated compression technologies such as H.264 which may potentially help SPs reduce the bandwidth required to deliver a standard or high-definition video stream by as much as 50% (compare MPEG-2). SPs receive often content via satellite or terrestrial broadcasts. The requirements of the content distribution system then dictate how the received content is encoded. Typically, encoding national content occurs only once at the master head-end, assuming the SP has the required infrastructure ready to distribute national content to regional head-ends or hubs. The master head-end then passes the encoded content to the regional hub. The regional hub also receives and encodes local content. This is a cost-effective way of managing the encoding process as it greatly reduces the need for high-capacity encoding at the regional level, thus minimizing the expense of purchasing costly encoding equipment.

*Rights management:* When SPs implement an IPTV solution, pay-TV content must be protected during transmission, from the head-end to the set-top box. The methods in which end-users consume content dictate how the content must be protected and the nature of the content protection. Cable and satellite operators have utilized conditional access security systems that serve to restrict content usage to only those authorized to view the content. As illegal copy and re-distribution of IPTV content now become easier and simpler, SPs should design proper digital rights management systems (DRM) with a look at the rising digital home. It is required to protect IPTV content or service. Consequently, flexible usage rules must accompany a part of video content throughout its usage lifespan, which may include storage within the *customer premise equipment* (CPE) itself (Local PVR), distribution to client devices within the home, and finally distribution outside the home to a mobile device as well. The work in [3] analyzes the security threats and requirements, and addresses interoperability issues among different content and service protection systems for IPTV.

### 2.2 Back-office billing, provisioning, activation, and monitoring

An IPTV system includes a master head-end and a host of regional hubs. IPTV network architectures need

back-office content management systems that handle billing for subscriptions to linear and on-demand content. For example, linear content is billed on a subscription basis according to the specific tier of service being subscribed to. On the other hand, on demand content is more complex in that different types of pricing and packaging models are often offered, including subscription and a la carte at various price points with potentially different usage rules. Content management systems must also handle service activation in such a way that subscribers get what they pay for. In an IPTV system, content must be accessed on the fly from wherever it resides. In addition, the system needs to keep track of content once it has been delivered to the consumer for billing and rights management purposes. From another perspective, the IPTV system must be intelligent with respect to targeted advertising campaign management and digital asset location and delivery.

Within an IPTV framework, the delivery of video content based on multicast and unicast over IP network (e.g., IP MPLS) is what really differentiate IPTV from legacy radio frequency (RF) cable systems. Most of IPTV providers use IPTV middleware with centralized or distributed server infrastructure connected over IP network operator core, aggregation and access network to IPTV end-device in home network.

IPTV providers or content providers can use also dedicated infrastructure *Content Delivery Network* (CDN) where content is replicated over several mirrored CDN servers in order to perform transparent and effective delivery of content to the end-users [4],[5]. In particular, content can be pre-recorded or retrieved from live sources; it can be persistent or transient data within the system [6]. CDNs provide services that improve network performance by maximizing bandwidth, improving accessibility, and maintaining correctness through content replication. Some of the most popular commercial CDNs (e.g., Akamai, Limelight, EdgeStream, Jetstream, etc.) that provide distribution of content (contained in web pages) are also delivering video content over the public Internet. CDNs actually handle significant part of overall Internet traffic and video content ratio growth (e.g., Netflix video service traffic has accounted more as 20% of total downstream traffic during peak period in North America in 2011). CDN provider uses its own CDN infrastructure on top of existing broadband connectivity over consumer cable or ADSL modem connections around the globe, be able deliver video streaming over paths that have no more as 20 router hops between their server and end-user (move content as close to end-user as possible). Generally, a CDN system is composed of three main entities: content provider, CDN provider, and end-users.

- A *content provider* (or customer) is one who delegates the Uniform Resource Locator (URL) name space of the video objects to be distributed.

- *CDN providers* use caching and/or replica servers located at different geographical locations to replicate content. CDN cache servers are also called *edge servers* or *surrogates*. The edge servers of a CDN are called *Web cluster* as a whole. CDNs distribute content to the edge servers in such a way that all of them share the same content and URL. Client requests are redirected to the nearby optimal edge server and it delivers requested content to the end-users.

The typical functionalities of a CDN include:

- *Request redirection and content delivery services*, to direct a request to the closest suitable CDN cache server by using mechanisms to bypass congestion. The request-routing component directs client's requests to appropriate edge servers and interacts with the distribution component to keep an up-to-date view of the content stored in the CDN caches. The content-delivery component delivers the content and consists of the origin server and a set of replica servers that deliver copies of content to the end users.
- *Content outsourcing and distribution services*, to replicate and/or cache content from the origin server to distributed Web servers. The distribution component moves content from the origin server to the CDN edge servers and ensures consistency of content in the caches.
- *Content negotiation services*, to meet specific needs of each individual user (or group of users).
- *Management services*, to manage the network components, to handle accounting, and to monitor and report on content usage. The accounting component maintains logs of client accesses and records the usage of the CDN servers. This information is used for traffic reporting and usage-based billing by the content provider itself or by a third-party billing organization. CDNs support an accounting mechanism that collects and tracks client usage information related to request-routing, distribution, and delivery [7].

## 2.3 End devices and home networks

The final part of IPTV architecture is home network connected to IPTV service provider access network with home gateway. The parts of the home network are also end-devices that receive and present IPTV service to subscriber TV. End devices, also called CPE (*customer premise equipment*), are usually set top boxes connected to home gateway and TV. In the future, we expect interconnection of the multiple type of devices with IPTV client software like for example connected TVs, game consoles, PCs, mobile or tablet devices.

### 3 Next Generation IPTV in Standardization

Most of the current IPTV middleware solutions have been developed before standardization bodies start working on specification for Next Generation of IPTV systems. In the following section, we describe the most important standard specifications that in the next years it is expected these to be implemented and bring new IPTV services and advanced user experience as it was provided with the first generation of proprietary IPTV systems.

It is noteworthy that ETSI TISPAN specifies (next-generation network) NGN-based IPTV as “*Multimedia system that provides IPTV services over the NGN architecture and may be implemented as an integrated subsystem with the NGN (NGN integrated IPTV) or may use the IMS subsystem (IMS-based IPTV) in the NGN*”.

#### 3.1 Overview about TISPAN, OIPF, ITU-T, ATIS specs

In the last years, IPTV has received significant attention from several organizations and standard bodies. Standardization activities on IPTV have resulted in different technical specifications covering the architecture and functions of an IPTV system, which have been made available to the telecommunication market. Hereafter, we describe the most relevant initiatives related with IPTV standardization.

One significant initiative corresponds to the standardization activities carried out by the ITU-T (International Telecommunication Union, Telecommunication Standardization Sector). ITU-T work on IPTV was initially addressed by the IPTV Focus group, which resulted in a first set of draft specifications. From 2008, IPTV Global Standards Initiative (IPTV-GSI) coordinates all the ITU-T activities related with IPTV. ITU-T recommendations for IPTV cover different topics related with home networking, applications and end-systems, architecture, QoE and security. The Recommendation Y.1910 [8] describes three IPTV architecture models: (a) non-NGN IPTV; (b) NGN without IMS IPTV; and (c) NGN with IMS IPTV.

ETSI TISPAN (European Telecommunications Standards Institute, Telecommunications and Internet converged Services and Protocols for Advanced Networking) is also conducting a relevant work on IPTV standardization. Release 2 of specifications of TISPAN NGN (April 2008) introduced IPTV as a service in the NGN architecture. This release describes an IMS-based and a non IMS-based IPTV system. In 2011, TISPAN work on IPTV has concluded to the IPTV specification in NGN release 3.

The Open IPTV Forum (OIPF) aims at developing open and interoperable end-to-end specifications for IPTV. Currently, OIPF has completed release 1 and 2 of specifications [9]. These specifications cover the

different aspects of IPTV such as media delivery, session management, service discovery and security.

The Alliance for Telecommunications Industry Solutions (ATIS) IPTV Interoperability Forum (IIF) develops an end-to-end solution for IPTV [10]. ATIS IIF specifications comprise a set of deliverables that describe the different aspects related with the delivery of IPTV service to the end-user, such as the IPTV architecture, QoS, security and interoperability.

#### 3.2 IMS vs. non-IMS IPTV

Release 2 of specifications of TISPAN NGN adds new features and services to the NGN such as IPTV. To leverage investments made on the IMS, this release addresses the specification of the IMS-based IPTV service architecture. This architecture implements a set of service layer requirements defined in [11]. On the other hand, NGN release 2 also defines an NGN integrated IPTV system, with the aim of supporting the aforementioned requirements and allowing the integration of existing IPTV solutions defined by other organizations, such as DVB, ATIS IIF or ITU.

TISPAN release 2 defines three main services for the IMS-based IPTV system: (a) Broadcast TV; (b) Content on Demand (CoD); and (c) Network-Personal Video Recorder (N-PVR). Release 3 adds a broad set of value-added services, such as push CoD, interactive TV, communications and messaging, interaction between users, user generated content, content recommendation, games, personalized channel, personalized service composition, service continuation fixed-mobile and remote control of IPTV services, to name some of them.

The NGN integrated IPTV subsystem (as shown in Figure 3) provides basic integration of IPTV functions to NGN architecture and other NGN subsystems especially with the User Profile Server Function (UPSF), and the transport control Resource and Admission Control Subsystem (RACS) and Network Attachment Subsystem (NASS) [12].

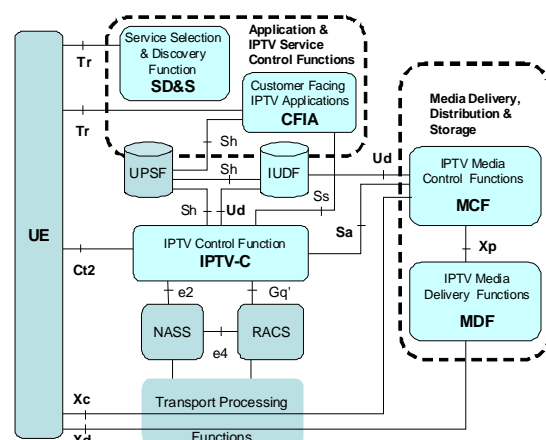


Figure 3: Simplified TISPAN NGN integrated IPTV subsystem functional architecture [12].



A user can access with his/her user equipment (UE like set-top-box STB) the service description via SD&S service, selection and discovery procedures that follow DVB IPI specification and use HTTP protocol. The same Tr interface can be used by UE for accessing the user interface and service selection over Customer Facing IPTV application (CFIA). CFIA provides this interface IPTV service provisioning, selection and authorization. IPTV control (IPTV-C) is enabled over HTTP or RTSP control. Media (e.g., content on demand – CoD) can be streamed by unicast or multicast over the Xd interface from a Media Delivery Function (MDF) and controlled via the Xc interface by a Media Control Function (MCF). UE can also access common services in NGN via interactions with NGN applications. Figure 4 shows an overview of the functional architecture defined by TISPAN for the IMS-based IPTV system in release 3 [13]. In Figure 4, the Service Discovery Function (SDF) and the Service Selection Function (SSF) assist the UE in the process of selecting an IPTV service. The SDF provides a set of SSF addresses, and the SSF provides the UE with service selection information. For each IPTV service, the SSF provides (a) the identifiers corresponding to the service; (b) the network parameters that may be required to establish the service; and (c) data related to the service for human consumption. Service selection information may be personalized by the SSF, or this entity may provide additional information to do this personalization.

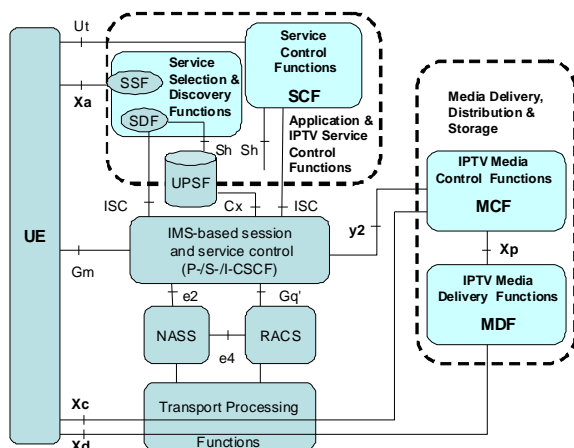


Figure 4: Simplified TISPAN NGN integrated IPTV subsystem functional architecture [13].

The IPTV Service Control Functions (SCF) is a SIP application server that performs tasks related with service authorization, during session initiation and modification, and with credit control. The UE interacts with the SCF via the Core IMS. This interaction (based on the SIP protocol) allows executing procedures related with the management of the media session. Additionally, the SCF can perform other tasks such as selecting the appropriate IPTV Media Functions, initiating the download of media content (e.g., CoD

content) to the UE or detecting IPTV service state information. An exhaustive list of SCF tasks can be found in [12].

The Media Control Function (MCF) handles the control of the media flows that are delivered by means of the Media Delivery Function (MDF). The MCF interacts with the SCF and may control a set of different MDFs. The MDF handles the delivery of media flows to the UE. This entity supports additional functionalities such as storing media, behaving as a source of IPTV streams, may also be implementing media processing and transcoding facilities and collecting QoE reports from the UE.

The User Profile Server Function (UPSF) maintains the IMS user profile and the profile information that is specific to IPTV. The SCF can access the UPSF and use the IPTV profile to personalize the user experience.

ETSI TISPAN specification for the IMS-based IPTV system and the NGN integrated IPTV system can be found in [12] and [13], respectively.

### 3.3 Unmanaged and hybrid scenarios

The specifications of Open IPTV Forum are partially based on TISPAN R2 specification for managed IPTV networks, but also specified IPTV for a non-managed networks (non-managed means without guarantee of QoS and Over Open Top – over Internet). The Open IPTV Forum focuses on standardizing the user-to-network interface (UNI) both for managed and non-managed network with their NGN based architecture [9].

The Hybrid broadcast broadband TV forum (HbbTV) specified [14] the HbbTV platform (in terminals) that combines a profile of the Open IPTV Forum specifications with a profile of the DVB specification for signalling and carriage of interactive applications and services in Hybrid Broadcast Broadband environments (broadcasted content delivered e.g., over DVB-T or DVB-S additional features or application provided via broadband Internet connection). In addition, HbbTV defines supported media formats, minimum terminal capabilities (e.g., browser capabilities based on CE-HTML), and the application life cycle.

Hybrid scenarios may be supported also by TISPAN NGN integrated IPTV platform, while more information about hybrid scenarios in TISPAN is described in [15].

In the Celtic Netlab project, a prototype has been implemented where IMS based IPTV has been integrated with DVB-H [16].

### 3.4 CDN evolution and standardization

As already mentioned, CDNs moved the content closer to the end-user, ensuring less delay. A CDN permits the optimization of the network use through a distribution of the content delivery servers in the

physical network, and the optimization of the storage resources through a popularity-based distribution of the content on the servers. Server and storage architecture is critical to a successful IPTV deployment as both the linear and on-demand content needs to be addressed. Telcos must occupy integrated storage systems that can be upgraded with additional storage, and that can intelligently move content from master head-ends out to the network's edge. This multi-tiered approach can help operators minimize the costs associated with needless redundancy, while providing first-class reliable service to subscribers. This is needed for programming content and advertising assets, which will be required for insertion within the on-demand environment.

From another viewpoint, SPs need to take into account the client's requirements in order to adopt best-in-class scalable solutions. On demand is growing in importance with consumers beginning to expect larger libraries of content as they change from viewing linear content to viewing what they want, when they want. Standard and high-definition on-demand content is also increasingly being used for competitive differentiation as a way to attract and retain subscribers and represents an all-important incremental revenue stream for SPs.

Menaï *et al.* [17] proved that an IPTV service provider could rely on a standard architecture to achieve load balancing and geo-targeted request routing. Moreover, they proved the feasibility of the standards developed in the Open IPTV Forum. According to the Open IPTV Forum [9], a standard CDN consists of three functions:

- *CDN Controller*: It analyses a client's location, media availability and CDN servers' load, and redirects the client's request to the appropriate *Cluster Controller* (CC) or to another CDN controller if no CC can be assigned.
- *Cluster Controller* (CC): It is a server that manages a set media servers placed in the same geographic location. When a CC receives a request from a CDN controller, it performs a second level of request filtering, to decide which *media server* will provide the media content to a client. When the choice is made, the request is assigned to a specific media server called Content Delivery Function (DCF).
- *Content Delivery Function (DCF)/Media Server*: A server where the media content is stored and from which it is delivered. It is the lowest element in the CDN's hierarchy, controlled by the CC and providing media flow directly to the client.

Existing IPTV systems are generally based on proprietary implementations that do not provide interoperability. Recently, many international standard bodies have published, or are developing a series of IPTV related standards. TISPAN defines the CDN architecture and its interconnection with TISPAN

IPTV architectures. ETSI's Media Content Distribution Technical Committee (TC MCD) is running a global study on the various CDN solutions and defines the use cases and requirements for CDN interconnection. It is noteworthy that Maisonneuve *et al.* [18] give an overview of the most significant recent and upcoming IPTV standards [19]. ETSI MCD and IETF CDNI are working on specifying architecture and interfaces for CDN interconnection.

## 4 New Approaches in Multimedia Services and Media Delivery

Telcos may adopt next-generation standards-based on-demand solutions that will offer both the power to deliver a competitive service today, while being scalable enough to migrate to an increasingly on-demand world tomorrow. As the NGN has gained attention for the IP multimedia service delivery platform, IPTV has been recognized as the way to provide the key value-added services. However, IPTV differs from typical NGN-based voice and data services by the fact that it combines three conceptually unfamiliar (until now) components: (a) streamed video; (b) Web services; and (c) NGN-based service control [20]. Another difference is in the sense of the quality-assured service delivery that in the case of IPTV is much stricter for two reasons:

- it is more challenging to meet an end-user's satisfaction in the case of television services.
- there are issues of quality assured provisioning of real-time multimedia services in an environment that is best-effort in its nature.

Volk *et al.* [21] present possible approaches to NGN-based IPTV services assurance from the QoE and QoS viewpoints. Current environments and an overview of the standardization efforts are also given. A proposal for a fully NGN-integrated quality-assured IPTV provisioning model is presented with an associated converged profile structure. The service-aware quality assurance approach is argued. Volk *et al.* investigate further NGN service delivery enhancements for quality-assured provisioning of IPTV services that until now remained unresolved. They present the design of a realistic quality assurance model, establish the associated framework for NGN-based IPTV services delivery, and contribute to discussions and research activities. The evolution of IP-based next-generation networks (NGN) will be largely driven by video service delivery requirements. Ahmad and Begen [22] review trends in the underlying technologies, extrapolating out to the 2015 timeframe, and drawing on the developments in standardization for IPTV, cable networks, and the IP NGN. These evolution trends lead to the notion of a *medianet* as a useful way to think of all of the enabling video and

multimedia technologies. A *medianet* is essentially an IP network that is optimized to deliver video services to any or multiple display devices, and uses any of optical, cable, wireline, and wireless networks for this purpose. Finally, Montpetit *et al.* [23] address the architecture, the value chain and the technical and business challenges of implementing the new connected mobile and social TV experience. To put the architecture into a context, they present a use case of the distributed community Digital Video Recorder (DVR) as an implementation of this vision.

#### 4.1 Enhancements in personalization, context awareness and social networking

The profiling and personalization capability that enables personalization of IPTV services based on user preferences and the user profile is crucial to further enhance user experience and to differentiate Next generation TV systems from usual non-NGN IPTV services. The provider can also use the information about the user behaviour and content consumption to improve interaction with user [24] and provide the targeted applications (e.g., personalized EPG, targeted advertising, content recommendation, etc.) by using the user's current presence state and service/content state to perform service personalization based on user service history, user preferences and content bookmarks or user indication of preference store in user profile.

If the IPTV system wants to personalize its services it has to react precisely on user action/expectation and any changes on actual service state based on identifying related changes in context (relation of user, preferences, actual expectation, changes in environment/situation, location, content, end device, etc). Therefore, the IPTV system has to be context awareness.

We can expect also improvements in the interaction of IPTV applications with social networks information to enable socialized watching TV and interaction among users.

#### 4.2 Advances in user interfaces and interactive apps

One of the most important features that affects user's perception and his/her quality of experience is the user interface and the way (s)he interacts with the IPTV system. The first generation of IPTV systems was using a user interface that was based on simple menus. In such user interfaces, user navigates by using remote control and also the interactivity is limited to "color" or function buttons (e.g., red button for start apps, guide, VOD buttons).

The first interactive applications have been simple information pages with weather forecast, news, etc. As IPTV technologies were evolved, interactive applications also were improved and became more

complex and more personalized. The new generation of STB also support OpenGL [25] specification for 3D graphics that can bring new way of designing IPTV user interfaces. Game consoles already use cameras and sensors to control game by player movements and in future we can expect that similar technologies will be available also for hi-end TV set or STBs.

#### 4.3 Over the top delivery, hybrid, web/tv and adaptive streaming

Advances in TV delivery are coming to unmanaged networks where the public Internet is used to deliver TV or multimedia content over the top (OTT) to end devices with Ethernet connection and browser. The connected TVs are the best example that shows dramatic changes in TV capabilities where a TV set is able to directly connect to a TV manufacture application store and access TV apps. Some of the TV use HbbTV [14] specified browser, while others use proprietary technologies. ITU-T H.760 [26] identified several important technologies that can be used for TV browsers (CEA-2014, DVB-HTML, SVG, etc.).

The W3C (that defines most of the Web standards) has established the Web/TV interest group [27] to consider issues related with the delivery of web content and apps to TV. HTML-5 [28] is expected as a further standard not just for Internet browser, but also for other devices like mobile or TV browsers.

The main issue for over the top content delivery is to assure that content will be not negatively affected with changeable condition of Internet connectivity (i.e., changed bandwidth, delay, packet loss, etc.). Such conditions can be overcome with some of the new developed technologies like CDN, packet retransmission, adaptive streaming, etc.

In the future, standardization of adaptive streaming may have significant impact on the availability of OTT on multiple devices. MPEG DASH (Dynamic Adaptive Streaming over HTTP) was finalized as ISO Standard (ISO/IEC 23009-1) [29]. We can expect that MPEG DASH has the potential to replace in the future existing proprietary technologies like Apple HTTP Live Streaming (HLS), Microsoft Smooth Streaming, Adobe Dynamic Streaming. The global application of unified adaptive streaming standard could have a dramatic impact on significant growth of OTT delivery since in this case content providers publishers can produce one set of files that play on all DASH-compatible devices.

The combination of a CDN infrastructure and the adoption of open DRM (or DRM interoperability technologies for right authentication and licence systems like *UltraViolet* [30]) demonstrate a lot of potential to these new ways of delivering video content to end-users.

#### 4.4 Hybrid P2P/Multicast media delivery based on popularity

Many existing commercial IPTV deployments provide a high-quality and reliable service by using dedicated network infrastructures, commonly referred to as *walled-gardens*. Due to their limited external access and custom capabilities, the streaming of the TV channels from the IPTV head-end server to the customers is done using IP multicast. Although the walled-garden design offers the best path toward a high-quality IPTV service, it has a number of drawbacks, especially in the context of continuously evolving digital services.

First, the approach comes at the cost of decreased flexibility in terms of supporting third-party providers for both economically (i.e., how to price multicast) and technical reasons. In addition, IP multicast does not scale easily with large number of TV channels. This effect is further amplified by newer service options, such as near-video-on-demand (NVoD) that delivers time-shifted programs multiple times. Finally, hypothetical future services introducing user-generated live video content simply do not make IP multicast an option due to the limited number of multicast addresses.

In order to mitigate these issues, Bikfalvi *et al.* [31] propose an alternative solution: using peer-to-peer streaming techniques between customer's IPTV set-top boxes to forward the unpopular TV channels, or to generalize, live video content. Towards this end, their study analyzes the efficiency of the peer-to-peer solution from three different perspectives: (a) overall bandwidth utilization; (b) video quality; and (c) scalability properties. Using extensive simulations, the findings from their study suggest that peer-to-peer offers a viable alternative for a selection of unpopular TV channels. While for low-popularity channels, the bandwidth utilization is similar to the IP multicast approach, the video quality approaches the multicast implementation.

#### 4.5 IETF peer to peer streaming

Peer to peer streaming protocol (PPSP) is an IETF working group with the main goal of standardizing the signalling and control in P2P streaming systems, for exchanging media content. The working group considers two types of nodes in the P2P system: *peers* and *trackers*. Peers are fixed and mobile terminals that exchange streaming media. Trackers are well-known nodes that record information about media content and peers, making it available to other peers. The PPSP working group is currently working on the draft specifications of a tracker protocol [32] (i.e., a control protocol between trackers and peers) and a peer protocol [33] (i.e., a control protocol between peers).

## 5 Open issues

Some aspects that have been identified [34] as future topics and open issues for NGN based IPTV are the following ones:

- Evolution of NGN based IPTV in the context of Future Networks requirements (re-design of Internet Architecture discussed in scientific community as Future Internet).
- New IPTV services (e.g., enhanced Release 3 services, TV communities, TV commerce, multi-screen approach, public interest services, fully personalized IPTV, etc.).
- Support for new media (Ultra-HD, 3D Content, Virtual Realities, Networking/Social Media).
- Hybrid IPTV models (partial delivery of TISPAN IPTV services over "non-TISPAN" networks e.g., DVB-H/T/S/C/SH, OMA BCAST, 3GPP MBMS/PSS, DOCSIS3.0, unmanaged networks/over-the-top).
- IPTV interconnections (roaming support, interconnection with Media Content Delivery/Content Providers/Media Sources) and integration with Content Delivery Networks, Peer to Peer.
- Home network support for managed/unmanaged models, integration of IPTV services with future home networks (smart homes services, metering, near field communication, etc.)
- Convergence of end devices (converged end devices for IMS/non-IMS IPTV or hybrid models).
- Enhanced IPTV security (Service & Content Protection in converged and open environment, content mobility).
- IPTV management (content distribution management, interconnection aspects).
- Interoperability issues, consolidation in standardization of NGN based IPTV architecture.

## 6 Conclusions

The evolution of IPTV architecture and services will depend on the acceptance of the NGN based IPTV concept by operators and vendors. However, satisfied end-users (as paying subscribers) will play a crucial role to any commercially successful service. Therefore, not only the used technology but also definitely simple usability and a rich set of IPTV services/content will be dominant on the market in the future. Such ideal IPTV services/content will constitute perfect solutions for satisfying the higher expectation of user by adopting personalization and context awareness capabilities. In this paper, we have provided an overview about current NGN based IPTV standards and several new technologies and developments that can impact on content services for the next years.

## References

- [1] Broadband Forum, (2011), *Q1 2011 statistics*. Available at: <http://www.broadband-forum.org/news/download/pressreleases/2011/Q1Stats.pdf>
- [2] A. Harris, *Enabling IPTV: What carriers need to know to succeed*, White Paper, May 2005, IDC Analyze the Future Series Report. Available at: [http://www.emc.com/analyst/pdf/IDC\\_IPTV\\_WhitePaper\\_Jun\\_9\\_05.pdf](http://www.emc.com/analyst/pdf/IDC_IPTV_WhitePaper_Jun_9_05.pdf)
- [3] S. O. Hwang, "Content and service protection for IPTV". *IEEE Transactions on Broadcasting*, 55(2): 425-436 (2009).
- [4] G. Pallis and A. Vakali, "Insight and perspectives for content delivery networks". *Communications of the ACM*, 49(1): 101–106. (2006) ACM Press, NY, USA.
- [5] A. Vakali and G. Pallis, "Content delivery networks: status and trends". *IEEE Internet Computing*, 7(6): 68–74 (2003).
- [6] T. Plagemann, V. Goebel, A. Mauthe, L. Mathy, T. Turletti and G. Urvoy-Keller, "From content distribution to content networks—issues and challenges". *Computer Communications*, 29(5): 551–562 (2006).
- [7] M. Day, B. Cain, G. Tomlinson, and P. Rzewski, *A model for content internetworking (CDI)*. Internet Engineering Task Force RFC 3466, (2003).
- [8] ITU-T, Telecommunication Standardization Sector of ITU; Series Y: Global Information Infrastructure, Internet Protocol Aspects and Next-Generation Networks; IPTV functional architecture, Recommendation ITU-T Y.1910 (Sep 2008).
- [9] Open IPTV Forum (2011), *Functional Architecture V2.1* Available at: <http://www.openiptvforum.org/>
- [10] Alliance for Telecommunications Industry Solutions (ATIS) IPTV High Level Architecture Standard (ATIS-0800007), ATIS IPTV Interoperability Forum (IIF), 2007.
- [11] ETSI, Telecommunications and Internet converged Services and Protocols for Advanced Networking (TISPAN); Service Layer Requirements to integrate NGN Services and IPTV, ETSI TS 181 016 v3.3.1 (Jul 2009).
- [12] ETSI, Telecommunications and Internet converged Services and Protocols for Advanced Networking (TISPAN); IPTV Architecture; IPTV functions supported by the IMS subsystem, ETSI TS 182 027 v3.5.1 (Mar 2011).
- [13] ETSI, Telecommunications and Internet converged Services and Protocols for Advanced Networking (TISPAN); NGN integrated IPTV subsystem Architecture, ETSI TS 182 028 v3.5.1 (Feb. 2011).
- [14] ETSI, ETSI TS 102 796 V1.1.1 (2010-06), Technical Specification, Hybrid Broadcast Broadband TV, 2010.
- [15] ETSI, ETSI TR 182 030 V3.1.1 (2011-05) TISPAN; NGN based IPTV mapping or interconnect between IPTV systems, 2011.
- [16] E. Mikoczy, S. Schumann, P. Podhradsky, T. Koski, and M. Heinikangas, "Hybrid IPTV Services with IMS: Integration of IMS based IPTV with Broadcast and Unicast Mobile TV Services using DVB-H". *5th International Conference on Next Generation Mobile Applications, Services and Technologies, NGMAST 2011*, Cardiff, United Kingdom, September 14-16, 2011. IEEE 2011, ISBN 978-1-4577-1080-3, pp.76-81
- [17] M.F. Menai, F. Fieau, A. Souk and S. Jaworski, "Demonstration of standard IPTV content delivery network architecture interfaces: prototype of standardized IPTV unicast content delivery server selection mechanisms". *Proceedings of the 6th IEEE Conference on Consumer Communications and Networking Conference*, IEEE Press Piscataway, NJ, USA, (2009).
- [18] J. Maisonneuve, M. Deschanel, J. Heiles, W. Li, H. Liu, R. Sharpe and Y. Wu, "An overview of IPTV standards development". *IEEE Transactions on Broadcasting*, 55(2): 315-328 (2009).
- [19] Draft ETSI TR 102 688-9 V.0.6.1 (2011-09), Media and Content Distribution, MCD Framework, Part 9: Content Delivery Infrastructures (2011).
- [20] O' Driscoll Gerard, (2008), *Next Generation IPTV Services and Technologies*, Wiley, Canada.
- [21] M. Volk, J. Guna, A. Kos and J. Bester, "Quality-assured provisioning of IPTV services within the NGN environment". *IEEE Communications Magazine*, 46(5): 118-126 (2008).
- [22] K. Ahmad and A. Begen, "IPTV and video networks in the 2015 timeframe: The evolution to media nets". *IEEE Communications Magazine*, 47(12): 68-74 (2009).
- [23] M.-J. Montpetit, N. Klym and T. Mirlacher, "The future of IPTV: Connected, mobile, personal and social". *Multimedia Tools and Applications*, 53(3): 519-532 (2011).

- [24] E. Mikóczy, S. Schumann, and P. Podhradsky, “Personalization of internet protocol television (IPTV) services in next-generation networks (NGN) architectures”. In *Proceedings of the 8<sup>th</sup> International Conference on Advances in Mobile Computing and Multimedia (MoMM '10)*. ACM, New York, USA, (2010) pp. 366-369.
- [25] The OpenGL Graphics System: A Specification.
- [26] ITU-T Recommendation H.760: "Overview of multimedia application frameworks for IPTV services".
- [27] W3C, Web/TV interest group <http://www.w3.org/2011/webtv/>
- [28] W3C, HTML5 - A vocabulary and associated APIs for HTML and XHTML, W3C Working Draft 25 May 2011.
- [29] MPEG, Dynamic Adaptive Streaming over HTTP, MPEG-DASH, ISO/IEC 23009-1, 2011
- [30] DECE, *Ultra Violet Alliance*, Available at: <http://www.uvvu.com/>
- [31] A. Bikfalvi, J. Garcia-Reinoso, I. Vidal, F. Valera, and A. Azcorra. “P2P vs. IP multicast: comparing approaches to IPTV streaming based on TV channel popularity”. *Computer Networks*, 55(6): 1310–1325, April 2011.
- [32] R.S. Cruz, M.S. Nunes, Y. Gu, J. Xia, D.A. Bryan, J.P. Taveira and D. Lingli. *PPSP Tracker Protocol*, draft-gu-ppsp-tracker-protocol-06, October 2011 (expires: May 2012).
- [33] A. Bakker. *Peer-to-Peer Streaming Protocol (PPSP)*, draft-ietf-ppsp-peer-protocol-00.txt, December 2011 (expires: June 2012).
- [34] E. Mikóczy, “Discussion on future topics” as ETSI TISPAN contribution 22bTD113, ETSI TISPAN 22bis meeting, 2.11.-6.11.2009, Sophia Antipolis, France, 2009.

# IPTV Services Personalization Using Context-Awareness

Songbo Song, Hassnaa Moustafa  
 France Telecom - Orange, France  
 E-mail: {songbo.song, hassnaa.moustafa}@orange-ftgroup.com

Hossam Afifi  
 Telecom & Management South Paris, France  
 E-mail: hossam.afifi@it-sudparis.eu

**Keywords:** NGN, IPTV, user-centric, personalization, QoE

**Received:** September 12, 2011

*The advances in IPTV (Internet Protocol Television) technology enable a new user-centric and interactive TV model, in which context-awareness is promising in making the user's interaction with the TV dynamic and transparent. Our research interest is how to achieve user-centric personalized IPTV services applying context-awareness. In this paper, we present a solution for IPTV services personalization introducing context-awareness on top of the IPTV architecture on one hand to gather different information about the user and his environment and on the other hand allowing each user to be distinguished to the system in a unique manner. Consequently, IPTV services personalization is achieved in a real-time manner for each user. We implemented the proposed solution on top of an IPTV platform considering the NGN IPTV architecture as a proof of concept and as a means to evaluate the performance.*

*Povzetek: Članek opisuje prilagajanje televizije IP posamičnim uporabnikom z uporabo konteksta.*

## 1 Introduction

IPTV (Internet Protocol TV) presents a revolution in digital TV. In which digital television services are delivered to users using Internet Protocol (IP) over a broadband connection. The ETSI/TISPAN [1] provides the Next Generation Network (NGN) architecture for IPTV as well as an interactive platform for IPTV services. However, IPTV services personalization is still in its infancy, where the consideration of the context of the user and his environment (devices and network) and the distinguishing of each user in a unique manner still presents a challenge.

IPTV services personalization is beneficial for users and for service and network providers. For users, more adaptive content could be provided as for example targeted advertisement, movie recommendation, provision of a personalized Electronic Program Guide (EPG). The format of the content could also be modified according to the users' devices the network conditions which in turn allows for better Quality of Experience (QoE). IPTV services personalization is promising for services providers in promoting new services and opening new business and allows network operators to make better utilization of network resources through adapting the delivered content according to the available bandwidth.

Context-awareness is promising in allowing services personalization through considering in real-time the context of the user and his environment (devices and network) as well as the context of the service itself [2]. Through context-awareness, users can transparently

interact with the IPTV system (users will no longer be required to give explicit instructions at every step while watching TV). In this paper we present a solution for IPTV services personalization that considers NGN IPTV architecture while employing context-awareness, allowing access personalization for each user and triggering service adaptation. We carried out the implementation of this solution on top of an NGN IPTV platform to verify its correct functioning and evaluate its performance through different performance metrics.

The remainder of this paper is organized as follows: Section 2 gives an overview on the related work. Section 3 presents the solution. Section 4 describes our implementation platform and Section 5 presents the performance analysis. We conclude the paper in Section 6 and present the future work.

## 2 Related Work

IPTV services personalization attracted the attention of the IPTV standards. Some services personalization are defined in TISPAN [1]: 1) Personal video recorder (PVR): is an end-user-controlled electronic device service that records linear TV and stores it in a digital storage facility, either in standalone Set-Top-Boxes (STB) or in the network. 2) Personal Broadcast Service: is a service providing the end-user with a way to advertise personal content (possibly including scheduling information) description so that other users can access such content. 3) Targeted Advertising: is a commercial

advertising or public promotion of goods, services, companies and ideas, usually personalized according to the end-user's preferences. 4) Content Recommendation Service: recommends the contents to users according to the user's preferences. Although these services bring new experience to users, without considering the general context information the service personalization is limited.

Several solutions for context-aware TV systems have been proposed for services personalization. [3] proposes a context-aware based personalized content recommendation solution that provides a personalized Electronic Program Guide (EPG) applying a client-server approach, where the client part is responsible for acquiring and managing the context information, and forwarding it to the server (residing in the network operator/service provider side) which collects the TV programs and determines the most appropriate content for the user. A recommendation manager in the client side notifies the user about the content recommended for him according to the acquired context, including the user context information (user identity, user preference, time) and the content context information (content description). This work does not consider the network context and does not describe the integration with the whole IPTV architecture, however focuses on the Set-Top-Box (STB) as the client and a server in the network operator/service provider side. In addition, services personalization is limited to the content recommendation without considering any other content adaptation means.

A personalization application for context-aware real-time selection and insertion of advertisements into live broadcast digital TV stream is proposed in [4]. This work is based on the aggregation of advertisement information (i.e. advertised products type, advertised products information) and its association with the current user context (identity, activity, agenda, past view) in order to determine the most appropriate advertisement to be delivered. This solution is implemented in STBs and includes modules for context acquisition, context management and storage, advertisement acquisition and storage and advertisement insertion. This solution does not consider the devices and network contexts and does not describe the integration with the whole IPTV architecture, however focusing on the STB side in the user domain.

In [5], we proposed integrating a context-awareness system on top of IPTV/IMS (IP Multimedia Subsystem) architecture aiming to offer personalized IPTV services. This solution relies on the core IMS architecture for transferring the different context information to a context-aware server. The main limitation of this solution is its dependency on IMS which necessitates employing the SIP (Session Initiation Protocol) protocol and using SIP-based equipments, which in turn limits the interoperability of different IPTV systems belonging to different operators and which requires also a complete renewal of the existing IPTV architecture (currently commercialized) which does not employ IMS. Furthermore, the dependency on the SIP protocol limits the possible integration of IPTV services with other rich internet applications (which is an important NGN trend)

and hence presents a shortcoming. Consequently, we aim by the solution presented in this paper to increase the integration possibilities with web services in the future and to ease the interoperability with the current IPTV systems. So we advocate the use of HTTP protocol, where the presented solution introduces context-awareness on top of NGN IPTV non-IMS architecture. In addition, a mechanism for personalized identification allowing to distinguish each user in a unique manner and a mechanism for content adaptation allowing the customization of the EPG (Electronic Program Guide) and the personalized recommendation are proposed.

### 3 Solution Description

#### 3.1 Overview of the solution

The presented solution in this paper relies on a Context-Aware System, which we proposed in [5] while extending it and integrating it in the NGN IPTV non-IMS architecture. This Context-Aware System has been also implemented in this paper. The necessary communication between the Context-Aware System and the other architectural entities in the core network and IPTV service platform is achieved through HTTP and DIAMETER protocols while extending them to allow for the transmission of the acquired context information in real-time. In addition, we propose and implement in this paper a mechanism to distinguish each user in a separate manner through providing a personal identity for each user. This personal identity is used as a part of the user context. Finally, the presented solution provides and implements a mechanism for IPTV personalization based on each distinguished user and on the different context information acquired in real-time. We considered the following personalization means: the customization of the EPG to match the users' preferences, recommending the content best matching the users' preferences, and adapting the content according to the device used by each user.

#### 3.2 Context information for the IPTV services

The International Telecommunication Union Telecommunication Standardization Sector (ITU-T) defines four main functional domains involved in the provision of an IPTV service [6]: a) Content Provider: owns or sells the content to be streamed to the Customer; b) Service Provider: provides The IPTV service; c) Network Provider: provides the connection between the Service Provider and the Customer; d) Customer: purchases and consumes the IPTV service.

From the IPTV function domains, four types of contexts can be defined for IPTV services:

*i) User Context:* includes information about the user, which could be static information, dynamic information and inferred information. Static information describes the user's personal information which is stored in the database (ex, name, age, sex, and input preference). Dynamic information is dynamically captured by sensors



or by other services (ex, user's location, agenda, and usage history). Inferred information is high-level information, which is inferred by other information (ex, user's action "user is going to the bed" is inferred by the changed location").

*ii) Device/Terminal Context:* includes information about the devices (terminals), which could be the device identity, status (turn on or off, volume), device capacity, and the device proximity with respect to the user.

*iii) Network Context:* represents the characteristics of the access link being used for accessing TV content. Network context information includes: a) Access network type: Information about available access networks enables selecting the most appropriate network; b) Available link bandwidth: this information is used by the IPTV system to select appropriate content format for example SD or HD, c) QoS information: this information is used to monitor the state of the network.

*iv) Service Context:* includes information about the service, which could be the content description, language, and format description.

### 3.3 Context-aware system

We follow a hybrid approach in the design of the Context-Aware System including a centralized server (Context-Aware Server "CAS") and some distributed entities to allow the acquisition of context information from the different domains along the IPTV chain (user, network, service and content domains) while keeping the context information centralized and updated in real-time in the CAS enabling its sharing (and the sharing of users' profiles) by different access networks belonging to the same or different operator(s).

The CAS is a central entity residing in the operator network and includes four modules: i) A Context-Aware Management (CAM) module, gathering the context information from the user, the application server and the network and deriving higher level context information through context inference. ii) A Context Database (CDB) module, storing the gathered and inferred context information and providing the query interface to the Service Trigger (ST) module iii) A Service Trigger (ST) module, triggering the personalized service for the user according to the different contexts stored in the CDB. iv) A Privacy Protection (PP) module, verifying the different privacy levels for the user context information before triggering the personalized service.

In addition, other distributed modules exist to gather different context information in real-time. In the user domain, the User Equipment (UE) includes a Client Context Acquisition (CCA) module and a Local Service Management (LSM) module. The CCA module discovers the context sources in the user sphere and collects the raw context information (related to the user and his devices) for transmission to the CAM module located in the CAS. Sensors are the frequently used context sources which can be present in the user sphere, in the environment or in the device and retrieve context information from them. While, the LSM module controls and manages the services personalization in local manner

within the user sphere (example, volume change, session transfer from terminal to terminal, etc). In the service domain, the Service Context Acquisition (SCA) module collects the service context information and transmits it to the CAM, and the Media Delivery Context Acquisition (MDCA) module monitors the content delivery and dynamically acquires the network context information during the content delivery and sends it to the CAM. In the network domain, the Network Context Acquisition (NCA) module acquires the network context (mainly the bandwidth information) during the session initiation and transmits it to the CAM.

### 3.4 NGN IPTV-non IMS architecture with context awareness

This subsection shows the integration of the Context-aware Server in the NGN IPTV non-IMS [1] architecture together with the different protocols (and the protocols extensions) used for the communication between the different architectural entities. During the different communications for context information transfer, we use the RPID (Rich Presence Extensions to the Presence Information Data Format) [11] to present the context information while enhancing it to include more attributes for context information. Figure 1 illustrates the context-aware NGN IPTV non-IMS architecture.

The NGN IPTV architecture standard includes the following functions: Service Discovery and Selection (SD&S), for providing the service attachment and service selection, Customer Facing IPTV Application (CFIA), for IPTV service authorization and provisioning to the user, User Profile Server Function (UPSF), for storing user's related information mainly for authentication and access control, IPTV Control Function (IPTV-C), for the selection and management of the media function; and Media Function (MF), for controlling and delivering the media flows to the User Equipment (UE).

In the service plane, the SCA (Service Context Acquisition) module is integrated in the SD&S IPTV function to dynamically acquire the service context information making use of the Electronic Program Guide (EPG) received by the SD&S from the content provider which includes content and media description. The MDCA (Media Delivery Context Acquisition) module is integrated in the MF to dynamically acquire the network context information during a media session through gathering the network information statistics (mainly on packet loss, jitter, round-trip delay) delivered by the Real Time Transport Control Protocol (RTCP) [7]. In the network plane, the NCA (Network Context Acquisition) module is integrated in the Resource and Admission Control Sub-System (RACS) [8] extending the resource reservation procedure during the session initiation to collect the initial network context information (available link bandwidth). In the user plane, we use the UPSF to store the static user context information including user's personal information ("age, gender ..."), subscribed services and preferences, and the CCA (Client Context Acquisition) and LSM (Local Service Management) modules extend the UE (User Equipment) to acquire

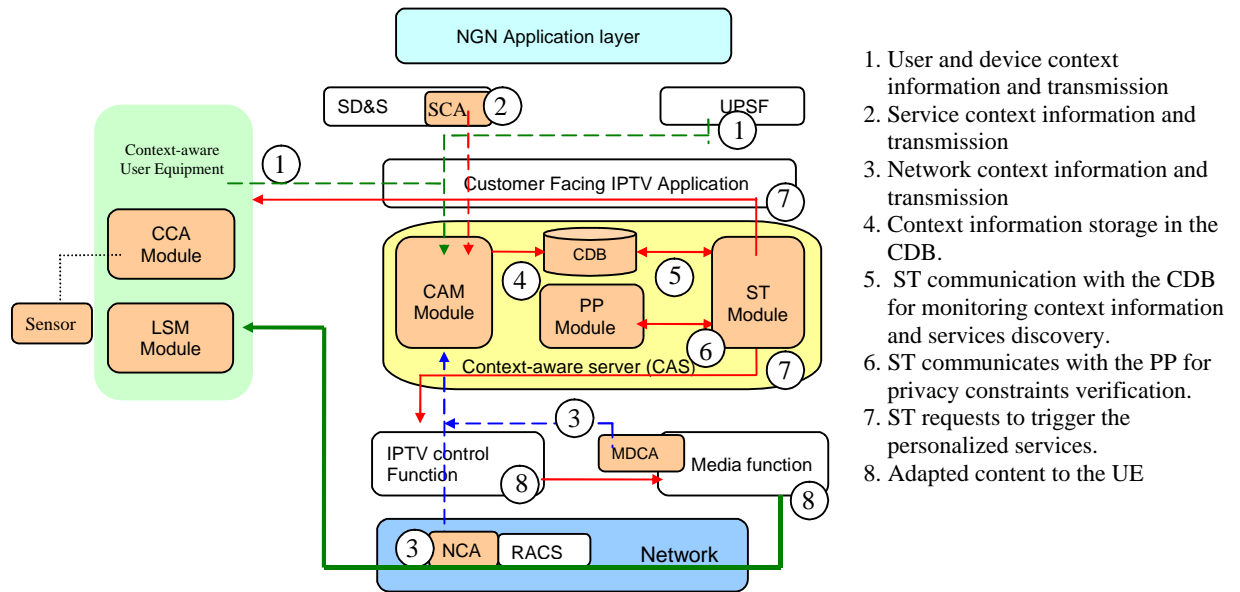


Figure 1: Context-Aware NGN IPTV non-IMS Architecture.

the dynamic context information of the user and his surrounding devices.

After each acquisition of the different context information (related to the user, devices, network and service), the CAM (Context-Aware Management) in the CAS (Context-Aware Server) infers the collected information and derives higher level context information which is stored in the CDB (Context Data Base). The ST (Service Trigger) module continuously communicates with the CDB module to monitor the context information, according to which the ST discovers the need for personalizing the established services or setting up a new services. Before triggering the service ST module communicates with the PP (Privacy Protection) module to verify if the corresponding service can fully use the existing context information. If there is no privacy constraint, the ST module activates the personalized services.

The communication and exchange procedures within the context-aware NGN IPTV non-IMS architecture take place as follows:

1) *Contextual Service initiation*: Figure 2 illustrates this procedure which is used to transfer the user's static profiles from the UPSF.

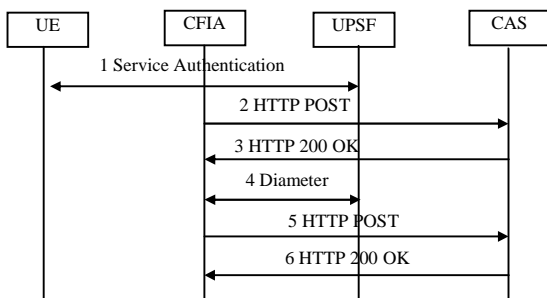


Figure 2: Contextual service initiation procedure.

After the classical IPTV service authentication (message 1), the CFIA registers the IPTV UE to the CAS service on behalf of the UE using the HTTP POST messages (messages 2-3), then downloads from the user's profile

his static context information using Diameter protocol [9] (message 4), extended to include a User-Static-Context Attribute. The CFIA then transmits the user static context information to the CAS through HTTP POST message (message 5). When the CAS receives the context information, it sends a 200 OK response message to the CFIA (message 6) for acknowledgement.

2) *Dynamic acquisition of context information of the user/device*: This procedure is proposed allowing the CCA module in the UE to transfer to the CAS and continuously update the user/device dynamic context information. Figure 3 illustrates the user and device dynamic context information transmission. We use the HTTP POST message to convey the context information.

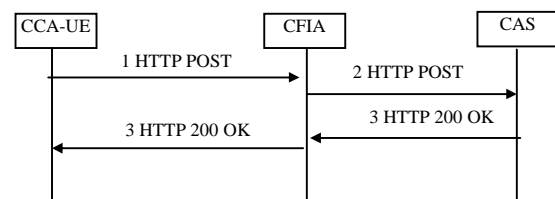


Figure 3: User/Device Dynamic Context Information Transmission.

3) *Dynamic acquisition of context information on service*: This procedure is proposed and is similar to the previously described one allowing the dynamic service context information transmission from the SCA to the CAS. The HTTP POST message is used to transmit the context information containing the service context.

4) *Network Context Information acquisition during the Session Initiation*: This procedure concerns the network context information transmission during the session initiation through extending the classical resource reservation process. In this latter, the IPTV-C receiving the service request sends a Diameter protocol AA-Request message to the Resource and Admission Control Sub-System (RACS) for the resource reservation. Based on the available resources (bandwidth information), the RACS will decide whether to do or not a resource

reservation for the service. An AA-answer message is sent by the RACS to the IPTV-C for informing the latter the results of the resource reservation (successful resource reservation or not). We extend this process in order to send the resource information (bandwidth) to the context-aware server. The bandwidth information is used by the IPTV system to select appropriate content format for example SD or HD. As illustrated in Figure 4, upon the reception of the service request (message 1), the IPTV-C sends a Diameter protocol AA-Request to RACS. Then the NCA generates a Context AA-Answer (CAA-Answer) message extending the AA-Answer message through adding a Network-Information Attribute Value Pair to include the bandwidth information. At last the IPTV-C forwards the received bandwidth information to the CAS using HTTP POST (message 4-5).

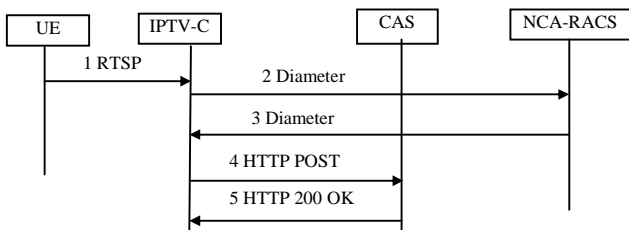


Figure 4: Network context information acquisition during the .Session Initiation

5) *Dynamic acquisition of network context information.* This procedure allows the MDCA to dynamically transmit the network context information related to the media session to the CAS, as illustrated in Figure 5. During the media session, the MDCA module acquires the network context information from the RTPC protocol statistic reports mainly indicating the jitter, packet loss and round-trip delay (message 1) and transmits to the CAS using HTTP POST (messages 2-3)

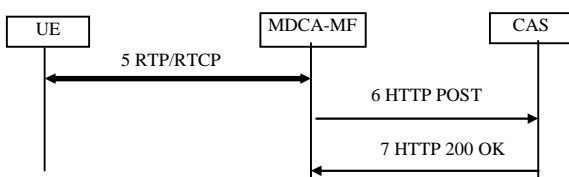


Figure 5: Network dynamic context information acquisition.

### 3.5 User's personal identification

Personal identification is an important step to achieve the service personalization and also the privacy protection. We propose the usage of a personal identity for user's access and service authentication, which allows distinguishing each user in a personal manner. Following user's authentication, the personal identity is stored in the CAS as a part of the user's context information. According to this personal identity, the CAS collects the user's static context information from the UPSF and also other dynamic context information from the user's environment (for example user's proximity to the device). Based on the collected information, the personalized service is provided to the user. The radio-identification

(RFID) [12] is a candidate technology to provide a personal identity and which we used in the implementation. However any other technology as NFC (Near Field Communication) or Bluetooth could be applied.

In order to personalize each user's access, each user holds a unique RFID tag and identifies himself to the system. In addition, each user device (for example TV or PC Screen) is connected to an RFID reader indicating the identity of the device which is associated to the device location. When the user comes next to the device, the RFID reader of this device reads the user's identity and sends it to the STB together with the device identity (presented by the identity of the RFID reader). The STB sends the user and device context information to the CAS, where the user location can be deduced with reference to the users' device location. When the user watching the TV through the same device changes (or when the same user changes the device during mobility "moves from one room to another for example"), the personalized access following the same previous approach takes place.

### 3.6 IPTV Content Personalization

Within the ETSI/TISPAN Standard, IPTV content personalization is carried out through the recommendation which may take the form of a text message (notification) from CFIA (Customer Facing IPTV Application) or video recommendation streamed from MDF (Media Delivery Function) [10]. Within the Content Recommendation Service, two basic functions are proposed:

- 1) Aggregation of metadata for recommendations:
  - (a) user's profile, preferences, consumed history, etc.
  - (b) Asset metadata for Content on Demand(CoD) and linear TV.
- 2) Generation of recommendations upon request of external triggers: for example a user sends service request or changes in the user's presence state.

We notice that this solution considers only the user's preference and consumed history as the criteria while the general context information is not considered (like device capacity, network states, location, time, user's activity...etc). However, the general context information takes an important role for the recommendation service because the user's preference on IPTV content is not fixed and it changes depending on certain context such as time, location, activity, etc. Furthermore, certain contents are accessible based on some conditions. For example only the device which supports the High Definition (HD) could display the HD film. If a user chooses to watch TV through a terminal which does not support HD, the recommendation system should not propose the HD content for him.

We propose a mechanism for content personalization to provide the content best matching the user. Our proposed mechanism selects the content for each user based on the preferences and context and adapts the content according to the other context information for example the used device characteristics

(supported resolution), and the content characteristics (High Definition “HD” or Standard Definition “SD” for instance). The content personalization mechanism employs the context information stored in the CAS following a Key-Value based model, in which each context type has a corresponded value. For example, each user can have several preferences, and each preference has a value (“preferred film type” of the user is “action”; “preferred sport” of the user is “football”). The Key-Value based model simplifies the content selection for each user.

The mechanism functions as follows: i) after the user identification and the context information collection, the ST module selects the user’s preferred contents through comparing the content context information and the user’s preference. Then it filters the selected contents according to the other context information (for example the resolution of the device through which the user watches TV, the content types “HD or SD”) and then generates a personalized EPG. ii) user-centric recommendation takes place through listing the contents on the EPG in the order of user’s preference, where the ST consults the CDB for the context information on the content and the user and then calculates the similarity between the content context information (content type, authors, etc) and the user’s preference. iii) After generating the personalized EPG, the ST module sends it to the user and the LSM module automatically displays the content of the first program listed in the EPG (however the user has the possibility to manually select another program). iv) The ST module also sends the necessary context information to the Media Function to make the latter adapt content according to the context information (for example change the format of the content according to the resolution of the device).

## 4 Implementation and Platform

We implemented our personalized IPTV system (the context-awareness system, the personal identification and content personalization mechanisms) and we integrated the implemented modules on top of NGN IPTV non-IMS architecture. We achieved a proof of concept through the correct functioning of the proposed system and we carried out a performance evaluation through several performance metrics that are presented below.

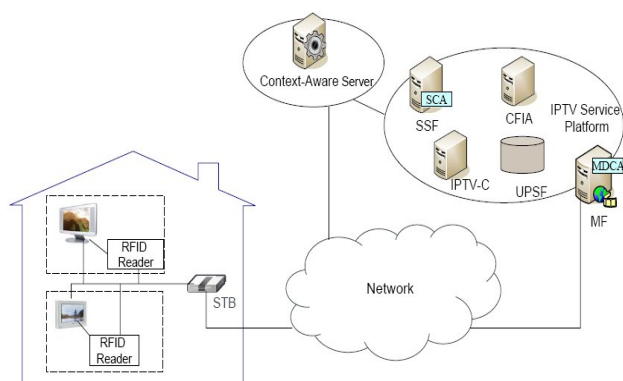


Figure 6: Implementation platform architecture.

Figure 6 illustrates the architecture of our implementation platform. The context-aware server (CAS) is developed based on a framework for HTTP services development. The NGN IPTV architecture is implemented on 5 machines for respectively the IPTV-C, CFIA, SD&SF, MF IPTV functional entities and the UPSF, while integrating the module implemented for context-awareness. At the end-user side, we implemented the STB containing three functional modules: traditional STB module (dealing with the authentication, session negotiation and service reception, etc.), Client Context Acquisition (CCA) and Local Service Management (LSM) modules of the context-aware system.

For the personal identification of each user and determining his proximity to the device, we employ the RFID technology to implement the solution proposed in Section 3.4. This technology is an implementation choice and not the absolute technology for our proposed mechanism.

## 5 System Evaluation and Performance Analysis

The platform with the implemented modules was firstly tested from the functional view and we observed the right functioning of the whole system with TV Live and VoD (Video on Demand) IPTV services. Then we evaluated the performance of the proposed solution for TV Live and VoD IPTV services and compared to the traditional IPTV case without personalization.

### 5.1 Performance metrics used

This subsection defines the metrics that we use in our performance analysis, which are: i) the delay of the personalized content selection (DPS), ii) the delay of the service initiation (DSI) and iii) the EPG Browsing time (EBT).

1) *Delay of personalized content selection (DPS)*: defined as the consumed time from sending the first context update request from user until receiving the personalized EPG from the CAS. This delay includes the new context information transmission, treatment and the personalized EPG generation. The DPS reflects the performance of the CAS. Since there could be large number of users request the personalized content at the same time, we analyze the increase of the DPS with the increase of the end users.

2) *Delay of the service initiation (DSI)*: defined as the consumed time from the start of the STB to the reception of the IPTV service (when the video starts playing). In traditional IPTV case, this delay includes the delay of user’s authentication, session negotiation and video display. In our proposed solution, besides the mentioned delay, DSI also includes the delay of personalized content selection (DPS).

3) *EPG Browsing Time (EBT)*: measures the user’s quality of experience (QoE) level in terms of how quickly a user can find his preferred program from the displayed EPG. EBT presents the consumed time from

the display of the EPG until finding the user's preferred content.

To derive the EBT for the traditional EPG browsing, we consider that there are  $n$  programs on the EPG and that the probability of the preferred program ( $i$ ) selection by the user is the same for all the programs (probability =  $1/n$ ). We also assume that during the EPG browsing, the user will watch the selected program for a duration ( $t'$ ) before switching to another program. So the expected traditional EBT can be calculated as:

$$E(EBT) = \sum_{i=1}^n \frac{1}{n} (i-1) t' = \frac{n-1}{2} t' \quad (1)$$

We then introduce another important parameter "accuracy probability  $P_a$ ", reflecting how well the personalized EPG programs meet the user's expectation. It is calculated as the ratio of the amount of recommended content in which the user is interested to the amount of recommended content [13].

$$P_a = \frac{\text{recommended} \cap \text{interested}}{\text{recommended}} \quad (2)$$

From the definition of the accuracy probability, the matching of the EPG programs to the user's expectation has the following probabilities: the first program matching probability is ( $P_a$ ); the second program matching probability is  $(1-P_a)P_a$  and the  $m^{\text{th}}$  program matching probability is  $(1-P_a)^{m-1}P_a$ . So the expected browsing time for our solution can be calculated as:

$$E(EBT) = \sum_{i=1}^n (1-p_a)^{(i-1)} p_a (i-1) t' \quad (3)$$

$$= \left( (1-p_a) \frac{1-(1-p_a)^{n-1}}{p_a} + (1-n)(1-p_a)^n \right) t'$$

### 5.2 Obtained results and analysis

From the Figure 7, we observe an increase of the delay of personalized content selection (DPS) when the number of end users increases. However the increase rate is small and slows down with the increase of users. When the number of user increases from the 50 to 400, the increase rate of the DPS is about 3%, and when the number of user increase from 400 to 1000, the increase rate of the DPS is about 0.5%.

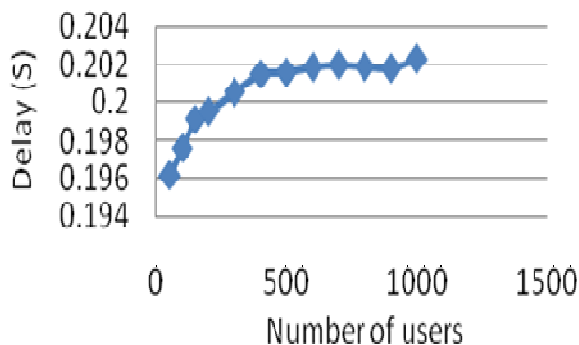


Figure 7: Delay of the personalized content selection.

Figure 8 presents the delay of the service initiation for both traditional IPTV and personalized IPTV cases. For traditional IPTV case, a delay of 1.75 and 2.12 seconds is observed respectively for the initiation of TV Live and VoD services. While, for the personalized IPTV case, a delay of 2.04 and 2.29 seconds is observed respectively for the initiation of TV Live and VoD services. This increase is mainly due to the consumed time for context information acquisition, treatment and personalization and service selection. Although we do not consider all the functions in the actual IPTV platform, compared with the average delay of service initiation which is about 2.9 seconds [14], the proposed solution does not impact the performance of the service.

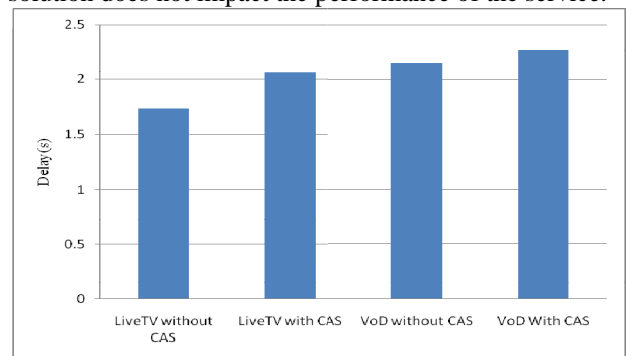


Figure 8: Delay of service initiation.

Figure 9 illustrates the obtained values for the EBT in the presence of various numbers of programs for the TV Live service. In the traditional IPTV case, the EBT linearly increases with the increase in the number of programs indicating poor QoE with the explosion of the number of programs. To calculate the EBT for our solution, we firstly measure the EPG accuracy probability ( $P_a$ ) through verifying the matching of the received personalized EPG to the users in the case of different users with different context information changes. Among several services requests, the  $P_a$  was found to be 0.8 (i.e. about 80% of the recommended contents correspond to users' interest). By substituting this obtained  $P_a$  value in equation (3) together with the change of the value of ( $n$ ) reflecting the number of programs, we obtained the EBT values illustrated in Figure 7 for the personalized TV case. We observe a slight increase at first with the increase in the number of programs then a constant EBT value is shown in spite of the number of programs, thus confirming the gain in terms of QoE with our proposed solution with the increase in the number of programs.

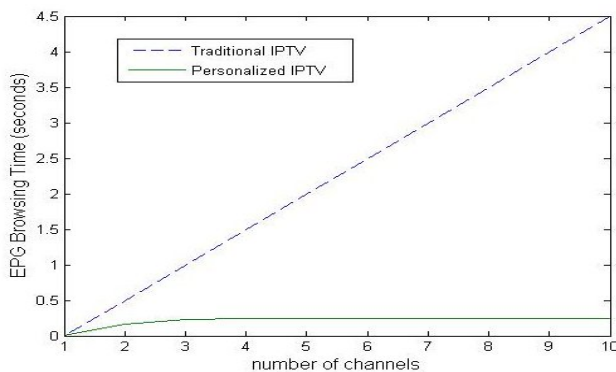


Figure 9: EPG Browsing time.

## 6 Conclusion

In this paper, we present a solution for IPTV services personalization that integrates a context-awareness system in the NGN IPTV non-IMS architecture allowing the consideration of different context information related to the user, his devices, the network and the service itself in real-time. We also provided a mechanism for user's personal identification to the IPTV system allowing distinguishing each user in a unique manner, and a mechanism for content personalization based on each individual user, his context and the context of his environment (devices and networks). We implemented our solution on top of an IPTV platform validating its correct functioning and evaluating its performance for TV Live and VoD IPTV service compared to classical IPTV case (with no personalization). We observed interesting results for our proposed solutions in terms of the personalized content selection and service initiation delays as well as the EPG browsing time. As a future work, we will analyze the performance of the whole system including more performance metrics and more scenarios for the user's interaction with the system and consider group personalization considering the contexts and preference of a group of users watching TV. Furthermore, we will communicate with real-end users to gather their requirements, know their constraints for this newly proposed personalization approach, and test the system ergonomics. We will then consider the end-users' feedbacks in our system.

## References

- [1] ETSI TS 182 028 v3.5.1 (2011-02) Telecommunications and Internet converged Services and Protocols for Advanced Networking

- (TISPAN); NGN integrated IPTV subsystem Architecture.
- [2] S. Song, H. Moustafa, H. Afifi, "Context-Aware IPTV," IFIP/IEEE Management of Multimedia & Mobile Network Service (MMNS 2009).
- [3] Santos Da Silva, F., Alves, L.G.P., Bressan, G. Personal TVware: A Proposal of Architecture to Support the Context-aware Personalized Recommendation of TV Programs. 7th European Conference on Interactive TV and Video (2009).
- [4] Thawni, A., Gopalan, S., Sridhar, V.: Context Aware Personalized Ad Insertion in an Interactive TV Environment. 4th Workshop on Personalization in Future TV (2004)
- [5] Songbo Song, Hassnaa Moustafa, Hossam Afifi, "Personalized TV Service through Employing Context-Awareness in IPTV/IMS Architecture," FMN (2010).
- [6] ITU-T IPTV Focus Group Proceeding 2008
- [7] H. Schulzrinne, S. Casner, R. Frederick, V. Jacobson.: RTP: A Transport Protocol for Real-Time Applications. In: IETF RFC 3550 (2003)
- [8] ETSI ES 282 003 V2.0.0 (2008-05) Resource and Admission Control Sub-System (RACS): Functional Architecture
- [9] ETSI TS 129 229 V10.2.0 (2011-10) Cx and Dx interfaces based on the Diameter protocol; Protocol details.
- [10] ETSI TS 183 064 V3.4.1 (2011-02) Telecommunications and Internet converged Services and Protocols for Advanced Networking (TISPAN); NGN integrated IPTV subsystem stage 3 specification
- [11] Schulzrinne, H., Gurbani, V., Kyzivat, P., Rosenberg, J. RPID Rich Presence Extensions to the Presence Information Data Format (RPID) in IETF RFC 4880 (2006).
- [12] Klaus Finkenzeller, *RFID Handbook*, Wiley, 1999
- [13] Chen, Y., Huang, H., Min, Y. Community-based Program Recommendation for the Next Generation Electronic Program Guide, IEEE Transactions on Consumer Electronics, Vol.55, No.2, pp.707-712.
- [14] Cruz, R.S.; Nunes, M.S.; Menezes, L.; Domingues, J.; SIP Based IPTV Architecture for Heterogeneous Networks, 10th International Conference on telecommunication, 2009

# Privacy Aware Recommender Service using Multi-agent Middleware – An IPTV Network Scenario

Ahmed M. Elmisery and Dmitri Botvich  
Telecommunications Software & Systems Group-TSSG  
Waterford Institute of Technology-WIT, Co. Waterford, Ireland  
E-mail: ahmedmohmed2001@gmail.com

**Keywords:** clustering, IPTV network, recommendation-services, multi-agent, secure multiparty computation

**Received:** October 28, 2011

*IPTV service providers are starting to realize the significant value of recommender services in attracting and satisfying customers as they offer added values e.g. by delivering suitable personalized contents according to customers personal interests in a seamless way, increase content sales and gain competitive advantage over other competitors. However the current implementations of recommender services are mostly centralized combined with collecting data from multiple users that cover personal preferences about different contents they watched or purchased. These profiles are stored at third-party providers that might be operating under different legal jurisdictions related to data privacy laws rather than the ones applied where the service is consumed. From privacy perspective, so far they are all based on either a trusted third party model or on some generalization model. In this work, we address the issue of maintaining users' privacy when using third-party recommender services and introduce a framework for Private Recommender Service (PRS) based on Enhanced Middleware for Collaborative Privacy (EMCP) running at user side. In our framework, PRS uses platform for privacy preferences (P3P) policies for specifying their data usage practices. While EMCP allows the users to use P3P policies exchange language (APPEL) for specifying their privacy preferences for the data extracted from their profiles. Moreover, EMCP executes a two-stage concealment process on the extracted data which utilize trust mechanism to augment the recommendation's accuracy and privacy. In such case, the users have a complete control over the privacy level of their profiles and they can submit their preferences in an obfuscated form without revealing any information about their data, the further computation of recommendation proceeds over the obfuscated data using secure multi-party computation protocol. We also provide an IPTV network scenario and experimentation results. Our results and analysis shows that our two-stage concealment process not only protect the users' privacy, but also can maintain the recommendation accuracy.*

*Povzetek: Članek obravnava priporočanje vsebin uporabnikom televizije IP, ki spoštuje uporabnikovo zasebnost.*

## 1 Introduction

Internet Protocol Television (IPTV) is a video service providing IP broadcasts and video on demand (VOD) over a broadband IP content delivery network (CDN) specialized in video services. The IPTV user has access to myriads of video content spanning IP Broadcast and VOD [1]. In this context, it is difficult for them to find content that matches their preferences from the huge amount of video content available. In order to attract and satisfy these users, IPTV service providers employ recommendation services to increase their revenues and offer added value to their patrons. In the same time, Recommender services can improve the overall performance of the current IPTV network by building up an overlay to increase content availability, prioritization and distribution.

Recommender service offers referrals to users by building up users' profiles (explicit or implicit) based on their past ratings, behaviour, purchase history or demographic information. In the context of this work, a

profile is a list comprises the video contents the user has watched or purchased combined with their meta-data extracted from the content provider (i.e. genres, directors, actors and so on) and the ratings the user gave to these contents. Maintaining the quality of offered referrals and quickly react to problems raised from merging data from different sources requires a lot of expertise, and not all IPTV providers have the ability to construct and interpret recommendation models. Therefore, there is a market for specialized firms on users' profiles storage and analysis. But there are some challenges face this business model, such as security and privacy. Because collected data from users cover personal information about different contents they have watched or purchased and these profiles might be stored at third-party providers that might be operating under different legal jurisdictions related to data privacy laws rather than the ones applied where the service is consumed. This is a serious threat to individual privacy since this data can be used for unsolicited marketing,

government surveillance, profiling of users, misused, furthermore it can be sold by providers when they face bankruptcy. As a matter of fact, users are more likely willing to give more truthful preferences if privacy measurements are provided or if they assured that the data does not leave their personal devices until it is properly desensitised.

The organization for economic co-operation and development (OECD) [2] formulated sets of principles for fair information practice that can be considered as the base for privacy laws. These principles allow the users to control the data they provide for recommender services operating at remote sites, they can be described as follows:

1. Collection limitation: data collection and usage for a recommender service should be limited only to the data it requires to offer appropriate service.
2. Data quality: data should be used only for the relevant purposes for which it is collected.
3. Purpose specification: data controllers should specify up front how they are going to use data and users should be notified up front when a system will use it for any other purpose.
4. Use limitation: data should not be used for purposes other than those disclosed under the purpose specification principle without user consent.
5. Security safeguards: data should be protected with reasonable security safeguards (encryption, secure transmission channels, etc).
6. Openness: the user should be up front notified when the data collection and usage practices started away.
7. Individual participation: users should have the right to insert, update, and erase data in their profiles stored at data controllers.
8. Accountability: data controllers are responsible for complying with the principles mentioned above.

In this work, we present an enhanced middleware for collaborative privacy (*EMCP*) that allows creating reasonable referrals without breaching user privacy. *EMCP* employs a set of mechanisms to allow users to share their data among each other in the network to form a group to attain collaborative privacy. The users' cooperation is needed not only to protect their privacy but also to allow the service to run properly. Highly reputable peers aggregate participants' preferences then encrypt these collected profiles using homomorphic encryption in order to permit particular operations to be performed on encrypted data without need for prior decryption then they submit these profiles to PRS in order to produce referrals. The encrypted profiles hide the identities of participants, and thus hamper the ability for the untrusted PRS to invade users' privacy by profiling or tracking them. However, participants cannot trust each other as well and hence the aggregation process should not expose their preferences. Hence, we proposed a trust based obfuscation mechanism, which designed especially to obfuscate items' ratings before their submission to these highly reputable peers.

This approach preserves the aggregates in the dataset to maximize the usability of information in order to

accurately predicate ratings for items that have not consumed before by the group members. In addition, *EMCP* employs interpersonal trust between users to enhance recommendation accuracy and preserve privacy. The enhancement in accuracy is achieved by employing trust based heuristics to propagate and spot users whom are trustworthy with respect to the target user. Moreover, trust based heuristics enhance privacy by transforming participants' data in different ways based on different trust levels to hide the raw ratings. Thus, In contrast to a single obfuscation level scenario where only one obfuscated copy is released for all users using fixed parameters for the obfuscation mechanism, now multiple differently obfuscated copies of the same data is released for different requests with different trust levels. The more trusted the target-user is the less obfuscated copy he can access. These different copies can be generated in various fashions. They can be jointly generated at different times upon receiving new request from target user, or on demand fashion. The latter case gives users maximum flexibility.

In rest of this work, we will generically refer to news programs, movies and video on demand contents as Items. This paper is organised as follows. In Section 2, related works are described. Section 3 presents the threat model assumed in this work. Section 4 introduces IPTV network scenario landing our private recommender service. The proposed solution based on *EMCP* is introduced in Section 5. In Section 6, the two-stage concealment process is described in details. Proof of security and correctness for the two-stage concealment process is demonstrated in Section 7. In Section 8, the Results from some experiments on the proposed mechanisms are reported. Finally, the conclusions and recommendations for future work are given in Section 9.

## 2 Related Works

The majority of the existing recommender services are based on collaborative filtering techniques that build users' profiles in two ways on ratings (explicit rating procedures) or on log archives (implicit rating procedures) [3]. These procedures lead to two different approaches for collaborative filtering including the rating based approaches and log based approaches. The majority of the literature addresses the problem of privacy on collaborative filtering techniques, due to it being a potential source of leakage of private information shared by the users as shown in [4]. In [5] a theoretical framework is proposed to preserve the privacy of customers and the commercial interests of merchants. Their system is a hybrid recommender system that uses secure two party protocols and public key infrastructure to achieve the desired goals. In [6, 7] a privacy preserving approach is proposed based on peer to peer techniques using users' communities, where the community will have a aggregate user profile representing the group as a whole but not individual users. Personal information will be encrypted and communication will be between individual users but not servers. Thus, the recommendations will be generated on



the client side. In [8, 9] another method is suggested for privacy preserving on centralized recommender systems by adding uncertainty to the data by using a randomized perturbation technique while attempting to make sure that the necessary statistical aggregates such as the mean do not greatly get disturbed. Hence, the server has no knowledge about the true values of the individual items' ratings for each user. They demonstrate that this method does not essentially decrease the accuracy obtained in the results. But recent research work [10, 11] pointed out that these techniques do not provide levels of privacy as was previously thought. In [11] it is pointed out that arbitrary randomization is not safe because it is easy to breach the privacy protection it offers. They proposed random matrix based spectral filtering techniques to recover the original data from the perturbed data. Their experiments revealed that in many cases, random perturbation techniques preserve very little privacy. Storing users' profiles on their own side and running the recommender system in a distributed manner without relying on any server is another approach proposed in [12], where the authors proposed only transmitting similarity measures over the network and keeping users' profiles secret on their side to preserve privacy. Although this method eliminates the main source of threat against user's privacy, it requires higher cooperation among the users to generate useful recommendations. The work in [13] stated that existing similarities are deemed useless as traditional user profiles are sparse and insufficient. Recommender systems need new ways to calculate user similarities. They utilize interpersonal trustworthiness to describe the relationship between two users. The authors in [14] show the correlation between similarity and trust and how it can elevate movie recommendation accuracy.

### 3 Threat Model

In this work, we assume that an adversary aims to collect users' preferences in order to identify and track users. Thus, we consider our main adversary to be an untrusted PRS to which users send their preferences. We do not assume the PRS to be completely malicious. This is a realistic assumption because PRS needs to accomplish some business goals and increase its revenues. PRS can construct the profiles of the users based on the requests sent. Hence, the problem we are tackling has two sides; we want to detain the ability of the adversary to identify users based on a set of identifying interests and thus track them by correlating these data with data from other publicly-accessible databases and in the same time we want to prevent the adversary from profiling the users through their network identity and therefore invade their privacy. Intuitively, the system privacy is high if PRS is not able to reconstruct the real users' preferences based on the information available to it. Other adversaries are malicious users trying to collect preferences information about others. Malicious users can eavesdrop and collect preferences while being aggregated. So, while hiding our identity from the recommender service, it should not be revealed to other users sniffing the network.

## 4 Private Recommender Service for IPTV Network Scenario

We extend the scenario proposed in [15-20], where a private recommender service (PRS) is implemented as an external third party server and users give their preferences to that server in order to receive referrals. *EMCP* preserves users' privacy by utilizing three mechanisms: trust based obfuscation, aggregation and threshold encryption. The basic idea for a recommendation based on *EMCP* is that the user who needs recommendation will form a group with other participants in the IPTV network who decided to join his recommendation process. Then, the group members elect highly reputable peers (that we call super-peer) to aggregate their preferences they are willing to share into profiles. The super-peers will cooperate to achieve privacy by encrypting collected profiles using threshold homomorphic encryption in order to permit particular operations to be performed on encrypted data without need for prior decryption and then submit these aggregates to PRS in order to produce referrals. The encrypted profiles hide the identities of the participants, and thus hamper the ability for the untrusted PRS to profile and track users that invade their privacy. However, participants cannot trust each other as well and hence the aggregation process should not expose their preferences. Hence, we proposed a trust based obfuscation mechanism to obfuscate preferences prior submission to super-peers.

Our solution relies on a two stage concealment process, the first stage is trust based obfuscation and it takes place at participant side to hide extracted preferences prior their submission to super-peers. Then the second stage is threshold homomorphic encryption and it takes place at super-peers to hide collected profiles prior their submission to PRS. The overall process might be described as follows: upon receiving a request from the target user, a group of participants is formed that is managed by an elected super-peer. Super-peers negotiate with both the target user and PRS to express their privacy practices for the data collection and usage via P3P policies which are XML statements that answers questions concerning purpose of collection, the recipients of these profiles, and the retention policy. After receiving P3P policy & request, *EMCP* ensures that the extracted preferences for specific request do not violate the privacy of its host by checking whether there is an APPEL privacy preference corresponding to that given P3P policy, and then it starts collecting preferences that fulfil the request and in the same time satisfies the extracted APPEL preferences. The extracted items' ratings are obfuscated using a trust based obfuscation mechanism provided by *EMCP*, such that each item's rating is obfuscated based on the privacy preferences of its owner and estimated trust level with the target user. Furthermore, items identifiers and meta-data are hashed using locality-sensitive hashing. This step prevents the super-peers from knowing each participant's raw ratings for different items identifiers. The super-peer collects

these obfuscated preferences and computes an aggregation on them, which does not expose individual ratings. Next, the collected profiles are encapsulated using threshold encryption and submitted to PRS to predicate ratings for the referred items that did not consumed before and will be offered in the end to the target-user and participants. The collaborative filtering task at PRS will be reduced to computing addition on aggregated ratings without exposing the raw ratings. Therefore, our solution ensures privacy in the relation between the participants and PRS and in between the participants themselves. In the following section we will describe some enhancements attained using *EMCP*:

1. **Usage of Pseudonymous for the Profiles:** The real user's identity is not always required to provide referrals. Users can be identified by anonymous pseudonyms or nicknames, so that the binding of nickname and the real life identity is not always manifested.
2. **User Private Data Store at the Client:** Shifting from the approach of storing the user profiles in the server side to the one of storing the profiles on the clients' STBs helps reducing the privacy concerns. One key aspect is keeping the profiles encrypted to avoid people having access to the client's machine or malware that looks for user profiles.
3. **Request-Oriented Collection:** Upon receiving a request from the target user, query rewriter and preference checker assures that learning agent extracts only the required preferences from user's profile for a particular request the user is engaged in. The key point relies on knowing what kind of data is required for a given request that can contribute to improve the performance of the recommendation, because the recommender service does not provide recommendation based on one user's full profile information (e.g.: other users' preferences might not be relevant to the request). Likewise, once a user completes a particular request, he/she may no longer be interested in receiving recommendations related to that request for a period.
4. **Communication through Anonymous Networks:** internet records containing IPs, etc stored at service providers, contain information that permit the identification of user when submitting their obfuscated preferences to the node that requested recommendation. *EMCP* employ anonymous communication to hide the network identity for the participants by routing the submission of their obfuscated preferences through relaying nodes in an anonymous communication network before sending them to Super-peers. The main challenge for *EMCP* is to tune up and optimize the performance of the anonymous network while maintaining the user anonymity, we employed the path selection algorithm presented in [18] to enhances the anonymous network performance. Figure (1) shows the architecture of our approach.

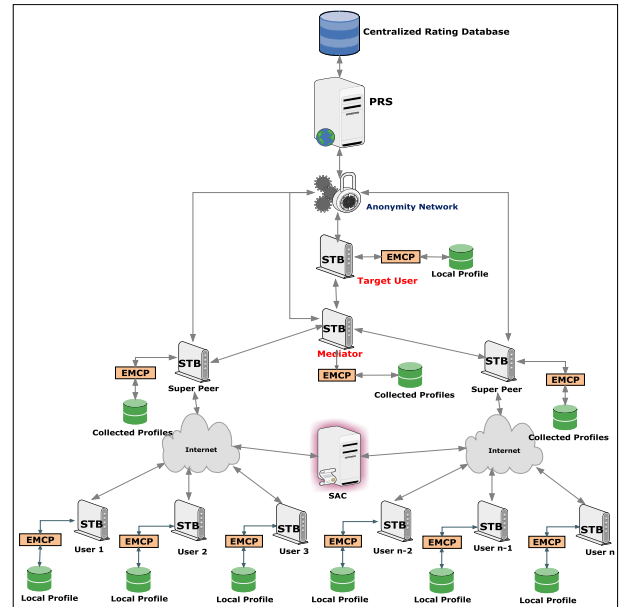


Figure 1: *EMCP* Middleware with Third Party Private Recommender Service.

Our solution relies on the hierarchical topology proposed in [21]; per each request participants are organized into peer-groups managed by super-peers. Electing super-peers is based on negotiation between the participants and security authority centre. The security authority centre (SAC) is a trusted third party responsible for making an assessment on those super-peers according to the participants' reports and periodically updating the reputation of each super-peer based upon it. Reputation mechanisms are employed to elect suitable super-peers based on estimating values for user-satisfaction, trust level, processing capabilities and available bandwidth, further details and information on complex reputation mechanisms can be found in [22]. When a problem occurs with a specific super-peer during the recommendation process, a participant can report it to SAC. After investigation, the assessment of the super-peer will be degraded. This will limit the chance for electing it as a super-peer in the future. On the other hand, successful recommendation processes will help upgrade the super-peer reputation. An IPTV provider can offer certain benefits (like free content, prizes,... etc) for those participants who have a sustained success rate as a super-peer.

Our solution depends upon the set top box (STB) device at the user side. STB is an electronic appliance that connects to both the network and the home television. With the advancement of data storage technology each STB is equipped with mass storage, e.g. Cisco STB. *EMCP* components are hosted on STB; Moreover STB storage stores the user profile. On the other hand, PRS maintains a centralized rating database that is used to provide referrals if the number of participants in group fall below a certain threshold. PRS is the third-party entity recruited by the IPTV network provider to operate referrals by consolidating the information received from multiple sources.

### 5 Proposed Solution

In the beginning, we want to introduce the notions of privacy and trust within our framework, we need to confirm what we mean by privacy and trust first. To define privacy and trust in our terms, we first approach the notion of privacy in following terms: “A participant who wants to join recommendation request in a network of users, does not has to reveal raw ratings in his/her profile during the recommendation process and elected super-peers does not wish PRS to learn any raw ratings in the collected profiles they provide”. While in the context of this paper, trust is interpreted as “a user’s expectation of another user’s competency in providing ratings to reduce its uncertainty in predicating new items’ ratings [23]”. In our framework, the notion of privacy surrounding the disclosure of users’ preferences and the protection of trust computation between different users are together the backbone of our solution. We apply a trust based obfuscation mechanism at participant side, which produces different copies of items’ ratings based on the various trust levels with target user. The trust computation is done locally over the obfuscated participant’s preferences, and then recommendation is served using secure multi-party computation protocol. Utilizing trust heuristic as input for both group formation and obfuscation process has been of great importance in mitigate some of malicious insider attacks such as infesting the trust computation results. As future work, we plan to investigate miscellaneous insider attacks and strengthen our framework against them.

In the next sub-sections, we will present our proposed middleware for protecting the privacy of users’ preferences. Figure (2) illustrates the *EMCP* components running inside user’s STB. *EMCP* consists of different co-operative agents. A Learning agent captures user interests about miscellaneous items explicitly or implicitly to build a rating database and meta-data database. The local obfuscation agent implements a trust based obfuscation mechanism to achieve user privacy while sharing his/her preferences with super-peers or PRS. The encryption agent is only invoked if the user is acting as a super-peer in the recommendation process; it executes *SR* protocol on the collected profiles. These mechanisms act as wrappers that conceal preferences before they are shared with any external entity. Since the database is dynamic in nature, the local obfuscation agent periodically desensitizes the updated preferences, and then a synchronize agent forwards them to the PRS upon owner permissions. Thus recommendation can be made on the most recent ratings. Moreover, synchronize agent is responsible for calculating & storing parameterized paths in anonymous network that attain high throughput[18], which in turn can be used in submitting preferences anonymously. The policy agent is an entity in *EMCP* that has the ability to encode privacy preferences and privacy policies as XML statements depending on the host role in the recommendation process. Hence, if the host role as a “super-peer”, the policy agent will has the responsibility to encode data collection and data usage practices as P3P policies via

XML statements which are answering questions concerning purpose of collection, the recipients of these profiles, and the retention policy. On the other hand, if the host role as a “participant” policy agent acquires the user’s privacy preferences and express them using APPEL as a set of preferences rules which are then decomposed into set of elements that are stored in a database called “privacy preferences” as tables called “privacy meta-data”. These rules contain both a privacy policy and an action to be taken for such privacy policy, in such way this will enable the preference checker to make self-acting decisions on objects that are encountered during data collection process regarding different P3P policies (e.g.: privacy preferences could include: Certain categories of items should be excluded from data before submission, Expiration of purchase history, Usage of items that have been purchased with the business credit card and not with the private one, Generalize certain terms or names in user’s preferences according to defined taxonomy, Using synonyms for certain terms or names in user’s preferences , suppressing certain items from the extracted preferences and insert dummy items that have same feature vector like the suppressed ones as described in [24], limiting the potentially output patterns from extracted preferences etc in order to prevent the disclosure of sensitive preferences in user’s profile). Query Rewriter rewrites the received request constrained by privacy preference for its host.

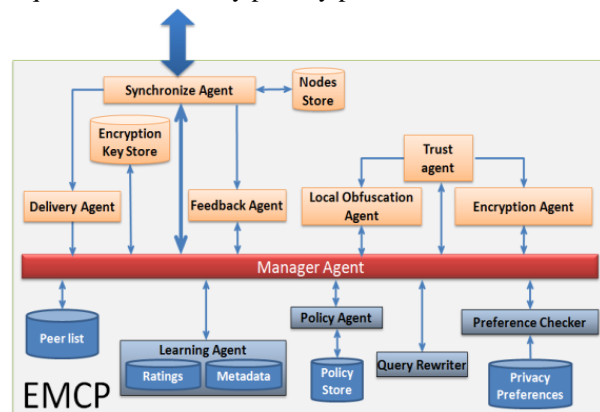


Figure 2: EMCP Components.

Figure (3) shows the participants interactions with super-peers and PRS. The recommendation process in our solution operates as follows:

1. The learning agent collects the user’s interest about different items which represent his profile. The local profile is stored on two databases, the first one is the rating database that contains (item\_id, rating) and the second is the meta-data database that contains the feature vector for each item [24] (item\_id, feature1, feature2, feature3). The feature vector can include genres, directors, actors and so on. Both implicit and explicit ways for information collection [25] are used to construct these two databases and maintain them.
2. As stated in [16], the target user broadcasts a message to other users in the IPTV network requesting recommendation for a specific genre or category of items. Thereafter, the target user selects a set of his

preferences to be used later in the computation of trust level at the participant side. So as to hide the items identifiers and meta-data from other participants, The local obfuscation agent uses locality-sensitive hashing (LSH) [26] to hash these values. One interesting property for LSH is that similar items will be hashed to the same value with high probability. Super-peers and PRS are still able to perform computation on the hashed values using appropriate distance metrics like hamming distance or dice coefficient. Simultaneously, local obfuscation agent sanitizes items' ratings using trust based obfuscation. Finally, the target user dispatches these obfuscated items' ratings along with their associated hashed values to the Individual users who have decided to participate in the recommendation process.

3. Each group of participants negotiates with SAC to select a peer with the highest reputation as a "super-peer" which will act as a communication gateway between the PRS and the participants in its underlying group.
4. Each super-peer negotiates with both the target user and PRS to express its privacy policies for the data collection and usage process via P3P policies which are XML statements that answers questions concerning purpose of collection, the recipients of these profiles, and the retention policy. Thereafter, super-peers engage in key distribution phase of the SR protocol, at the end of this phase each super-peer will possess a share of the decryption key along with the complete encryption key to encrypt the collected profiles. The encrypted profiles can only be decrypted only if any subset consisting of a threshold  $t$  of super-peers cooperate.
5. At the participant side, the manager agent receives the request from the target user along with the P3P policy form the elected super-peer; then it forwards P3P policy to preference checker and the request to query rewriter. The preference checker ensures that the extracted preferences for a specific request do not violate the privacy of its host by checking whether there is an APPEL preference corresponding to the given P3P policy and sends it to the query rewriter. The user's preferences can be transferred or collected only if the purpose of statement for the collectors satisfies the privacy preferences. The query rewriter will have knowledge about privacy preferences related to current request via APPEL preference then it rewrites the received request constrained by the privacy preference for its host in order to only retrieve the preferences that the host agrees to share as well as prevent the disclosure of confidential preferences in the participant's profile. This step enable the participant to decide when the recommendation takes place, which information should be collected and for which purpose. This step will ensure the privacy principles compliance and put the user in control the information that is part of their profiles. The modified request is directed to the learning agent to start collecting preferences that could satisfy the modified query. The manager agent ensures that the collected

preferences compliance with the collection data principle, as only the required preferences for the particular request the user is engaged in, is extracted for the local obfuscation process.

6. In the meanwhile, the trust agent calculates approximated interpersonal trust between its host and the target user based on the received preference. It is done in a decentralized fashion using the entropy definition proposed in [23] at each participant side. The entropy value becomes lower as the users' ratings are more consistent, which is similar to the definition of trust previously stated.  $\forall_{j=1}^n T(u_a, u_{b_j})$  is the estimated trust between the target user  $u_a$  and participant  $u_{b_j}$ . the whole process can be described using the following steps:

- i. Each participant  $\forall_{j=1}^n u_{b_j}$  determines a subset of his/her items' ratings that will be required for recommendation process. Then the participant utilizes shared items rated by both of  $u_a, u_{b_j}$  for the trust computation. Determining shared rated items is done by matching the received items' hash values from target user  $u_a$  with his/her local items' hash values.
- ii. Participant  $u_{b_j}$  computes the trust level using

$$T(u_a, u_{b_j}) = \frac{\text{Entropy}(u_a) - \text{Entropy}(u_a | u_{b_j})}{\text{Entropy}(u_a)} \quad (1)$$

$$= \frac{\left(1 - \frac{\log N}{\log ZN}\right) + \frac{1}{N \log ZN} (\sum_{i=1}^Z \sum_{j=1}^Z n_{ij} \log n_{ij} - \sum_{i=1}^Z n_i \log n_i)}{1 - \frac{1}{N \log ZN} \sum_{i=1}^Z n_i \log n_i}$$

Equation (1) is an adapted formalization of trust as proposed in [23] where  $Z$  denotes the number of states of rated values and  $N$  is the total number of rating times. For example if  $Z=6$  and  $N=20$  when 20 ratings are made with 1 to 6 integer valued scores. Employing entropy to select trustworthy neighbours achieves an improvement in the group formation and rating predication. The enhancement in rating predication is stemmed from trust propagation, so if  $u_{b_j=x}$  is selected as a trustworthy user and he/she does not have a rating for the item to be predicted, a trustworthy user  $u_{b_j=y}$  of user  $u_{b_j=x}$  can also be used for the predication.

- iii. Each participant  $\forall_{j=1}^n u_{b_j}$  sends his/her calculated trust value to the super-peer. The Estimated trust values are forwarded to both the super-peers and PRS.
- iv. Each participant  $\forall_{j=1}^n u_{b_j}$  sends this trust value to the local obfuscation agent to adjust the obfuscation level with trust level, in other words, we correlate the obfuscation level with different levels of trust, so the more trusted a target user is, the less obfuscated copy of users' preference he can access. The local obfuscation agent executes enhanced value-substitution (EVS) algorithm on items' ratings that are required in the recommendation process. Moreover the local

obfuscation agent hashes their identifiers and meta-data using LSH. The level of obfuscation is determined using the trust level with the target user, and then participants submit their obfuscated preferences to the super-peers of their group. Anonymous communication [18] utilized to hide the network identities of group members when submitting their obfuscated preferences to the super-peers.

- v. Finally, the policy agent audits the original and modified requests plus estimated trust level and P3P policy with previous requests; this step allows *EMCP* to prevent multiple requests that might extract sensitive preferences. In such a case, if the target user requests same data twice, its trust level will be reduced, in which will increase the level of the obfuscation in the extracted preferences. This step will cause extracted preferences appear as a completely different set of preferences to the target user.

7. Upon receiving the obfuscated preferences from the participants, each super-peer filters the received preferences based on the trust level of their owners such that  $T(u_a, u_{b_j}) > \theta$  where  $\theta$  is a minimum trust threshold value defined by the target user or PRS. Then, each super-peer collects the participants' pseudonyms and builds a group rating profile such that all the <hashed value, rating> elements belonging to similar items are grouped together. This allows the computing of the items popularity curve at each super-peer. The super-peer can seamlessly interact with the PRS by posing as an end-user and has a group profile as his own profile. Each super-peer  $\forall_{x=1}^k SP_x$  calculates the following intermediate values for each user in the  $N$ -neighbourhood of target user  $\forall_{j=1}^n u_{b_j} \in Neighbor(u_a)$ ,

$$\text{Then } \forall q = 1 \dots T \quad \widehat{r_{u_{b_j},q}} = r_{u_{b_j},q} - \bar{r}_q$$

$$\widehat{r_{q,u_{b_j}}^x} = \frac{T(u_a, u_{b_j}) * \widehat{r_{u_{b_j},q}}}{T(u_a, u_{b_j})} \quad (2)$$

Where  $r_{u_{b_j},q}$  is the rating value of participant  $u_{b_j}$  for item  $q$ .  $\bar{r}_q$  is the average rating for item  $q$  in each items' cluster. Next, each super-peer encrypts these intermediate ratings  $\widehat{r_{q,u_{b_j}}^x}$  using the encryption key  $pk$ . Finally, the super-peer submits these ratings along with their associated hashed values to PRS, which in turn collects them to produce final referrals.

8. Upon receiving the encrypted ratings  $\forall_{x=1}^k \forall_{j=1}^n \varepsilon_{pk}(\widehat{r_{q,u_{b_j}}^x})$  from all super-peers, PRS stores them along with their participants' pseudonyms and hashed values in the centralized rating database. The rating predication phase is performed using the additive homomorphic property of the threshold paillier encryption as the required computations are additive. Thus, PRS executes an additive operation on the encrypted rating profiles without decrypting them so the private data of multiple super-peers can be preserved during the computation. Calculating the

predicted rating for referrals done as shown in equation (3):

$$\begin{aligned} p_{u_a,q} &= \varepsilon_{pk}(\bar{r}_{u_a}) * \left( \prod_{j=1}^n \varepsilon_{pk}(\widehat{r_{q,u_{b_j}}^x}) \right) \\ &= \varepsilon_{pk} \left( \bar{r}_{u_a} + \left( \sum_{j=1}^n \widehat{r_{q,u_{b_j}}^x} \right) \right) \end{aligned} \quad (3)$$

Notice that the result will be equal to the weighted sum of the participants' rating plus the average rating of the target user  $r_{u_a}$ . Super-peers uses the reblinding property of the paillier encryption to prevent PRS and target user from obtaining any knowledge of  $\widehat{r_{q,u_{b_j}}^x}$  values by trying a few possible values.

9. PRS forwards the encrypted referrals list along with their predicated ratings to super-peers which in turn perform threshold decryption on these results. The threshold decryption process requires that at least  $t$  of the super-peers are honest. Only when the required number of super-peers cooperates, they can perform decryption using their local share of the private key, and then they will be able to have the final referrals list for the entire group. Super-peers publish the final list to the target user and participants. Finally, each participant report scores about the elected super-peer of his group and target-user to SAC, which helps to determine reputation of each entity involved in referrals generation.

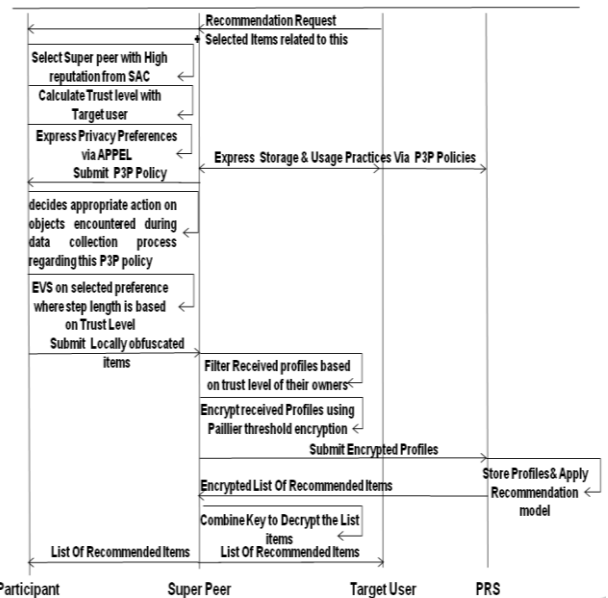


Figure 3: Interaction Sequence Diagram.

## 6 Proposed Two Stage Concealment Process

In the next subsections, we present our two stage concealment process used in *EMCP* to disguise the user items' ratings in way that secure the user's preferences in the untrusted PRS with minimum loss of accuracy. In our framework, each user has two datasets representing his/her profile. A local profile: represents the actual ratings of the user for different items; it is stored on his STB. Each user disguises this local profile before sending it to super-peer. An encrypted centralized

profile: this is the output of the two-stage concealment process that stored at PRS, the user gets recommendation directly from the PRS based on the previously collected profiles. We perform experiments on real datasets to illustrate the applicability of our mechanisms and the privacy and accuracy levels achieved by using them.

## 6.1 Cryptography Tools

Using additively homomorphic cryptosystem permit the computation of linear combinations of encrypted data without need for prior decryption, such that PRS can combine received encrypted rating profiles into a new ciphertext that is the encryption of the sum of the ratings of the original ratings. Formally, an encryption schema  $\varepsilon_{pk}(\cdot)$  denotes the encryption function with encryption key  $pk$  and  $D_{sk}(\cdot)$  denotes the decryption function with decryption key  $sk$ . Additive homomorphic cryptosystem possesses the following properties:

1. Given the encryption of plaintexts  $m_1$  and  $m_2$ ,  $\varepsilon_{pk}(m_1)$  and  $\varepsilon_{pk}(m_2)$ . The sum  $m_1 + m_2$  can be directly computed as  $\varepsilon_{pk}(m_1 + m_2) = \varepsilon_{pk}(m_1) * \varepsilon_{pk}(m_2)$ .
2. Given a constant  $k$  and the encryption of  $m_1$ ,  $\varepsilon_{pk}(m_1)$ . The multiplication of  $k$  with the plaintext  $m_1$  can be directly computed as  $\varepsilon_{pk}(k.m_1) = \varepsilon_{pk}(m_1)^k$ .

Paillier [27] proposed a probabilistic asymmetric algorithm for public key cryptography that is an example of an efficient additively homomorphic cryptosystem, this scheme is further extended by [28] with a threshold versions, but required the use of a trusted dealer to distribute the keys to the participants. The reliance on a trusted dealer was lifted in [29] to ensure that no single party or coalition of less than specific participants can recover the encrypted values. In designing *SR* protocol, we require a fully distributed key generation protocol. In particular, the coalition between PRS or target user with any super-peer within the group should not be able to decrypt the whole collected profiles submitted to PRS, but it only reveals the obfuscated profiles collected by this super-peer. Therefore neither can be used as a trusted “dealer” for key generation. Thus, we employ a fully distributed threshold cryptosystem, Since it is desirable to distribute trust between numerous super-peers and no single super-peer is assumed to be fully trusted, then the decryption key  $sk$  is shared among a number  $P$  of super-peers, and encrypted profiles can only be decrypted only if any subset consisting of a threshold  $t$  of super-peers cooperate but no subset smaller than  $t$  can perform decryption. Moreover, with the additively homomorphic property of Paillier schema it permits *SR* protocol to perform secure aggregation and predication over encrypted rating profiles. We assume a semi-honest model for the super-peers. Hence, we do not require zero-knowledge proofs (ZKPs) for the various cryptographic operations from the participants. We will briefly present the distributed paillier threshold cryptosystem below.

### Key Generation

In this step, each super-peer  $\forall_{i=1}^n SP_i$  generates  $n$  additive shares of two  $\kappa/2$ -bit strong primes, such that each super-peer have share  $p_i$  and  $q_i$ . Then use the method proposed in [29] to compute  $N = pq$ ,  $\lambda = LCM(p-1, q-1)$ ,  $g = N + 1$  such that  $p = \sum_{i=1}^n p_i$ ,  $q = \sum_{i=1}^n q_i$ , also  $d$  such that  $d \equiv 1 \pmod{N}$  and  $d \equiv 0 \pmod{\lambda}$ . The public key  $pk = (N, g)$  and the private key  $sk = d$ . Note that, super-peers perform biprimality test in [30] for checking if  $N$  is a product of two primes in a distributed way. If the test fails, the protocol is restarted

### Key Sharing

The private key  $sk$  is shared among  $n$  super-peers with the Shamir scheme as  $t-1$  degree polynomial where each party obtain  $(t, n)$  share of  $d$ : Let  $a_0 = d$ , and randomly choose  $a_i$  in  $\{0, \dots, N-1\}$  and set  $f(X) = \sum_{i=0}^{t-1} a_i X^i$ . The share  $s_i$  of the  $i$ th super-peer  $SP_i$  is  $f(i) \pmod{N}$ .

### Encryption

To encrypt a message  $M \in \mathbb{Z}_N$  with public key, randomly choose  $r \in \mathbb{Z}_N^*$  and compute  $C = g^{Mr} \pmod{N^2}$ .

### Share Decryption

To decrypt  $C$ , each super-peer  $SP_i$  computes the decryption share  $c_i = c^{2\lambda s_i} \pmod{N^2}$ , where  $\Delta = t!$  using his/her secret share  $s_i$ . And finally, if  $t+1$  valid shares are available, they can be combined to recover  $M$  as described in End decryption.

### End Decryption

Let  $S$  be a set of  $t+1$  valid shares. Compute

$$M = L \left( \prod_{i \in S} c_i^{2\lambda_i} \pmod{N^2} \right) \frac{1}{4\Delta^2} \pmod{N}$$

Where  $\lambda_i = \Delta \prod_{i \in S \setminus i} \frac{-i}{i-i}$ , See [29] for more details on the correctness of the scheme and for proofs of security.

## 6.2 Local Obfuscation using Enhanced Value-Substitution (EVS) Algorithm

We propose a novel algorithm for obfuscating the users' ratings before sending them to the super-peers. This algorithm is called *EVS*, which has been designed especially for the sparse data problem we have here. Moreover the algorithm tunes its obfuscation parameters based on trust level. The available anonymisation algorithms perform single obfuscation levels for all participants and release one obfuscated copy for all of them which result in increasing data distortion and construction of inaccurate recommendation model. The key idea for *EVS* is based on the work in [31] that uses Hilbert curve as a dimensionality reduction tool to create a cloaking regions to attain privacy for users. Hilbert curve also has the ability to maintain the association between different dimensions. In this subsection, we extend this idea as following, we also use Hilbert curve to map  $m$ -dimensional profile to 1-dimensional profile then *EVS* discovers the distribution of that 1-dimensional profile. Finally, we perform perturbation based on that distribution in such a way to preserve the profile range that led to providing accurate results when performing

rating predication. The output of our obfuscation algorithm should satisfy two requirements:

- Reconstructing the original profile from the obfuscated profile should be difficult, in order to preserve privacy.
- Preserving the distances of the data to achieve accurate results for the recommendation.

The steps for EVS algorithm consists of the following:

1. We denote the collected m-dimensional user preferences as dataset D of c rows, where each row is a sequence of m dimensions  $A = A_1, A_2, A_3, A_4 \dots \dots, A_m$ .
2. Trust level values are divided to a number of intervals defined by the user, associated with each interval an order k value. EVS chooses a value for order k according to the trust level associated with the target user.
3. EVS divides the m-dimensional dataset D into grids of order k as shown in [31, 32]. For order k, the range for each dimension divided into  $2^k$  intervals.
4. For each dimension  $\forall_{i=1}^m A_i$  of the collected preferences D:
  - Compute the k-order Hilbert value for each data point  $\forall_{x=1}^c a_{ix}$ . This value represents the index of the corresponding interval where it falls in.
  - EVS sort the Hilbert values from smallest to biggest, then use the step length (a user defined parameter) to measure whether any two values are near from each other or not. If these values are near, they are placed in the same partition  $\forall_{v=1}^k k_{iv}$ .

These two steps iterates for all m-dimensions. The final result from these steps is k partitions for each dimension denoted as  $\forall_{i=1}^m \forall_{v=1}^k C_{iv}$

5. EVS constructs a N shared nearest neighbour sets  $S_r$  where  $r = 1 \dots N$  as in [33] from different partitions with a new modified similarity function as following, two partitions in different dimensions  $C_{iv}, C_{i+1v}$  form a shared nearest neighbour set  $S_r$  if they share k-number of common elements such that  $S_r = C_{iv} \cup C_{i+1v}$
6. For each newly created set  $S_r$ , EVS calculates its interquartile range. Then, for each point  $a_i \in S_r$  generate a uniform distributed random point n in that range that can substitutes  $a_i$ .
7. Finally, the new set  $D' = \cup_{r=1}^N S_r$  is sent to Super-peer.

### 6.3 Secure Recommendation Protocol (SR)

We proposed a protocol that enables PRS to calculate predicted ratings from the encrypted rating profiles. We called this protocol secure recommendation protocol (SR). SR protocol starts with the selection of super-peers using SAC as it is heavily relies on the underlying network topology; also it requires a set of super-peers to aggregate all participants' preferences at the bottom of the hierarchy into profiles in order to remove any possibility of a single super-peer being the bottleneck. To achieve reasonable efficiency, super-peers reserve the

ability to independently reweight items' ratings based on trust values and omit the ones with low trust values, where such centralized computation can make the most significant difference. Moreover, they compute aggregated items' ratings from the obfuscated ratings received from their participants. Thereafter, super-peers engage in distributed key generation process using distributed threshold cryptosystem to generate public key to encrypt these profiles before submitting them to PRS. This key generation process will leave each super-peer with a share of the private key along with the complete public key. This makes sure that no single super-peer to able decrypt the profiles taken from different super-peers or the final referrals list retrieved from PRS. After all the super-peers collect preferences from participants and compute the aggregated ratings profiles, they engage independently in encrypting these results. Then, each super-peer will forward the ciphertext corresponding to the ratings profile over the entire group to the PRS. The PRS starts the rating predication phase on the ciphertext then forward back the results to the super-peers. The super-peers will then perform threshold decryption of these results. Only when the required number of super-peers cooperates, they can perform decryption using their local share of the private key, and then they will be able to have the final referrals list for the entire group. Note that we have focused on the decryption process to make sure that no single super-peer can get the profiles over a subset of super-peers in the group and malicious super-peers in the network are unable to compromise the security of the protocol. Moreover, utilizing fully distributed threshold cryptosystem ensures that all collected profiles become useless after the termination of recommendation process even if an attacker obtains the collected profiles. EMCP automatically destroys key shares directly after decrypting the received referrals list, without any explicit action by the participants or any party storing or archiving that data

Protocol \_SecureRecommendation

*Do forever*

*/\* Applied in cases where super-peers are not already defined, Electing super-peers is based on negotiation between participants and SAC to select peers with the highest reputations\*/*

SuperPeer = selectSP ();

*/\* Find out who are other super-peer from SAC \*/*

SPList = find SuperPeer ();

*/\* Check if I am super-peers to start collecting participants' preferences & generate keys for encryption agent \*/*

If (me == SuperPeer)

*/\* Delivery agent listens to receiver channel to collect obfuscated preferences from participants associated with this super-peer \*/*

ListenToReceiverChannel (CollectChannel,

ReceivedObfuscatedPreferences  $r_{u_b, q}, T(u_a, u_b)$  );

*/\* Delivery agent combine the obfuscated preferences on the receiver channel if trust level for its participant higher than specific threshold value  $\theta$  set by the target user \*/*

```

If  $T(u_a, u_{b_j}) > \theta$  then store  $r_{u_{b_j}, q}$ 
/* Delivery agent combine the obfuscated preferences
on the receiver channel, if there is a change in the local
preferences or if there is a new preferences received */
if (LocalObfuscatedPreferences == true ||
NewReceivedObfuscatedPreferences == true)
/* Calculates the normalized rating for item q from
rating of each participant  $u_{b_j}$  */
 $\forall_{j=1}^n u_{b_j} \in Neighbor(u_a)$  calculate  $\widehat{r_{u_{b_j}, q}}$ 

$$= r_{u_{b_j}, q} - \bar{r}_q$$

/* Combine the Received ratings with previously
collected ratings for each item q */
 $\forall q = 1 \dots T$ 
CombinedPreferences  $\widehat{r_{q, u_{b_j}}}$  = CombinePreferences

$$\left( \begin{array}{c} \text{LocalObfuscatedPreferences} \left( \frac{T(u_a, u_{b_j}) * (\widehat{r_{u_{b_j}, q})}}{T(u_a, u_{b_j})} \right), \\ \text{ReceivedObfuscatedPreferences} \widehat{r_{u_{b_j}, q}} \end{array} \right)$$

End if
End if
/* Generate public/private Key pair using a distributed
protocol employing all other super-peers. The function
SPDKG () leaves every super-peer with the entire public key
and a share of the private key */
(PublicKey, PrivateKey) = SPDKG(SPList);
/* Initiate the encryption agent to encrypt my combined
preferences with the public key and submit it to PRS */
Submit  $r_{q, u_{b_j}}$  (Enc(PublicKey, CombinedPreferences  $\widehat{r_{q, u_{b_j}}}$ ))
To PRS;
/* PRS receive collected preferences for different super-peers
*/
PRS receives  $r_{1, u_{b_1}^1}, r_{1, u_{b_1}^2}, \dots, r_{2, u_{b_1}^1}, r_{2, u_{b_1}^2}, \dots, r_{T, u_{b_j}^k}$ 
such that  $r_{q, u_{b_j}^k}$  is the encrypted rating for item  $q \in \{1, \dots, T\}$ 
by user  $u_{b_j}$  ( $\forall_{j=1}^n u_{b_j} | T(u_a, u_{b_j}) > \theta$ ) from super-peer
 $SP_x (\forall_{x=1}^k SP_x)$ 
/* PRS Calculates Predicated ratings for each unrated
Item in the collected profiles */
For each item  $q = 1$  to  $T$  do
PRS Calculates  $\forall_{x=1}^k p_{u_a, q} = (\text{Enc}(\text{PublicKey}, \bar{r}_{u_a}) * (\prod_{j=1}^n \text{Enc}(\text{PublicKey}, \text{CombinedPreferences } r_{q, u_{b_j}^x})))$ 
 $\forall_{j=1}^n u_{b_j}$ 
End for
/* Upon receiving the list of predicated ratings for referred
items, Target user request super-peers to start decrypt the
entire list */
if (me == SuperPeer) Reclist =
thresholdDecrypt(encryptedratingslist ( $p_{u_a, q}$ ), SPList)
End Protocol_SecureRecommendation
-----
Algorithm selectSP
/* Each participant contact SAC to obtain list of peers of
highest reputation to be elected as super-peer for the group */
Requst(HR_Peerlist);
/* Each super-peer broadcast to the neighbors indicating its
existence as their neighbor */
broadcast(SP_id);

```

```

/* if participant receives more than super peer id it
compare P3P policies for each adjacent super-peer &
select the one with suitable P3P policy to his privacy
preferences */
listenToReceiverChannel(defaultChannel, SP_id);
if (ReceiverPeerId(SP_id)  $\neq$  1)
Compare(SP_CollectionPolicies);
PeerGroupJoinRequest(SP_id);
End if
/*Each super-peer Listens to the receiver channel to form a
group */
listenToReceiverChannel(defaultChannel, numNeighbors);
broadcast(numNeighbors);
listenToReceiverChannel(defaultChannel, PeerGroupCountPair[
]);
superpeer= PeerGroupCountPair[maxIndex].getNeighborID();
return selectSP;

```

## 7 Proof of Security and Correctness

The proof of security for both EVS algorithm and SR protocol depends on how much information is leaked during the execution of the prediction phase. At the same time, our proposed mechanisms should output accurate results.

### 7.1 Privacy Breach Evaluation for EVS Algorithm

Privacy breach can be described in terms of how well the original user's ratings can be estimated from the submitted obfuscated ratings. Unlike other techniques, our method generates new data points, whose interpoint distances approximate the original distances. Consequently, points which lie close to one another in the original space mostly remain close to each other in the transformed space. Therefore, it seems theoretically to be more resilient to some potential attacks [34] that exploit the properties of the released data. These attacks are based on how much information about original data is available to the attacker that is obtained through either known input-output and known sample. In the known input-output, attacker knows collection of linearly dependant original data points and points they map in perturbed data. While in known sample, assumes that original data arose as independent samples of multidimensional random vector with unknown probability density function, and the attacker has access to a collection of these independent samples. In EVS algorithm, the linear ordering based on Hilbert curve retains the proximity and neighboring aspects of the original data. We define  $H_d^N$  for  $N \geq 1$  and  $d \geq 2$  as the  $N^{\text{th}}$  order Hilbert curve (defined values based on trust level) for a  $d$ - dimensional space.  $H_d^N: [0, 2^{Nd} - 1] \rightarrow [0, 2^N - 1]^d$  as follows : Hilbert value  $H = \epsilon(P)$  for  $H \in [0, 2^{Nd} - 1]$ , where  $P$  is coordinate of each point in  $[0, 2^N - 1]^d$ . Thereafter, we cluster nearby Hilbert values based on step length (a user-defined parameter) then EVS substitutes each point in the group with uniform distributed random point in the same



interquartile range for that cluster. Therefore we can consider  $\epsilon$  as a one-way function if the curve parameters are unknown. These parameters include (starting point, N, step length) are defined at the participant side and any external entity only know the final perturbed data that participant agree to release. As a result, the statistical information from the perturbed data are inconsistent with that from the original data. Therefore, attacks such as those described before would be inefficient in breaching privacy. In addition to that, clustering Hilbert values and substituting each point with random point introduces uncertainty about exact distance between data points, thus will make any distance based attack ineffective.

### 7.2 Proof of Security & Correctness for SR Algorithm

**Theorem 1:** additive operation performed by PRS in SR protocol is correct and accurate without the need of decryption keys.

**Proof:** based on the first property of additive homomorphic cryptosystem, we can determine that additive operations for encrypted data are correct as follows: given the encryptions  $\epsilon_{pk1}(m_1) = a$  and  $\epsilon_{pk1}(m_2) = b$  where  $\forall m_1, m_2 \in \mathbb{Z}_n$ , given encryption key

$$\begin{aligned} & \epsilon_{pk2}(\epsilon_{pk1}(m_1)) \cdot \epsilon_{pk2}(\epsilon_{pk1}(m_2)) \bmod N^2 = \\ & \epsilon_{pk2}(a) \cdot \epsilon_{pk2}(b) = (g^a r_1^N) \cdot (g^b r_2^N) \bmod N^2 = \\ & g^{a+b} (r_1 r_2)^N \bmod N^2 = \\ & \epsilon_{pk2}(\epsilon_{pk1}(m_1 + m_2)) \bmod N^2 \end{aligned}$$

Based on that, the PRS does not require any decryption key in order to aggregate all encrypted data.

**Theorem 2:** SR protocol computes predicated ratings for each referred item based on similar users' ratings without revealing extra information to any party.

**Proof:** Since each participant obfuscates his items' ratings and hashes their meta-data before submitting them to the super-peers. Moreover, each super-peer encrypts the collected profiles with the common encryption key and computation is performed on encrypted data and the decryption key is distributed between different super-peers. This makes sure that no single party will be able to decrypt these encrypted profiles taken from different super-peers or the final referrals list retrieved from PRS. This particular property is possible because of the threshold nature of the employed cryptosystem. In our two stage concealment process, the super-peer aggregates all obfuscated preferences then performs intermediate-computations on the obfuscated ratings for each item without having to know their real ratings or identifiers. No party can see extra information during the execution of the SR protocol. As for participants, they participate in the recommendation process without knowing other participants' identity. Since not all the participants have the same super-peer nor do they have direct communication with each other. The local profile is secured and can only be viewed by its owner before applying the trust based obfuscation mechanism. In

addition, employing reputation techniques to select super-peers with a high success rate in previous recommendation processes ensures the selection of reliable peers that will perform the required phases. PRS cannot see the received profiles as the decryption key is unknown. Furthermore, the decryption process requires a subset consisting of a threshold  $t$  of super-peers to cooperate. After PRS generates the final referrals list, PRS submits it to super-peers which in turn perform threshold decryption process. Then they publish this list to all participants.

**Theorem 3:** Assuming that all parties follow the protocol, SR protocol can correctly compute the predicated rating for each referred item.

**Proof:** When each super-peer encrypts collected rating profiles with encryption key  $\epsilon_{pk}(r_{q,u_b_j^x})$ . PRS performs the additive operation on encrypted rating profiles based on paillier's homomorphic cryptosystem as follows:

$$\begin{aligned} p_{u_a,q} &= \epsilon_{pk}(\overline{r_{u_a}}) + (\epsilon_{pk}(\widehat{r_{q,1}}) + \epsilon_{pk}(\widehat{r_{q,2}}) + \epsilon_{pk}(\widehat{r_{q,3}}) + \dots \\ & \quad + \epsilon_{pk}(\widehat{r_{q,u_b_j^x}})) \\ p_{u_a,q} &= \epsilon_{pk}\left(\overline{r_{u_a}} + \left(\sum_{j=1}^n \widehat{r_{q,u_b_j^x}}\right)\right) \\ p_{u_a,q} &= \epsilon_{pk}(\overline{r_{u_a}}) \left(\prod_{j=1}^n \epsilon_{pk}(\widehat{r_{q,u_b_j^x}})\right) \end{aligned}$$

After the threshold decryption by super-peers, we will obtain the final predicated rating as in equation

$$p_{u_a,q} = \overline{r_{u_a}} * \left(\prod_{j=1}^n \widehat{r_{q,u_b_j^x}}\right)$$

So the result from SR protocol is correct.

## 8 Experiments

In this section, we describe the implementation of our proposed solution. The experiments are run on 2 Intel® machines connected on local network, the lead peer is Intel® Core i7 2.2 GHz with 8 GB Ram and the other is Intel® Core 2 Duo™ 2.4 GHz with 2 GB Ram. We used MySQL as data storage for the users' preferences that acquired by learning agent. The proposed two stage concealment process is implemented in C++ using the MPICH implementation of the MPI communication standard for distributed memory implementation of the SR protocol to mimic a distributed reliable network of peers. To implement Paillier encryption scheme, the Number Theory Library (NTL) was used. One practical issue that must be dealt with when using the Paillier cryptosystem is the fact that it cannot naturally encrypt floating-point numbers. Floating-point numbers must be converted to a fixed-point representation. This is done by multiplying them by a large constant  $C$  and then truncating the result to an integer. In these experiments,  $C = 100000$ . Other methods in [35] can also be used. The experiments presented here were conducted using the Jester dataset provided by Goldberg from UC Berkley [36]. The dataset contains 4.1 million ratings on jokes using a real value between (-10 and +10) of 100 jokes from 73.412 users. The data in our experiments consists

of ratings for 36 or more items by 23.500 users. We evaluated the proposed solution from different aspects: privacy achieved, accuracy of results and performance. We used the mean absolute error (MAE) metric proposed in [37]. MAE is one of most famous metrics for recommendation quality. As it measures the predication verity between the predicated ratings and the real ratings, so smaller MAE means better recommendation provided by PRS. To measure the privacy or distortion level achieved using our mechanism, we used the variation of information metric VI [38] to estimate data error. A higher value of VI means a larger distortion between the obfuscated and original dataset, which means a higher level of privacy. The experiments involve dividing the data set into a training set and testing set. The training set is obfuscated then used as a database for PRS. Each rating record in the testing set is divided into rated items  $t_i$  and unrated items  $r_i$ . The set  $t$  is presented to PRS for making predication  $p_i$  for the unrated items  $r_i$ . For the representation process of the trust calculation, we add the default value 0 for items not rated. In our dataset, the first column of every raw stores how many items are rated by the user, which is necessary for the trust estimation process. We divided trust levels into three intervals [highest, moderate, and lowest] and associated hilbert curve order for each interval. The experiments were performed while keep the number of super-peers  $n = 9$ , as described earlier they will be responsible for aggregating the data of 23.496 participants. We assume the trust level for all participants to be above the minimum trust threshold  $\theta$ , which is required for the inclusion in the prediction process. The recommendation process can be initiated by any user that will act as the target-user for the referrals list. The trust level between participants and target-user is calculated locally on their STB devices.

In the first experiment, we want to measure the elapsed time for distributed key generation by varying the encryption key length and number of participants. Therefore we run the function SPDKG in SR protocol on 9 super-peers with different key length, and then we measure the elapsed time for distributed key generation and plot results in figure (4). Moreover, we set the key length to 1024 bits with varying number of super-peers (3 to 17) and we plot the elapsed time results in figure (5).

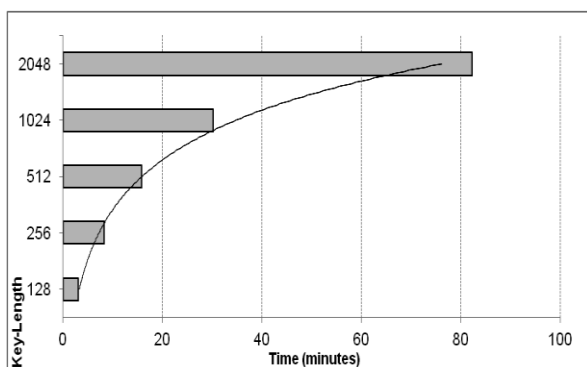


Figure 4: Key generation time for different Key Length.

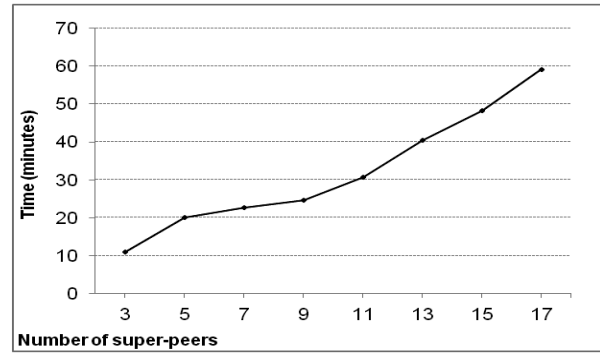


Figure 5: Key generation time for various numbers of super-peers.

In the second experiment, we want to measure the elapsed time for calculating the predicated ratings in SR protocol by varying the encryption key length and number of participants. We run the predication phase several times by encrypting all 12.674 records with different key length and distribute them equally on 9 super-peers, then PRS start collecting these records to perform predication phase. The results for elapsed time are shown in figure (6). Moreover, we set the key length to 1024 bits with varying number of super-peers (3 to 17) and we plot the elapsed time results in figure (7).

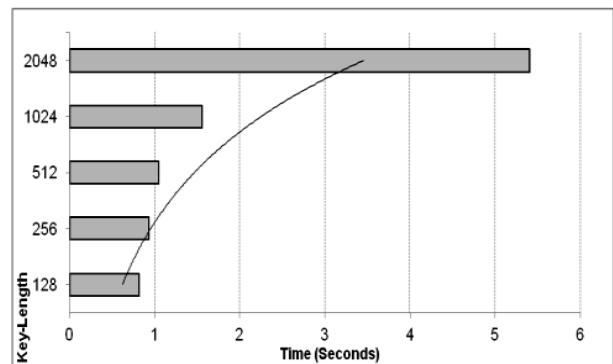


Figure 6: Ratings predication time for different Key Length.

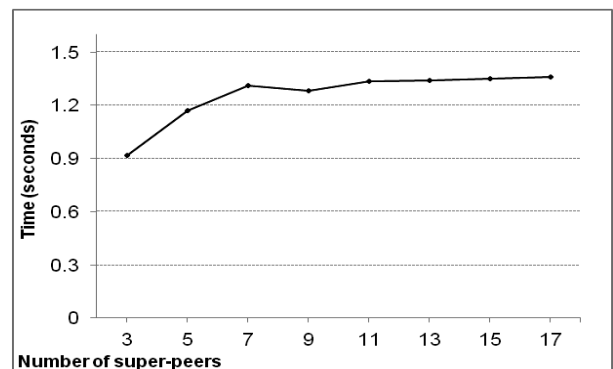


Figure 7: Ratings predication time for various numbers of super-peers.

In the third experiment, we aim to analyze execution time for SR protocol for varying set of data sizes. Therefore, we vary the minimum trust threshold to obtain a different number of participants' records in the recommendation process, then we run the SR protocol on these aggregated records in sizes of 7.249, 10.572,

12.674, 17.685, and 23.496. As shown in figure (8), the results indicate the elapsed time to perform (encrypt, calculate ratings and decrypt) with 1024 bits key length. The curve scales linearly as it represents the increase of execution time by increasing the data size.

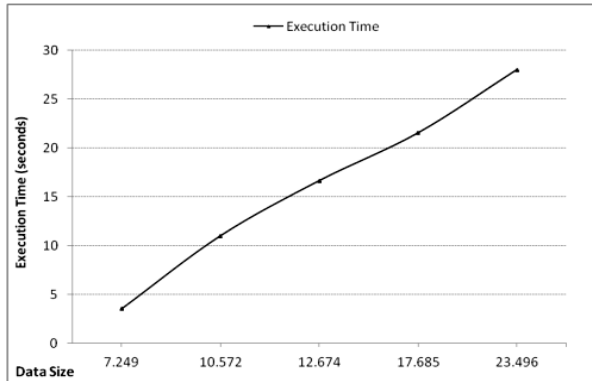


Figure 8: Execution time for different data sizes.

In the first experiment performed on *EVS* algorithm, we measured the relation between different Hilbert curve parameters (order and step length) on the accuracy and privacy levels attained. We mapped the participant’s dataset to Hilbert values using orders 3, 6 and 9. We gradually increased the step length from 10 to 80. Figure (9) shows the accuracy of recommendation based on different step length and curve order. We can see that as the order increases, the obfuscated data can offer better predictions for the ratings. Since, with higher values for the curve order, the granularity of the Hilbert curve becomes finer. So, the mapped values can preserve the data distribution of the original dataset. On the other hand, selecting larger step length increases MAE values as large partitions are formed with higher range to generate random values from it, such that these random values substitute real values in the dataset.

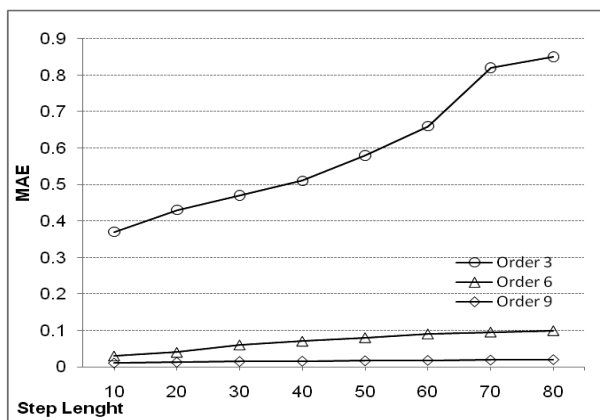


Figure 9: Accuracy level for different step length and orders for *EVS*.

As for the privacy as shown in figure (10), when the order increases a smaller range is calculated within each partition which introduces less substituted values compared with lower orders that attain higher VI values. The reason for this is that larger order divides the *m*-dimensional profile into more grids, which makes Hilbert curve to better reflect the data distribution. Moreover, we

can see that for the same Hilbert curve order the VI values are generally the same for different step length except for order 3, in which VI values has a sharp increase when step length grows from 50 to 60. The effect of increasing step length on VI values is more sensible in lower curve orders as fewer grids are formed and the increase of step length covers more portions of them, which will introduce a higher range to generate random values from it. Based on that, the trust agent employs trust value as an input to tune-up *EVS* parameters in such a way to achieve a trade off between privacy and accuracy.

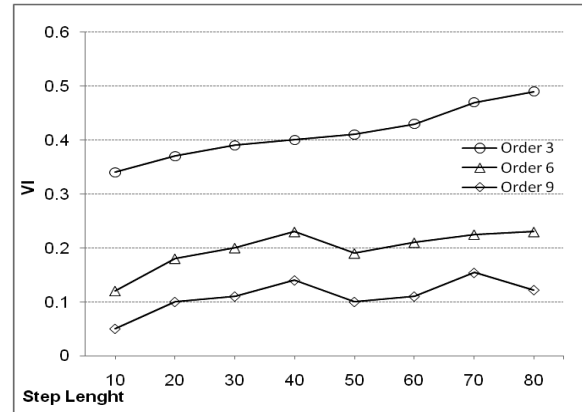


Figure 10: Privacy level for different step length and orders for *EVS*.

We continued our experiments with *EVS* algorithm; we measured the execution time for *EVS* as it is executed locally at the participant’s STB box on his profile. The execution time for *EVS* is composed of the time to get partitions based on Hilbert curve and the time to generate random noise. The results for the execution time are shown in figure (11). We can see that as the order of Hilbert curve goes higher, the execution time generally increases than that for a lower order. This growth because of the time consumed in mapping data points to different Hilbert values is dependent on curve order. For different step lengths, the executions time various without substantial trend. As the step length only determines the size of partitions in each dimension; finding these partitions are only dependant on the number of dimensions.

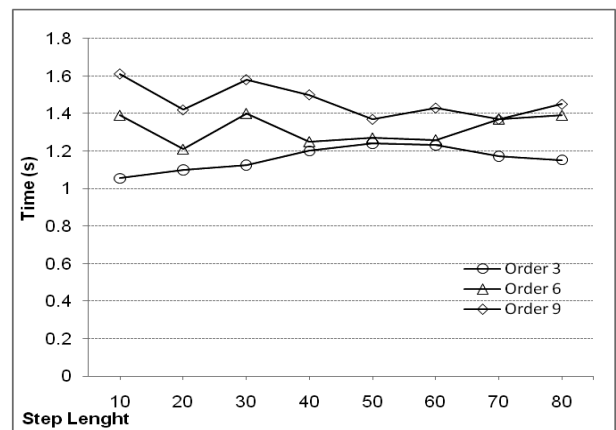


Figure 11: Execution time for different step length in *EVS*.

Finally, we measured the overall recommendation accuracy of our two stage concealment on the same dataset. For *EVS* algorithm, we set the curve order to be 3 (lowest trust level) and the step length to be 10. We first obfuscate different datasets using *EVS* algorithm, then super-peers apply *SR* protocol on these datasets and submit them to PRS. At PRS side, it calculates referrals list then return results back to super-peers which in turn decrypt and publish them. The graph in figure (12) plot MAE values for different data sizes, it clearly shows that the proposed two stage concealment process is very effective in making recommendation and that its privacy preserving nature has marginal impact on the accuracy of recommendation, since *SR* protocol employ homomorphic cryptosystem that preserves the accuracy characteristics of *EVS* algorithm on the dataset. These results indicate two features of the two stage concealment process:

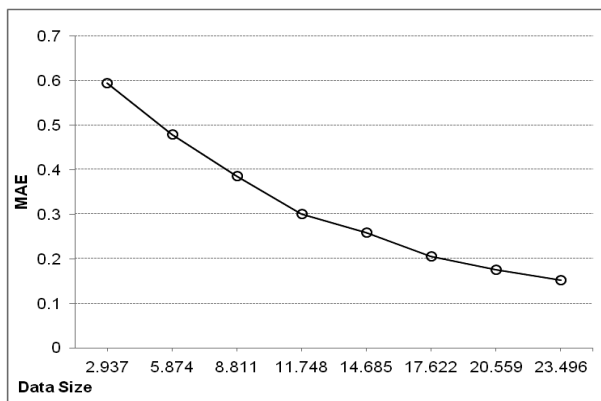


Figure 12: Accuracy of our proposed approach with for different data sizes.

1. Accuracy of the recommendation improves with the increase of collected data, as more diverse ratings produce a reasonable explanation and rank from a reliable sources.
2. Accuracy of the recommendation is reasonable for small datasets, which is highly desirable feature in a dynamic environment like IPTV networks where users' profiles are not large enough.

## 9 Conclusion and future work

In this paper, we presented our attempt to develop an enhanced middleware for collaborative privacy based on Multi-agent with application to recommender service for IPTV providers. We gave a brief overview of *EMCP* architecture, components and recommendation process. We presented a novel two stage concealment process which provides complete privacy control to participants over their preferences. The concealment process utilizes hierarchical topology, where participants are organized into groups, from which super-peers are elected based on their reputation. Super-peers & PRS use platform for privacy preferences (P3P) policies for specifying their data usage practices. While Participants describe their privacy constraints for the data extracted from their profiles in a dynamically updateable fashion using P3P policies exchange language (APPEL). *EMCP* allows fine grained enforcement of privacy policies by allowing

participants to ensure that the extracted preferences for specific request do not violate their privacy by automatically checking whether there is an APPEL preference corresponding to the given P3P policy. Super-peers aggregate the preferences obtained from underlying participants and then encapsulate intermediate values computed on these profiles and then send them to PRS. Trust based obfuscation mechanism is used in the course of participant preferences collection, while the *SR* protocol is used to protect the privacy of collaborative filtering by distributing the participants' preferences between multiple super-peers and encrypting a subset of the aggregated ratings profiles which is useful for the recommendation. We tested the performance of the proposed mechanisms on a real dataset. We evaluated how the overall accuracy of the recommendation depends on data sizes and trust level. The experimental and analysis results show that privacy increases under the proposed middleware without hampering the accuracy of the recommendation. In particular the mean absolute error can be reduced with proper tuning of the trust based obfuscation parameters for a large data sizes. Moreover, utilizing trust levels for obfuscation is an optimization to maintain the utility of the items' ratings. Thus adding the proposed middleware does not severely affect the accuracy of the recommendation based on collaborative filtering techniques.

We realized that there are many challenges in building a privacy enhanced middleware for recommender services. As a result we focused on middleware in a collaborative privacy scenario. A future research agenda will include utilizing game theory to better formulate user groups, sequential preferences release and its impact on privacy of whole profile. We will consider reducing transmission time and the load on the network traffic by adding a secure filtering phase to the *SR* protocol that will allow PRS to exclude items with low predicated rating from the final referrals list. Furthermore it is included to strengthen our middleware against shilling attacks, extending our scheme to be directed towards multi-dimensional trust propagation and distributed collaborative filtering techniques in a P2P environment. Moreover, we need to investigate weighted features vector methods and its impact on released ratings. Such that, the participant not only obfuscates his items' ratings based on the trust level of target-user, but he can also express specific items to be diversely obfuscated with each trust level. We need to perform extensive experiments on other real datasets from the UCI repository and compare our performance with other techniques proposed in the literature. Finally we need to consider different data partitioning techniques as well as identify potential threats and add some protocols to ensure the privacy of the data against those threats.

## 10 Acknowledgment

This work has received support from the Higher Education Authority in Ireland under the PRTL I Cycle 4 Programme, in the FutureComm Project (Serving

Society: Management of Future Communications Networks and Services).

## References

- [1] K. Kawazoe, et al., "Platform Application Technology Using the Next Generation Network," NTT2007.
- [2] L. F. Cranor, "I didn't buy it for myself: privacy and Ecommerce personalization," in *Designing personalized user experiences in eCommerce*, ed: Kluwer Academic Publishers, 2004, pp. 57-73.
- [3] M. d. Gemmis, et al., "Preference Learning in Recommender Systems," presented at the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML/PKDD), Slovenia, 2009.
- [4] F. McSherry and I. Mironov, "Differentially private recommender systems: building privacy into the net," presented at the Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, Paris, France, 2009.
- [5] A. Esmā, "Experimental Demonstration of a Hybrid Privacy-Preserving Recommender System," 2008, pp. 161-170.
- [6] J. Canny, "Collaborative filtering with privacy via factor analysis," presented at the Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval, Tampere, Finland, 2002.
- [7] J. Canny, "Collaborative Filtering with Privacy," presented at the Proceedings of the 2002 IEEE Symposium on Security and Privacy, 2002.
- [8] H. Polat and W. Du, "Privacy-Preserving Collaborative Filtering Using Randomized Perturbation Techniques," presented at the Proceedings of the Third IEEE International Conference on Data Mining, 2003.
- [9] H. Polat and W. Du, "SVD-based collaborative filtering with privacy," presented at the Proceedings of the 2005 ACM symposium on Applied computing, Santa Fe, New Mexico, 2005.
- [10] Z. Huang, et al., "Deriving private information from randomized data," presented at the Proceedings of the 2005 ACM SIGMOD international conference on Management of data, Baltimore, Maryland, 2005.
- [11] H. Kargupta, et al., "On the Privacy Preserving Properties of Random Data Perturbation Techniques," presented at the Proceedings of the Third IEEE International Conference on Data Mining, 2003.
- [12] B. N. Miller, et al., "PocketLens: Toward a personal recommender system," *ACM Trans. Inf. Syst.*, vol. 22, pp. 437-476, 2004.
- [13] C.-N. Ziegler, et al., "Improving recommendation lists through topic diversification," presented at the Proceedings of the 14th international conference on World Wide Web, Chiba, Japan, 2005.
- [14] J. Golbeck and J. Hendler, "FilmTrust: movie recommendations using trust in web-based social networks," in *Consumer Communications and Networking Conference, 2006. CCNC 2006. 3rd IEEE*, 2006, pp. 282-286.
- [15] A. M. Elmisery and D. Botvich, "An Agent Based Middleware for Privacy Aware Recommender Systems in IPTV Networks," in *Intelligent Decision Technologies*. vol. 10, J. Watada, et al., Eds., ed: Springer Berlin Heidelberg, 2011, pp. 821-832.
- [16] A. Elmisery and D. Botvich, "Private Recommendation Service For IPTV System," in *12th IFIP/IEEE International Symposium on Integrated Network Management*, Dublin, Ireland, 2011.
- [17] A. Elmisery and D. Botvich, "Agent Based Middleware for Maintaining User Privacy in IPTV Recommender Services," in *3rd International ICST Conference on Security and Privacy in Mobile Information and Communication Systems*, Aalborg, Denmark, 2011.
- [18] A. Elmisery and D. Botvich, "Privacy Aware Obfuscation Middleware for Mobile Jukebox Recommender Services," in *The 11th IFIP Conference on e-Business, e-Service, e-Society*, Kaunas, Lithuania, 2011.
- [19] A. Elmisery and D. Botvich, "Privacy Aware Recommender Service for IPTV Networks," in *5th FTRA/IEEE International Conference on Multimedia and Ubiquitous Engineering*, Crete, Greece, 2011.
- [20] A. Elmisery and D. Botvich, "Enhanced Middleware for Collaborative Privacy in IPTV Recommender Services " *Journal of Convergence*, vol. 2, p. 10, 2011.
- [21] [21] W. Nejdl, et al., "Super-peer-based routing and clustering strategies for RDF-based peer-to-peer networks," presented at the Proceedings of the 12th international conference on World Wide Web, Budapest, Hungary, 2003.
- [22] [22] J. Carbo, et al., "Trust management through fuzzy reputation. Int," *Journal in Cooperative Information Systems*, vol. 12, p. 135—155, 2002.
- [23] [23] H. D. Kim, "Applying Consistency-Based Trust Definition to Collaborative Filtering," *KSII TRANSACTIONS ON INTERNET AND INFORMATION SYSTEMS*, vol. 3, pp. 366-374, 2009.
- [24] [24] A. Elmisery and D. Botvich, "Agent Based Middleware for Private Data Mashup in IPTV Recommender Services," in *16th IEEE International Workshop on Computer Aided Modeling, Analysis and Design of Communication Links and Networks*, Kyoto, Japan, 2011.
- [25] [25] D. Kelly and J. Teevan, "Implicit feedback for inferring user preference: a bibliography," *SIGIR Forum*, vol. 37, pp. 18-28, 2003.
- [26] [26] P. Indyk and R. Motwani, "Approximate nearest neighbors: towards removing the curse of dimensionality," presented at the Proceedings of the thirtieth annual ACM

- symposium on Theory of computing, Dallas, Texas, United States, 1998.
- [27] P. Paillier, "Public-Key Cryptosystems Based on Composite Degree Residuosity Classes."
- [28] I. Damgård and M. Jurik, "A Generalisation, a Simplification and Some Applications of Paillier's Probabilistic Public-Key System Public Key Cryptography." vol. 1992, K. Kim, Ed., ed: Springer Berlin / Heidelberg, 2001, pp. 119-136.
- [29] I. Damgård and M. Kopolowski, "Practical Threshold RSA Signatures without a Trusted Dealer Advances in Cryptology — EUROCRYPT 2001." vol. 2045, B. Pfitzmann, Ed., ed: Springer Berlin / Heidelberg, 2001, pp. 152-165.
- [30] D. Boneh and M. Franklin, "Efficient generation of shared RSA keys Advances in Cryptology — CRYPTO '97." vol. 1294, B. Kaliski, Ed., ed: Springer Berlin / Heidelberg, 1997, pp. 425-439.
- [31] G. Ghinita, et al., "PRIVE: anonymous location-based queries in distributed mobile systems," presented at the Proceedings of the 16th international conference on World Wide Web, Banff, Alberta, Canada, 2007.
- [32] A. Reaz and B. Raouf, "A Scalable Peer-to-peer Protocol Enabling Efficient and Flexible Search," ed, 2010.
- [33] R. A. Jarvis and E. A. Patrick, "Clustering Using a Similarity Measure Based on Shared Near Neighbors," IEEE Trans. Comput., vol. 22, pp. 1025-1034, 1973.
- [34] K. Liu, et al., "An Attacker's View of Distance Preserving Maps for Privacy Preserving Data Mining Knowledge Discovery in Databases: PKDD 2006." vol. 4213, J. Fürnkranz, et al., Eds., ed: Springer Berlin / Heidelberg, 2006, pp. 297-308.
- [35] P.-A. Fouque, et al., "CryptoComputing with Rationals Financial Cryptography." vol. 2357, M. Blaze, Ed., ed: SpringerBerlin / Heidelberg, 2003, pp. 136-146.
- [36] D. Gupta, et al., "Jester 2.0 (poster abstract): evaluation of an new linear time collaborative filtering algorithm," presented at the Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, Berkeley, California, United States, 1999.
- [37] J. L. Herlocker, et al., "Evaluating collaborative filtering recommender systems," ACM Trans. Inf. Syst., vol. 22, pp. 5-53, 2004.
- [38] C. Kingsford, "Information Theory Notes," 2009.

# An RTSP Proxy for Implementing the IPTV Media Function Using a Streaming Server

Zelalem S. Shibeshi, Alfredo Terzoli and Karen Bradshaw  
 Department of Computer Science  
 Rhodes University, P.O. Box 94, Grahamstown 6140  
 Tel: +27 46 6038247, Fax: +27 46 6361915  
 E-mail: zelalems@rucus.ru.ac.za; {A.Terzoli, K.Bradshaw}@ru.ac.za

**Keywords:** IPTV, RTSP Proxy, IPTV media function, IPTV services

**Received:** October 15, 2011

*Multimedia content delivery in IMS, including IPTV, is handled by a separate unit, the Media Function (MF), made up of media control and media delivery units, which in the case of IPTV are the Media Control Function (MCF) and Media Delivery Function (MDF), respectively. According to the different specifications of an IMS based IPTV architecture, the User Equipment (UE) is expected to use the RTSP protocol as a media control protocol to interact with the MCF, and obtains delivery of media from the MDF using the RTP protocol. This also means that the streaming session needs to be initiated from the media controller on behalf of the user but the delivery of media is sent to the UE from the media deliverer (media server). Due both to the lack of free and open source Media Servers and the availability of free and open source Streaming Servers, the ideal choice for the delivery of multimedia services, including IPTV, by the research community is Streaming Servers. Nevertheless, because of denial of service attacks and other issues, most streaming servers do not allow a different location for the session setup request and the delivery of media in the streaming session. In other words, most streaming servers are not designed to be controlled by some other entity other than the RTSP client that consumes the media. This makes it difficult to have a separate media control unit for IPTV service in IMS if one wanted to use a streaming server as an MDF unit. So, while waiting for streaming servers to work in this manner, it is better to find a work around in order to use streaming servers to develop and test IPTV services in IMS environments. For this purpose we propose another component (an RTSP proxy and relay unit) as part of the IPTV MF and to mediate between the MCF and MDF. This unit correctly relays media control commands from the MCF to the MDF and RTP packets from the MDF to the UE. It also helps in the implementation of other streaming functionalities that are required for IPTV service delivery, but which are not implemented in the current open source streaming servers. Additional services can also be easily implemented with the help of this unit. This will facilitate the development of an IPTV service using readily available open source streaming servers and help researchers to evaluate their proposals on new services they would like to develop. In this paper we show how this RTSP proxy unit can be integrated into the Media Function of the IPTV architecture to ease the media delivery process of an IMS based IPTV service.*

*Povzetek: Članek predstavi posredniški strežnik za televizijo IP, ki uporablja standard RTSP.*

## 1 Introduction

The popularity of YouTube and other Internet based video services shows the potential of these services in the telecom world<sup>1</sup>. Companies involved in video service delivery are reaping the benefits of this huge demand. According to [2], the growth of on-line video spending surpassed \$2.12 billion in 2008, up 36% from 2007 and has been forecast to continue double-digit increases through the years to come. A recent report on “The Mobile TV Market” from the ABI Research Group also revealed that the mobile TV market has tremendous long-term promise as a next-generation infotainment

experience and will grow to a value of more than \$50 billion by 2013 [3]. On the other hand, users are moving beyond viewing short, low-quality clips of user-generated content on YouTube and increasingly seeking out TV shows, films, and other professionally created, high-quality video content. Nevertheless, because of the inherent characteristics of the Internet, quality of service (QoS) cannot be guaranteed with Internet based services, and here lies the advantage for Telecoms to engage themselves in the delivery of video oriented services. As video is one service that requires large bandwidth, users will be even keener if they can obtain the service with the required quality of service. IMS (IP Multimedia System), as an implementation of NGN (Next Generation Networks), provides the required QoS for users and is the right environment in which to deliver the IPTV service.

<sup>1</sup> This work is being carried out in the Distributed Multimedia Centre of Excellence at Rhodes University, with financial support from Telkom, Stortech, Tellabs, Amatole Telecom Services, Bright Ideas 39, and THRIP.

Apart from granting users the ability to access their services using different devices and access technologies, the major goal of IMS is the delivery of multimedia services, like IPTV. There are different proposals of implementation standards for IMS by different standard bodies, each with particular emphasis on a specific service type. The major ones include: the 3rd Generation Partnership Project (3GPP) [4], European Telecommunications Standards Institute (ETSI) Telecommunications and Internet Converged Services and Protocols for Advanced Networking European Telecommunications Standards Institute (TISPAN) [5], and the Telecommunication Standardization Sector of International Telecommunications Union (ITU-T) [6]. Similarly, there are also different specifications for the delivery of IPTV, again from different bodies. However, the specification that has received the greatest interest from the research community for the development and testing of IPTV services is the one proposed by ETSI-TISPAN [7]. As such, we have adopted their standard in this paper.

Multimedia session delivery involves the use of a session control protocol to control the session and a media control protocol to control the media delivery. The media delivery in a standard IMS architecture, for example, is carried out by what is known as the Media Resource Function (MRF), consisting of two distinct parts, namely the MRFC (MRF Controller) and MRFP (MRF Processor). The IPTV specification also has a similar component for media delivery and control, which is called IPTV Media Function (MF). The control unit of the IPTV MF is the Media Control Function (MCF) and the delivery unit is the Media Delivery Function (MDF). The media delivery unit, MDF, is supposed to be implemented by a fully-fledged Media Server. However, because of the lack of free and open source media servers, researchers mostly use open source streaming servers to develop and test media services.

In addition to the media delivery and control units, IMS services, like IPTV, are controlled by a service controller unit, which in the case of IPTV is the IPTV Service Control Function (SCF). Basically, this unit is a SIP application server (AS). So, if a user knows the service description of a given IPTV service, s/he contacts the SCF to obtain the desired service. The SCF, in turn, will contact the MCF to initiate the delivery of media. The MCF then initiates the media delivery by instructing the MDF to send the requested stream directly to the user. In general, the MCF initiates the media request on behalf of the user and the media server, MDF, delivers the stream to the user (not to the initiator of the session). As all media requests pass through the MCF, this means that if one were to follow the specification directly, all media requests including session initiation should pass through MCF.

As mentioned above, open source streaming servers are being used for the delivery of streaming media by IPTV researchers. However, most streaming servers do not allow the delivery of media to a destination that is not the client that initiated the streaming session. For this reason, researchers tend to combine the MCF and MDF

units into one unit (the streaming server) and initiate the media delivery and control from the UE, instead of from the MCF. The most popular open source IPTV application server used by the research community is the one developed by researchers at the University of Cape Town (UCT) [8]. Because of the problems with the current open source streaming servers mentioned above, the application server has been developed in such a way that the UE sends media requests directly to the streaming server without any involvement by the MCF (which is contrary to the specification). The UCT IPTV AS has become the de facto standard for developing IPTV services by the research community. Nevertheless, because the UE directly contacts the streaming server, neither the AS nor the MCF has control over the session and thus, it would be difficult to develop services that involve media session control. The Convergence Research Group at Rhodes University also makes use of the UCT IMS client [8] to test IPTV services and has followed this approach all along. This, however, is not in accordance with the specification of the IPTV service. Actually, not only does it contradict the specification, but it also does not allow the control of media to be done by a controlling unit because the MCF is not involved in the media setup process. A recent article by researchers from UCT [25] referred to the development of a media server that can work with the MCF, but it is currently not available to the research community. As a result, a work around is required if streaming servers are to be used for media delivery.

In this paper, we describe in detail how including the new lightweight component introduced in [1], that is, the Streaming Server Proxy and Relay (SSPR) unit, in the IPTV MF can help a service developer to use streaming servers and also assist the development of advanced IPTV services. The paper also introduces new functionalities such as “media switching” and “bookmarking” that are included in the proxy. This new unit is integrated into the MF to overcome the problem mentioned above. The remainder of this paper is organized as follows. Section 2 gives background information. Section 3 describes related works, while Section 4 explains the new proposed architecture of an IPTV service. Section 5 presents implementation and discussion, and finally, Section 6 gives our conclusions and future work.

## 2 Background

As mentioned in the Introduction, the IPTV architecture proposed by ETSI-TISPAN is the one followed by many researchers. We also use this architecture to present and discuss our proposal. The major components of this architecture are: Service Discovery and Selection Functions (SDF and SSF, respectively), Service Controller Function, and Media Control and Delivery Functions. Figure 1 shows these functional units of the IMS-based IPTV architecture as proposed by the ETSI-TISPAN standards body.



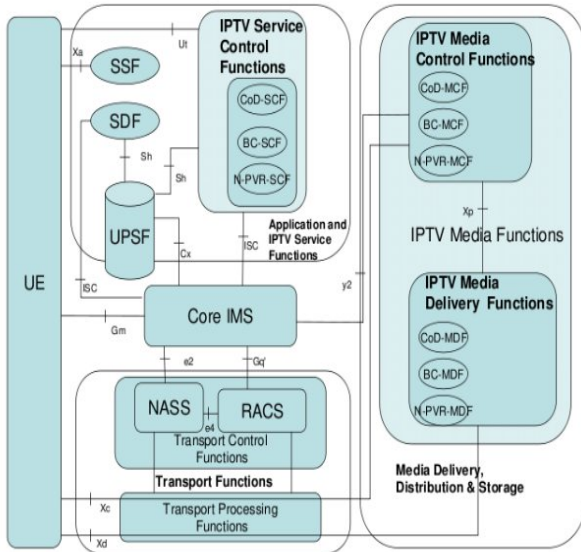


Figure 1: Functional Architecture of IPTV.

Table 1 gives the protocols used by the different reference points or interfaces.

The specification also makes it clear that the UE uses RTSP commands to communicate with the MCF for the purpose of media control. With regard to setting up media for an IPTV service, the specification specifies two methods for media initiation, referred to as Method 1 and Method 2. The distinction relates to where the RTSP session setup commands (specifically DESCRIBE and SETUP) are initiated from. In the Method 1 session setup technique, the session is initiated and the setup originates from the MCF. However, in Method 2, it is initiated by the UE, but the UE sends all RTSP commands (including session initiation and setup) to the MCF and not to the MDF. So even with the media initiation of Method 2, the media initiation request should be sent to the MCF.

| Interface          | Protocol used        |
|--------------------|----------------------|
| ISC                | SIP/SDP              |
| X <sub>d</sub>     | RTP/RTCP             |
| X <sub>a</sub>     | HTTP/DVBSTP or Flute |
| sh, C <sub>x</sub> | Diameter             |
| X <sub>c</sub>     | RTSP                 |
| U <sub>t</sub>     | HTTP                 |
| X <sub>p</sub>     | not defined          |

Table 1: Protocols used on reference points (adapted from ETSI TS 183063 V3.5.2).

In the following, we describe the steps taken by the UE to access an IPTV service using Method 2:

- The UE, like any IMS client, has to register with the IMS core before it attempts to request any service. After registration, the first step in accessing the IPTV service is the identification and selection of the service that the user desires. This is done by contacting the Service Discovery and Selection Function (SDF and SSF, respectively) units.
- Service discovery, also called service attachment, is accomplished by contacting the SDF, which provides information about the user's IPTV services and where the user can select the services. Basically,

this is information about the address of the service server or portal that will provide the user with a description of the available service. In general the service attachment information consists of SSF addresses in the form of URIs and/or IP addresses.

- Once the UE obtains the service description, it contacts the SSF to retrieve relevant information about the IPTV service, like the URL of the media (content identifier), to initiate the IPTV session.
- After a service has been selected, the relevant content identifier is inserted in the SIP session initiation message sent to the IPTV Service Control Function (SCF) that provides access to this service. The UE does this by sending an INVITE request to the IMS core.
- The IMS core then forwards the request to the SCF that is responsible for controlling the service.
- The SCF then performs service authorization and credit control, selects the relevant IPTV media control function, and forwards the request to the MCF that is responsible for controlling the media for this particular user.
- The MCF is responsible for initiating the media session by contacting the MDF that is supposed to serve the particular user. Once the media is set up correctly, the MCF notifies the UE of the status of the media session and the UE can send the RTSP PLAY media control command to start the media session.
- The delivery of media then starts from the MDF to the UE.

All media control commands from the UE are sent to the MCF, which then forwards the request to the MDF using the media control protocol. The media control protocol that the MCF uses to control the MDF is not specified in the specification and it is up to the implementers to choose an appropriate protocol.

As mentioned before, both media initiation (SETUP) and other media control commands are handled by the MCF, but the media is sent directly to the UE through the X<sub>d</sub> reference point (see Fig. 1). For the MDF all media initiation requests come from the MCF. This also implies that the MDF should be able to handle media requests from a different location other than the UE and deliver the media to the UE.

The RTSP protocol [9] specifies a “*destination*” parameter that needs to be used in the transport section of a SETUP request to set up the destination of the media. The different versions of RTSP specification refer to this parameter by different names. Version 1.0, for example, specifies it as “*destination*”, while version 2.0 of the RTSP protocol specification, which is an Internet draft, specifies it as *dest\_addr* [10]. This parameter (field) needs to be included in the transport section of each SETUP request for which a different destination is needed. If the server supports this feature, it then sends the media to the specified destination when the media delivery begins, but continues to send the RTSP responses to the location that initiated the RTSP session. This could have been used by the MCF to initiate an

RTSP session from streaming servers on behalf of the user; however, based on our investigation of the available open source streaming servers, VLC [11], Darwin Streaming Server (DSS) [12], and the Mobicents Streaming Server [13] do not support the use of this parameter. Darwin returns an “Invalid Code” error code, while VLC and the Mobicents Streaming Server just ignore it and continue sending the media to the media session initiator. Live555 [14], on the other hand, allows the use of this parameter (using the name “destination”) by modifying the RTSPServer.cpp file. Because of the possibility of denial of service attacks, this feature is disabled by default but can be enabled by inserting a “#define

RTSP\_ALLOW\_CLIENT\_DESTINATION\_SETTING 1” statement at the beginning of the above mentioned file. Nevertheless, Live555 only plays video files that are encoded with the MPEG video codec, which is not supported by most UEs because it is not the default codec suggested by the IPTV specification. Consequently, we cannot make use of this media server either and the only option left to enable the use of a streaming server as the MDF is to include an RTSP proxy. The work presented in this paper aims to solve this problem.

The problem with open source streaming servers is not only related to the support for “destination” parameter, but also there are other features that IPTV services require, but the current open source streaming servers do not support. A bookmarking service, for example, requires that the current position of the media that is being played be recorded and kept together with detailed media information. As a result, the application server needs to request the media server to obtain this information in order to store bookmark information of the media that is being played. The RTSP protocol has a command that can be used for this purpose. The specification defines a *get\_parameter* command for the purpose of querying a streaming server to obtain media related information including the current play time. Specifically one can use this command together with a *range* parameter to obtain the current media position. In fact, the Open IPTV Forum (OIPF) also suggests the use of this parameter for the purpose of bookmarking. However, both VLC and DSS do not support this. DSS responds with a 500 error code, while VLC again merely ignores it. Consequently, we have implemented this functionality in the proxy, with the details given in the Implementation section.

### 3 Related Work

Various researches have been carried out on the media processing aspects of IPTV, particularly with regard to the type of media control protocol to be used for IPTV. In this regard, an evaluation of SIP for the use of streaming control instead of the RTSP protocol has been presented in different IETF Internet drafts [15][16][17]. Taking this idea a bit further, various researchers have also reported their experiences with regard to implementing SIP as a media control protocol. The

authors in [18] showed how a new SIP header (called SIP-MEX) and new SIP bodies (an XML document in the SIP INFO message) can be used to send media control commands to the MCF. On the other hand, other researchers have also suggested the integration of SIP and RTSP to create a comprehensive media control protocol [19]. However, as mentioned in [20], to avoid the IMS signaling procedures causing extra delays, it is always necessary to define a clear separation between service/session control performed at IMS level and media flow control handled end-to-end between user equipment and the content service. Actually, this could be one of the reasons that ETSI-TISPAN proposed a different media control protocol other than SIP in the standard specification. In general, those who have proposed SIP as a media control protocol have tried to justify their proposal from the point of view of media control requirements that cannot be handled by RTSP and also for the purpose of handling bandwidth reservation requests and responses. Nevertheless, as to the support of bandwidth negotiation, the IETF has developed extensions to SDP [21] and it should no longer be a problem to use RTSP. In general, both approaches have their own strengths and weaknesses, but these are not considered in detail here, as this is beyond the scope of this paper. However, the advantage of using RTSP as a media control protocol is that the MCF is not required to translate media control commands received from the UE when it forwards them to the MDF.

On the other hand, with regard to having a separate MCF and MDF, some researchers have also proposed the integration of the service selection function with the media function. In [22], for example, the authors proposed a comprehensive service function, called the Multimedia Service Control Function (MSCF). The MSCF combines the functionality of the SDF, SSF and MF functions of the IMS based IPTV units. The authors also proposed a Media Distribution Function consisting of three components, namely, Interconnection (similar to the IPTV MDF), Serving (IPTV MDF), and Primary (IPTV MDF), abbreviated as I-IMDF, S-IMDF, and P-IMDF, respectively. According to the authors, the function of the P-IMDF is to serve as the primary contact point, and also to handle the streaming function.

The concept of an RTSP gateway is also presented in [23], where the authors proposed a gateway that converts RTSP messages to SIP messages and vice versa. On the other hand, the use of an RTSP proxy for the delivery of streaming service for UEs without RTP support is presented in [24].

As mentioned before, researchers tend to use streaming servers for media delivery in IPTV services. However, with regard to the implementation of a proper IPTV media function, Ref. [25] discusses an initiative for the development of the UCT IPTV testbed and mentions the current work on the MCF and MDF. The authors have not, however, clearly specified what media control protocol they used, nor explained how the MDF is implemented. The UCT IPTV client and AS are very popular open source IMS components in the research community, but this particular project was new and not

available at the time of conducting the research presented in this paper. As a result, until the issue of the media control protocol settles down, and an open source IPTV MF is commonly available to the research community, we hope that our proposal will be helpful to researchers wishing to develop an IPTV service using streaming servers particularly as it conforms to the specification. In fact, the proxy also implements new functionalities such as easy media switching and bookmarking services for use by service developers.

### 4 Proposed Architecture

The aim of this paper is to describe how streaming servers can be used as an MDF unit by incorporating the proxy explained below.

The proposed architecture is basically the same as the TISPAN architecture presented in Fig. 1, except that the SSPR unit is added within the MF. Thus, the focus of this section is on the MF unit.

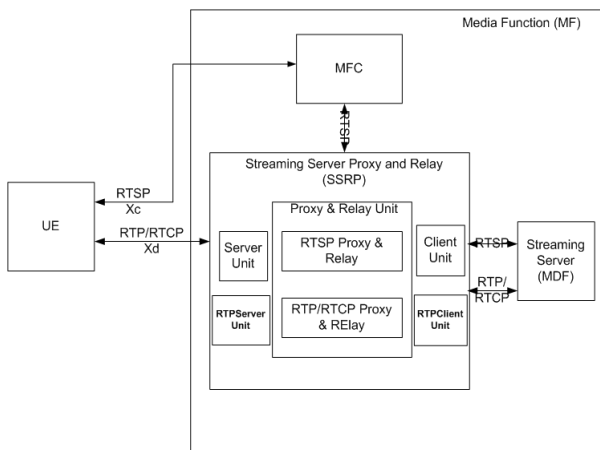


Figure 2: Block diagram of the modified Media Function.

As can be seen from Figure 2, which shows a block diagram of the modified Media Function, the SSPR has five main components or units: the proxy and relay, server, and client units. The server and client units are responsible for handling RTSP traffic. The proxy and relay unit also has two distinct components: the RTSP proxy and relay, and the RTP/RTCP proxy and relay units. We have also RTPServer and RTPClient units which are responsible for relaying RTP/RTCP packets from the server to the client and also back to the server. The following paragraphs briefly describe the function of each of these units.

- The server unit handles all RTSP requests coming from the MFC. Upon receipt of an RTSP request, it forwards the request to the proxy and relay unit, which is responsible for forwarding the request to the client unit.
- The client unit acts like an RTSP client to the streaming server and sends the request that it receives from the proxy and relay unit to the streaming server, and also forwards the response it receives from the server back to the proxy and relay unit.

- The RTPServer unit sends RTP/RTCP packets to the client and also sends RTCP packets back from the client to the server unit. It communicates with the proxy and relay unit to do this.
- The RTPClient unit relays RTP/RTCP packets that come from the streaming server to the client through the proxy and relay unit. It also forwards RTCP packets that come from the client to the server unit.
- The proxy and relay unit is responsible for forwarding requests from the server units to the client unit and also forwards responses from the server unit to the client unit. It also relays RTP/RTCP packets from the streaming server to the client and vice versa. The proxy and relay unit must change the request that comes from the client (MCF) so that the streaming server can return the responses and media delivery to it. For this purpose, it records the address information of the client and generates or uses its own address before forwarding the request to the client unit. It does the same thing when forwarding responses that come from the streaming server back to the client (MCF).

Basically, for the streaming server (MDF), the request comes from the SSRP and the response is also sent back to the SSRP. As a result, the problems mentioned in the previous section do not arise in this scenario. The SSRP

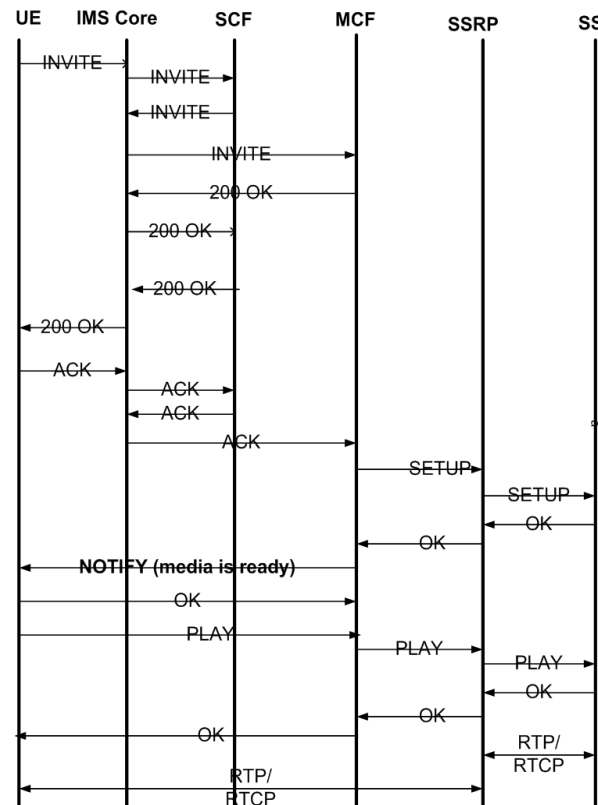


Figure 3: IPTV service access in the proposed architecture (using media access according to Method 1).

is designed to manage streaming sessions and can also handle the proper proxy and relay functions to process stream control commands and deliver the stream. The flow diagram in Fig. 3 shows the IPTV service initiation and access using this architecture.

## 5 Implementation and Discussion

### 5.1 Implementation

The Open IMS Core testbed from the *Fraunhofer Institute FOKUS* [26] is the most popular IMS testbed in the research community. The Convergence Research Group at Rhodes University has been using this testbed to develop and test IPTV services. We also used this testbed to test the functionality of the proxy. Regarding an IMS user agent, we used the UCT IMS client, described earlier. The client is designed to work with Method 2 of the IPTV media access methods. As a result, to avoid excessive work on the client side we used the client with this setting. The client has the capability of sending all RTSP commands.

Since there is no open source Service Discovery and Selection component, we learn from others' experiences and used the technique provided in [27] to deliver the URL of the media to the UE. As mentioned in the paper, the AS upon receipt of an INVITE from UE, includes the URL of the media in the SIP OK message that is sent to the UE. Although we adopted this technique, in our case we transferred this functionality to the MCF, instead of the AS. As a result, as can be seen in Fig. 4, upon receipt of an INVITE from the UE, the AS forwards the INVITE to the MCF and the MCF then matches the requested channel to a URL and sends it to the UE including the URL in the SIP OK message. If there is no MCF, the approach taken is that after establishing the SIP session with the AS, the UE then sends the DESCRIBE and SETUP commands to the streaming server to initiate and set up an IPTV media session. However, this time around, the client sends these commands to the MCF instead of the streaming server. When the MCF gets the RTSP command from the UE, it forwards the request to the RTSP proxy. The MCF uses the *destination* parameter of the RTSP protocol, discussed in an earlier section, to pass on the destination of the media, i.e., the address of the UE. This parameter is included in the SETUP command of the request. The MCF obtains client address information from the SDP payload of the SIP INVITE command. Accordingly, the proxy forwards the request to the streaming server and upon receipt of the media, delivers it to the UE. The proxy also uses a configuration file to obtain information about the streaming server, such as its address and port.

The proxy has different handlers on the client and server sides for both the RTSP and RTP/RTCP sessions. On the client side, the RTSP session is created with the MCF while the RTP/RTCP session is created with the client (UE). On the streaming server side, both sessions (RTSP and RTP/RTCP) are created with the streaming server.

Session handling is one important aspect of media servers. An RTSP session initiation request (for example, DESCRIBE) may not necessarily end up in an RTSP session. The client may not be able to play the media (video) if it does not support the codec that the media is encoded in. As a result, even though there is an I/O (network) session between client and server, an RTSP session is basically created when the client sends a SETUP request to the server. This tells the server that the client can play the media and the server generates and sends a unique session ID within the response. Both the client and server use this id to refer to the session in subsequent communication. The proxy is also designed in a similar manner. As is the case with any proxy system, a session that is supposed to be established between a client and a server is divided into two sessions: one on the client side and another on the server side. Similarly, if we consider the RTSP session, we have two separate RTSP sessions (one on the client side – MCF to proxy's client side and another from the server side – proxy's server side to streaming server). Similar to the RTSP principle of establishing an RTSP session mentioned above, a proxy session object is created when the client sends a SETUP request to the server. The proxy then creates a unique random session ID for the client side RTSP session and records it in the proxy session object that it created for this particular session. When a response comes from the server with the server's session ID, that session ID is also recorded in the proxy

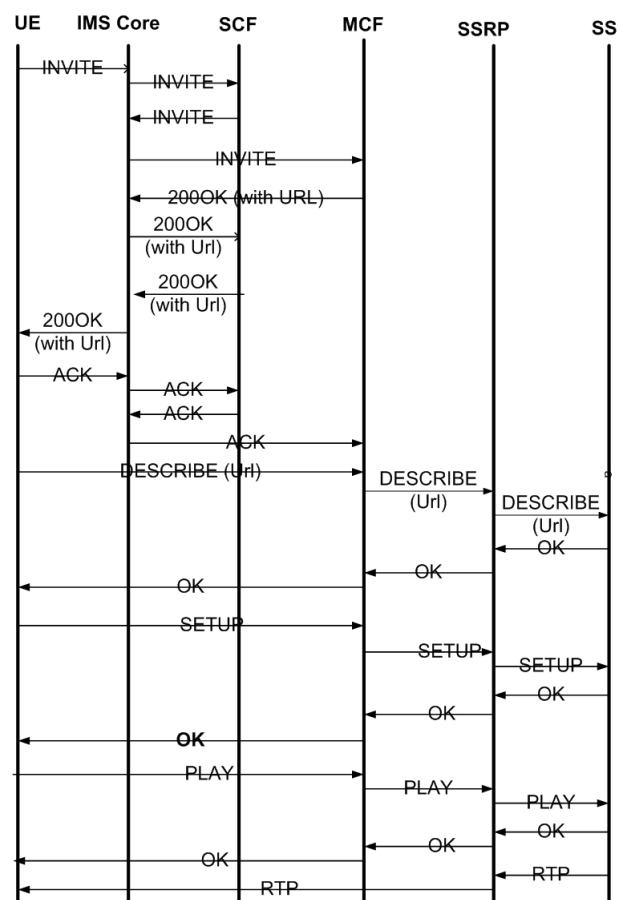


Figure 4: IPTV service access in the proposed architecture (using media access according to Method 2).

session object. In other words, the proxy session contains the client and server RTSP sessions. The proxy session object also contains a track list.

Multimedia sessions may contain more than one media (track). For this purpose we also defined an object called a Track object that contains detailed information about a track, including client address, server address, client RTP/RTCP ports, server RTP/RTCP ports, and client and server RTP/RTCP sessions, which are useful for forwarding requests and responses from the client to the server and vice versa. The track object has methods to forward requests to the server and responses to the client. The track class contains a hash map of the different track objects related to different sessions. As a result, a particular track object is identified using the session id of a request or response.

In a similar way to RTSP sessions, RTP sessions are also identified by unique ids called the SSRC (Synchronization Source) identifier. As a result, the proxy also creates a “proxy SSRC” id to identify the RTP/RTCP session between the proxy and the client (MCF). The server sends its own SSRC id when it starts sending RTP packets. The RTPClient unit is responsible for handling the RTP/RTCP packets that come from the server and forwards them to the client (MCF) through the proxy and relay unit. Similarly the RTCPServer unit handles the relay of RTP and RTCP packets to the MCF. The proxy modifies the SSRC id before forwarding the RTP/RTCP packets to the MCF. For the client (MCF), the RTP/RTCP packets come from the proxy.

The ProxySession class also has a hash table to match client and server RTP/RTCP sessions together with the proxy object mentioned before. In general, RTSP sessions are identified using the “Session ID” and RTP/RTCP sessions are identified using the “SSRC id”.

Advantages of media sessions being handled by the proxy can be seen in the creation of session related services. For example, if a media switch is requested, an efficient way of doing this would be to use the existing connection on the client side and create a new connection on the server side. One advantage could be to continue feeding media to the user until the new media setup is ready on the server side. Another advantage is the reduction of processing time because there is no session setup on the client side. In addition to this, interesting services like “Switch with Pause” can be developed with this type of approach. Switch with Pause involves pausing the current media and switching to the new media. Once the new media finishes, the proxy can resume playing the previous media. In general, using this technique the proxy creates different RTSP and RTP/RTCP sessions for the new media on the server side by sending and receiving the RTSP requests/responses itself automatically until the media setup is ready. When the new media setup has been completed, it uses the same session on the client side to deliver the media. In other words, a new connection is only created on the server side (from the proxy to the streaming server.)

Figure 5 shows a proxy session containing both active and paused sessions. The “media switch” command is defined and sent to the proxy using the

RTSP OPTION command. This command is extended by defining a field named “switch” that can take parameters like “immediately” or “number of seconds” after which the switch is sought. The URL of the media to be switched is also included in the RTSP OPTION command.

We have also implemented a bookmarking feature in the proxy. The RTP protocol includes a feature whereby each packet contains a timestamp of the packet being delivered. The timestamp is the sampling frequency of the packet being delivered relative to a running clock. As a result, the timeframe of the first packet does not start from zero and it is always good to record the first timestamp as a reference point to obtain the relative position of future packets. So, having recorded the start time, we obtain the final time (when bookmarking is requested) and subtract the two to get the difference. We divide this by the clock rate of the media, which is included in the SDP of the media. This information is recorded in the Track class discussed in a previous section.

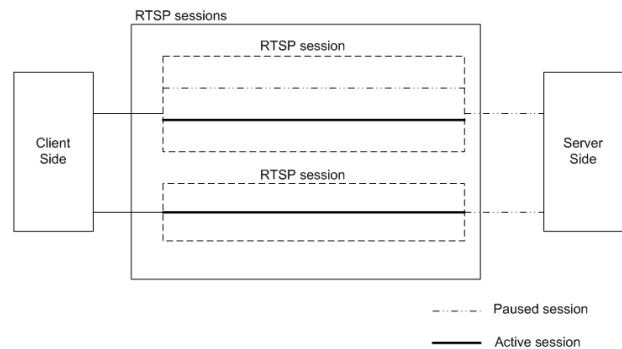


Figure 5: Proxy session handling.

As mentioned above, the RTSP proxy and relay component can be easily extended to include a variety of functions required by IPTV developers.

The RTSP proxy and relay is implemented using the Java programming language, and particularly the Apache MINA framework [28]. The MCF unit is implemented as a SIP AS and the RTSP proxy and relay is included as a separate class within the MCF and initiated from within the MCF.

## 5.2 Discussion

We tested the proxy on a dual Core Intel 2.66 GHz PC with 2 GB RAM. The machine also ran the streaming server. We used the Java System API (the nanoTime() method) to record the delay introduced by the proxy. The time was recorded when the proxy received a request and also when the same request left the proxy to the streaming server. The difference was calculated to determine the delay through the proxy. The same approach was used for the responses. The average delay introduced by the proxy was found to be negligible (close to 40 nano seconds). As a result, we believe that the proxy can be considered a good solution for those who wish to use a streaming server for media delivery in IMS based IPTV services, as it does not introduce much delay into the whole system.

As mentioned previously, among the many advantages of having a separate RTSP proxy in place is the ability to change media within an existing session (e.g., inserting an advert), which can be done easily without the involvement of the UE. This is important functionality because the UE does not require a different connection to obtain the new stream as the proxy can handle that itself. The AS can initiate the modification of media within an existing session based on different sets of rules and the proxy can deliver the modified stream without involving the UE. We believe this will help researchers to implement new and innovative services that can be implemented by any standard UE. Because of the nature of the RTSP protocol, in the event that the media source disappears for whatever reason, there is no way that the MCF can know about the situation. On the other hand, if we use a proxy, because the proxy also handles RTP packets, it knows if the session is alive or dead and can take the appropriate action immediately when a problem arises.

The proxy can also be easily extended to include other features. For example, a transcoder unit can be integrated into the proxy and it can be used to carry out transcoding based on the capabilities of the UE, if such functionality is required.

## 6 Conclusion and Future Work

The IPTV research community mainly uses streaming servers for the delivery of media for IPTV services. On the other hand, the IMS-based IPTV specification specifies that a streaming session is initiated by the MCF on behalf of the user. Nevertheless, most open source streaming servers do not allow the initiation of a streaming session by a different client to the RTSP client intending to consume the stream. To overcome this limitation, a proxy as described in this paper, can be used as a work around thereby enabling researchers to use the available streaming servers while adhering to the standard. The RTSP proxy can be integrated into the MCF or be deployed as a separate entity. According to timing experiments conducted, the proxy does not introduce a significant delay to the service delivery process and as such the authors believe it to be a good solution. In addition to allowing the use of streaming servers as MDF units for the delivery of IPTV services, the paper also presented some of the additional benefits arising from use of the proxy.

As a future work we plan to include a transcoding capability in the proxy so that the stream can be transcoded or translated on the fly based on devices' capabilities.

We also intend extending this work to include load balancing functionality in the proxy, so that the proxy can choose different streaming servers based on their status. The proxy will also be packaged as an API by abstracting the streaming servers and providing interfaces that service developers can use.

## References

- [1] Z. S. Shibeshi, A Terzoli, K Bradshaw (2010). Using an RTSP Proxy to implement the IPTV Media Function via a streaming server. In ICUMT'10: Proceedings of the International Congress on Ultra Modern Telecommunications and Control Systems. Moscow, Russia, 18 to 20 October 2010
- [2] Accustream media research. Online Video Spend Tops \$2.12B in 2008. <http://www.marketingcharts.com/interactive/online-video-spend-tops-212b-in-2008-225-growth-forecast-in-2009-7955/>. Accessed: July 1, 2010.
- [3] ABI Research – Technology Research Market. <http://www.abiresearch.com/home.jsp>. Accessed on July 1, 2010.
- [4] The 3rd Generation Partnership Project. <http://www.3gpp.org/>
- [5] ETSI Telecommunications and Internet converged Services and Protocols for Advanced Networking (ETSI- TISPAN). <http://www.etsi.org/tispan/>
- [6] The Telecommunication Standardization Sector of ITU (ITU-T). <http://www.itu.int/ITU-T/>
- [7] ETSI TS 182 027: IPTV Architecture; IPTV functions supported by the IMS subsystem, March 2011.
- [8] The UCT IMS Client. <http://uctimsclient.berlios.de/>. Accessed November 15, 2011.
- [9] H. Schulzrinne, A. Rao, and R. Lanphier (1998). Real Time Streaming Protocol (RTSP). <http://tools.ietf.org/html/rfc2326>
- [10] Internet draft Real Time Streaming Protocol 2.0 (RTSP). <http://tools.ietf.org/html/draft-ietf-mmusic-rfc2326bis-28>. Expires: April 30, 2012.
- [11] VLC. VideoLAN, Free streaming and multimedia solutions for all OS. <http://www.videolan.org/>. Accessed: November 15, 2011.
- [12] Darwin, “Open Source Streaming Server,” <http://developer.apple.com/opensource/>
- [13] Mobicents-Public. <http://groups.google.com/group/mobicents-public/web>
- [14] Live555. Internet Streaming Media, Wireless, and Multicast technology, services, & standards. <http://www.live555.com/>
- [15] Internet draft. Framework of media control for IPTV services draft-siva-iptv-media-01.txt. <http://tools.ietf.org/html/draft-siva-iptv-media-01>. Expires: September 2010.
- [16] Internet draft. An Evaluation of Session Initiation Protocol (SIP) for use in Streaming Media Applications draft-whitehead-sip-for-streaming-media-00.txt. <http://tools.ietf.org/html/draft-whitehead-sip-for-streaming-media-00>. Expires: April, 2006.
- [17] Internet draft. Media Playback Control Protocol Requirements draft-whitehead-mmusic-sip-for-streaming-media-03. <http://tools.ietf.org/html/draft-whitehead-mmusic-sip-for-streaming-media-03>. Expires: August 2008.

- [18] S. Sivasothy, G. M. Myoung, and N. Crespi (2009). A unified session control protocol for IPTV services. In *Proceedings of the 11th International Conference on Advanced Communication Technology*. pp. 961-965, February 15-18, 2009, Gangwon-Do, South Korea.
- [19] R. G. Shiroor (2007). IPTV and VoD services in the context of IMS. In *International Conference on IP Multimedia Subsystem Architecture and Applications*, pp. 1-5, December 6-8, 2007.
- [20] B. Chatras, M. Saïd. Delivering Quadruple Play with IPTV over IMS. [www.icin.biz/files/programmes/Session8A-1.pdf](http://www.icin.biz/files/programmes/Session8A-1.pdf). Accessed: July 1, 2010.
- [21] Internet draft. SDP media capabilities Negotiation draft-ietf-mmusic-sdp-media-capabilities-09. <http://tools.ietf.org/html/draft-ietf-mmusic-sdp-media-capabilities-09>. Expires: August 2010.
- [22] E. Mikoczy. IMS based IPTV services: architecture and implementation. In *Proceedings of the 3rd International Conference on Mobile Multimedia Communications*. pp. 1-7. 2007. Brussels, Belgium.
- [23] C. Riede, A. Al-Hezmi and T. Magedanz (2008). Session and media signaling for IPTV via IMS. In *Proceedings of the 1st International Conference on Mobile Wireless Middleware, Operating Systems, and Applications*. February 13 - 15, 2008. Brussels, Belgium.
- [24] Z. Shibeshi, A. Terzoli, and K. Bradshaw (2010). Streaming Session Transfer between Registered User Agents. In *SATNAC'10: Proceedings of the 13th Southern African Telecommunications Networks and Applications Conference*. Spier Estate, Stellenbosch, South Africa, 5 to 8 September 2010.
- [25] R. Spiers, R. Marston, R Good and N. Ventura (2009). The UCT IMS IPTV Initiative. In *Proceedings of the 2009 Third International Conference on Next Generation Mobile Applications, Services and Technologies*. pp. 503-508. 2009.
- [26] The Open Source IMS Project. <http://www.openimscore.org/>
- [27] P. R. Wilson and N. Ventura (2009). A Direct Marketing Platform for IMS-Based IPTV. In *SATNAC'09: Proceedings of the 12th Southern African Telecommunications Networks and Applications Conference*. Ezulwini, Swaziland, 30 August to 2 September 2009.
- [28] The Apache Mina Software Foundation. <http://mina.apache.org/>

Mr. Zelalem S. Shibeshi holds an MSc in Information Science, Diploma in Computer Science, and BSc in Physics, all from Addis Ababa University, Ethiopia, and is currently working towards his PhD in the Computer Science Department at Rhodes University.

Alfredo Terzoli is a Professor of Computer Science at Rhodes University, where he heads the Telkom Centre of Excellence in Distributed Multimedia. He is also Research Director of the Telkom Centre of Excellence in ICT for Development at the University of Fort Hare. His main areas of academic interest are converged telecommunication networks and ICT for development.

Dr. Karen Bradshaw is a Senior Lecturer in the Computer Science Department at Rhodes University. Her research interests lie in Distributed Systems and Parallel Programming.





# Secure Key Exchange Scheme for IPTV Broadcasting

Ravi Singh Pippal and Shashikala Tapaswi  
 ABV-Indian Institute of Information Technology and Management, Gwalior, India  
 E-mail: {ravi, stapaswi}@iiitm.ac.in

Jaidhar C. D.  
 Defence Institute of Advance Technology, Girinagar, Pune, India  
 E-mail: jaidharc@diat.ac.in

**Keywords:** cryptography, IPTV, mutual authentication, nonce, smart card, STB

**Received:** October 7, 2011

*In Internet Protocol Television (IPTV) broadcasting, service providers charge subscription fee by scrambling the program in Conditional Access System (CAS). This avoids unauthorized users to receive the programs. A smart card (CA card) is used to decrypt the Control Words (CWs) and transfer them back to Set-Top Box (STB) in order to descramble the scrambled program. This paper presents a secure mutual authentication and key exchange scheme between STB and smart card for IPTV broadcasting. Its security is based on one way hash function and the discrete logarithm problem. It allows subscribers to choose and change the password freely, provides dynamic session key agreement and mutual authentication between STB and smart card. Security analysis proves that the scheme is strong against subscriber and STB impersonation attacks, replay attack, stolen verifier attack, smart card loss attack, man-in-the-middle attack and attack on perfect forward secrecy which are considered as common threats in IPTV environment. Moreover, the scheme also prevents serious attacks such as smart card cloning and McCormac Hack attack particular to authentication using smart cards.*

*Povzetek: Članek opisuje način šifriranja vsebine za televizijo IP.*

## 1 Introduction

There are several security issues that must be considered before transmitting confidential data over a public network. In order to prevent unauthorized access, first step of the communication is legitimacy verification. In other words, authentication is vital requirement which identifies the legitimate user in order to prevent unauthorized access. Verities of authentication schemes have been proposed in the literature [1, 2, 3, 4, 5, 6, 9]. Most widely used one is password based authentication scheme.

Using one way hash function, Peyravian and Zunic [1] proposed a secure method for protecting passwords while being transmitted over insecure channel. Further, secure password change phase has also been proposed. In addition, they claimed that their schemes do not require any symmetric key or public key cryptosystem. However, Tseng *et al.* [2] found that Peyravian-Zunic's scheme is insecure against dictionary attack and fails to provide mutual authentication. To overcome these flaws, they proposed improved schemes based on Diffie-Hellman key exchange scheme and claimed that their improved schemes not only provide secure password transmission and password change, but also generate a session key between user and the server. Yang *et al.* [3] pointed out that Tseng *et al.*'s protected password changing scheme is susceptible to

modification attack. Further, they suggested an improved scheme without using symmetric or asymmetric cryptosystem to overcome the weakness of Tseng *et al.*'s scheme. They claimed that their scheme is secure against replay attack, guessing attack, server spoofing and modification attack. Nevertheless, Yoon *et al.* [4] and Ku and Tsai [5] found that Yang *et al.*'s scheme is still vulnerable to Denial-of-Service attack and stolen verifier attack. To overcome these security pitfalls, they proposed their modified schemes.

In all the schemes discussed so far, server maintains a database or verification table for the registered users to authenticate the legitimate users. However, there is a threat in such a process as an intruder can penetrate the server and steal or modify the contents of the verification table. To resist these possible attacks on the verification tables, smart card based password authentication scheme has been proposed. In this scheme, server authenticates the legitimate user without maintaining a verification table.

In this context, Hwang and Li [6] proposed a remote user authentication scheme based on ElGamal's cryptosystem. They claimed that their scheme does not maintain any password or verification table and it is secure against replay attack. However, Chan and Cheng [7] proved that Hwang-Li's scheme is vulnerable to impersonation attack. Chang and Hwang [8] found that Chan-Cheng's attack does not

work well and they suggested different ways to cryptanalyze Hwang-Li’s scheme. Based on symmetric key cryptography and one way hash function, Song [9] suggested an efficient smart card authentication scheme and claimed that the scheme is secure against impersonation attack, parallel session attack, replay attack and modification attack. Moreover, it provides mutual authentication and shared session key. Though, Pippal *et al.* [10] pointed out that Song’s scheme is inadequate to resist Denial-of-Service attack and fails to provide perfect forward secrecy.

The remainder of this paper is organized as follows: Section 2 briefly describes the work related to secure communication in IPTV broadcasting. The proposed key exchange scheme is presented in Section 3. Section 4 discusses security analysis of the proposed scheme and finally, section 5 concludes the paper.

## 2 Secure Communication in IPTV Broadcasting

Internet Protocol Television (IPTV) is a next generation television capable of transmitting, receiving and displaying a video stream. Gist of IPTV structure is shown in Figure 1. It provides access to on-demand gaming, home security, data services and digital music. IPTV is capable of providing a single stream to multiple users simultaneously and also to a single user such as Video on-Demand.

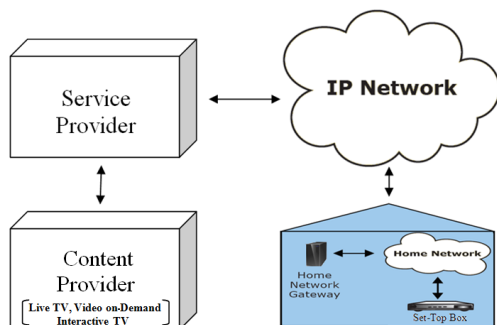


Figure 1: Overview of IPTV Structure

In IPTV broadcasting, service providers charge subscription fee by scrambling the program in CAS. A smart card is used to decrypt CWs and transfer them back to STB in order to descramble the scrambled program. STB receives encoded digital signals and decodes these signals to convert them back to analog signals so that the analog television can understand. Therefore, secure key exchange with mutual authentication between STB and smart card is needed to improve the security of the system. Without this, single smart card can be used in different STBs of the same type which results smart card cloning and McCormac Hack attacks [11].

Figure 2 shows a typical CAS, it operates as follows [12]. The server chooses a random variable CW which is used to

initialize the Pseudo Random Generator (PRG) to generate a pseudo random sequence for scrambling the Transport Stream (TS). Simultaneously, for each subscriber, CW is encrypted by Authorization Key (AK) to form Entitlement Control Message (ECM). A Master Private Key (MPK) is used to encrypt AK and other entitlement message together in order to form Entitlement Management Message (EMM). These ECM, EMM and the scrambled TS stream are multiplexed into a new TS stream and broadcasted to subscribers over an insecure channel. Subscriber Management System (SMS) is used to deliver the smartcard, which contains MPK and other account information, to authorized subscriber.

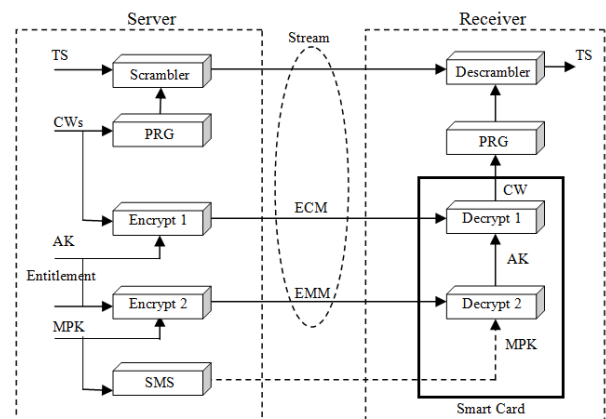


Figure 2: Conditional Access System

The receiver can descramble the program by following the same steps performed by the server in reverse order with the collaboration of smart card and STB [13]. In STB, an ECM/EMM filter is used to filter out the ECM and EMM sections and a descrambler is used to descramble the program. After receiving radio frequency (RF) signal, tuner and demodulator process the signal to bring back the TS stream. The ECM/EMM filters filter out the ECM and EMM sections and sent to the smart card to be decrypted for CW with Decrypt 1 and Decrypt 2. CW is encrypted by using Session Key (SK) and it is sent back to STB. This CW is used by descrambler to descramble the TS stream which is then de-multiplexed and decoded. The STB takes copyright protection before outputting program to subscriber.

To provide secure communication between STB and smart card, various elegant key exchange schemes have been proposed [14, 15, 16, 17]. Based on one way hash function and Schnorr’s digital signature protocol, Jiang *et al.* [14] first proposed a key exchange scheme for DTV broadcasting. They claimed that their scheme allows users to freely choose the password, provides mutual authentication and session key agreement between STB and smart card. Moreover, it has lower computation cost. However, Yoon and Yoo [15] found that Jiang *et al.*’s scheme is susceptible to impersonation attack and fails to provide per-

fect forward secrecy. They also suggested a new key exchange scheme to overcome these security weaknesses and claimed that their scheme is free from replay attack, impersonation attack and provides perfect forward secrecy.

Based on symmetric and asymmetric key cryptosystems, Hou *et al.* [16] proposed a secure authentication scheme for DTV broadcasting and claimed that their scheme allows users to freely choose the password, provides security against replay attack, impersonation attack, offers mutual authentication and session key generation. However, Kim [17] found that the message transmitted during mutual authentication phase of Hou *et al.*'s scheme can be forged by the attacker. To overcome this security flaw, an improved scheme has also been suggested.

Secure IP multicast can be used to implement IPTV services, but still, it has problems that need to be addressed. These issues were addressed and a centralized form of secure group communication was proposed to transmit group cryptographic material [18]. However, Pinto and Ricardo [19] found that there are other issues also, like access control and network management, which were left. They proposed a secure and efficient IPTV solution which enforces individual access control to groups of real-time IPTV video channels, IP multicast admission control for both multicast senders and receivers, supports user generated videos and generates low signalling overheads. Moreover, it does not introduce perceivable delays, particularly in video channel zapping circumstances.

### 3 The Proposed Key Exchange Scheme

This section describes the proposed key exchange scheme for IPTV. The notations used throughout this paper are summarized as follows.

- $U_i$  : subscriber
- $ID_i$  : identity of  $U_i$
- $PW_i$  : password chosen by  $U_i$
- $SMS$  : Subscriber Management System
- $STB$  : Set-Top Box
- $ID_s$  : identity of  $STB$
- $PW_i^*$  : password guessed by an attacker
- $x$  : secret key of  $STB$
- $d$  : secret number of  $STB$
- $p$  : large prime number
- $g$  : primitive element
- $h(\cdot)$  : secure one way hash function
- $E_k(\cdot)$  : symmetric encryption with key ' $k$ '
- $D_k(\cdot)$  : symmetric decryption with key ' $k$ '
- $\oplus$  : bitwise XOR operation
- $N_1$  : random nonce generated by  $U_i$
- $N_2$  : random nonce generated by  $STB$
- $S_{Key}$  : common shared session key
- $\dashrightarrow$  : secure channel
- $\rightarrow$  : insecure channel

The proposed scheme consists of five phases: Registration phase, Login phase, Mutual Authentication phase, Key Agreement phase and  $CW$  Transmission phase. The detailed description of the proposed scheme is shown in Figure 3. This scheme works as follows.

#### 3.1 Registration Phase

This phase is invoked when a new subscriber  $U_i$  wants to subscribe the subscribed program. In this phase,  $U_i$  selects  $ID_i$  and  $PW_i$ , computes  $h(PW_i)$  and submits  $\{ID_i, h(PW_i)\}$  to  $SMS$ . Upon receiving the registration request from  $U_i$ ,  $SMS$  computes

$$x_i = g^{h(PW_i)} \times d \text{ mod } p$$

$$y_i = h(ID_i, x)$$

$$z_i = y_i \oplus h(PW_i)$$

and issues a smart card over secure channel to  $U_i$  by storing  $\{x_i, y_i, z_i, ID_s, p, g, h(\cdot), E_k(\cdot), MPK\}$  along with other account information into smart card memory.

#### 3.2 Login Phase

This phase is invoked when  $U_i$  wants to receive the subscribed program.  $U_i$  inserts the smart card to  $STB$  and keys in  $ID_i$  and  $PW_i$ . The smart card generates a random nonce  $N_1$ , computes

$$a_i = g^{y_i} \text{ mod } p$$

$$b_i = a_i^{y_i \times N_1} \text{ mod } p$$

$$c_i = a_i^{h(PW_i) \times N_1} \text{ mod } p$$

$$d_i = (h(PW_i) + y_i \times \lambda) \text{ mod } (p - 1)$$

$$e_i = g^{h(PW_i)} \text{ mod } p$$

$$f_i = b_i \oplus c_i$$

where  $\lambda = h(ID_i, ID_s, x_i, a_i, b_i, c_i, N_1)$ .  $U_i$  sends the login request  $\{ID_i, d_i, e_i, f_i, N_1\}$  to  $STB$ .

#### 3.3 Mutual Authentication Phase

Upon receiving the login request  $\{ID_i, d_i, e_i, f_i, N_1\}$ ;  $STB$  first checks the validity of  $ID_i$  to accept/reject the login request. If true,  $STB$  computes

$$x_i = e_i \times d \text{ mod } p$$

$$y_i = h(ID_i, x)$$

$$a_i = g^{y_i} \text{ mod } p$$

$$b_i = a_i^{y_i \times N_1} \text{ mod } p$$

$$c_i = b_i \oplus f_i$$

and checks whether

$$g^{d_i} = e_i \times a_i^\lambda \text{ mod } p \tag{1}$$

is true or not.

$$g^{d_i} = (g^{(h(PW_i) + y_i \times \lambda)}) \text{ mod } p$$

$$= (g^{h(PW_i)} \times g^{(y_i \times \lambda)}) \text{ mod } p$$

$$= (g^{h(PW_i)} \text{ mod } p) \times ((g^{y_i})^\lambda \text{ mod } p)$$

$$= e_i \times a_i^\lambda \text{ mod } p$$

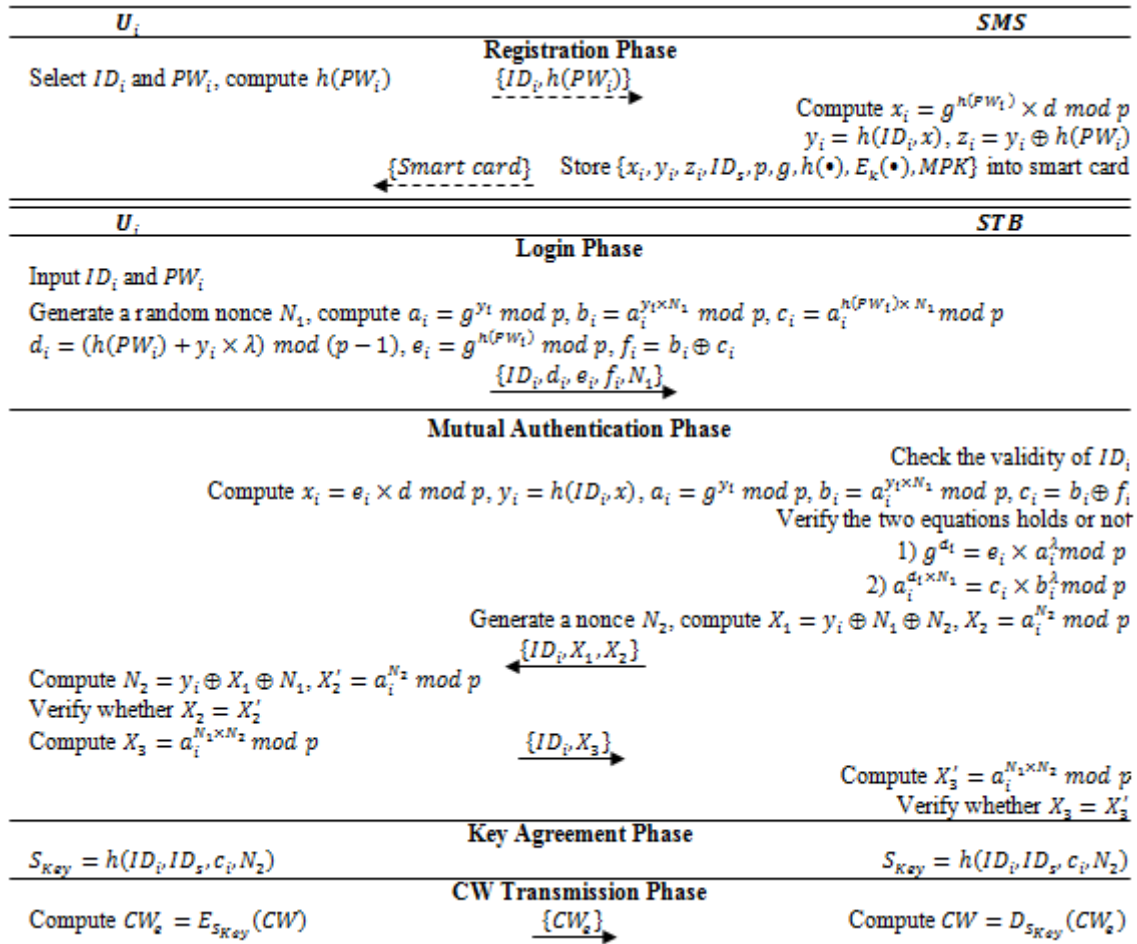


Figure 3: Proposed Key Exchange Scheme

If eq. 1 holds,  $STB$  checks whether

$$a_i^{d_i \times N_1} = c_i \times b_i^\lambda \text{ mod } p \quad (2)$$

is true or not.

$$\begin{aligned} a_i^{d_i \times N_1} &= (a_i^{(h(PW_i) + y_i \times \lambda) \times N_1}) \text{ mod } p \\ &= (a_i^{h(PW_i) \times N_1} \times a_i^{(y_i \times \lambda \times N_1)}) \text{ mod } p \\ &= (a_i^{h(PW_i) \times N_1} \text{ mod } p) \times ((a_i^{y_i \times N_1})^\lambda \text{ mod } p) \\ &= c_i \times b_i^\lambda \text{ mod } p \end{aligned}$$

If both the equations, (eq. 1 and eq. 2), hold,  $STB$  generates a nonce  $N_2$ , computes  $X_1 = y_i \oplus N_1 \oplus N_2$ ,  $X_2 = a_i^{N_2} \text{ mod } p$  and sends the message  $\{ID_i, X_1, X_2\}$  to  $U_i$ 's smart card. After getting the message  $\{ID_i, X_1, X_2\}$  from  $STB$ , smart card computes  $N_2 = y_i \oplus X_1 \oplus N_1$ ,  $X_2' = a_i^{N_2} \text{ mod } p$  and checks whether  $X_2$  and  $X_2'$  are equal or not. If it holds,  $STB$  is authentic otherwise terminate the session. Subsequently, smart card computes  $X_3 = a_i^{N_1 \times N_2} \text{ mod } p$  and sends  $\{ID_i, X_3\}$  to  $STB$ . Once the message  $\{ID_i, X_3\}$  is received,  $STB$  computes  $X_3' = a_i^{N_1 \times N_2} \text{ mod } p$  and checks whether  $X_3$  and  $X_3'$  are equal or not. If it holds, mutual authentication between  $U_i$ 's smart card and  $STB$  is achieved.

### 3.4 Key Agreement Phase

If mutual authentication is achieved successfully, both  $U_i$ 's smart card and  $STB$  compute common session key  $S_{Key} = h(ID_i, ID_s, c_i, N_2)$ . It consists of identities ( $ID_i$  and  $ID_s$ ) as well as the random nonces ( $N_1$  and  $N_2$ ) chosen by  $U_i$  and  $STB$ .

### 3.5 CW Transmission Phase

After decrypting  $CW$ , smart card uses the session key  $S_{Key}$  to encrypt it as  $CW_e = E_{S_{Key}}(CW)$  and sends  $CW_e$  back to  $STB$  to descramble the program. After receiving,  $STB$  decrypts it as  $CW = D_{S_{Key}}(CW_e)$ .

## 4 Security Analysis and Discussion

This section describes an in-depth security analysis of the proposed key exchange scheme for IPTV broadcasting. Since a smart card is a temper-resistant device, it is assumed that no one can extract any information stored in the smart card memory.

#### 4.1 Subscriber Impersonation Attack

In the proposed scheme, the login request contains  $\{ID_i, d_i, e_i, f_i, N_1\}$ , where  $d_i = (h(PW_i) + y_i \times \lambda) \bmod (p - 1)$ ,  $e_i = g^{h(PW_i)} \bmod p$  and  $f_i = b_i \oplus c_i$ . To impersonate the subscriber, attacker has to generate a forged login request by guessing the correct values of  $PW_i$ ,  $y_i$  and  $d$ . Let us suppose that the attacker is successful in guessing the correct password  $PW_i^*$ . The correct values of  $y_i$  and  $d$  are still required to forge the login request. In addition, it is difficult to derive  $h(PW_i)$  from  $e_i$  because of discrete logarithm problem. Moreover, if an attacker modifies any of the login request parameters,  $STB$  easily detects them as both the equations, (eq. 1 and eq. 2), are unsatisfied. Hence, this scheme provides security against subscriber impersonation attack.

#### 4.2 STB Impersonation Attack

To impersonate  $STB$ , the attacker has to generate valid response message  $\{ID_i, X_1, X_2\}$  corresponding to the login request  $\{ID_i, d_i, e_i, f_i, N_1\}$ . However, without the knowledge of  $y_i$  and  $N_2$ , no one can compute the correct value of  $X_1$  and  $X_2$ . Moreover, attacker is unable to get  $N_2$  from the eavesdropped response message as the value of  $y_i$  is unknown. Therefore, the scheme is secure against  $STB$  impersonation attack.

#### 4.3 Replay Attack

An attacker may try to act as an authentic subscriber by resending previously intercepted messages. This scheme uses random nonces,  $N_1$  and  $N_2$ , which are different from session to session. As a result, attackers cannot enter the system by resending the previously transmitted messages to impersonate legal subscribers. Suppose that the intercepted login request  $\{ID_i, d_i, e_i, f_i, N_1\}$  is replayed to pass the authentication phase. Attacker is unable to retrieve  $N_2$  correctly from the response message  $\{ID_i, X_1, X_2\}$  to compute the correct message  $\{ID_i, X_3\}$  for mutual authentication. Consequently,  $STB$  rejects the request by comparing  $X_3$  with  $X_3'$ .

#### 4.4 Stolen Verifier Attack

In order to verify the legitimacy of subscribers, use of verification table at  $STB$  is not efficient. Moreover, if  $STB$  stores  $U_i$ 's secret information, it will be always under the risk. In the proposed scheme,  $STB$  keeps long term secret key ' $x'$ ' and secret number ' $d'$ ' to avoid maintaining verification table used to verify subscriber login request. Hence, the scheme avoids stolen verifier attack.

#### 4.5 Man-in-the-Middle Attack

If an attacker intercepts the communicating messages exchanged between the subscriber and  $STB$ , it does not generate any useful information as they are dissimilar from session to session due to property of randomness of  $N_1$  and  $N_2$ . Moreover, to alter  $N_1$ , one needs to recalculate  $b_i$ ,  $c_i$ ,  $d_i$  and  $f_i$ . Similarly,  $y_i$  is needed to alter  $N_2$ . Hence, the scheme is able to resist man-in-the-middle attack.

#### 4.6 Smart Card Cloning and McCormac Hack Attack

In the proposed scheme, if an attacker uses the cloned smart card to another  $STB$ , it will not pass the mutual authentication phase as there is no  $STB$ 's  $ID_s$  in the cloned smart card memory.

If an attacker redirects one smart card's communication message to another  $STB$ , the  $STB$  cannot decrypt the message as it has no information about the session key  $S_{Key}$ .

#### 4.7 Smart Card Loss Attack

If accidentally, subscriber's smart card is lost or stolen; the scheme must be strong enough so that no one can impersonate the smart card owner. In this scheme, attacker is unable to receive the program without knowing the correct  $ID_i$  and  $PW_i$  of the subscriber even if he or she got subscriber's smart card.

#### 4.8 Attack on Perfect Forward Secrecy

In the proposed scheme, the session key is computed as  $S_{Key} = h(ID_i, ID_s, c_i, N_2)$ . The attacker is unable to find out the present session key or any of the previously used session keys from the eavesdropped messages as the values of  $ID_s$ ,  $c_i$  and  $N_2$  are unknown to the attacker and it is infeasible to guess all these values simultaneously.

#### 4.9 Subscriber can change the Smart Card Password Securely

This phase is invoked whenever  $U_i$  wants to change the current password  $PW_i$  with a new password  $PW_{inew}$ .  $U_i$  inserts the smart card to  $STB$  and keys in  $ID_i$  and  $PW_i'$ . The smart card computes  $z_i' = y_i \oplus h(PW_i')$  and checks whether computed  $z_i'$  equals stored  $z_i$  or not. If true,  $U_i$  is prompted to enter a new password  $PW_{inew}$ . The smart card computes  $z_{inew} = y_i \oplus h(PW_{inew})$ ,  $x_{inew} = (x_i/g^{h(PW_i)}) \times g^{h(PW_{inew})} \bmod p$  and stores  $x_{inew}$ ,  $z_{inew}$  instead of  $x_i$ ,  $z_i$  respectively, in the smart card memory. Thus,  $U_i$  can change the smart card password.

It can be clearly seen that the given scheme keeps all the previous advantages and achieves the necessary security requirements.

### 5 Conclusion

In IPTV services, content is crucial that needs strong protection from unauthorized entities. In order to provide secure communication, dynamic session key generation and mutual authentication between smart card and  $STB$  is imperative. Considering all the common threats in IPTV environment, this paper proposes secure key exchange scheme for IPTV broadcasting. Security analysis section shows that the proposed scheme is robust against impersonation attacks, replay attack, stolen verifier attack, smart card loss attack and man-in-the-middle attack.

In addition, it is secure against two serious attacks in IPTV broadcasting such as smart card cloning and McCormac Hack attack. Proposed scheme allows the subscribers to choose and change their smart card password freely. It ensures perfect forward secrecy as well as dynamic session key generation with mutual authentication.

## Acknowledgement

The authors would like to thank ABV-Indian Institute of Information Technology and Management, Gwalior, India for providing academic support.

## References

- [1] Peyravian, M. and Zunic, N. (2000). Methods for protecting password transmission. *Computers and Security*, 19(5), pp. 466–469.
- [2] Tseng, Y.M., Jan, J.K. and Chien, H.Y. (2001). On the security of methods for protecting password transmission. *Informatica*, 12(3), pp. 469–476.
- [3] Yang, C.C., Chang, T.Y. and Hwang, M.S. (2003). Security of improvement on methods for protecting password transmission. *Informatica*, 14(4), pp. 551–558.
- [4] Yoon, E.J., Ryu, E.K. and Yoo, K.Y. (2005). Attacks and solutions of Yang *et al.*'s protected password changing scheme. *Informatica*, 16(2), pp. 285–294.
- [5] Ku, W.C. and Tsai, H.C. (2005). Weaknesses and improvements of Yang-Chang-Hwang's password authentication scheme. *Informatica*, 16(2), pp. 203–212.
- [6] Hwang, M.S. and Li, L.H. (2000). A new remote user authentication scheme using smart cards. *IEEE Transactions on Consumer Electronics*, 46(1), pp. 28–30.
- [7] Chan, C.K. and Cheng, L.M. (2000). Cryptanalysis of a remote user authentication scheme using smart cards. *IEEE Transactions on Consumer Electronics*, 46(4), pp. 992–993.
- [8] Chang, C.C. and Hwang, K.F. (2003). Some forgery attacks on a remote user authentication scheme using smart cards. *Informatica*, 14(3), pp. 289–294.
- [9] Song, R. (2010). Advanced smart card based password authentication protocol. *Computer Standards and Interfaces*, 32(5-6), pp. 321–325.
- [10] Pippal, R.S., Jaidhar, C.D. and Tapaswi, S. (2010). Comments on symmetric key encryption based smart card authentication scheme. In *Proceedings of 2<sup>nd</sup> International Conference on Computer Technology and Development*, Cairo, Egypt, pp. 482–484.
- [11] Kanjanarin, W. and Amornraksa, T. (2001). Scrambling and key distribution scheme for digital television. In *Proceedings of IEEE International Conference on Networks*, Bangkok, Thailand, pp. 140–145.
- [12] Jiang, T., Zheng, S. and Liu, B. (2004). Key distribution based on hierarchical access control for conditional access system in DTV broadcast. *IEEE Transactions on Consumer Electronics*, 50(1), pp. 225–230.
- [13] Kamperman, F. and Rijnsoever, B.V. (2001). Conditional access system interoperability through software downloading. *IEEE Transactions on Consumer Electronics*, 47(1), pp. 47–53.
- [14] Jiang, T., Hou, Y. and Zheng, S. (2004). Secure communication between set-top box and smart card in DTV broadcasting. *IEEE Transactions on Consumer Electronics*, 50(3), pp. 882–886.
- [15] Yoon, E.J. and Yoo, K.Y. (2009). Robust key exchange protocol between set-top box and smart card in DTV broadcasting. *Informatica*, 20(1), pp. 139–150.
- [16] Hou, T.W., Lai, J.T. and Yeh, C.L. (2007). Based on cryptosystem secure communication between set-top box and smart card in DTV broadcasting. In *Proceedings of TENCON 2007*, IEEE Region 10 Conference, Taipei, Taiwan, pp. 1–5.
- [17] Kim, H. (2008). Secure communication in digital TV broadcasting. *International Journal of Computer Science and Network Security*, 8(9), pp. 1–5.
- [18] Pinto, A. and Ricardo, M. (2010). Secure multicast in IPTV services. *Computer Networks*, 54(10), pp. 1531–1542.
- [19] Pinto, A. and Ricardo, M. (2011). On performance of group key distribution techniques when applied to IPTV services. *Computer Communications*, 34(14), pp. 1708–1721.

# ‘The Frozen Accident’ as an Evolutionary Adaptation: A Rate Distortion Theory Perspective on the Dynamics and Symmetries of Genetic Coding Mechanisms

James F. Glazebrook  
 Department of Mathematics and Computer Science  
 Eastern Illinois University  
 600 Lincoln Avenue, Charleston IL 61920–3099, USA  
 E-mail: jfglazebrook@eiu.edu

Rodrick Wallace  
 Division of Epidemiology  
 The New York State Psychiatric Institute  
 Box 47, 1051 Riverside Drive, New York NY 10032, USA  
 E-mail: wallace@pi.cpmc.columbia.edu

**Keywords:** frozen accident, rate distortion function, protein folding, free energy density, spin glass, groupoid, Onsager relations, holonomy

**Received:** October 8, 2011

*We survey some interpretations and related issues concerning ‘the frozen accident’ hypothesis proposed by Francis Crick and how it can be explained in terms of several natural mechanisms involving error-correction codes, spin glasses, symmetry breaking and the characteristic robustness of genetic networks. The approach to most of these questions involves using elements of Shannon’s rate distortion theory incorporating a semantic system which is meaningful for the relevant alphabets and vocabulary implemented in transmission of the genetic code. We apply the fundamental homology between information source uncertainty with the free energy density of a thermodynamical system with respect to transcriptional regulators and the communication channels of sequence/structure in proteins. The collective outcome of these processes supports previous suggestions that ‘the frozen accident’ may in fact have been a temporal evolutionary adaptation.*

*Povzetek: Članek obravnava izvor genetskega kodiranja.*

## 1 Introduction

Examining and predicting the geometric/topological structures of the genetic coding network is essential to understanding its (co)evolution as a complex communications system, employing a vocabulary of a given genetic code that determines the family of proteins encodable by the genes themselves. The architecture of this network developed from a coevolution of genes and of genetic structures that were progressively conditioned to shield against translation and replication errors. Crick’s hypothesis [30, 31](surveyed in e.g. [4]), in broad terms, says that on reading the mRNA script, the coding strategy determines the amino acid sequence of the evolved proteins, as is the case for most organisms. So in a post-translational phase any kind of alteration to the size of the code would have dire consequences owing to a global impact on proteins created by new amino acids subject to the likelihood of nonsensical messaging. Crick gave flexible rules for pairing the third base of the codon with the first base of the anticodon, to the extent that a single tRNA type would be

able to recognize up to three codons. More complex protein structures arise when there is an enrichment and expansion of the vocabulary while any ambiguity in the code is minimized, so restricting the content of information. When the codon meaning is altered, the information selected would condition that codon to some advantage. In this way the ‘freezing’ was professed to be an outcome of such selective restrictions and this would put the brakes on further evolvability.

While over the years there has been much debate and challenge concerning these rules, and to establish a concrete mechanism for the companion ‘wobble hypothesis’, we outline here several scenarios from the point of view of coevolutionary rate distortion dynamics in graphs that represent ‘robustness’ while admitting ‘meaningful’ signalling paths which are susceptible to vocabulary enrichment, and furthermore, give rise to structure preserving patterns that evolve towards optimizing error-correction. These collective mechanisms can be formulated in the context of a spin-glass model (cf [12, 21, 25]), that incorporates the Onsager relations of statistical physics applied to networks of

mutating sequences and error-correction in the presence of rate distortion dynamics, then leading to phase transitions through which symmetry breaking occurs and hence causes a change in topological structure of the graph. These observations are supported by a number of relatively recent theoretical findings, and thus it seems reasonable to provide some of the necessary background material. Related are the approaches to evolutionary (population) biology employing Boltzmann statistics, Fisher and Kolmogorov diffusion equation methods, and stochastic evolution for which there is already a large amount written (see e.g. [78]).

A position often maintained is that evolution influences the emergence of the genetic code by selecting an amino acid map that is error-minimizing and the subsequent competition between organisms is determined by the overall capability of their respective codes. Following this line of thought, Tlusty [73, 74, 75, 76], implementing a topological graph-theoretic approach, has developed a model for the emergence of the genetic code as a supercritical phase transition occurring within noisy information channels as traced by maps between nucleotides and amino acids with error bounds in place. The proposed paradigm is that these processes are indeed ‘cognitive’ [80, 81, 82, 85] following the immunology/language perspective of Atlan and Cohen [6] (see also [26, 27]) that human and biological organizations at all scales are cognitive in so far that once patterns of threat and opportunity are perceived, these patterns are compared with an internal image of the environment, and then a choice of responses from a vast repertoire of possibilities is initiated.

This present paper continues with this theme to establish one of several possible corollaries derived from [80, 82] by addressing the question of how coevolutionary robustness against errors, error-correction, and phase transitions modeled by the topological dynamics of graphs that can be represented by certain spin glass/error correcting structures that are susceptible to thermodynamic spontaneous symmetry breaking; these factors shed further light to explaining what exactly was the ‘accident’ that did occur. Such symmetry breaking of the genetic code has been considered in the context of Lie algebra representations in [10, 11, 46]. Our perspective using rate distortion dynamics, is that such a sequence of broken symmetries corresponds to phase transitions in the underlying error correcting networks through which the codon allocation to amino acids is mainly the outcome of error-correction minimization and efficiency (see [10] and references therein), a scenario that appears relevant to the approach of Ardell and Sella [4, 66, 67].

While on the mathematical-physical side of things, several explanations for ‘freezing’ and ‘wobbling’ can be given in terms of error-correction and the structural theory of Lie algebras, which we survey. A novel technique introduced here involves showing how the dynamics governing the underlying mechanisms can be represented in terms of a ‘covariant differentiation’ of the Shannon entropy along ‘meaningful paths’ embedded in a (genetic)

coding graph that also includes a correlation with error-correction and folding rates. This operation over which the various ‘directions’ are taken<sup>1</sup> subsequently determines the *holonomy* of the system through an error-correction network—a broader scale geometric representation of transitional phases in which the broken symmetries may be expressed in terms of holonomy groups that collectively, via disjoint union, form a holonomy groupoid, a structure which in principle can be given explicitly.

## 2 ‘The Frozen Accident’— or Not Quite

We start by putting matters into perspective by surveying some basic observations. Recall that genes can be represented by molecular words written in terms of the nucleotide bases *U*(Uracil/Thymine), *C*(Cytosine), *G*(Guanine) and *A*(Adenine), whereas proteins are written in a language of 20 letters corresponding to the amino acids in which each of the latter is encoded by specific triplets of the basis members, known as *codons*, so connecting hereditary characteristics to vital units. In theory there are  $64 = 4^3$  codons with the number of possible observables lying somewhere between 48 and 64 (see e.g. [50, 73]). However, it is claimed in [50] that the code mapping the 64 codons to the 20 amino acids is anything but random. There are at least 48 discernable codons but only 20 amino acids available (and 3 stop codons), so the code is degenerate in so far that several codons can represent the same amino acid. Entropy analysis [1, 55] reveals that the information content of a random protein structure can occupy  $\log_2(20) \simeq 4.32$  bits of entropy per amino acid residue in a primary sequence.

In the presence of topological changes there would have been alterations of an excessive amount of (protein) structures, and those frequently observed tend to be the ones that have managed to remain intact as the structures became more complex. The ‘wobble rules’ assume that only 48 codons can be distinguished owing to the physiochemical limitations of the translational mechanism and the resulting codon graph converges to 20 amino acids. The question is: does a single sRNA molecule recognize several codons? The ‘wobble’ effect aside, there exist 64 distinguishable codons and the maximal number of amino acids increases to 25, which is not a dramatic amount by any means, though it has been a puzzling matter as to why evolution did freeze prior to improving the translational mechanism to single out all 64 codons. Once the meaning of a codon had changed, again, selectivity would apply that codon to a site for a new amino acid to serve to some advantage, or otherwise simply to replace it.

The traditional approach to producing more tRNAs

<sup>1</sup>A reader with some acquaintance with differential geometry will understand this as ‘covariant differentiation over (or along) a vector field’— an operation specified by choice of ‘connection’. This we implement on graphs in §6.



would have been to change the anticodons of existing ones, giving rise to a new class of amino acids proliferating across the code while systematically reshuffling a large number of codons in the process. To an extent the ‘wobble hypothesis’ concerns stereochemical limitations on the actual tRNA capacity to single-out codons [38]. In more basic terms, interfering with the genetic code would change the meaning of a codon, hence from our viewpoint, reducing the fidelity of information when the rate distortion estimate is violated (see §3.2).

As was recalled in the introduction, Crick’s hypothesis had suggested that no new amino acids could arise without disrupting a large number of proteins, hence stalling evolution – a claim that has since been challenged from many fronts (see e.g. [4, 68]). A product of the coevolutionary dynamics gives rules for load minimization and diversification for regulating patterns of the code that were robust to both error and redundancy, the degrees of which are influenced by the code’s topology that would have been alterable through sequences of stochastic fluctuations. Codons interchanged through error may subsequently be assigned to compatible amino acids so minimizing the possible detrimental effects. At the same time, an enrichment of the vocabulary provided a broader scope for the encoding of proteins [66, 67].

In [77] there is claimed a ‘communality’ and ‘universality’ to be established out of a tournament between a variety of innovative sharing protocols which may include several non-Darwinian mechanisms. Relative to time scales, the long-term reduces ambiguity, whereas in the short-term the code has to be fortified to tolerate a higher degree of ambiguity in assimilating new types of genes. More specifically [77]:

A protein that is robust to translational errors *a fortiori* is also more tolerant to translation with a different code. Conversely, the less optimized the recipient code, the more error-tolerant its proteins, and therefore the less harmful the effect on the established genes of a code change in the direction of the donor code. This has the important consequence that in the initial stages of the genetic code evolution, when the diversification tendency of codes was strongest, HGT (horizontal gene transfer) was possible and must have been extensive despite the presence of many different codes ... Once the optimization of the genetic code is complete, there is no pressure to maintain compatibility. Therefore, the “freezing” of the universal genetic code could trigger the radiation of the underlying translational machineries...

We may reasonably assume that transmission errors eventually corrupt code patterns and those codes that can withstand and manipulate errors possess natural advantages over those that do not. In concluding differently to Crick’s assertion, code-messaging evolution is perceived in [4] as

producing structure preserving codes which have near optimal error-correcting properties, with the selection of mutations and translational error inducing a bias in the codon distribution to amino acids which in the long-term favors optimal error-correction patterns. Crick’s claim of ‘freezing’ makes some sense because the errors themselves condition evolution to some sort of frozen state of an error-correcting code. Specifically, the claim is that an evolutionary constraint on messages with respect to selective pressures, may actually induce the error-correcting codes to evolve rather than to have erased them altogether. Thus, in this evolutionary context the allied and relevant mechanisms of protein synthesis, folding and mutations, provide suitable clues.

An underlying assumption proposed in [1] is that an organism’s complexity reflects upon that of its genome and therefore has evolutionary consequences. So one may ask what actually is the information provided by DNA beyond a road map for the structure of an organism? The current perspective sees this as a blueprint for constructing an organism that can survive within its native environment and then pass on that information to its progeny (cf [33]). In this respect, an organism’s DNA catalogs not only information concerning its structure, but to some extent information concerning its environment and the coevolution of its species as well. In keeping with this basic principle, one may propose an explanation of genomic complexity within the information-theoretic framework of Shannon’s basic principles (see [1, 2] and references therein for related work). It is in this respect that the fundamental theorems of information transmission are sufficiently general to the extent that biological systems can sustain a Shannon-based coding scheme to facilitate the transmission of genomic information within a range of mechanisms, provided that semantics can be incorporated as a functional component (see §4.1 and cf [35]).

## 3 Encoding and Decoding

### 3.1 Basic genetic messaging

The transmission of genetic messaging follows a sequence starting from a source alphabet via a channel code to a target alphabet. The source messaging in the DNA alphabet is relayed to the encoding DNA alphabet to the mRNA alphabet with certain reciprocation. Leading on from mRNA messaging in the RNA alphabet is a channel to point mutation through which (genetic) noise may enter, thence a channel to decoding into which amino acylated tRNA and mischarged tRNA, with further genetic noise, enter via translation. Subsequent to decoding is the protein messaging in the target protein alphabet. This is a basic sequence of events that is schematically represented in [92, Figure 2].

At the same time, evidence suggests that primordial tRNAs along with their various companion types and the overall translation mechanism have coevolved in some de-

gree of compliance with the genetic code, rather than the reverse, and possibly the assignment of amino acids to nucleotides may have been pre-translational. If the code were to be pre-translational in nature, then how it was originally imprinted within tRNAs could be researched in the quest of the so-called ‘RNA world idea’ [63, 72].

### 3.2 The rate distortion function

For the sake of self-containment in this paper, we next briefly recall some elementary facts from the Shannon theory. As it is commonly understood, distortion arises when there is a fast relay of information through some channel which exceeds the latter’s capacity. One of the guiding principles asserts that in order to reproduce a message transmitted from a source to a receiver, it is necessary to know what sort of information should be transmitted, and how. These facts along with specifying the nature of the communicating channel are essential ingredients for engineering a reliable encoding/decoding system. Following [14] we briefly recall some of the basic operations.

**Source encoder:** We may consider some output  $x(t)$  emanating from the source as projected to a finite set of preselected images; namely, the space of possible source outputs is partitioned into a set of *equivalence classes*, and the source encoder informs the channel encoder of that class containing the particular source output observed. Once the channel encoder is informed that the source output belongs to say, the  $m$ -th equivalence class, it transforms the corresponding waveform  $\tilde{x}_m(t)$  across the channel. These equivalence classes as schematically represented by a graph (network), are manifestly the main computational procedures as described in this paper.

**Source decoder:** Within the system is a cascade of a channel encoder and a source decoder. The channel decoder receives a waveform  $\tilde{y}(t)$  of a corresponding function  $y(t)$  over some time interval and decides upon the nature of the message as transmitted. Then it sends its approximation  $m'$  of the message number to the source decoder which in turn creates  $y_{m'}(t)$  to register the system’s estimate of  $x(t)$  over that time interval. Initially, we may think of  $x(t)$  and  $y(t)$  as ‘waveforms’, but in our case, we consider these as consisting of a language with its own intrinsic grammar/syntax, as well as ‘meaning’ – to be made more specific in §4.1. Analogous considerations apply to the channel signals  $\tilde{x}(t)$  and  $\tilde{y}(t)$ .

One of Shannon’s notable results was that a communication system can be designed such that it achieves a level of fidelity  $D$  once the *rate distortion*  $R(D) \leq C$ , where  $C$  denotes the channel capacity. Putting it another way, if the receiver can tolerate an average amount of distortion  $D$ , the rate distortion  $R(D)$  is the effective rate at which the source can relay information with that level of tolerance, and the estimate  $R(D) \leq C$  is a necessary condition for effective communication. More specifically,  $R(D)$  can be defined in terms of *average mutual information* as follows. Firstly, for  $k, j$  running over a suitable alphabet, let us write

a given conditional probability assignment as  $Q(k|j)$  such that in the usual way, we have an associated joint distribution  $P(j, k) = P(j)Q(k|j)$ . We express *the average distortion* as

$$d(Q) = \sum_{j,k} P(j)Q(k|j) d(j, k), \quad (3.1)$$

where  $d(, )$  denotes the distortion measure. A conditional probability assignment  $Q(k|j)$  is said to be  $D$ -admissible if and only if  $d(Q) \leq D$ . The set of all  $D$ -admissible conditional probability assignments we denote by

$$Q_D = \{Q(k|j) : d(Q) \leq D\}. \quad (3.2)$$

Along with an average distortion  $d(Q)$ , we also have an *average mutual information*

$$I(Q) = \sum_{j,k} P(j)Q(k|j) \log \left[ \frac{Q(k|j)}{Q(k)} \right]. \quad (3.3)$$

Then for fixed  $D$ , the rate distortion function is defined as

$$R(D) = \min_{Q \in Q_D} I(Q). \quad (3.4)$$

The rate at which a source produces information subject to insisting upon perfect reproduction, is the *source entropy*  $H$ . Given a distortion measure such that perfect reproduction is assigned zero distortion, then we have  $R(0) = H$ . As  $D$  increases,  $R(D)$  becomes a monotonically decreasing (convex) function which eventually is zero, typically at a maximum value for  $D$  (see [14, Ch. 1]). This is a very basic observation, and typically in rate distortion theory one seeks a reduction of  $H$  by either slowing down the emission of coding, or encoding the relevant languages at a lower rate. In view of Shannon’s theorem, as long as  $H < C$ , there will be suitable fidelity in transmission. In the case of genetic coding considered here, conditions of *discrete memoryless information source* (DMI) and *discrete memoryless channels* (DMC) [57, 92] are usually assumed, but in any event, how well a communicating system can evolve in order to satisfy such an estimate is a common problem for communications engineering since in practice the source rate may be corrupted due to low memory and coding congestion; for protein folding and mutations; references [2, 32, 55, 73, 74, 80, 81] address such issues.

### 3.3 The Groupoid Free Energy Density

Recall that for a thermodynamic state of a given system at fixed temperature  $T$  with energy  $E$  and entropy  $S$ , the *free energy density*  $F$  is defined to be

$$F = E - TS. \quad (3.5)$$

In the Hamiltonian formalism one takes the volume  $V$  and the partition function  $Z(K)$  derived from the system’s

Hamiltonian at inverse temperature  $K$  [51, 52]. The free energy density is then defined to be

$$F[K] = \lim_{V \rightarrow \infty} -\frac{1}{K} \frac{\log[Z(K, V)]}{V} \tag{3.6}$$

$$= \lim_{V \rightarrow \infty} \frac{\log[\widehat{Z}(K, V)]}{V}, \text{ where } \widehat{Z} = Z^{-\frac{1}{K}}.$$

At this stage we introduce the *groupoid* concept (generalizing the algebraic concept of a ‘group’) in relationship to *equivalence classes* which can be based upon a network with concatenation of edges, as explained in Appendix 8.1 (see also [40, 41]). Thus, consider an information source  $H_{G_\alpha}$  over a corresponding groupoid  $G_\alpha$ ; heuristically, we can consider  $H$  as parametrized by  $G_\alpha$ . The probability of  $H_{G_\alpha}$  is given by:

$$P(H_{G_\alpha}) = \frac{\exp[-H_{G_\alpha} K]}{\sum_\beta \exp[-H_{G_\beta} K]}, \tag{3.7}$$

where the normalizing sum is over all possible subgroupoids of the largest available symmetry groupoid. On setting

$$Z_G = \sum_\alpha \exp[-H_{G_\alpha}], \tag{3.8}$$

the *groupoid free energy density (GFE)* of the system  $F_G$  at inverse normalized equivalent temperature  $K$  is then defined as

$$F_G[K] = -\frac{1}{K} \log[Z_G(K)]. \tag{3.9}$$

With each such groupoid  $G_\alpha$  we can associate a dual information source  $H_{G_\alpha}$ . We recall the rate distortion function between the message sent by the cognitive process and the observed impact, while noting that both  $H_{G_\alpha}$  and  $R(D)$  may be considered as free energy density measures. In a sense,  $R(D)$  constitutes a sort of ‘thermal bath’ for the process of cognition. Then the probability of the dual information source can be expressed by

$$P(H_{G_\alpha}) = \frac{\exp[-H_{G_\alpha}/\kappa R(D)\tau]}{\sum_\beta \exp[-H_{G_\beta}/\kappa R(D)\tau]}, \tag{3.10}$$

where  $\kappa$  denotes a suitable dimensionless constant characteristic of the system in the context of a fixed ‘machine response time’  $\tau$ . Associated with (3.10) is a *free energy Morse Function*

$$F_R = -\lambda R(D) \log\left[\sum_{\alpha=1}^n \exp[-H_\alpha/\lambda R(D)]\right], \tag{3.11}$$

whose critical point behavior determines certain topological characteristics of an underlying manifold that can be expressed in terms of its Morse-theoretic indices [56, 58]. In each case the sum is over all possible subgroupoids of the largest available symmetry groupoid (see Appendix 8.1). Accordingly, the term  $R(D)\kappa$  in (3.10) represents a rate distortion energy, in this case, a kind of temperature analog. In the context of a fixed response time  $\tau$ , a decline in

$R(D)$  (on increase in average distortion), acts to ‘lower the machine temperature’ and thus driving it to more simple, albeit less enriched signalling. Observe that if a range over all possible  $\alpha$  is taken, the groupoids  $G_\alpha$  and corresponding relationships such as (3.10), create an even larger picture which reveals the structure of a *groupoid atlas* [9], a concept that has been applied to several descriptive cognitive mechanisms as we have demonstrated in [40, 41, 42].

### 3.4 Phase transition and symmetry breaking

The relationship between phase transitions in physical systems and topological changes has become a central topic of research across a broad range of subdisciplines. One can see that phase transitions in physical systems are ubiquitous, following Landau’s group symmetry shifting arguments [52, 59]. Higher temperatures enable higher system symmetries, and as temperature changes, punctuated shifts to different symmetry states occur in characteristic manners. The claim in [37] is that the standard way of studying phase transitions in a physical system is to consider how the empirical values of thermodynamic states, vary with temperature, volume, or an external field, and then to associate the experimentally observed discontinuities at a phase transition to the occurrence of a singularity. In such a case analyticity may fail in the mathematical sense, though it remains to be seen whether this is the ultimate level of an analytic understanding of such transitional phenomena, or if indeed some reduction to a more basic level is possible. It is observed that non-analyticity is the ‘shadow’ of a more fundamental phenomenon occurring in a given model space: *a topology change*, and that the latter is a *necessary* condition for a phase transition to occur. Such topology changes can be studied within the framework of Morse theoretic influenced topological structures such as the case, say, for certain handle-body decompositions [56], an essential observation that may be consequential for protein functions (cf [82]). Note however, that the converse of the main result of [37] does not hold, thus ruling out a one-to-one correspondence between phase transitions and topology changes. An open problem is that of *sufficiency* conditions; that is, to determine which kinds of topology changes can influence a phase transition, and how this might be achieved. There are other approaches such as demonstrated in relatively straightforward models, where as in [64], a fuzzy clustering system based of annealing through a probabilistic process leads to phase transitions with critical (non-zero) vectors for the free energy at each temperature.

Extension of such transitional arguments in terms of rate distortion and metabolic measures appear direct, particularly in the setting of the groupoids constructed by the disjoint union of the homology groups representing the different coding topologies identified in [73] (see also [80]). To clarify matters, let us recall that in many thermodynamic systems, the associated Hamiltonian may be invariant under a symmetry transformation due to certain parameter

changes, in contrast to the lowest energy state which is not. In subsequent phase transitions the overall symmetry is lost (*spontaneous symmetry breaking*) and consequently, lower temperature states will admit lower symmetries, and due to the randomization of higher temperatures, the higher states will become more accessible to the system as a result of their modified symmetries and energy levels [52]. In the informational context of error-correction, we will need to turn to the fundamental homology between the Shannon entropy and the free energy density of the system as outlined in §4.1.

This scenario becomes more apparent when we look at the symmetries of the genetic code and how these are broken (cf. [71]). For instance, in [46] it is recalled from [15] that the computation of at least  $10^{71}$  to  $10^{84}$  possible genetic codes entails permuting the 64 codons and distributing them over 20 amino acids. By considering those Lie algebras admitting 64 dimensional irreducible representations, [10, 11, 46] initiate a chain of sub-representations commencing from the Lie algebra  $\mathfrak{sp}(6)$ , and postulate a sequence of symmetry breaking in accordance with that chain:

$$\begin{aligned} \mathfrak{sp}(6) &\supset \mathfrak{sp}(4) \oplus \mathfrak{su}(2) \\ &\supset \mathfrak{su}(2) \oplus \mathfrak{su}(2) \oplus \mathfrak{su}(2) \\ &\supset \mathfrak{su}(2) \oplus \mathfrak{u}(1) \oplus \mathfrak{su}(2) \\ &\supset \mathfrak{su}(2) \oplus \mathfrak{u}(1) \oplus \mathfrak{u}(1). \end{aligned} \tag{3.12}$$

At any stage the number of representations occurring corresponds to the number of amino acids that were then incorporated into the code and those currently observed are the net outcome of broken symmetries. In this analysis, four amino acids (phenylalanine, serine, argine and cysteine) seemingly do not divide under the  $U(1)$ (circle)-action. If they had subdivided they would have created a ‘symmetry perfect code’ with 26 amino acids (hence a redundancy of 6) and a stop code (see [46, Figure 1]). Such a claim may be compared with the combinatorial-geometric arguments based on the topology of codon space in [73] (see also §6.1) suggesting that further evolutionary measures may expand the code’s expression from 20 to possibly 25 amino acids.

The observations of [10, 11] reflect back upon an earlier claim of [48] that the ‘freezing’ of the code would have been the result of partial symmetry breaking achieved by the aforementioned parameter choices in the Hamiltonian. The work of [10, 11] differs in its approach by opting for codon-anticodon pairings in place of codon-amino acid assignments and then applying combinatorial-branching techniques commencing from the Lie algebra  $\mathfrak{sl}(6, 1)$ . Besides identifying possible ‘wobble-effects’ due to reshuffling through combinatorial symmetries, they investigate the structure of eukaryotic and vertebrate mitochondrial codes along branching chains and introduce a  $\mathbb{Z}_2$ -grading on codon space (just as there is a grading into bosonic and fermionic types in quantum mechanics) thus extending matters towards representations of super Lie algebras. Along with these codes are variants such as the metabacteria and chloroplast codes with exchange symme-

tries and branching rules for which such patent intricacy may eventually necessitate using groupoid techniques.

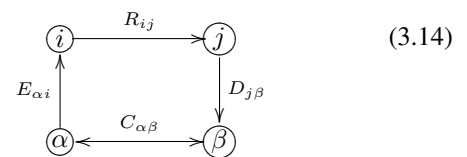
An alternative approach to Lie algebra representations due to [47] is to consider representations on hypercubes as based on Gray coding structures (for a survey of the latter in genetic error-correction, see [45]). Already some known group structures show up here for various assortments of codon doublets, and since sub-symmetries of these representations involve cubical methods, patterns of groupoid symmetries can be expected to appear. Thus we approach increasingly complex situations involving *groupoid representations* (see e.g. [18]) and *groupoid symmetry breaking*, techniques that can be computationally highly non-trivial, since even for relatively straightforward symmetries such as those appearing in certain ‘windmill patterns’, constraints do apply in order to facilitate current programming capabilities [39]. Other questions may arise, such as the possibility of breaking ‘mirror symmetry’ states in the genetic code caused by biochemical perturbations of chiral fields at the molecular level [8].

### 3.5 Amino acid encoding–codon decoding and error load

In order for free energy and error load to fit into the picture, we follow part of the framework of error-correction network analysis of [73, 74] (cf [66]). We take an amino acid  $\alpha$  to be encoded by a unique codon  $j$  represented in the encoder matrix  $[E_{\alpha j}]$ , satisfying  $\sum_j E_{\alpha j} = 1$ , and similarly, the decoder matrix  $[D_{j\beta}]$ , satisfying  $\sum_\beta D_{j\beta} = 1$ , means that each codon is translated into a unique amino acid  $\beta$ , given a number  $\mathcal{N}_c$  of protein chains for  $c$  codons. Next we set

$$R_{ij} = P(\text{the probability that codon } i \text{ may be read correctly as or misread as } j), \tag{3.13}$$

and then let  $[R_{ij}]$  denote the *reading matrix* and  $C_{\alpha\beta}$  the *chemical distance* between the original amino acid  $\alpha$  and the one that is read as  $\beta$ . As adapted from [73, Figure 2] the passage of encoding/decoding then follows as:



On setting  $P_\alpha = P(\text{amino acid } \alpha \text{ is required})$ , the *error load*  $H_{ED}$  (the average distortion in an  $R(D)$  problem) of the map specified by *encoding/decoding* can be expressed in terms of paths  $P_{\alpha ij\beta}$ , specifically by

$$\begin{aligned} H_{ED} &= \sum_{\alpha \rightarrow i \rightarrow j \rightarrow \beta} P_{\alpha ij\beta} C_{\alpha\beta} \\ &= \sum_{\alpha, i, j, \beta} P_\alpha E_{\alpha i} R_{ij} D_{j\beta} C_{\alpha\beta}. \end{aligned} \tag{3.15}$$

This leads to a 'take-over' probability given by  $P_{ED} \sim \exp(-H_{ED}T^{-1})$  and to the *average error load*  $\langle H \rangle$  as follows. If we take  $S$  to denote entropy due to random drift, and  $T$  to be inversely proportional to average error size (the strength of the random drift relative to the selection force that pushes towards maximization), then this probability can be seen to minimize a functional analogous to the Helmholtz free energy  $F$  in terms of the average error load  $\langle H \rangle$  as in (3.5):

$$F = \langle H \rangle - TS = \sum_{ED} H_{ED} P_{ED} + T \sum_{ED} P_{ED} \ln P_{ED}, \quad (3.16)$$

which effectively averages out the difference between the genetic message relayed by a codon statement and that which is actually expressed by the genetic/epigenetic translation machinery itself.

## 4 Meaningful Paths, Robustness and Error Correction

### 4.1 Meaningful paths

We now specify our observations in a more general context. Suppose we consider a pattern of signalling input  $S_i$  describing the state of the protein with initial codon stream  $S_0$  to be mixed in an unspecified but systematic algorithmic manner with a pattern of an otherwise unspecified ongoing activity, including cellular, epigenetic and environmental signals  $W_i$  to create a path of combined signals  $x = (a_0, a_1, \dots, a_n, \dots)$ . Each  $a_k$  thus represents some functional composition of internal and external signals in an iterative form according to which

$$S_{i+1} = f([S_i, W_i]) = f(a_i), \quad (4.1)$$

for some unspecified function  $f$ . Comparing this with the situation in §4.2, the above  $S$  would be a vector,  $W$  a matrix, and  $f$  a product of their function at some time stage  $i$ . This path is fed into a highly nonlinear, but otherwise similarly unspecified, decision oscillator  $h$  which generates an output  $h(x)$  that is an element of one of two disjoint sets  $B_0$  and  $B_1$  of possible system responses, as follows. Let

$$\begin{aligned} B_0 &\equiv b_0, \dots, b_k, \\ B_1 &\equiv b_{k+1}, \dots, b_m. \end{aligned} \quad (4.2)$$

Then:

- (1) assume a graded response, supposing that if

$$h(x) \in B_0, \quad (4.3)$$

the pattern is not recognized, and

- (2) if

$$h(x) \in B_1, \quad (4.4)$$

the pattern is recognized, and some action  $b_j, k + 1 \leq j \leq m$ , takes place.

Expecting the coding signals to be filtered appropriately (cf [4]), we can further assume that  $B_0$  and  $B_1$  admit countable filtrations of the sort:

$$\begin{aligned} B_0 &= B_0^0 \subseteq B_0^1 \subseteq B_0^2 \subseteq \dots \\ B_1 &= B_1^0 \subseteq B_1^1 \subseteq B_1^2 \subseteq \dots \end{aligned} \quad (4.5)$$

where at level  $j$  we have set  $B_0^j \equiv b_0^j, \dots, b_k^j$ , and  $B_1^j \equiv b_{k+1}^j, \dots, b_m^j$ . Note that these oscillators may be influenced by 'forcing' when a signal is subjected to some impulse such that its frequency, and hence the response, adjusts accordingly with respect to an applied impulse. More familiar oscillating physical systems may react accordingly by exhibiting beats and resonance, for instance.

The principal objects of formal interest are paths  $x$  which, through information flow, trigger patterns of recognition-and-response. That is, given a fixed initial state  $a_0 = [S_0, W_0]$ , we examine all possible subsequent paths  $x$  beginning with  $a_0$  and leading to the event  $h(x) \in B_1$ . Thus  $h(a_0, \dots, a_j) \in B_0$  for all  $0 < j < m$ , but  $h(a_0, \dots, a_m) \in B_1$ . We can view  $B_1$  then as the set of final possible states  $S_f \cup \{S_{\text{path}}\}$  that includes both the final physical states and the set of all possible pathological conformations (see [80, Figure 3]).

For each positive integer  $n$ , let  $N(n)$  be the number of high probability grammatical/syntactical paths of length  $n$  which begin with some particular  $a_0$ , and further leading to the condition  $h(x) \in B_1$ . These are paths of combined signals as above, that are structured to some language. For short, we call such paths 'meaningful', assuming, not unreasonably, that  $N(n)$  will be considerably less than the number of all possible paths of length  $n$  leading from  $a_0$  to the condition  $h(x) \in B_1$ .

One critical assumption which permits an inference on the necessary conditions constrained by the asymptotic limit theorems of information theory, is that the entropy, as defined by the finite limit

$$H \equiv \lim_{n \rightarrow \infty} \frac{\log[N(n)]}{n}, \quad (4.6)$$

both exists and is independent of the path  $x$ . The rate distortion principle applies as follows [79]: *the restriction to meaningful sequences of symbols increases the rate at which information can be transmitted with arbitrary small error, and that the grammar/syntax of the path can be associated with a dual information source.*

Besides the DMI and DMC properties introduced in §3.2, we may also assume a typical information source  $X$  to be 'adiabatic', 'piece-wise stationary' and 'ergodic' (APSE), and that the relevant systems engaging in a bio-cognitive process is describable as such. Specifically, the essence of 'adiabatic' is that given the information source is parametrized according to some appropriate scheme, then within continuous 'pieces' of that parametrization, alterations in parameter values occur slowly enough so that the information source  $X$  remains as close to stationary and ergodic as necessary in order to implement the specific

limit theorems. In this way, ‘structure’ is subsumed within the sequential grammar and syntax of the dual information source, rather than within the sets of developmental paths as considered in [85].

In view of (4.6), the Shannon entropy of  $\mathbf{X}$  can be stated more specifically by (see e.g. [5, 14, 29, 49]):

$$H[\mathbf{X}] = \lim_{n \rightarrow \infty} \frac{\log[N(n)]}{n}. \quad (4.7)$$

With respect to e.g. the robustness criteria of §4.2, the time dependent information sources  $\mathbf{X}_i(t)$  are identified with the  $i$ -th component of the expressional pattern  $\mathbf{S}(t)$ ; that is, we assign  $\mathbf{X}_i(t) \mapsto \mathbf{S}_i(t)$ , where as before  $\mathbf{S}_i(t) = f(a_{i-1})$ .

Recalling how the information source uncertainty was defined as in equation (4.6), an essential observation is a *fundamental homology* with the free energy density of a thermodynamical system such as that displayed in equation (3.6). Such a homology arises from Feynman’s observations [36] reflecting in part on Bennett’s work [13] where this homology is effectively an identity, at least for very simple systems. From a more general perspective, [36] postulates the information contained in a message as proportional to the amount of free energy density needed to erase it. This simply amounts to the fact that computing in any form takes work and the more complicated a coding or signalling process so measured by its source uncertainty, the greater its energy consumption. Putting it another way, the less information available to us concerning an event the higher its entropy, and information retrieved is not without a cost in expenditure (of energy), where ‘cost’ is interpreted as the necessary number of bits needed to encode a message (the thermodynamic minimum of energy in terms of bits of information is  $k_B T \log_2 e$  erg/bit, or  $= k_B T$  erg/nat). So the efficiency in an information system essentially happens when there is the minimum amount of energy expended in retrieving information. Specifically, if  $F$  is taken to denote the free energy, then setting  $\Lambda$  equal to the minimum number of nats/sec, the efficiency of the system is given by  $\eta = k_B T F^{-1} \Lambda$  (see e.g. [14]).

## 4.2 Transcriptional regulators and robustness

There are certain evolutionary innovations resulting from an interplay of mutations and natural selections whereby, in a descriptive sense, a genotype corresponds to a regulatory network with a given topology and a phenotype to that of a steady state genetic pattern. This mechanism is constrained by certain conditions requiring processes to sustain a degree of robustness, meaning here a resilience towards environmental perturbations and thermodynamic effects, while at the same time admitting some ‘diversity’ in the process of messaging reception. Such a function of evolution and environment is to ensure that proteins can continue their catalyzing role in the presence of amino acid mutations, that the regulatory networks can continue to function in a

noisy environment, and that embryos can develop normally in the presence of such perturbations. In any case, these regulatory networks, (protein) synthesis and the mutational operations can be seen as part and parcel with the question of folding (misfolding), while observing that error-minimization permits the appropriate codon allocation to amino acids through sequences of broken symmetries in terms of tRNA mutations (see [10, 11]).

Thinking back to the context of §4.1, we next turn to an analogous, but closely related sequence of  $N$  transcriptional regulators represented by their expressional patterns  $\mathbf{S}(t) = (\mathbf{S}_1(t), \mathbf{S}_2(t), \dots, \mathbf{S}_N(t))$ , in network form, at some time  $t$ , that can influence expressions between themselves via cross-regulatory and auto-regulatory interactions as expressed by a matrix  $W = [w_{ij}]$ , where  $w_{ij}$  represents a signaled regulatory influence  $w_{ij} : \text{gene } i \Rightarrow \text{gene } j$ , given the rules (1)  $w_{ij} > 0$ , means activating, (2)  $w_{ij} < 0$ , repressing, and (3)  $w_{ij} = 0$ , absence.

In [25] such regulatory interactions describe the expressional state of the network  $\mathbf{S}(t)$  akin to a typical spin-glass model [21, 69, 91] (see also Appendix 10), as specified by

$$\mathbf{S}_i(t + \tau) = \sigma \left[ \sum_{j=1}^N w_{ij} \mathbf{S}_j(t) \right], \quad (4.8)$$

where  $\tau$  is a constant and  $\sigma(\cdot)$  is a sigmoidal function  $\sigma : \mathbf{S}(t) \rightarrow (-1, 1)$ . For instance, with strong cooperation we may have  $\sigma = \text{sgn}$ , giving  $\mathbf{S}_i = \pm 1$ . Here  $\mathbf{S}(t)$  can be taken as an *incoming input*, mixed in a systematic way relative to  $W = [w_{ij}]$ , to create a path of combined signals  $x = (a_0, a_1, \dots, a_n, \dots)$  as to be seen in §4.1, homologous to the sequence  $\mathbf{S}(t + \Delta t)$ , with  $n = t(\Delta t)^{-1}$ , where on recalling expression (4.1), we set  $\mathbf{S}_{i+1} = f([\mathbf{S}_i, \mathbf{W}_i]) = f(a_i)$ . Accordingly, the structure becomes as much of a function of the sequential grammar and syntax of the dual information source as it is for the cross-sectional intervals of the space of the  $W = [w_{ij}]$  (see [87]). Typically, one would denote by  $\mathbf{S}(0)$  an initial state and by  $\mathbf{S}_\infty$  a stable equilibrium state, with a distance measure  $\mathcal{D}$  for graph topologies  $W, W'$  taken to be

$$\mathcal{D}(W, W') = \frac{1}{2M_+} \sum_{i,j} |\text{sgn}(w_{ij}) - \text{sgn}(w'_{ij})|, \quad (4.9)$$

where  $M_+$  denotes the maximum number of regulatory interactions.

In essence this construction reveals that genotype space, for instance, can be traversed in small increments without changing the phenotype which has evolutionary significance for genetic patterns: randomly selected pairs of networks of the same phenotype may have very different structure and may be subject to varying selective pressures. One may imagine that a large overall ‘diameter’ of the network may be a critical feature for diversity of phenotype, and because some lengthy travel across the graph may be necessary to find all new phenotypes [25], a distance measure of two phenotypes  $\mathbf{S}_\infty, \mathbf{S}'_\infty$  is given by the Hamming

distance  $d_H$  in the form

$$d_H = d_H(\mathbf{S}_\infty(j), \mathbf{S}'_\infty(j)) = 1 - \sum_j \frac{\delta}{N} [\mathbf{S}_\infty(j) - \mathbf{S}'_\infty(j)], \quad 0 \leq d_H \leq 1, \quad (4.10)$$

where Kronecker  $\delta = 1$  should both arguments be equal, and  $\delta = 0$  otherwise. Note that for such Hamming codes it is a basic fact that decoding all patterns of length  $\leq k$  is equivalent to  $(d_H)_{\min} \geq 2k + 1$  (see e.g. [57, 92]).

Related is how, in the statistical mechanics formulation, genetic algorithms based on spin glass models can reveal optimal selectivity as increasing with evolution. In [61] it is shown how selecting those solutions that are at a higher level of fitness, can be paired (through a crossover operation say) and then tested. This is performed iteratively through an algorithm up to the point where there is no further improvement in the examined population. Using spin glass states, [61] apply a chain as represented by vectors of the spins  $\sigma^{(\alpha)}$  (where  $\alpha = 1, \dots, P$ ) indexed by different members of the population; this spin vector is then implemented in the genetic algorithm. In such a case new spins  $\tau_i^\alpha = \sigma_i^\alpha \sigma_{i+1}^\alpha$  are created. Selectivity on the basis of mutation and crossover follows from the energy levels of the Ising spin glass (which is described later in Appendix 10).

## 5 Rate Distortion Coevolutionary Dynamics

### 5.1 The basic equations

Understanding the time dynamics of cognitive systems away from phase transition critical points thus requires a phenomenology similar to the thermodynamic Onsager relations. If the dual source uncertainty of a cognitive process is parametrized by some vector of quantities  $\mathbf{K} \equiv (K_1, \dots, K_m)$ , then in view of the analogy with nonequilibrium thermodynamics, the gradients in the  $K_j$  of the *disorder*, defined as

$$S \equiv H(\mathbf{K}) - \sum_{j=1}^m K_j \partial H / \partial K_j, \quad (5.1)$$

are of central interest. Note that equation (5.1) is analogous to the definition of entropy in terms of the free energy density of a physical system, as suggested by the homology between the latter and the information source uncertainty. Pursuing the homology further, the generalized Onsager relations defining temporal dynamics become

$$dK_j/dt = \sum_i L_{ji} \partial S / \partial K_i, \quad (5.2)$$

where the kinetic coefficients  $L_{ji}$  are, in first order, constants interpreted as reflecting the nature of the underlying cognitive phenomena (without requirement of the symmetry condition  $L_{ij} = L_{ji}$ ). The partial derivatives  $\partial S / \partial K$

are analogous to thermodynamic forces in a chemical system, and may be subject to override by external physiological driving mechanisms as shown in [79, 88] along with further extensions of these dynamical procedures.

Induced by the fundamental homology between the Shannon entropy and free energy density, the rate distortion  $R(D)$  follows a homologous path relation to the latter, thus suggesting that the dynamics of any bio-cognitive module interacting in characteristic real-time  $\tau$ , will be constrained by the system as described in terms of  $R(D)$ . This can be seen more generally [85, 86] by producing a vector-valued function  $R(\mathbf{Q})$  where in the vector  $\mathbf{Q} = (Q_1, \dots, Q_k)$ , the first component is defined to be the average distortion, and then (cf (5.1)), we have

$$S_R \equiv R(\mathbf{Q}) - \sum_{i=1}^m Q_i \partial R / \partial Q_i, \quad (5.3)$$

which leads to the deterministic and stochastic systems of equations analogous to the Onsager relations of nonequilibrium thermodynamics

$$dQ_j/dt = \sum_i L_{ji} \partial S_R / \partial Q_i, \quad (5.4)$$

together with

$$dQ_t^j = L^j(Q_1, \dots, Q_k, t) dt + \sum_i \sigma^{ji}(Q_1, \dots, Q_k, t) dB_t^i, \quad (5.5)$$

where the  $dB_t^i$  represents often highly structured stochastic noise whose properties may be described in terms of Brownian motion and quadratic variation (see e.g. [60]).

### 5.2 The phenomenological Onsager relations

Here we turn to different developmental subprocesses of gene expression characterized by information sources  $H_m$  interacting via chemical or other types of signals, and assume that different processes become each other's principal environments. This is a working hypothesis within a broad coevolutionary context that underscores the cognitive element. Let

$$H_m = H_m(K_1, \dots, K_s, \dots, H_j, \dots), \quad (5.6)$$

where the  $K_s$  represent other relevant parameters, and  $j \neq m$ . We regard the dynamics of this system as driven by a recursive network of stochastic differential equations. Letting the  $K_j$  and  $H_m$  all be represented as parameters  $Q_j$  (with the caveat that  $H_m$  does not depend on itself), we follow the generalized Onsager formulation of [85] in terms of the equation

$$S^m = H_m - \sum_i Q_i \partial H_m / \partial Q_i, \quad (5.7)$$

to obtain a recursive system of *phenomenological Onsager relations*, in terms of a system of stochastic differential equations

$$dQ_t^j = \sum_i [L_{ji}(t, \dots, \partial S^m / \partial Q^i, \dots) dt + \sigma_{ji}(t, \dots, \partial S^m / \partial Q^i, \dots) dB_t^i], \quad (5.8)$$

in which, for ease of notation, both the terms  $H_j$  and the external  $K_j$ 's are expressed by the same symbol  $Q_j$ . As  $m$  ranges over the  $H_m$ , we could allow different kinds of 'noise'  $dB_t^i$  having particular forms of quadratic variation which may represent a projection of environmental factors within the scope of what may be viewed as a *rate distortion manifold* [41]. The noise factor is significant in view of the findings of [7] where it was observed that perturbations of the network parameters inducing stochastic fluctuations in the molecular patterns, may in turn influence regulatory mechanisms, and in a similar way to how the presence of stochastic resonance may amplify certain signals, noise-spectral measurements may then uncover further mechanisms which could be potentially beneficial to the code's evolution.

We remark that equation (5.8) can be generalized somewhat [85] with respect to crosstalk, its distortion, the inherent time constants of the various bio-cognitive modules, and in particular, the overall available free energy density. As shown in [42], analysis of the rate distortion dynamics on a case-by-case basis, motivates integration to a multidimensional Itô process as given by

$$Q_t^\alpha = Q_0^\alpha + \sum_{\beta=\{ij\}} \left[ \int_0^t L_\beta(s, \dots, \partial S_R^\beta / \partial Q^\alpha, \dots) ds + \int_0^t \sigma_\beta(s, \dots, \partial S_R^\beta / \partial Q^\alpha, \dots) dB_s^\beta \right], \quad (5.9)$$

and this in turn leads to a stochastic flow on a suitable topological manifold which in this present context could serve as a more general model for the codon space. In fact, such a flow property had already been observed in [73], namely, that the standard genetic code and its variants evolve as a flow within the codon space. However, given that 'freezing' of some sort is likely to re-occur in the quest for optimal error-correction, we expect such a flow to be stalled at certain time intervals, thus creating singularities in the flow in a dynamical systems sense (an analytic technicality to be finessed here).

### 5.3 A metric on a space of languages

Let us note that equations (5.1) and (5.2) can be derived in a simple parameter-free covariant manner which relies on the underlying topology of the information source space that is implicit to the processes as envisaged. Different bio-cognitive phenomena have, according to our development, dual information sources, and we are interested in the local properties of the system near a particular reference state.

We impose a topology on the system, so that near to a particular language  $A$  dual to an underlying bio-cognitive process, there is an open set  $U$  of closely similar languages  $\hat{A}$ , such that  $A$  and  $\hat{A}$  are subsets of  $U$ .

Since the information sources dual to the processes are similar, for all pairs of languages  $A, \hat{A}$  in  $U$  within a given embedding alphabet, we define a metric on the latter by

$$\mathcal{M}(A, \hat{A}) = \left| \lim_{\int_{A, \hat{A}} d(Ax, \hat{A}x) / \int_{A, A} d(Ax, A\hat{x})} - 1 \right|, \quad (5.10)$$

with respect to a distortion measure  $d(Ax, \hat{A}x)$ , and apply standard integration arguments over the high probability paths, where the usual metric properties apply, as in e.g.[22]. In the context of [4], we may see such a metric as derived from an informational driven physico-chemical distance function with respect to the analogous  $A$  and  $\hat{A}$  coding. Also, since  $H$  and  $\mathcal{M}$  are both scalars, a covariant derivative can be defined directly as

$$dH/d\mathcal{M} = \lim_{\hat{A} \rightarrow A} \frac{H(A) - H(\hat{A})}{\mathcal{M}(A, \hat{A})}, \quad (5.11)$$

where  $H(A)$  is the source uncertainty of language  $A$ .

A relatively straightforward case is the following. Suppose the system is set in some reference configuration  $A_0$ . To obtain the unperturbed dynamics of that state, impose a Legendre transform using this derivative, defining another scalar

$$S \equiv H - \mathcal{M}dH/d\mathcal{M}. \quad (5.12)$$

The simplest possible Onsager relation – here seen as an empirical, fitted, equation like a regression model, becomes

$$d\mathcal{M}/dt = LdS/d\mathcal{M}, \quad (5.13)$$

where  $t$  is the time and  $dS/d\mathcal{M}$  represents an analog to the thermodynamic force in a chemical system (cf [14, §6.4]).

### 5.4 Mutations: mutual entropy between sequence-structure

As analogous to the expressional patterns of §4.2, the previous techniques are applied to the following case of mutations which are themselves functions of evolution, and together with selection and translational error, can influence the distribution of codons to the extent that the latter favor patterns of error-correction that drift to some optimal level and can ameliorate mutation effects [4, 66, 67]. For instance, let us consider as in [55] a series of amino acid sequences

$$\{ \dots, \text{Seq}_{t-1}, \text{Seq}_t, \text{Seq}_{t+1}, \dots \} = \{ \text{Seq}_t \}_{t \in \mathbb{Z}}, \quad (5.14)$$

where each  $\text{Seq}_t$  applies to one protein chain, ordered by a discrete temporal order  $t \in \mathbb{Z}$  of corresponding tertiary structures

$$\{ \dots, \text{Str}_{t-1}, \text{Str}_t, \text{Str}_{t+1}, \dots \} = \{ \text{Str}_t \}_{t \in \mathbb{Z}}. \quad (5.15)$$



Such a chain can be represented as a noisy digital communication channel with an output probability of at least  $\sim 30\%$ , and with a Shannon limit at  $10^{-2}$  bits/amino acid, where at each level  $t$  of sequence-structure we have the coding sequence

$$\begin{aligned} \text{Seq}_t &\Rightarrow \text{Encoder} \Rightarrow \text{Folding channel} \\ &\Rightarrow \text{Decoder} \Rightarrow \text{Str}_t \end{aligned} \quad (5.16)$$

as depicted in [55, Figure 1].

In [4] it is claimed that codes evolving with messages that mutate under such a process, tend to freeze with redundancy. This situation can be reduced to analyzing three different possibilities: the coevolution of genetic codes with:

- (1) transitional-biased message mutation and no translation misreading;
- (2) translational misreading and no transition bias in mutation;
- (3) transition-biased message mutation and translational misreading.

An example in [55] considers concatenated primary sequences  $\{\text{Seq}_t\}_{t \in \mathbb{Z}}$  resulting in a stream of letters from the amino acid alphabet  $A$  with (alphabetical) size  $|A| = 20$ . The encoder is a map that uses a block code of fixed length  $n$ , say, to encode the source through the code book; in other words, a map for every sequence

$$\text{Seq}_t \longrightarrow (\text{single code word}) X^n(\text{Seq}_t), \quad (5.17)$$

represented by an  $n$ -vector  $(X_1, \dots, X_n)$  of integers. The code word in turn belongs to the book of 20 possible structure symbols  $A^* = \{a_1^*, \dots, a_{20}^*\}$ , the finite set of all code words corresponding to the 20 amino acid symbols  $\{A, G, \dots\}$ , where  $a_j^* \in A^*$  are contact vectors determining the amino acid sequence. The message input term  $X^n(\text{Seq}_t)$  from (5.17) is relayed over a noisy channel which then outputs an  $n$ -vector  $\Upsilon^n(\text{Str}_t) = (Y_1, \dots, Y_n)$  representing the folded protein chain  $\text{Str}_t$ , following which a single use of the channels is the transmission of a single amino acid sequence subject to the *channel capacity*

$$C = \max_{p(A)} I(A, A^*). \quad (5.18)$$

In view of §5.3, we modify the role of  $\hat{A}$  via the assignment  $\hat{A} \mapsto A^*$ , and for times stages  $t, t'$ , take as above the metric  $\mathcal{M}(\text{Str}_t, \text{Str}_{t'})$ . At each side of the communication channel we have for the symbol sequences  $|S_A| = 7702314$  amino acid symbols and  $|S_{A^*}| = 31609$  corresponding structural symbols [55].

As for the code rate, we have  $R(D) = H(A)/n$ , where  $H(A)$  is interpreted as the Shannon entropy of the amino acid sequence, where  $n$  is the code block length implemented by the encoder. Assuming the code rate  $R(D)$  and channel capacity  $C$  are known, then in accordance with the Rate Distortion Theorem, we have  $R(D) < C$ , leading to,

for every block size,  $n > n_{\min} = H(A)/C$ , and the codes exist, and no such code when  $R(D) \geq C$ . The Shannon entropy  $H(A) = 3.90$  bits for the amino acid alphabet  $A$ , and  $H(A^*) = 3.76$  bits for the structural code words in  $A^*$  [55]. Further, the mutual entropy between structure and sequence following [2] is given by

$$I(\text{Seq}_t : \text{Str}_t) = H(\text{Seq}_t) - H(\text{Seq}_t | \text{Str}_t), \quad (5.19)$$

and should the environment directly influence the structure, then we would have

$$H(\text{Str}_t | \text{Seq}_t) \simeq H(\text{Seq}_t | \text{Env}_t). \quad (5.20)$$

When taking  $H(\text{Str}_t | \text{Seq}_t) = 0$ , we can re-formulate (5.19) as

$$\begin{aligned} I(\text{Seq}_t : \text{Env}_t) &\simeq I(\text{Seq}_t : \text{Str}_t) \\ &= H(\text{Str}_t) - H(\text{Str}_t | \text{Seq}_t) \\ &= H(\text{Str}_t), \end{aligned} \quad (5.21)$$

which in view of the mutual entropy between sequence and structure, expresses to what extent the thermodynamical entropy of possible protein structures can be constrained by information about the environment as it is coded by the sequence. For instance, excessive noise and random inputs of symbols in  $S_{A^*}$  would most probably corrupt a corresponding code in  $A^*$ , and once again the Shannon estimate serves as a threshold should errors exceed a critical bound. Empirically, the Protein Data Bank (PDB) provides sequence-structure data giving  $H(A) = 3.90$  bits, with block length  $n = 400$ , with transmission rate  $R(D) = 0.010$  bits per amino acid symbol followed, with channel capacity estimated at  $C = 0.016$  bits (per amino acid symbol). When restricted to  $\mathcal{N}_{25} = 2372$  protein chains with mutual sequence identity of  $< 0.25$ , the estimated  $C_{(25)} = 0.016$  bits, was attained (see [55, Figure 4]).

## 6 The Topological Hypothesis and Phase Transitions

### 6.1 The codon space as a graph

The carrier for the dynamics surveyed here is modeled on a rate distortion manifold which has wide-scale overlap with those codon spaces structured in such a way that evolution can be influenced by mapping out those regions which can accommodate load minimization and diversification so that site type, coding fitness, targets, etc. can be correlated as in [4]. One expects the rate distortion manifold to have (in an analytic sense) some degree of differentiability, though here we will finesse this technical issue and elect to consider the underlying combinatorial structure. Specifically, we let  $\Gamma = (V, E)$  denote a graph with  $V$  denoting a finite vertex set,  $E$  an edge set with an oriented edge  $e = (u, v)$  (accordingly,  $e^{-1} = (v, u)$ ) such that  $u = i(e)$  is the initial vertex and  $v = t(e)$  is the terminal vertex, and let  $F$  be

the number of enclosed faces. As seen in [73, 74] there is a formulation of the code that emerges at the phase transition appears in the form of a mode  $e_{\alpha i}$  that minimizes the free energy  $F$ . The codon space can be described as such a graph  $\Gamma$  whose vertices are the codons and two codons  $i, j$  are linked by an edge if (see §3.5) there exists an associated  $R_{ij} (\neq 0)$  in the reading matrix, under the following conditions/observations:

- (1) The vertex set  $V$  consists of codons whereby two codons are linked by an edge in the likelihood they may be confused by misreading.
- (2) Two codons are most likely to be confused if all their letters, except for one, agree and then they are connected by an edge. The resulting graph  $\Gamma$  is natural for considering the impact of translation errors on mutations because such errors almost always involve a single letter difference, that is, a movement along an edge of the graph to a neighboring vertex.
- (3) The native state of the protein has the lowest available free energy induced by the interaction of the amino acid sequence with the embedding environment.
- (4) Recall that there is an embedding  $\Gamma \rightarrow S$  into a surface  $S$ , and the topology of  $\Gamma$  is characterized by its genus  $\gamma(S)$  which is the minimal number of holes required for  $\Gamma$  to be embedded in  $S$  such that no two edges cross. For the underlying network we have the well-known combinatorial formula  $\gamma = 1 - \frac{1}{2}(V - E - F)$ .

Thus the greater the number of connected components in the graph, the higher the genus becomes for a minimal embedding. In [73] the interconnected 64-codon graph can be embedded in a surface with genus  $\gamma(S) = 41$ . If only 48 effective codons are considered, then the genus is reduced to  $\gamma(S) = 25$ .

In light of these observations, it is claimed that the evolution of the code is determined by the underlying topology of its graph and in a transitional phase, it is only those modes with the least error-bound that can emerge and are subjected to alteration by the topology. From the perspective of [59], a free energy argument serves as a Morse function whose critical points characterize just such a topology. More specifically, [73] considers the topology of the code as imposing an upper limit to the number of low modes – critical points – of the corresponding free energy-analog functional, and this is also the number of amino acids. The low modes define a partition of the codon surface into domains, and in each domain a single amino acid is encoded. The partition optimizes the average distortion by minimizing the boundaries between the domains as well as the dissimilarity between neighboring amino acids. This bound on the number of low nodes (and thus as claimed, the number of amino acids) arises as an application of the well-known *chromatic number* as given by Heawood’s formula

[62]:

$$\text{chr}(\gamma(S)) = \text{int}\left[\frac{1}{2}(7 + \sqrt{1 + 48\gamma(S)})\right], \quad (6.1)$$

where  $\text{chr}(\gamma(S))$  is the number of color domains of a surface  $S$  with genus  $\gamma(S)$ , and  $\text{int}[x]$  denotes the integer value of  $x$ . Recall also that the Euler characteristic  $\chi(S) = 2 - 2\gamma(S)$ . In particular, in [73, 75] it is the genus that represents the number of holes in the protein folding error network associated with the code and the chromatic number  $\text{chr}(\gamma(S))$  is a measure of the number of protein symmetries (see Tables 1 and 2.)

**Example 6.1.** Several topological configurations for doublet and triplet codes of 3-letter alphabets drawn from the mRNA alphabet  $\{U, C, G, A\}$  are exhibited in [73, Fig. 3] and are enumerated by (6.1). The topological limit to the number of amino acids (AA’s) for different codes as given by the chromatic number  $\text{chr}(g(S))$  is also given. For instance, a code of 48 codons gives rise to  $g = g(S) = 25$  and  $\text{chr}(g(S)) = 20$ , the maximal number of amino acids. Other cases are listed in [73, Table 1]. Further calculations for pairs  $(g(S), \text{chr}(g(S)))$  are presented in [82] where the chromatic number  $\text{chr}(g(S))$  gives the number of protein symmetries: (0, 4), (1, 7), (2, 8), (3, 9), (5, 10), (6, 11), (7, 11), (8, 12), (9, 12).

More generally, for a topological manifold  $M$  having a Morse function  $F$ ,  $\chi(M)$  can be expressed as the alternating sum of the function’s Morse indices  $\mu_i$  ( $i = 0, 1, \dots, m$ ) of  $F$  on  $M$ , defined as the number of critical points ( $dF(x_c) = 0$ ) of index  $i$ , that is, the number of negative eigenvalues of the matrix  $H_{i,j} = \partial^2 F / \partial x_i \partial x_j$ . Then by the Poincaré-Hopf theorem,

$$\chi(M) = \sum_{i=0}^m (-1)^i \mu_i, \quad (6.2)$$

which holds true for any Morse function on  $M$  (see e.g. [56] and Appendix 9.2 here).

**Remark 6.1.** Applying a spontaneous symmetry breaking argument to  $F_R$  generates topological transitions in the codon graph structure as the ‘temperature’  $R(D)$  increases; that is, as the average distortion  $D$  declines, via the inherent convexity of the rate distortion function. In other words, as the channel capacity connecting codon machines with amino acid machines increases, the more complex coding schemes become possible. In this respect, we recall that for the surface  $S$ , the Euler characteristic  $\chi(S) = 2 - 2\gamma(S)$  as in (9.4) can be expressed in terms of the cohomology structure of  $S$  (e.g. [53, Theorem 13.38]) where by the Poincaré Duality Theorem, the homology groups of a manifold are related to the cohomology groups in the complementary dimension (e.g. [19, p.348]) and thus points to the ‘fundamental homology’ described earlier. One can then envisage the (co)homology groupoid to be taken as the disjoint union of the (co)homology groups of the embedding manifold.

### 6.2 Spectrum of the graph Laplacian

Next we consider the Laplacian  $\Delta$  of  $\Gamma$ . If a pair of vertices  $(i, j) \in E$  are adjacent, then in terms of e.g. the reading matrix  $[R_{ij}]$  (with  $R_{ij} > 0$ ), we have

$$\Delta_{ij} = \Delta_{ji} = -R_{ij} < 0, \tag{6.3}$$

otherwise  $\Delta_{ij} = 0$ , and  $\Delta_{ii} = -\sum_{i \neq j} \Delta_{ij}$  (see Appendix §9.3). For instance, if  $\Gamma$  is taken to be the error graph of §6.1, then  $\Delta$  is the operator that measures the effect of errors and so regulates any phase transition.

Corresponding to the  $n$ -th eigenvalue  $\lambda_n$ , the eigenfunction  $u_n$  admits at most  $n$  weak sign graphs; in particular, for  $n = 2$ , the eigenfunction  $u_2$  divides  $\Gamma$  into precisely two weak sign graphs (see §9.3). Thus it is of interest to determine the dimension of the corresponding eigenspace and multiplicity  $m$  of  $\lambda_2$ . The quantity  $m$  is a measure of the first energy excitation being the primal mode for types of continuous (or second order) phase transitions. The chromatic number  $\text{chr}(\gamma(S))$  of (6.1) identifies the maximal number of first excited modes of the  $\Delta$ .

Letting  $\bar{m}(S)$  denote the supremum of  $m$  over all possible  $\Delta$  on  $S$ , there is the estimate of Colin de Verdière stating that  $\bar{m}(S) \geq \text{chr}(\gamma(S)) - 1$  (see e.g. [74]). In the case of functions, the graph  $\Gamma$  is a reliable 'spectral' model for  $S$  in the sense that from [34, Theorem 5.7], the eigenvalues of all orders of  $\Delta$  on  $\Gamma$  converge to those of the *continuous* Laplacian on functions as defined on  $S$  (see Appendix 9.3).

### 6.3 Phase transitions and holonomy

Given the graph  $\Gamma = (V, E)$ , the *star of a vertex*  $\text{st}(v)$  is the set of edges emanating from  $v$ , that is

$$\text{st}(v) = \{e : i(e) = v\}. \tag{6.4}$$

The various components of the graph may be thought of a comprising a cell network in which the coupling and equivalence of cells leads to a natural groupoid structure having a system of specific equivalence classes  $[v]_V$  and  $[e]_E$ , for vertices and edges, respectively (see Appendix 8.1). With the inclusion of this extra structure we then append  $\Gamma$  to  $\Gamma = (V, E, \sim_v, \sim_e)$ . Here the vertices (nodes) of the network are representative of certain cells where the synchrony of the system depends on groupoid symmetries that in a sense is broken by an impinging rapid crosstalk internal to the system while the latter attempts to manage a slower external crosstalk.

Next, we implement some general procedures based upon the idea of a *connection*  $\nabla$  on  $\Gamma$ , relative to *the stars (st) of vertices* which following [17], is explained with some details in Appendix 9.1 as the combinatorial analog of covariant differentiation (a principle familiar to students of calculus). We take vertices  $(e_1, e_2, \dots, e_{k+1})$  interpreted as  $k+1$  information sources  $(\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_{k+1})$  in accordance with the APSE condition of §4.1, where the  $\mathbf{X}_i$  act with the set of tuning parameters. A connection  $\nabla$

is considered as an operation

$$\nabla(\mathbf{X}_i, \mathbf{X}_j) : \text{st}(\mathbf{X}_i) \longrightarrow \text{st}(\mathbf{X}_j), \tag{6.5}$$

for  $1 \leq i, j \leq k+1$ , satisfying certain properties (see Appendix 9.1). With respect to the metric  $\mathcal{M} = \mathcal{M}(\mathbf{X}_i, \mathbf{X}_j)$  applied to these information sources, the above connection in (6.5) implements on the underlying network, the covariant differentiation along the path  $\mathbf{X}_i \longrightarrow \mathbf{X}_j$ , just as in (5.11):

$$dH/d\mathcal{M} = \lim_{\mathbf{X}_j \rightarrow \mathbf{X}_i} \frac{H(\mathbf{X}_j) - H(\mathbf{X}_i)}{\mathcal{M}(\mathbf{X}_i, \mathbf{X}_j)}. \tag{6.6}$$

Corresponding to each  $\mathbf{X}_i$ , a maximized channel capacity  $C_i$  is assigned, in accordance with the Shannon estimate  $H(\mathbf{X}_i) \leq C_i$ , for  $1 \leq i \leq k+1$ , thus respecting the Rate Distortion Theorem along paths  $\mathbf{X}_j \longrightarrow \mathbf{X}_i$ . If necessary, we can view  $(\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_{k+1})$  as comprising a closed geodesic, and as explained in Appendix 9.1, the set of these in a given graph will thus specify  $\nabla$ . Once we have a handle on  $\nabla$  it is then possible to apply to  $\Gamma$  certain operations analogous to the more familiar differential-geometric setting in order to explore the structural geometry of the various graphs as described (cf [40]).

This technique of the network geometry can be applied to the entropy rates occurring in the various cases we have considered so far. For the sequence-structure-environment in the noisy communication channels along with the data of §5.4, we assign  $\text{Str}_t$  (at time  $t$ ) to a corresponding sensory input  $S_t$ , further combined with environmental signals  $W_t$ , and combined signals  $a_t$  just as in (4.1):

$$\begin{cases} \text{Str}_{t+1} & = f([\text{Str}_t, W_t]) = f(a_t) \\ I(\text{Seq}_t : \text{Env}_t) & = H(\text{Str}_t) \end{cases} \tag{6.7}$$

(here we have made replacements  $i \mapsto t$  and  $j \mapsto t'$ ), where we make a straightforward assignment from the vertex information source, at time  $t$ :

$$\mathbf{X}_t \mapsto \text{Str}_t, \tag{6.8}$$

(and likewise for  $\text{Seq}_t$ ). Using the principle of (6.5) applied to the mutual information

$$I(\text{Seq}_t : \text{Env}_t) = H(\text{Str}_t), \tag{6.9}$$

in (5.21), leads to considering the covariant derivative

$$dH/d\mathcal{M} = \lim_{\text{Str}_t \rightarrow \text{Str}_{t'}} \frac{H(\text{Str}_t) - H(\text{Str}_{t'})}{\mathcal{M}(\text{Str}_t, \text{Str}_{t'})}, \tag{6.10}$$

as implementing the graph connection

$$\nabla(\text{Str}_t, \text{Str}_{t'}) : \text{st}(\text{Str}_t) \longrightarrow \text{st}(\text{Str}_{t'}), \tag{6.11}$$

where again at each time stage  $t$ , the Shannon estimate  $H(\text{Str}_t) \leq C_t$  is observed. Likewise, the error load  $H_{\text{ED}}$  of §3.5 expressed in terms of paths  $P_{\alpha i j \beta}$  in (3.15) and their concatenation, now become the meaningful paths of §4.1.

In this present graph formalism these paths are considered as determined by edges  $e_\nu \in E$ , where each  $\nu = \nu(\alpha i j \beta)$  is a multi-index of the path subscripts.

A property of the connection  $\nabla$  in (6.5) is its *holonomy* which can be best described by considering how, in the traditional differential-geometric sense, a smooth connection implements the parallel translation of vectors around closed paths, and the induced representation of the space of the latter into a group of global symmetries is essentially the holonomy (of the connection). The classic example is the *Poincaré first–return map* of a dynamical system that incorporates typical phase transitions. In the combinatorial setting of [17] the holonomy of  $\nabla$  can be described formally in terms of permuting the ‘stars’ of vertices towards a *spatiotemporal reorientation*, as follows. Let  $\mathcal{C} = \{e_1, \dots, e_n\}$  be any cycle in the graph  $\Gamma$ , for which the terminal and initial vertices satisfy  $t(e_\alpha) = i(e_{\alpha+1})$  modulo  $n$ . Then the connection around  $\mathcal{C}$  leads to a permutation

$$\nabla_{\mathcal{C}} = \nabla_{e_n} \circ \dots \circ \nabla_{e_1} \circ \nabla_{e_0}, \quad (6.12)$$

of the star set  $\text{st}(u)$ . The *holonomy group*  $\text{Hol}(\Gamma, \nabla)_u$  at a vertex  $u$  of  $\Gamma$ , is the subgroup of the *permutation group* of  $\text{st}(u)$  generated by the permutations  $\nabla_{\mathcal{C}}$  over all such cycles  $\mathcal{C}$  that pass through the vertex  $u$ . A phase transition may then be represented by a permutation through vertices in  $\Gamma$ , and such a ‘geometric phase’ accounts for how the various bio-cognitive modules shift gear and create a reorientation of the system.

Now let us return to equivalence classes and the role of groupoids. This implements the above permutation groups of  $\text{st}(u)$ . A holonomy groupoid is obtained via the disjoint union

$$\text{Hol}(\Gamma, \nabla) = \bigvee_{u \in \Gamma} \text{Hol}(\Gamma, \nabla)_u, \quad (6.13)$$

which pieces together the local operations, and at the same time produces an equivalence class representation of the phase transition and its internal amplitudes. We summarize this as follows: *the holonomy groupoid represents a globalization of the local dynamic iterates by providing what is essentially a representation of the graph’s path components onto some prevailing group of symmetries*. In the presence of symmetry breaking, it would be reasonable to consider the groups  $\text{Hol}(\Gamma, \nabla)_u$  as commensurable to some degree with, for instance, the corresponding Lie groups featuring in the  $\mathfrak{sp}(6)$  chain in (3.12), or that of the  $\mathfrak{sl}(6, 1)$  chain as enumerated [10, 11].

## 7 Discussion and Conclusions

The code’s development passed through ‘accidental phases’ created by probabilistic events that could be both regulated and manipulated by an evolving error-correction mechanism. Here we have viewed the latter within the framework of Shannon entropy and the context of the fundamental homology relative to the free energy density of

a thermodynamical system. A common thread to this and other works suggests that increased selection forces may have been significantly enhanced by rate distortion dynamics in regard to the critical behavior of the free energy Morse function and varying topology, a function which would have induced an order of redundancy so mandated by coevolution. Thermodynamic parameter changes in turn induced spontaneous symmetry breaking, which we have shown can be captured by several techniques of representation theory. One can then invert Landau’s arguments and apply them to the (co)homology groupoid in terms of the rising ‘temperature’  $R(D)$ , to obtain a punctuated shift to increasingly complex genetic codes with increasing channel capacity. Our development here realizes mappings **codon space**  $\rightarrow$  **amino acid space** quite explicitly in the context of rate distortion manifolds.

Such arguments can be supported by the known mechanisms occurring in the case of protein folding. The latter originating from an amino acid string is not an entirely random process, but may be the consequence of an evolved structured statement by an information source’s uncertainty, and the occurrence of mutations which may not have been all random but were subject to environmental forces. Thus our present survey, besides regarding the functioning of gene expression as a cognitive process, has a link to the theme of the thermodynamic free energy landscape picture as a function of information sequences [3, 91](cf [54]), evolution as a problem in non-equilibrium statistical physics, and the self-referential character of evolutionary processes at large [43] (cf [83, 84]). We certainly acknowledge (though details are beyond the scope of this survey) that the evolution of organisms has evolved through environmentally sensitive biochemical processes. The phylogenetic analysis of sequence data and branching events suggests that amino acid sequences alter at almost a constant rate which is purported to depend on the functional nature of each class of protein. Thus the changing mechanism has been hypothesized in terms of an evolutionary, stochastic ‘molecular clock’ whereby minor fluctuations can alter the evolutionary rate of certain protein classes [90]. At the same time we have seen in the cognitive paradigm that some organisms may increase their rates of potentially deleterious mutation in response to environmental stress, and such occurrences afford a parallel interpretation in terms of rate distortion analysis as was previously surveyed.

Returning to the redundancy issue, the corresponding evolutionary processes may be capable of extending the code’s expression from 20 to 25 amino acids with the possibility of there being many other protein folding codes [73] (cf [10, 11, 46]). Having said this, we add that there remain a number of open questions concerning the role of the rate distortion function  $R(D)$ , since this in turn drives punctuated changes in the genetic code and further exploration will be necessary. But what seems to follow from the collective processes we have described in explaining ‘the frozen accident’, is that certain *adaptation effects* are

in play (just as one finds in various neurocognitive and biosociological phenomena), and in this respect it seems fitting to quote from [4]:

... Our work has been motivated by the belief that the patterns of the standard genetic code may be explicable as adaptations of a system of information processing. If this turns out to be plausible and correct, we may say that adaptations have reduced the deleterious consequences of genetic and physiological error at a very fundamental level of biological organization ...

So the ‘frozen accident’ by any reasonable account, may have arisen as an evolutionary ‘adaptation’ against a temporary unreadiness (or an enforced over-robustness) to assimilate a barrage of highly complex genetic messaging, in a noisy and not so user-friendly biological environment, during which time error-correction patterns strived to crystallize and to evolve accordingly in order to withstand ongoing selective pressures. It is perhaps from this point of view that advocates of the ‘RNA world idea’ are likely to view a given adaptation at one stage as simply providing a pre-adaptation at another [63, 72].

We point out that holonomy and symmetry breaking are essentially geometric concepts that arise from the iterates of local-to-global procedures, and one such product of this is indeed the holonomy groupoid, a concept that has been introduced in this paper for the purpose of analyzing genetic networks in a novel setting. Further, the question of groupoid representations may uncover deeper conceptual issues in view of representation spaces that are spaces of operators (‘fields of Hilbert or Banach spaces’ as in e.g. [18]), a setting that may be compared the ‘supersymmetric’ model of [10, 11], but one that is likely to be highly non-trivial and costly in a computational sense. Thus in view of the various methods we have brought to the forefront, we cannot fail to acknowledge the remarkable insight of Erwin Schrödinger who claimed that classical physics was insufficient for understanding fundamental life processes. In particular, Schrödinger [65] had envisaged the potential importance of information theory in evolutionary genetics, how living systems can be alterable under thermodynamic effects that are often the results of adverse biological contagion and that quantum mechanical effects might catalyze potential mutations, revealing the organization and evolutionary drive of the genetic code all the more extraordinary.

**Acknowledgements** We wish to thank the reviewers for their various comments and the editors for their management of this paper. JFG wishes to thank Dr. Patrick Coulton for discussions concerning Heawood’s formula. We are also grateful to Tracy Grauman for some production assistance.

## 8 Appendix: Groupoids and Their Atlases

### 8.1 Concept of a groupoid

Many bio-cognitive processes are naturally dynamical systems (see e.g. [40]). One aim in these systems is to unify the internal and external symmetries, and to be able to reduce vast myriad-like network configurations into manageable schemes involving the corresponding equivalence classes analogous to those already mentioned in source encoding/decoding, etc. in §3.2 (see also §5.3 below). A precise way of doing this lies within the categorical concept known as a *groupoid* (see e.g. [20, 28, 89]). In essence, a groupoid  $G$  consists of both a set of objects  $X$  and a set of morphisms, or ‘arrows’, each of which project to an object in  $X$ , and all such morphisms are invertible.

**Remark 8.1.** The most familiar example of a groupoid, as known to students of algebra, is that of a ‘group’ where there is a single object (‘the identity’). Hence groupoids can be viewed as extensions of the ‘group’ concept to sets of *multiple identities* thus providing a wide scope of applications to the dynamics of neurocognitive, socio-bioinformatic and cellular networks (see e.g. [40, 71]).

A groupoid can be depicted by

$$\alpha, \beta : G \begin{array}{c} \xrightarrow{\alpha} \\ \xrightarrow{\beta} \end{array} X \quad (8.1)$$

where the groupoid morphisms  $(\alpha, \beta)$  onto objects, are called the *range* and *source maps*, respectively. Informally, the groupoid represents a feature of built in reciprocity between its algebraic structures, internalizing and externalizing the prevailing symmetries. The morphisms  $\alpha, \beta$  satisfy certain algebraic relations of associativity, existence of two-sided identities, etc. (for details, see [20, 28, 89]). A groupoid can here be understood in relationship to a linkage by a meaningful path of an information source dual to a cognitive process for which the underlying principle is that: *states  $a_j, a_k$  in a set  $A$  are related by the groupoid morphism if and only if there exists a high probability grammatical path connecting them to the same base point, and the tuning across the various possible ways in which that can happen – the different cognitive languages – parametrizes the set of equivalence relations and creates the groupoid.*

**Example 8.1.** Since we have already mentioned equivalence classes in the context of source encoding/decoding, it seems appropriate to see how an equivalence relation  $\mathcal{R}$  defined on (a set)  $X$  takes shape as a groupoid. Here we have the two projections  $\alpha, \beta : \mathcal{R} \rightarrow X$ , and a product  $(x, y)(y, z) = (x, z)$  whenever  $(x, y), (y, z) \in \mathcal{R}$  together with an identity, namely  $(x, x)$ , for each  $x \in X$ . Moreover, the essential equivalence relations and equivalence classes derived from a systems space (network) arise from the orbit equivalence relation of some groupoid  $G$  acting on that space (see e.g. [89]). In the context of con-

nected (sub)networks/graphs with path concatenation, representable in terms of equivalence classes, natural groupoid structures arise in accordance with equivalence classes of relations  $\mathcal{R}(xy)$ , as above, that is simply interpreted as there exists an edge linking node  $x$  to node  $y$  (thus  $x\mathcal{R}y$ ). Conversely, a groupoid (of equivalence relations) admits an underlying graph structure via its implicit scheme of objects and morphisms between objects (for details, see e.g. [20, 44]). Thus we have the two-way associations whereby ‘objects’ can be identified with ‘nodes’, and ‘morphisms’ identified with ‘edges’ in groupoids (of equivalence relations) and networks, respectively:

$$\begin{array}{ccc} \text{Network} & \xrightarrow{\text{equivalence relation}} & \text{Groupoid} \\ \text{Network} & \xleftarrow{\text{underlying graph}} & \text{Groupoid} \end{array}$$

## 9 Appendix: Some Geometry of the Network Architecture: Geodesics and Phase Transitions

### 9.1 Connections on graphs and geodesics

Firstly, for graph-theoretic models there are certain combinatorial notions which can be used to replicate a ‘differential’ structure as realized on a standard differentiable manifold (such as a sphere or a torus). Let  $\Gamma = (V, E)$  be a graph with  $V$  denoting a finite vertex set,  $E$  an edge set with an oriented edge  $e = (u, v)$  (accordingly,  $e^{-1} = (v, u)$ ) such that  $u = i(e)$  is the initial vertex and  $v = t(e)$  is the terminal vertex. The *star of a vertex*  $st(v)$  is the set of edges emanating from  $v$ , that is

$$st(v) = \{e : i(e) = v\}. \tag{9.1}$$

In principle, we would like to handle on both the groupoid and geometric dynamics of a given network. One point is that the star of a vertex may be viewed as the combinatorial version of the tangent space to a manifold at a point, rather similar to how the latter may be regarded as an equivalence class of curves through that point. In [17] there is defined the notion of a *connection*  $\nabla$  on a graph  $\Gamma$  expressed in terms of a set of one-to-one functions  $\nabla(u, v)$ , one for each oriented edge  $e = (u, v)$  of  $\Gamma$  satisfying the following relationships:

- (1)  $\nabla(u, v) : st(u) \rightarrow st(v)$
- (2)  $\nabla(u, v)(u, v) = (v, u)$
- (3)  $\nabla(v, u) = (\nabla(u, v))^{-1}$

Given a graph  $\Gamma$  admits a connection  $\nabla$ , [17] define the notion of a *3-geodesic* as a sequence of four vertices  $(u, v, w, z)$  with edges  $\{u, v\}$ ,  $\{v, w\}$  and  $\{w, z\}$  for which

$$\nabla(v, w)(v, u) = (w, z). \tag{9.2}$$

**Remark 9.1.** In differential calculus, a ‘connection’ is simply a generalized gradient implementing covariant differentiation. We have already encountered a form of this in (5.11). The notion of a graph/network connection introduced here is a more manageable concept, particularly for bio-cognitive modules, and does not involve applying the advanced techniques of calculus.

A *k-geodesic* is defined inductively across a sequence of  $(k + 1)$  vertices. The three consecutive edges  $\{d, e, f\}$  of a 3-geodesic is referred to as *an edge chain*. A *closed geodesic* can then be specified as a sequence of edges  $e_1, \dots, e_n$  such that each consecutive triple  $(e_\alpha, e_{\alpha+1}, e_{\alpha+2})$  is an edge chain for each  $1 \leq \alpha \leq n$ , modulo  $n$ . The geodesic returns to the same pair of edges in the same order. Thus one finds a unique closed geodesic through each pair of edges in the star of the vertex, and as pointed out in [17], the set of all closed geodesics completely determines the connection on the graph.

In terms of the geometric evolution of our networks, the family  $(G_{\mathcal{A}}, \nabla_{\mathcal{A}})$  of local groupoids with connection satisfies:

- (1) Once  $\nabla_{\mathcal{A}}$  is given, then the graph geodesics can be derived iteratively from (9.2).
- (2) Conversely, given the underlying graph of each  $G_{\mathcal{A}}$ , the connection  $\nabla_{\mathcal{A}}$  is determined by the set of all closed geodesics as specified.

We also have the following useful characterization [17]: given  $(\Gamma, \nabla)$ , a subgraph  $\Gamma_0 = (V_0, E_0) \subset \Gamma$  is said to be *totally geodesic* if all geodesics commencing at  $E_0$  remain within  $E_0$ . In other words, for every two adjacent vertices  $u, v$  in  $\Gamma_0$ , we have

$$\nabla(u, v)(st(u) \cap E_0) \subseteq E_0. \tag{9.3}$$

Note that the above concepts have been formulated graph-theoretically, and as mentioned in Remark 9.1, they do not require the usual manipulations of advanced differential calculus.

### 9.2 The graph Betti numbers

By analogy with finding the dimensions of the homology groups of a topological manifold, [17] specify the notion of *Betti numbers* associated with  $\Gamma$ . This involves the using certain concepts such as an *axial function*  $\varphi$  and *generic direction*  $\xi$ . Thus we regard  $(\Gamma, \nabla)$  as having an axial function  $\varphi$  and write this as  $(\Gamma, \varphi)$  when  $\nabla$  is understood. In which case the *index* of a vertex  $u \in V$  is the number of edges  $e \in st(u)$  such that the product  $\varphi(e) \cdot \xi < 0$ . Let  $\beta_i(\xi)$  denote the number of vertices  $u$  such that the index at  $u$  is exactly  $i$ . When these values do not depend on the choice of direction  $\xi$ , they are called the *Betti numbers of  $(\Gamma, \varphi)$* , and satisfy a combinatorial duality condition  $\beta_i(\Gamma, \varphi) = \beta_{d-i}(\Gamma, \varphi)$ , for  $1 \leq i \leq d$ . In certain cases, they can shown to be similar to the indices of a standard

Morse function (see [17, 58, 56]) such as  $F_R$  in (3.11). Thus on the underlying graph of the groupoid on which  $F_R$  is defined, we identify  $F_R$  with a Morse function compatible with a generic direction on  $(\Gamma, \varphi)$  whose index is essentially a measure of the homology of information relay within the graph, where at level  $i$ , we have  $\mu_i = \beta_i(\Gamma, \varphi)$ .

In fact, to clarify the role of the topological invariants of  $\Gamma$  to those of the surface  $S$ , we need the following description. Firstly,  $S$  taken to be a compact surface permits seeing  $S$  also as a (connected) compact, one-dimensional complex manifold (viz. a Riemann surface) on which a certain analytic group action takes place. The standard way of representing  $\Gamma$  (see e.g. [17, §4]) is to identify  $V$  as the (finite) fixed point set, and  $E$  as the (finite) set of one-dimensional orbits of this action. Consequently, the  $\beta_i(\Gamma, \varphi)$  coincide with the usual Betti numbers  $\beta_i(S)$  of  $S$ , and by the Poincaré-Hopf Theorem we have

$$\chi(S) = \sum_i (-1)^i \beta_i(\Gamma, \varphi). \tag{9.4}$$

### 9.3 The graph Laplacian

Suppose now that  $\Gamma = (V, E)$  is an undirected loop-free graph. If the vertices(nodes) are indexed  $1 \leq i \leq N$ , then the *graph Laplacian*  $\Delta$  can be viewed as a symmetric  $N \times N$  matrix defined as follows (see e.g. [16, 74]):

- (1) If vertices  $(i, j) \in E$  are adjacent, then the corresponding entry in the matrix  $\Delta_{ij} = \Delta_{ji} < 0$ .
- (2) Otherwise,  $\Delta_{ij} = 0$ , and the diagonal terms imply that the sum over rows and columns vanishes, leading to  $\Delta_{ii} = -\sum_{i \neq j} \Delta_{ij}$ .

Note the term 'weighted Laplacian' is sometimes used for the operator  $\Delta$ , whereas in other cases 'Laplacian' is used for when the negative entries are all  $\Delta_{ij} = -1$ . Specifically, if  $f : V \rightarrow \mathbb{R}$  is a vector function induced by the vertices of  $\Gamma$ , and  $x \sim y$  denotes there is an edge linking  $x$  and  $y$ , then from [16]:

$$(\Delta f)(x) = -\sum_{x \sim y} [f(x) - f(y)]. \tag{9.5}$$

Of particular interest are the eigenvalues of  $\Delta$  ordered as  $0 = \lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_N$ , obtainable through the spectrum of an associated operator  $L$ , for which

$$\langle f, Lf \rangle = \sum_{x, y \in V} L_{xy} f(x) f(y) = \sum_{xy \in E} [f(x) - f(y)]^2. \tag{9.6}$$

Also, we have the *Rayleigh Quotients* [16], given by

$$\begin{cases} \mathcal{R}_\Delta &= \frac{\langle f, \Delta f \rangle}{\langle f, f \rangle} \\ \mathcal{R}_L(f) &= \frac{\sum_{xy \in E} [f(x) - f(y)]^2}{\sum_{x \in V} f(x)^2}. \end{cases} \tag{9.7}$$

In [34, Theorem 5.7] estimates on (9.7) lead to showing that, in the case of functions, the eigenvalues of the graph

Laplacian converge to those of the continuous Laplacian. Further, in [34] it is shown that the zeta functions of the former converge to those of the latter, where

$$\zeta^{(n)}(s) = \sum_{\lambda_k^n \neq 0} (\lambda_k^n)^{-s}. \tag{9.8}$$

In the continuous case, sets that are the zero-level sets of the eigenfunctions are called *nodal sets*, and *nodal domains* are those sets in which a corresponding eigenfunction takes on one sign and they are separated by nodal sets. Courant's nodal line theorem (see e.g. [24]) states that if the eigenfunctions of a continuous Laplacian on a domain are ordered according to increasing eigenvalues, then the nodes of the  $n$ -th eigenfunction divide the domain into no more than  $n$  nodal domains. In the combinatorial case, for the graph Laplacian, the nodal domains become *sign-graphs*: maximal connected subgraphs on which an eigenfunction carries the same sign. On *weak sign-graphs* the eigenfunction is either  $\geq 0$  or  $\leq 0$ , while on *strong sign-graphs*, the sign of the eigenfunction is either  $> 0$  or  $< 0$ . This leads to an analogue of Courant's nodal line theorem in the combinatorial case [16]: *On a connected graph  $\Gamma$ , the  $n$ -th eigenfunction  $u_n$  of the Laplacian  $\Delta$  admits at most  $n$  weak sign graphs.* The case  $n = 2$  is significant because the corresponding eigenfunction  $u_2$  then splits  $\Gamma$  into exactly two weak sign graphs and  $\lambda_2$  is significant for Brownian motion on the graph and to its first excited energy level.

## 10 Spin Glasses in Brief

Spin glass models, as discrete structures, may be based on combinatorial decompositions of surfaces usually in some square lattice configuration which can be modified (e.g. from square to triangular). The basic idea leading to the prototypical *2-dimensional Ising model* goes as follows (we follow [23, 69]). Firstly, consider a sequence of symbols  $a_i = 0, 1$  and a signal  $v_i$  transmitted across some time interval. Set  $v_i = v$  if  $a_i = 1$ , and  $v_i = -v$  if  $a_i = 0$ . Then let  $a(i, j)$  (for  $1 \leq i, j \leq m$ ) denote the  $m^2$  bits of information transmitted. These are subject to redundancy relations

$$\begin{aligned} a(i, m+1) &= \sum_{j=1}^m a(i, j), \\ a(m+1, j) &= \sum_{i=1}^m a(i, j), \end{aligned} \tag{10.1}$$

with addition mod 2. The quantity  $m^2/(m+1)^2$  is the rate of the code and measures the redundancy. With noise terms  $y(i, j)$  included, the modified signal is then taken to be  $u(i, j) = v(i, j) + y(i, j)$ . This leads to a simple error-correcting code that is of the Hamming type [57]. Further, the correspondence  $u(i, j) = \frac{1}{2}(\sigma(i, j) + 1)$  between information bits and Ising spins or qubits  $\sigma(i, j)$  in mod 2 addition and spin multiplication respectively, are equivalent.

More specifically, let a qubit  $\sigma(i, j)$  be attached to each edge of some lattice which is to be viewed as a configuration space  $P = \{\pm\}^{\mathbb{Z}^2}$ . On taking  $J_1$  (horizontal) and  $J_2$  (vertical) to be interaction constants, the Hamiltonian  $\mathcal{H}(\sigma)$  is given by

$$\mathcal{H}(\sigma) = - \sum J_1 \sigma(i, j) \sigma(i+1, j) + J_2 \sigma(i, j) \sigma(i, j+1), \quad (10.2)$$

for the appropriate ranges of summation. Suppose we consider  $H(\sigma)$  over a finite lattice given by  $\Lambda_{LM} = \{(i, j) : |i| \leq M, |j| > L\}$ , and then take the thermodynamic limit. If  $J_1, J_2 > 0$ , there are interactions in which the energy is minimized on alignment of all of the spins. Then either:

- i) all are  $\uparrow$  or  $\downarrow$ , or,
- ii)  $\sigma(i, j) \equiv 1$ , or  $\sigma(i, j) \equiv -1$ , respectively.

For absolute temperature  $T$ , the equilibrium state is that which minimizes [internal energy]  $- T \cdot$  [entropy]. Within the model two competing forces can be realized by the following:

1. One minimizes the internal energy by attempting to align the signs either  $\uparrow$  or  $\downarrow$  to create order: it wins if  $T$  is small.
2. The other, on maximizing entropy, attempts to produce as much chaos as possible: it wins if  $T$  is large.

At finite critical temperature  $T_c$ , chaos wins if  $T \geq T_c$ , and order wins if  $T < T_c$ .

## References

- [1] Adami, C., Ofria, C., and Collier, T., 2000, Evolution of biological complexity, *Proc. Natl. Acad. Sci. USA* **97**, 4463–4468.
- [2] Adami, C., 2004, Information theory in molecular biology, *Physics of Life Reviews* **1**, 3–22.
- [3] Anfinsen, C.B., 1973, Principles that govern the folding of protein chains, *Science* **181** (96), 223–230.
- [4] Ardell, D.H., and Sella, G., 2002, No accident: genetic codes freeze in error-correcting patterns of the standard genetic code, *Phil. Trans. R. Soc. Lond. B* DOI 10.1098/rstb.2002.1071
- [5] Ash, R., 1990, *Information Theory*, Dover Publications, New York.
- [6] Atlan, H., and Cohen, I., 1998, Immune information, self-organization and meaning *Int. Immunology*, **10**, 711–717.
- [7] Austin, D.W., Allen, M.S., McCollum, J.M., Dar, R.D., Wilgus, J.R., Saylor, G.S., Samatova, N.F., Cox, C.D., and Simpson, M.L., 2006, Gene network shaping of inherent noise spectra, *Nature* **439**, doi:10.1038/nature04194
- [8] Avetisov, V., and Goldanskii, V., 1996, Mirror symmetry breaking at the molecular level, *Proc. Natl. Acad. Sci. USA* **93**, 11435–11442.
- [9] Bak, A., Brown, R., Minian, G., and Porter, T., 2006, Global actions, groupoid atlases and related topics, *J. Homotopy and Related Structures* **1**, 1–54.
- [10] Bashford, J.D., Tsochantjis, I., and Jarvis, P.D., 1998, A supersymmetric model for the evolution of the genetic code, *Proc. Natl. Acad. Sci. USA* **95**, 987–992.
- [11] Bashford, J.D., and Jarvis, P.D., 2008, Spectroscopy of the genetic code, in (Abbott, D. et al. eds) *Quantum Aspects of Life*, Imperial College Press, London.
- [12] Belongie, M.L., 1994, Spin glasses and error-correcting codes, *TDA Progress Report 42–118*, 26–36.
- [13] Bennett, C.H., 1982, The thermodynamics of computation: a Review, *Internat. J. Theor. Phys.* **21** (12), 905–940.
- [14] Berger, T., 1971, *Rate Distortion Theory: A mathematical basis for data compression*, Prentice–Hall Inc., Englewood Cliffs, NJ.
- [15] Bertman, M.O., and Jungck, J.R., 1978, Some unresolved mathematical problems in genetic coding, *Notices Amer. Math. Soc.* **25** A-174.
- [16] Biyikoğlu, T., Leydold, J., and Stadler, P.F., 2007, *Laplacian Eigenvectors of Graphs*, Lecture Notes in Math. **1915**, Springer-Verlag, Berlin Heidelberg.
- [17] Bolker, E.D., Guillemin, V.W., and Holm, T.S., 2006, How is a graph like a manifold?, to appear. <http://arxiv.math.CO/0206103>
- [18] Bos, R., 2011, Continuous representations of groupoids, *Houston J. Math.* **37**(3), 807–844.
- [19] Bredon, G., 1993, *Topology and Geometry*, Springer, New York.
- [20] Brown, R., 2006, *Topology and Groupoids* (3rd Ed.), BookSurge LLC, Charleston, S. Carolina.
- [21] Bryngelson, J.D., and Wolynes, P.G., 1987, Spin glasses and the statistical mechanics of protein folding, *Proc. Natl. Acad. Sci. USA* **84**, 7524–7528.
- [22] Burago, D., Burago, Y., and Ivanov, S., 2001, *A Course in Metric Geometry*, American Mathematical Society, Providence, RI.



- [23] Carey, A., and Evans, D. E., 1988, The operator algebras of the two-dimensional Ising model, in ‘Braids’ (Santa Cruz, CA, 1986), 117–165, *Contemp. Math.* **78**, Amer. Math. Soc., Providence, RI.
- [24] Chavel, I., 1984, *Eigenvalues in Riemannian Geometry*, Academic Press, New York.
- [25] Ciliberti, S. Martin, O., and Wagner, A., 2007, Innovation and robustness in complex regulatory genetic networks, *Proc. Natl. Acad. Sci. USA* **104**, 13591–13596.
- [26] Cohen, I., 2000, *Tending Adam’s Garden: Evolving the Cognitive Immune Self*, Academic Press, New York.
- [27] Cohen, I., and Harel, D., 2007, Explaining a complex living system: dynamics, multiscaling and emergence, *J. Royal Soc. Interface* **4**, 175–182.
- [28] Connes, A., 1994, *Noncommutative Geometry*, Academic Press, San Diego, CA.
- [29] Cover, T., and Thomas, J., 1991, *Elements of Information Theory*, John Wiley and Sons, New York.
- [30] Crick, F., 1966, Codon-anticodon pairing: the wobble hypothesis, *J. Mol. Biol.* **19**, 548–553.
- [31] Crick, F., 1968, The origin of the genetic code, *J. Mol. Biol.* **38**, 367–379.
- [32] Crooks, G.E. and Brenner, S., 2004, Protein secondary structure: entropy, correlations and prediction, *Bioinformatics* **20**(10), 1603–1611.
- [33] Dawkins, R., 1976, *The Selfish Gene*, Oxford Univ. Press, London.
- [34] Dodziuk, J., 1976, Finite-difference approach to the Hodge theory of harmonic forms, *Amer. J. Math.* **98**(1), 79–104.
- [35] Dretske, F., 1981, *Knowledge and the Flow of Information*, MIT Press, Cambridge, MA.
- [36] Feynman, R., 1996, *Feynman Lectures on Computation*, Addison-Wesley, Reading, MA.
- [37] Franzosi, R., and Pettini, M., 2004, Theorem on the origin of phase transitions, *Phys. Rev. Lett.* **92**:060601.
- [38] Freeland, S.J., Wu, T., and Keulmann, N., 2003, The case for an error minimizing standard genetic code, *Origins of Life and Evolution of the Biosphere* **33**, 457–477.
- [39] Gent, I.P., Kelsey, T., Linton, S., Pearson, J., and Roney-Dougal, C.M., 2010, Groupoids and conditional symmetry, preprint, Univ. St. Andrews, UK.
- [40] Glazebrook, J.F., and Wallace, R., 2009, Small Worlds and Red Queens in the Global Workspace: an information-theoretic approach, *Cognitive Systems Research* **10**, 333–365.
- [41] Glazebrook, J.F., and Wallace, R., 2009, Rate distortion manifolds as model spaces for cognitive information, *Informatica* **33** (2009), 309–345.
- [42] Glazebrook, J.F., and Wallace, R., 2010, Rate distortion coevolutionary dynamics and the flow nature of cognitive epigenetic systems, arXiv:1101.4984v1 [q-bio.OT]
- [43] Goldenfeld, N., and Woese, C., 2011, Life is physics: evolution as a collective phenomenon far from equilibrium, *Annu. Rev. Condens. Matter Phys.* **2**, 375–399.
- [44] Golubitsky, M., and Stewart, I., 2006, Nonlinear dynamics and networks: the groupoid formalism, *Bull. Amer. Math. Soc.* **43**, 305–364.
- [45] Gupta, M.K., 2006, The quest for error correction in biology, *IEEE Engineering in Medicine and Biology Magazine*, 46–53.
- [46] Hornos, J.E., and Hornos, Y.M., 1994, A search for symmetries in the genetic code, *J. Bio. Phys.* **20**, 289–294.
- [47] Jiménez-Montaña, M.A., de la Mora-Basáñez, C.R., and Pöschel, T., 1996, The hypercube structure of the genetic code explains non-conservative aminoacid substitutions *in vivo* and *in vitro*, *BioSystems* **39**, 117–125.
- [48] Jukes, T.H., 1983, Evolution of the amino acid code, pp. 191–207 in (Nei, M. et al. eds.) *Evolution of Genes and Protein*, Sinauer, Sunderland, MA.
- [49] Khinchin A., 1957, *The Mathematical Foundations of Information Theory*, Dover Publications, New York.
- [50] Koonin, E., and Novozhilov, A., 2009, Origin and evolution of the genetic code: the universal enigma, *Life* **61**, 99–111.
- [51] Kurzynski, M., 2006, *The Thermodynamic Machinery of Life*, Springer-Verlag, Berlin Heidelberg New York.
- [52] Landau, L., and Lifshitz E., 2007, *Statistical Physics (I)* (3rd Ed.), Elsevier, New York.
- [53] Lee, J., 2000, *Introduction to Topological Manifolds*, Springer, New York.
- [54] Levinthal, L., 1968, Are there pathways for protein folding?, *J. Chim. Phys. PCB* **65**, 44–45.

- [55] Lisewski, A. M., 2008, Random amino acid mutations and protein misfolding lead to Shannon limit in sequence-structure communication, *PloS ONE*, **3**(9), e3110 doi:10.371/journal.pone.0003110
- [56] Matsumoto, Y., 2001, *An Introduction to Morse Theory*, Translations Amer. Math. Soc. **208**, Providence, RI.
- [57] McEliece, R.J., 2004, *The Theory of Information and Coding*, Encyclopedia of Mathematics and its Applications, Vol. 86, Cambridge University Press.
- [58] Milnor, J., 1963, *Morse Theory*, Princeton University Press, Princeton, NJ.
- [59] Pettini, M., 2007, *Geometry and Topology in Hamiltonian Dynamics*, Springer, New York.
- [60] Protter, P., 1995, *Stochastic Integration and Differential Equations: A New Approach*, Springer, New York.
- [61] Prügel-Bennett, A. and Shapiro, J. L., 1994, Analysis of genetic algorithms using statistical mechanics, *Phys. Rev. Lett.* **72**(9), 1305–1309.
- [62] Ringel, G., and Youngs J., 1968, Solutions of the Heawood map-coloring problem, *Proc. Natl. Acad. Sci. USA* **60**, 438–445.
- [63] Rodin, A.S., Szathmáry, E., and Rodin, S.N., 2011, On origin of genetic code and tRNA before translation, *Biology Direct* **6**:14, 1–24.
- [64] Rose, K., Gurewitz, E., and Fox, G.C., 1990, Statistical mechanics and phase transitions in clustering, *Phys. Rev. Lett.* **65** No. 6, 945–948.
- [65] Schrödinger, E., 1967, *What is Life?*, Cambridge University Press.
- [66] Sella, G. and Ardell, D.H., 2002, The impact of message mutation on the fitness of the genetic code, *J. Mol. Evol.* **54**, 638–651.
- [67] Sella, G. and Ardell, D.H., 2006, The coevolution of genes and genetic codes: Crick's frozen accident revisited, *J. Mol. Evol.* **63**, 297–313.
- [68] Söll, D., and RajBhandary, U.L., 2006, The genetic code—Thawing the 'frozen accident', *J. Biosci.* **31**(4), 459–463.
- [69] Sourlas, N., 1989, Spin-glass models as error-correcting codes, *Nature* **339**, 693–695. doi:10.1038/339693a0
- [70] Stewart, I., Golubitsky, M., and Pivato, M., 2003, Symmetry groupoids and patterns of synchrony in coupled cell networks, *SIAM J. Appl. Dynam. Sys.* **2**, 609–646.
- [71] Stewart, I., 1994, Broken symmetry in the genetic code?, *New Scientist* **1915**, 16.
- [72] Szathmáry, E., 1999, The origin of the genetic code, *Trends in Genetics* **15**(6), 223–229.
- [73] Tlusty, T., 2007, A model for the emergence of the genetic code as a transition in a noisy information channel, *J. Theor. Bio.* **249**, 331–342.
- [74] Tlusty, T., 2008, Rate-distortion scenario for the emergence and evolution of noisy molecular codes, *Phys. Rev. Lett.* **100**, 048101 (1-4).
- [75] Tlusty, T., 2007, A relation between the multiplicity of the second eigenvalue of a graph Laplacian, Courant's nodal line theorem and the substantial dimension of tight polyhedral surfaces, *Elect. J. Linear Algebra* **16**, 315–324.
- [76] Tlusty, T., 2008, A simple model for the evolution of molecular codes driven by the interplay of accuracy, diversity and cost, *Physical Biology* **5**, 016001.
- [77] Vetsigian, K., Woese, C., and Goldenfeld, N., 2006, Collective evolution and the genetic code, *Proc. Natl. Acad. Sci. USA* **103**, 10696–10701.
- [78] de Vlarar, H.P., and Barton, N.H., 2011, The contribution of statistical physics to evolutionary biology, *Trends in Ecology and Evolution* **26**(8), 424–432.
- [79] Wallace, R., 2005, *Consciousness: A Mathematical Treatment of the Global Neuronal Workspace Model*, Springer, New York.
- [80] Wallace, R., 2010, A rate distortion approach to protein symmetry, *BioSystems* **101**, 97–108.
- [81] Wallace, R., 2011, Metabolic constraints on the evolution of genetic codes: Did multiple 'preaerobic' ecosystem transitions entrain richer dialects via Serial Endosymbiosis, <http://precedings.nature.com/documents/4120/version/4>.
- [82] Wallace, R., 2011, Structure and dynamics of the 'protein folding code' inferred using Tlusty's topological rate distortion approach, *Biosystems* **103**, 18–26.
- [83] Wallace, R., and Wallace R.G., 1998, Information theory, scaling laws, and the thermodynamics of evolution, *J. Theor. Bio.* **192**, 545–559.
- [84] Wallace, R., and Wallace R.G., 1999, Organisms, organizations and interactions: an information theory approach to biocultural evolution, *BioSystems* **51**, 101–119.
- [85] Wallace, R., and Wallace, D., 2009, *Gene Expression and its Discontents: The social production of pandemic chronic disease*, Springer, New York.

- [86] Wallace, R., and Wallace, D., 2008, Punctuated equilibrium in statistical models of generalized coevolutionary resilience: how sudden ecosystem transitions can entrain both phenotype expression and Darwinian selection, *Trans. Comp. Systems Biology IX*, LNBI 5121, 23–85.
- [87] Wallace, R., and Wallace, D., 2011, Cultural epigenetics: on the heritability of complex diseases, *Trans. Comp. Systems Biology XIII*, LNBI 6575, 131–170.
- [88] Wallace, R., and Fullilove M., 2008, *Collective Consciousness and Its Discontents: Institutional distributed cognition, racial policy, and public health in the United States*, Springer, New York.
- [89] Weinstein, A., 1996, Groupoids: unifying internal and external symmetry, *Notices Amer. Math. Soc.* **43**, 744–752.
- [90] Wilson, A.C., Carlson, S.S., and White, T.J., 1977, Biochemical Evolution, *Ann. Rev. Biochem.* **46**, 573–639.
- [91] Wolynes, P.G., 1996, Symmetry and the energy landscapes of biomolecules, *Proc. Nat. Acad. Sci. USA* **93**, 14249–14255.
- [92] Yockey, H.P., 2005, *Information Theory, Evolution and the Origin of Life*, Cambridge University Press.



# Times Limited Accountable Anonymous Online Submission Control System from Single-Verifier $k$ -times Group Signature

Xingwen Zhao and Fangguo Zhang

School of Information Science and Technology, Sun Yat-Sen University, Guangzhou 510275, P.R.China

Guangdong Key Laboratory of Information Security Technology, Guangzhou 510275, P.R.China

E-mail: sevenzhao@hotmail.com, isszhfg@mail.sysu.edu.cn

**Keywords:** privacy, anonymous authentication, accountability, group signature

**Received:** May 26, 2010

*People in authority may want to submit some messages anonymously on a famous website, while the maintainers may want to limit the times each person can submit messages on the website so as to save the storage space. More over, when people abuse the system, the maintainers want to find ways to identify their identities. To realize such a system, what we need are some methods that can protect users' privacy, control their access times, and at the same time can identify malicious users when abuses are found. Current signature schemes or credential systems cannot fully achieve above purpose. A single-verifier  $k$ -times group signature scheme is proposed, adding times limited property to the group signature scheme. It allows a user to issue group signatures to the only verifier up to  $k_i$  times for period  $T_i$ . We use online tracing method to restrict each user to  $k_i$  signatures strictly, and use the tracing ability of group signature to identify those who abuse the system. Based on it, we can construct times limited accountable anonymous online submission control system for websites. Within allowed times, people can submit articles anonymously, even website maintainers cannot identify two articles are from the same person. When a person posts more than the allowed times, his submission will be rejected. When abuse is found, website maintainers can send the signature to the corresponding open authority to find out the identity.*

*Povzetek: Članek predlaga metodo podpisovanja spletnih sporočil, ki zagotavlja anonimnost le pri omejenem številu uporab.*

## 1 Introduction

Nowadays people may want to post articles anonymously. Imaging there is a famous website, and people from several authority organizations are allowed to post articles on it. People read the articles and know they are from a person of the authority organization, but they do not know who indeed posts the messages. Even the website maintainers cannot tell two articles are from the same person. At the same time, the maintainers of the website may want these authorities to be concise to save the storage space, so they may want to limit the times each person can post messages on the website. When a person attempts to post more than allowed times, he will be found immediately and his submission is rejected, while his anonymity is still protected. Moreover, when people abuse the system, the maintainers can find ways to identify those abusers. Can we have some methods to protect users' privacy, control their access times, and at the same time identify malicious users when abuses are found?

**Related Works.** Group signature [7], ring signature [15] and anonymous credential [6] can be used to protect people's privacy. However, users in these protocols can show their signatures and credentials as many times as they want.

In the linkable ring signature [12], signatures from the

same signer can be linked so as that multiple signing behaviors can be controlled. However, the signers are always anonymous and abusers can never be found.

The  $k$ -times anonymous authentication ( $k$ -TAA)[16, 18, 11, 17] was introduced to protect the privacy while limiting the authentications. Each user can only authenticate anonymously up to  $k$  times, with  $k$  determined by each application provider (AP) and fixed for all users. When a user authenticates more than  $k$  times to an AP, its privacy is compromised. Some articles [14, 13, 1] described dynamic  $k$ -TAA, enabling APs to grant or revoke users independently. However, the property that enables AP to grant users may compromise users' privacy in a sense, because AP knows the identity of each granted user. Again, the value  $k$  is determined by each AP and fixed for all users. Camenisch et al. [4] brought forward a periodic  $n$ -times anonymous authentication scheme. In their scheme, each user can authenticate anonymously up to  $n$  times in each period, no matter how many APs there exist. However, in schemes listed above, the authentications are fully anonymous if they are issued within allowed times. If some users abuse the system within allowed times, they cannot be identified and punished. Therefore, the existing variants of  $k$ -TAA are not applicable to our case.

Emura et al. [10] presented a selectable  $k$ -TAA scheme,

allowing each user has different allowed number of authentication for different AP. In their scheme, the user chooses the allowed number anonymously, and AP decides to accept or reject. However, the computation cost for granting phase is high, which is linear to the allowed number  $k$ . The anonymity is weakened since authentications between the same user and the same AP are linkable to AP (AP knows they are from the same user).

**Our Contribution.** In this paper, a single-verifier  $k$ -times group signature scheme is proposed as building block, where all the group signatures are verified by the only verifier, and the signatures from the same person are limited to  $k_i$  times during time period  $T_i$  times. Based on it, times limited accountable anonymous online submission control system for websites is constructed. Within allowed times, people can post articles anonymously with their signatures, even website maintainers cannot identify two articles are from the same person. When a person posts more than the allowed times, his post will be rejected. When abuse is found, website maintainers can send the signature to open authority of the group to find out the identity.

**Paper Outline.** The rest of this paper is organized as follows: In Section 2 we introduce some preliminaries. In Section 3, we describe the proposed single-verifier  $k$ -times group signature scheme, the building tool for the submission control system. In Section 4, we briefly describe the times limited accountable anonymous online submission control scheme for websites. In Section 5, system attributes and comparison with related previous schemes are presented. Finally, conclusion is given in Section 6.

## 2 Preliminaries

### 2.1 Bilinear Pairings and q-SDH Problem

We first review a few concepts related to bilinear pairings. We follow the notation of [3]:

**Definition 1** (Bilinear Pairings). *Let  $\mathbb{G}$  be a (multiplicative) cyclic group of prime order  $p$  and  $g$  is a generator of  $\mathbb{G}$ . A one-way map  $e : \mathbb{G} \times \mathbb{G} \rightarrow \mathbb{G}_T$  is a bilinear pairing if the following conditions hold.*

- **Bilinear:** For all  $u, v \in \mathbb{G}$ , and  $a, b \in \mathbb{Z}_p$ ,  $e(u^a, v^b) = e(u, v)^{ab}$ .
- **Non-degeneracy:**  $e(g, g) \neq 1$ , i.e., if  $g$  generates  $\mathbb{G}$ , then  $e(g, g)$  generates  $\mathbb{G}_T$ .
- **Computability:** There exists an efficient algorithm for computing  $e(u, v)$ ,  $\forall u, v \in \mathbb{G}$ .

The q-SDH problem was introduced by Boneh and Boyen [2] to construct short signatures without random oracles.

**Definition 2** (q-SDH Problem). *The q-SDH problem in  $(\mathbb{G}_1, \mathbb{G}_2)$  is defined as follows: given a  $(q+2)$ -tuple  $(g_1, g_2, g_2^\gamma, g_2^{(\gamma^2)}, \dots, g_2^{(\gamma^q)})$  as input, output a pair*

$(g_1^{\frac{1}{\gamma+x}}, x)$  where  $x \in \mathbb{Z}_p^*$ . We say that the q-SDH is  $(q, t, \epsilon)$ -hard if for all  $t$ -time adversaries  $A$ , we have

$$Pr \left[ A(g_1, g_2, g_2^\gamma, g_2^{(\gamma^2)}, \dots, g_2^{(\gamma^q)}) = (g_1^{\frac{1}{\gamma+x}}, x) \right] < \epsilon.$$

### 2.2 Proofs of Knowledge of Discrete Logarithms

We will use the notation introduced by Camenisch and Stadler [5] for various proofs of knowledge of discrete logarithms. For instance,

$$PK\{(\alpha, \beta, \gamma) : y = g^\alpha h^\beta \wedge z = g'^\alpha h'^\gamma\};$$

is used for proving the knowledge of integers  $\alpha, \beta$  and  $\gamma$  such that  $y = g^\alpha h^\beta$  and  $z = g'^\alpha h'^\gamma$  holds. Here  $y, g, h, z, g'$  and  $h'$  are elements of some groups  $\mathbb{G} = \langle g \rangle = \langle h \rangle$  and  $\mathbb{G}_T = \langle g' \rangle = \langle h' \rangle$ .

## 3 The Building Tool: Single-verifier $k$ -times Group Signature

The building tool is a single-verifier  $k$ -times group signature scheme, in which authorized people can issue group signatures up to  $k_i$  times for time  $T_i$ . Signatures are all verified by a same verifier so that each user are limited to  $k_i$  signatures strictly.

### 3.1 The Model

A single-verifier  $k$ -times group signature scheme is similar to group signature scheme, while the signing times are limited during each period. It consists of the following algorithm:

- **Key Generation:** The algorithm generates the secret key for the group manager, the secret key for the open authority, and the public parameters. There may be several groups.
- **Member Joining:** User registers with the group manager to join the group. After that, user obtains group member certificate, while group manager obtains user's identification and tracing information, which will be used to identify the abuser.
- **Times Announcing:** The verifier announces the signing times allowed for each future period. They can be the same or different, depending on the applications.
- **Sign:** The user signs the message to show that he is a member of a certain group. Each member certificate can be used to sign up to  $k$  times during each period.
- **Verify:** The verifier checks if the signature is valid and it is within the allowed times. If accepted, some tracing information is recorded into a log file. If not, the signature will be rejected.

- **Open:** When necessary, the open authority finds out user's identification from the signature.

### 3.2 Security Notions

A single-verifier  $k$ -times group signature scheme should fulfill the following security notions:

- **Correctness:** The signature from an honest user within the allowed times should be accepted by the honest verifier. And the open algorithm should correctly identify the signer.
- **Unforgeability:** It is computationally impossible to produce a valid signature, without the knowledge of a membership certificate.
- **Anonymity:** Only the open authority can identify which user provided the signatures.
- **Traceability:** It must be hard to produce a valid signature such that either the honest open authority is unable to identify the signer, or the open authority believes it has identified the origin but is unable to produce a correct proof of its claim.
- **Detectability:** Suppose  $k$  is the number of times the verifier allows each user to sign during a single period. Detectability means that an adversary, who colludes with  $w$  users, is unable to issue more than  $kw$  signatures without being detected the same verifier.

### 3.3 Online $k$ -times limitation

Our idea of online  $k$ -times limitation is developed from Damgård et al. [8]. Each user holds a secret key  $SK$  which is different from the others. A hash function  $H_1$  maps string  $str$  to elements in a group where decisional Diffie-Hellman (DDH) problem is hard. If each user is allowed to sign only once during period  $T$ , then the user needs to show  $H_1(T)^{SK}$ , the commitment to  $SK$ , as the tracing tag, along with the proof of knowledge of owning the group member certificate. If  $H_1(T)^{SK}$  appears twice, the verifier rejects the signature. If each user is allowed to sign  $k$  times during period  $T$ ,  $H_1(T, k_i)^{SK}$  is used, with  $k_i = 1, \dots, k$ .

### 3.4 Single-verifier $k$ -times Group Signature

Our single-verifier  $k$ -times group signature is built upon the group signature scheme by Delerablée and Pointcheval [9].

- **Key Generation:** Selects bilinear pairings  $e : \mathbb{G}_1 \times \mathbb{G}_2 \rightarrow \mathbb{G}_T$  as required in Section 2. Randomly selects generator  $g_2 \in_R \mathbb{G}_2$ , so  $g_1 \leftarrow \psi(g_2)$  is the generator of  $\mathbb{G}_1$ . Randomly selects another generator  $h \in_R \mathbb{G}_1$ , and  $\xi_1, \xi_2 \in_R \mathbb{Z}_p^*$ , then calculates  $u \in \mathbb{G}_1$ , satisfying  $u^{\xi_1} = h$  and  $u^{\xi_2} = g_1$ .  $\xi_1, \xi_2$  are the secret keys for the open authority. Selects  $\gamma \in_R \mathbb{Z}_p^*$  as the secret key

for the group manager, and calculates  $w = g_2^\gamma$ . Two collision resistant hash functions  $H_1 : \{0, 1\}^* \rightarrow \mathbb{G}_T$ ,  $H_2 : \{0, 1\}^* \rightarrow \mathbb{Z}_p$  are selected. The public parameters for the group are  $gpk = (\mathbb{G}_1, \mathbb{G}_2, \mathbb{G}_T, e, \psi, u, h = u^{\xi_1}, g_1 = u^{\xi_2}, g_2, w = g_2^\gamma, H_1, H_2)$ . User generates public key  $upk$  and private key  $usk$  for itself.

- **Member Joining:** User interacts with the group manager to join the group. The steps are described as follows.

1. User  $U$  selects  $y \in_R \mathbb{Z}_p$  to calculate  $C = h^y$ , and generates non-interactive zero knowledge proof  $\pi$  to prove knowing  $y$ .  $U$  sends  $C$  and  $\pi$  to GM.
2. GM checks if  $C$  is ever used by other users before. If used, GM requires  $U$  to run the Member Joining algorithm again. If  $C$  is never used, GM checks whether  $\pi$  is valid. If invalid, GM rejects the joining.
3. GM selects  $x \in_R \mathbb{Z}_p$  and calculates  $A = (g_1 C)^{\frac{1}{\gamma+x}}$ ,  $B = e(g_1 C, g_2) / e(A, w)$ ,  $D = e(A, g_2)$ . GM generates non-interactive zero knowledge proof  $V$  to prove knowing the discrete logarithm of  $B$  in basis  $D$ . GM sends  $A$  and  $V$  to  $U$ . In fact,  $B = e(A^{\gamma+x}, g_2) / e(A, w) = e(A^\gamma, g_2) e(A^x, g_2) / e(A, g_2^\gamma) = e(A^x, g_2) = e(A, g_2)^x$ , which means GM only needs to prove knowing  $x$ .
4. After User  $U$  obtains  $A$  and  $V$ , he calculates  $B = e(g_1 C, g_2) / e(A, w)$ ,  $D = e(A, g_2)$ , checks if  $V$  is valid. If valid,  $U$  signs  $A$  with his key  $usk$  to obtain  $S$ , and sends  $S$  to GM.
5. GM verifies  $S$  with  $upk$  and  $A$ . If valid, GM records  $(upk, A, x, S)$ , and sends  $x$  to  $U$ . The joining records are sent to open authority.
6. User obtains  $x$  and verifies the following equation

$$e(A, g_2)^x e(A, w) e(h, g_2)^{-y} = e(g_1, g_2).$$

If the equation holds, the user joins the group successfully and the group member certificate is  $(A, x, y)$ . The equation is expressed as followed:

$$\begin{aligned} & e(A, g_2)^x e(A, w) e(h, g_2)^{-y} \\ &= e(A, g_2)^x e(A, g_2^\gamma) e(h, g_2)^{-y} \\ &= e(A, g_2^{x+\gamma}) e(h, g_2)^{-y} \\ &= e((g_1 h^y)^{\frac{1}{x+\gamma}}, g_2^{x+\gamma}) e(h, g_2)^{-y} \\ &= e(g_1, g_2) e(h^y, g_2) e(h, g_2)^{-y} \\ &= e(g_1, g_2). \end{aligned}$$

- **Times Announcing:** The verifier publishes a list to show the signing times allowed for each future period. The list is as follows,  $(T_1, k_1), \dots, (T_n, k_n)$ . Or the verifier can publish only a  $k$  indicating the users can sign messages up to  $k$  time during each period.

– **Sign:** Suppose user  $U$  with certificate  $(A, x, y)$  wants to sign a message  $m$  during period  $T$  and each user is allowed to sign  $k$  times during period  $T$ . Suppose it is the  $i$ th ( $0 < i \leq k$ ) time user  $U$  shows to the verifier, he behaves as follows. User  $U$  calculates  $h_i = H_1(T, i)$  with the commitment  $E_i = h_i^y$ , and generates a standard non-interactive zero-knowledge proof  $\pi_i = ZKP\{(A, x, y) : E_i = h_i^y \wedge e(A, g_2)^x \cdot e(A, w) \cdot e(h, g_2)^{-y} = e(g_1, g_2)\}$ , which is also the signature on message  $m$ . User  $U$  sends  $(i, E_i, \pi_i)$  to the verifier. Technical details of zero-knowledge proof are as follows.

1. User  $U$  selects  $\alpha, \beta \in_R \mathbb{Z}_p$ , calculates  $T_1 = u^\alpha$ ,  $T_2 = Ah^\alpha$ ,  $T_3 = u^\beta$ ,  $T_4 = Ag^\beta$ .
2. In order to sign the message  $m$ , user  $U$  selects  $r_\alpha, r_\beta, r_x, r_y, r_{x\alpha} \in_R \mathbb{Z}_p$ , calculates  $R_1 = u^{r_\alpha}$ ,  $R_2 = e(T_2, g_2)^{r_x} \cdot e(T_2, w) \cdot e(h, g_2)^{-r_{x\alpha}} \cdot e(h, w)^{-r_\alpha} \cdot e(h, g_2)^{-r_y}$ ,  $R_3 = u^{r_\beta}$ ,  $R_4 = h^{r_\alpha} g^{-r_\beta}$ ,  $h_i = H_1(T, i)$ ,  $E'_i = h_i^{r_y}$ , and  $c = H_2(m, T_1, T_2, T_3, T_4, R_1, R_2, R_3, R_4)$ .  $R_2$  can be written as  $R_2 = e(A, g_2)^{r_x} \cdot e(A, w) \cdot e(h, g_2)^{\alpha r_x - r_{x\alpha} - r_y} \cdot e(h, w)^{\alpha - r_\alpha}$ , so that user  $U$  can obtain  $R_2$  with fewer computations, since all these pairings can be pre-computed.
3. User  $U$  calculates  $s_\alpha = r_\alpha + c\alpha$ ,  $s_\beta = r_\beta + c\beta$ ,  $s_x = r_x + cx$ ,  $s_y = r_y + cy$ ,  $s_{x\alpha} = r_{x\alpha} + cx\alpha$ .
4. so  $\pi_i = (T_1, T_2, T_3, T_4, E'_i, c, s_\alpha, s_\beta, s_x, s_y, s_{x\alpha})$

– **Verify:** The verifier maintains a tracing log  $TLOG_T$  for time period  $T$ . On receiving  $(i, E_i, \pi_i)$ , the verifier checks if  $1 \leq i \leq k$ , and makes sure that  $E_i$  does not exist in  $TLOG_T$ . Else, the verifier rejects the execution. If both of them hold, the verifier checks whether the zero-knowledge proof is valid as follows.

1. The verifier calculates  $R_1 = u^{s_\alpha} T_1^{-c}$ ,  $R_3 = u^{s_\beta} T_3^{-c}$ ,  $R_4 = h^{s_\alpha} g^{-s_\beta} T_2^{-c} T_4^c$ ,  $R_2 = e(T_2, g_2)^{s_x} \cdot e(T_2, w) \cdot e(h, g_2)^{-s_{x\alpha}} \cdot e(h, w)^{-s_\alpha} \cdot e(h, g_2)^{-s_y} \cdot e(T_2, w)^c \cdot e(g_1, g_2)^{-c} = e(T_2, g_2)^{s_x} \cdot e(h, g_2)^{-s_{x\alpha} - s_y} \cdot e(h, w)^{-s_\alpha}$ . We notice that the pairing computation we need to obtain  $R_2$  is only one, and other pairings can be pre-computed.
2. The verifier calculates  $h_i = H_1(T, i)$ , and checks if  $h_i^{s_y} = E'_i \cdot (E_i)^c$  holds.
3. The verifier checks if  $c$  is equal to  $H_2(m, T_1, T_2, T_3, T_4, R_1, R_2, R_3, R_4)$ . If all the verifications pass, the verifier accepts the signature, and records  $E_i$  into tracing log  $TLOG_T$ . Else, the verifier rejects the execution.

– **Open:** When necessary, the open authority obtains  $T_1, T_2, T_3, T_4$  from the signature, and calculates  $A = T_2(T_1)^{\xi_1} = T_4(T_3)^{\xi_2}$ . By searching  $A$  in Member Joining records, the open authority can find out the corresponding  $upk$ , the public key of the user who has issued that signature.

### 3.5 Security Analysis

**Theorem 3.1.** *The proposed scheme is correct, assuming GM, user, verifier and open authority are all honest.*

**Proof.** The proof is straightforward. If user  $U_i$  and GM are honest,  $U_i$  will carefully select a secret key  $y_i$  and GM will make sure that  $g^{y_i}$  is different from others' public keys. So the secret key  $y_i$  is also different from the others. In the Member Joining protocol,  $U_i$  obtains a member certificate  $(A = (g_1 h^{y_i})^{\frac{1}{\gamma+x}}, x, y_i)$  from the honest GM. For each unused  $k_i \in [1, k]$  during period  $T$ , there will not exist a value the same as  $E_i = h_i^{y_i} = (H_1(T, k_i))^{y_i}$  for index  $k_i$  during period  $T$ , because  $y_i$  is different from others' secret keys. Then the verifier will not reject such an execution. On the knowledge of  $(A = (g_1 h^{y_i})^{\frac{1}{\gamma+x}}, x, y_i)$ ,  $U_i$  is able to generate a valid proof  $\pi_i = ZKP\{(A, x, y_i) : E_i = h_i^{y_i} \wedge e(A, g_2)^x \cdot e(A, w) \cdot e(h, g_2)^{-y_i} = e(g_1, g_2)\}$ , which will then be accepted by the verifier as a successful execution. Since  $\pi_i$  is honestly generated, the open authority can extract  $A$  and find out the corresponding user public key, when it is necessary.  $\square$

**Theorem 3.2.** *If the group signature scheme by Delerablée and Pointcheval [9] is unforgeable, the proposed single-verifier  $k$ -times group signature is unforgeable.*

**Proof. (Sketch)** Suppose that an algorithm  $\mathcal{A}$  can forge the proposed group signature with non-negligible probability. Our scheme is combination the online  $k$ -times limitation with Delerablée et al's group signature [9], both of which are linked by a zero-knowledge proof of a secret value  $y$ . Given an instance of forged single-verifier  $k$ -times group signature, we can extract from it an instance of Delerablée et al's group signature.

The technique is briefly described here. The single-verifier  $k$ -times group signature is of this form:  $(i, E_i, (T_1, T_2, T_3, T_4), E'_i, (R_1, R_2, R_3, R_4), c, (s_\alpha, s_\beta, s_x, s_y, s_{x\alpha}))$ . When the algorithm  $\mathcal{A}$  is about to generate the forged signature, we take control of the random oracle for the challenge and rewind the process. Then we can extract two related signatures, with the same hash-query but different challenges.  $(i, E_i, (T_1, T_2, T_3, T_4), E'_i, (R_1, R_2, R_3, R_4), c, (s_\alpha, s_\beta, s_x, s_y, s_{x\alpha}))$  and  $(i, E_i, (T_1, T_2, T_3, T_4), E'_i, (R_1, R_2, R_3, R_4), c', (s'_\alpha, s'_\beta, s'_x, s'_y, s'_{x\alpha}))$ . Thereafter, simply applying the same technique as the one used to prove the soundness of zero-knowledge proof, one gets a valid certificate  $(A, x, y)$  and then generates a successful forgery for Delerablée and Pointcheval's group signature scheme.  $\square$

**Theorem 3.3.** *If the group signature scheme by Delerablée and Pointcheval [9] is anonymous, DDH problem is hard in  $\mathbb{G}_T$ , and the proof of knowledge technique is zero-knowledge, the proposed single-verifier  $k$ -times group signature is anonymous.*

**Proof.** A user's signature issued in the Sign protocol can be divided into two parts. One is tracing tag  $E_i = h_i^{y_i}$ , with  $h_i \in \mathbb{G}_T$ , the other part is zero knowledge proof of knowing  $(A = (g_1 h^{y_i})^{\frac{1}{\gamma+x}}, x, y_i)$  which satisfying  $E_i = h_i^{y_i}$



and  $e(A, g_2)^x \cdot e(A, w) \cdot e(h, g_2)^{-y^i} = e(g_1, g_2)$  at the same time. Suppose there exists an adversary  $\mathcal{A}$  can determine which user is running the signing execution in the Anonymity game, then we can use  $\mathcal{A}$  to solve DDH problem in  $\mathbb{G}_T$ . In this case,  $\mathcal{A}$  can serve as an algorithm that shows connections between elements in  $\mathbb{G}_T$  and  $\mathbb{G}$ . Suppose  $\mathcal{A}$  inverts elements in  $\mathbb{G}_T$  to those in  $\mathbb{G}$  with probability at least  $\epsilon$ , we denote it as  $\mathcal{A}(g, x)$ , with  $x \in \mathbb{G}_T$ . We are given a DDH problem instance, that is a quadruple  $\{y, y^a, y^b, y^c\}$  of elements of  $\mathbb{G}_T$ , and we are asked to determine if  $c = ab \pmod{p}$ . Define algorithm  $\mathcal{B}$  as follows.

1. Choose a random  $g \in \mathbb{G}$ , and compute  $q_1 = \mathcal{A}(g, y)$ ,  $q_2 = \mathcal{A}(g, y^a)$ ,  $q_3 = \mathcal{A}(g, y^b)$ ,  $q_4 = \mathcal{A}(g, y^c)$ .
2. Compute  $e(q_1, q_4)$  and  $e(q_2, q_3)$ . If the two are equal output 1; else output 0.

Suppose all four outputs of algorithm  $\mathcal{A}$  are correct. Then  $q_2 = q_1^a$ ,  $q_3 = q_1^b$ , and  $q_4 = q_1^c$ . We therefore have  $e(q_1, q_4) = e(q_1, q_1)^c$  and  $e(q_2, q_3) = e(q_1, q_1)^{ab}$ . The two elements are equal if and only if  $c = ab \pmod{p}$ . Thus if all four outputs are correct  $\mathcal{B}$  gives a correct output to the Decision Diffie-Hellman problem. The probability that all four outputs are correct is at least  $\epsilon^4$ .

We use standard proof of knowledge skill for the signature. If the adversary  $\mathcal{A}$  can figure out which user is running the signing execution, we derive a contradiction for the zero-knowledge proof of knowledge.

We notice that if two users simultaneously select the same index  $k_i$  during period  $T$ , the verifier can determine that two signing executions are from two different users. However, this will not weaken the anonymity of the users, because the verifier still cannot determine which user is running the execution and cannot find out that two signatures are from the same user.  $\square$

**Theorem 3.4.** *The proposed single-verifier  $k$ -times group signature is traceable assuming the group signature scheme by Delerablée and Pointcheval [9] is traceable.*

**Proof. (Sketch)** Our single-verifier  $k$ -times group signature is of this form:  $(i, E_i, (T_1, T_2, T_3, T_4), E'_i, (R_1, R_2, R_3, R_4), c, (s_\alpha, s_\beta, s_x, s_y, s_{x\alpha}))$ . With any instance of single-verifier  $k$ -times group signature, one can extract an instance of Delerablée et al's group signature, which is of the form  $(T_1, T_2, T_3, T_4), (R_1, R_2, R_3, R_4), c, (s_\alpha, s_\beta, s_x, s_{x\alpha})$ . If single-verifier  $k$ -times group signature is untraceable, the extracted group signature is untraceable for Delerablée et al's scheme. Then we obtain a contradiction for the traceability of Delerablée et al's group signature scheme.  $\square$

**Theorem 3.5.** *Suppose an adversary colludes with  $w$  users and  $k$  is the number of times the verifier allows each user to sign on period  $T$ . If the adversary issues signatures more than  $kw$  times, it must be detected by the honest verifier.*

**Proof.** For each user on period  $T$ , only  $k$  bases can be used for generating traceable tags during the execution, namely

$h_1 = H_1(T, 1), \dots, h_k = H_1(T, k)$ . If the user uses additional base, it can be easily detected by the verifier, because the user has to tell the verifier which base he is using in the execution.

Using these  $k$  bases, each user can perform only  $k$  times successful executions, and  $w$  users can perform  $kw$  times. After  $kw$  times normal executions, if they collude together and want to sign the messages for one more time, they need a new secret key to create a different tracing tag other than the  $kw$  used tags. And they also need to prove knowing a message-signature pair for the secret key, which is not obtained from normal interaction with the GM. It cannot be fulfilled due to the unforgeability of our scheme.  $\square$

## 4 Times Limited Accountable Anonymous Online Submission System

Our system can be divided into two parts: organization and website. To ensure the security of our system, the security of both parts should be considered.

- **Hardware Infrastructure Security.** The Hardware infrastructures that support the system, such as servers and backup disks, should only be accessible to trusted persons.
- **Secure Channels.** There are secure channels between organization and website, in order for website to obtain and update public keys securely.
- **Reliability.** Robust design should be provided to support for backup, load balancing, and failover.

The security considerations described above are out of scope of this paper.

With the building tool provided in Section 3, we can easily construct times limited accountable anonymous online submission system suitable for various websites. Each website needs only to announce the allowed times for future periods and act as the only verifier.

### 4.1 System Preparation Phase

The website signs agreements with several organizations that the website allows people from these organizations to post messages anonymously on it. In addition, the organizations agree to reveal the identities of abusers when the website requires. The agreements should show what kinds of messages are allowed to be submitted. A trusted party in each organization generates the public parameters and the corresponding secret keys for GM and OA, as described in Section 3. GM and OA in each organization can be trusted by the websites. The website prepares storage space for recording the authenticated submissions. It can be a database on a hard disk, with different tables for different periods and organizations.

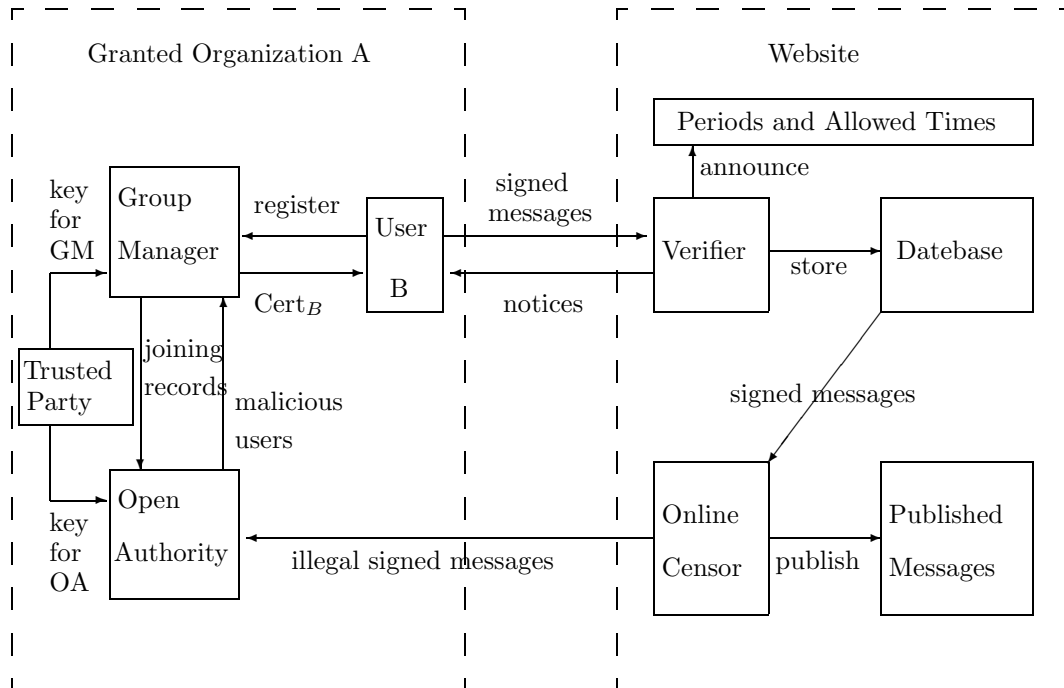


Figure 1: Times Limited Accountable Anonymous Submission System with Online Censor

## 4.2 Member Joining Phase

Members in each organization register with GM to obtain certificates, by running **Member Joining** algorithm as described in Section 3. Each organization should notify its members of the allowed kinds of messages to be submitted.

## 4.3 Times Announcing Phase

The verifier publishes a list to indicate submission times allowed for each future period. The list is as follows,  $(T_1, k_1), \dots, (T_n, k_n)$ . Alternatively, the verifier can choose to publish a  $k$  for all future periods. In addition, the website should notify the users of the current period.

## 4.4 Message Submission Phase

When a user from the agreed organization wants to submit a message to the website, he selects a random unused number less than the allowed time for current period, and generates the single-verifier  $k$ -times group signature as described in Section 3. He writes down the message in a pre-designed form, indicating his organization and attaching the signature.

## 4.5 Signed Message Verification Phase

On receiving the submission, the website can verify it as described in Section 3. If the submission is from the granted organization and its signature passes the verification, the message and its signature are recorded into the

database table for current period and the coming organization. If not, the submission is rejected and a notice is sent to the user immediately.

## 4.6 Message Review Phase

These verifications and recording can be automatically done by software programs. The website can choose online or offline review for the messages, according to its running policy. When online review is chosen, some people (censors) are deployed to read the messages before they can be published. Figure 1 illustrates a submission system with online censor. When in the case of offline review, the messages can be published immediately. Later, some people are deployed to browse through these published messages and remove those inappropriate. No matter which kind of review is used, if a message is found inappropriate for the website, the message and corresponding signature are sent to open authority for further treatment.

# 5 Security Attributes and Performance Comparison

## 5.1 Security Attributes

Our system inherits the security attributes from group signature scheme [9], in addition to times limited authentication.

- **Anonymity.** Given signatures produced by a user, no one except the open authority should be able to find

out the signer's identity. The website cannot decide two signatures are from the same user.

- **Traceability.** Given a valid signature, the open authority is bound to identify the signer.
- **Non-frameability.** Anybody, even group manager and open authority, is not able to wrongly accuse someone for having signed a message.
- **Concurrent Join** The system allows for several users to register at the same time.
- **Dynamic Revocation.** As indicated in [9], the group manager can remove a user from the organization, by publishing new public parameters and some information of the revoked user. The unrevoked users can update their certificates accordingly. Mass revocations are done one by one.
- **Times Limited Authentication.** No one can authenticate more than announced number to the honest verifier.

## 5.2 Performance Comparison

We evaluate our scheme in Table 1 by comparing it with several related works, including the group signature scheme by Delerablée and Pointcheval [9], dynamic  $k$ -TAA by Au et al. [1], periodic  $k$ -times anonymous authentication by Camenisch et al. [4], and the selectable  $k$ -TAA scheme by Emura et al. [10]. Because schemes in [1, 4] did not provide detailed computations of zero-knowledge proof, so the computation cost and signature length are given by us approximately.

We notice that in scheme of  $k$ -TAA and its variants, such as dynamic  $k$ -TAA [1] and periodic  $n$ -TAA [4], the signing and verifying cost is huge and the signature length is not constant, since they need to prove that the committed signing index lies in an interval  $[1, k]$ . Users are fully anonymous when they authenticate no more than the allowed times, or else their identities are exposed.

The selectable  $k$ -times relaxed anonymous authentication by Emura et al. [10] is efficient in signing and verifying phase. However, the computation cost for user and AP is huge in granting phase, which is linear to the allowed number  $k$ . The storage cost for user is also linear to  $k$ . The weakened anonymity is suitable for their application since the linkable authentications are needed for AP to adjust marketing strategy.

Our scheme is almost as efficient as the group signature scheme [9]. We need only a few more computations and some extra length to realize the times limited property.

## 6 Conclusion and Discussion

We propose single-verifier  $k$ -times group signature scheme to allow each user to authenticate up to  $k_i$  times during period  $T_i$ , without leaking the privacy of the users,

while maintaining the ability of revealing the identities of abusers. We show that the scheme can be used to construct flexible anonymous online submission control system for websites.

We notice that if two users from the same organization using the same index  $k_i$  for signing during the same period, the website can know the two submissions are from two different users. However, this will not weaken the anonymity of the users, because the website still cannot determine which user is submitting and cannot tell two submissions are from the same user.

Our single-verifier  $k$ -times group signature scheme can be turned into a generic scheme, in which we can employ other group signature scheme so as to achieve different efficiency.

## Acknowledgement

This work is supported by the National Natural Science Foundation of China (No. 60773202, 61070168).

## References

- [1] M. H. Au, W. Susilo, and Y. Mu, (2006). Constant-size dynamic  $k$ -TAA. *SCN 2006*, LNCS 4116, Springer-Verlag, Maiori, Italy, pp. 111-125.
- [2] D. Boneh, and X. Boyen, (2004). Short signatures without random oracles. *EUROCRYPT 2004*, LNCS 3027, Springer-Verlag, Interlaken, Switzerland, pp. 56-73.
- [3] D. Boneh, B. Lynn, and H. Shacham, (2001). Short signatures from the Weil pairing, *ASIACRYPT 2001*, LNCS 2248, Springer-Verlag, Gold Coast, Australia, pp. 514-532.
- [4] J. Camenisch, S. Hohenberger, M. Kohlweiss, A. Lysyanskaya, and M. Meyerovich, (2006). How to win the clone wars: efficient periodic  $n$ -Times anonymous authentication. *ACM CCS 2006*, ACM Press, Alexandria, VA, USA, pp. 201-210.
- [5] J. Camenisch, and M. Michels, (1997). Efficient group signature schemes for large group. *CRYPTO 1997*, LNCS 1296, Springer-Verlag, Santa Barbara, California, USA, pp. 410-424.
- [6] D. Chaum, (1985). Security without identification: transaction systems to make big brother obsolete. *Communications of the ACM*, 28(10), pp. 1030-1044.
- [7] D. Chaum, and E. V. Heyst, (1991). Group signatures. *EUROCRYPT 1991*, LNCS 547, Springer-Verlag, Brighton, UK, pp. 257-265.
- [8] I. Damgård, K. Dupont, and M. O. Pedersen, (2006). Unclonable group identification. *EUROCRYPT 2006*, LNCS 4004, Springer-Verlag, St. Petersburg, Russia, pp. 555-572.

Table 1: Performance Comparison with Previous Works

|                                       | [9]               | [1]                                   | [4]                                    | [10]  | Our scheme                 |
|---------------------------------------|-------------------|---------------------------------------|--|---|----------------------------|
| Public Key Size                       | constant          | constant                              | constant                               | constant                                      | constant                   |
| Credential Size                       | $(1G_1+2p)$       | $1G_1+1G_2+4p$                        | $(1G_1+2p)$                            | $(k+2)G_1+2P$                                 | $(1G_1+2p)$                |
| Granting Computations                 | constant          | constant                              | constant                               | $4E_T+3M_T+(3k+13)E+(k+3)M$                   | constant                   |
| Signing Computations                  | $8M+3E_T+2M_T$    | $(\log k+16)E+9M$                     | $(\log k+21)E+14M+3P$                  | $3E$  | $9E+2M+3E_T+3M_T$          |
| Verifying Computations                | $9M+3E_T+3M_T+1P$ | $(\log k+27)E+19M+3P+9E_T+9M_T$       | $(\log k+37)E+24M+3P+1M_T$             | $4E+2M$                                       | $8E+5M+3E_T+3M_T+1P$       |
| Signature Length                      | $(4G_1+4p+1c)$    | $(21+3\log k)p+(\log k+2)G_p+4G_1+1c$ | $(13+3\log k)p+\log kG_p+8G_1+2G_T+3c$ | $3G_1+1p+1c$                                  | $4G_1+2G_T+5p+1c$          |
| Action when authenticated $> k$       | -                 | identity can be extracted             | identity can be extracted              | reject  | reject                     |
| Anonymity when authenticated $\leq k$ | -                 | fully anonymous                       | fully anonymous                        | linkable to AP, anonymous to others except OA | anonymous to all except OA |
| Different Limits for Each User        | -                 | no                                    | no                                     | yes   | no                         |
| AP Can Choose Users                   | -                 | yes                                   | no                                     | no  | no                         |

$G_1$ : element in  $\mathbb{G}_1$ ;  $G_2$ : element in  $\mathbb{G}_2$ ;  $G_T$ : element in  $\mathbb{G}_T$ ;

$G_p$ : element in  $\mathbb{G}_p$  where DDH problem is difficult;  $p$ : element in  $\mathbb{Z}_p$ ;  $P$ : pairing in  $\mathbb{G}_1 \times \mathbb{G}_2$ ;

$M$ : multiplication (or division) in  $\mathbb{G}_1$ ;  $M_T$ : multiplication (or division) in  $\mathbb{G}_T$ ;

$E$ : exponentiation in  $\mathbb{G}_1$ ;  $E_T$ : exponentiation in  $\mathbb{G}_T$ ;  $c$ : a small integer for zero-knowledge proof.

- [9] C. Delerablée, and D. Pointcheval, (2006). Dynamic fully anonymous short group signatures. *VIETCRYPT 2006*, LNCS 4341, Springer-Verlag, Hanoi, Vietnam, pp. 193-210.
- [10] K. Emura, A. Miyaji, and K. Omote, (2009). A Selectable k-Times Relaxed Anonymous Authentication Scheme. *WISA 2009*, LNCS 5932, Springer-Verlag, Busan, Korea, pp. 281-295.
- [11] M. Layouni, and H. Vangheluwe, (2007). Anonymous k-Show credentials. *EuroPKI 2007*, LNCS 4582, Springer-Verlag, Palma de Mallorca, Spain, pp. 181-192.
- [12] J. K. Liu, V. K. Wei, and D. S. Wong, (2004). Linkable spontaneous anonymous group signature for ad hoc groups (extended abstract). *ACISP 2004*, LNCS 3108, Springer-Verlag, Sydney, Australia, pp. 325-335.
- [13] L. Nguyen, (2006). Efficient dynamic k-Times anonymous authentication. *VIETCRYPT 2006*, LNCS 4341, Springer-Verlag, Hanoi, Vietnam, pp. 81-98.
- [14] L. Nguyen, and R. Safavi-Naini, (2005). Dynamic k-Times anonymous authentication. *ACNS 2005*, LNCS 3531, Springer-Verlag, New York, NY, USA, pp. 318-333.
- [15] R. Rivest, A. Shamir, and Y. Tauman, (2001). How to leak a secret. *ASIACRYPT 2001*, LNCS 2248, Springer-Verlag, Gold Coast, Australia, pp. 552-565.
- [16] I. Teranishi, J. Furukawa, and K. Sako, (2004). k-Times anonymous authentication (Extended Abstract). *ASIACRYPT 2004*, LNCS 3329, Springer-Verlag, Jeju Island, Korea, pp. 308-322.
- [17] I. Teranishi, J. Furukawa, and K. Sako, (2009). k-Times Anonymous Authentication. *IEICE Transactions*, 92-A(1), pp. 147-165.
- [18] I. Teranishi, and K. Sako, (2006). k-Times anonymous authentication with a constant proving cost. *PKC 2006*, LNCS 3958, Springer-Verlag, New York, NY, USA, pp. 525-542.

# Multiple Attribute Decision Making Method Based on the Trapezoid Fuzzy Linguistic Hybrid Harmonic Averaging Operator

Peide Liu and Yu Su

School of Management Science and Engineering, Shandong University of Finance and Economics

No.7366 Erhuandong Road, Lixia District, Ji'nan 250014, Shandong Province, P.R. China

E-mail: Peide.liu@gmail.com

**Keywords:** the trapezoid fuzzy linguistic variables (*TFLVs*), the *TFLHHA* operator, multiple attribute decision making (MADM)

**Received:** October 18, 2010

*A new method is proposed to solve the multiple attribute decision making (MADM) problems with the trapezoid fuzzy linguistic variables (TFLVs) based on the trapezoid fuzzy linguistic hybrid harmonic averaging (TFLHHA) operator. To begin with, this paper reviews the concept and operational rules of the TFLVs, the calculation method of the possibility degree with TFLVs, and the comparison method of TFLVs. Then, some operators are proposed, in order to aggregate the TFLVs, such as the trapezoid fuzzy linguistic weighted harmonic averaging (TFLWHA) operator, the trapezoid fuzzy linguistic ordered weighted harmonic averaging (TFLOWHA) operator, and the trapezoid fuzzy linguistic hybrid harmonic averaging (TFLHHA) operator. Furthermore, based on the TFLHHA operator, a new method solving the MADM problems with the TFLVs is proposed. Finally, an illustrative example is given to show the decision making steps, and it verifies the effectiveness of the developed method.*

*Povzetek: Članek opisuje metodo za podporo odločanju, ki uporablja mehko logiko.*

## 1 Introduction

In the process of the multiple attribute decision making (MADM), the decision making information, given by the decision makers, often takes the form of the linguistic variables, because of the complexity and uncertainty of the objective things, and the ambiguity of human thinking. Therefore, the MADM under the linguistic context is an interesting research topic which has been receiving more and more attention in recent years [1-4]. Some operators were widely used to aggregate the decision making information in the process of the MADM. Bordogna et al. [5] developed a model within fuzzy set theory by the linguistic ordered weighted average (*LOWA*) operators for the group decision making in the linguistic context. Xu [6] proposed an approach to solve the multiple attribute group decision making problems with the uncertain linguistic information, based on the uncertain linguistic ordered weighted averaging (*ULOWA*) operator and the uncertain linguistic hybrid aggregation (*ULHA*) operator. Wu and Chen [7] introduced the linguistic weighted arithmetic averaging (*LWAA*) operator to aggregate the decision making information which took the form of the linguistic variables. Xu [8] developed some operators for aggregating the triangular fuzzy linguistic variables, such as the fuzzy linguistic averaging (*FLA*) operator, the fuzzy linguistic weighted

averaging (*FLWA*) operator, the fuzzy linguistic ordered weighted averaging (*FLOWA*) operator, and the induced *FLOWA* (*IFLOWA*) operator.

But in the real situation, the decision-makers sometimes can only provide the decision making information in the form of the trapezoid fuzzy linguistic variables (*TFLVs*). The trapezoid fuzzy linguistic variable (*TFLV*) generalizes the linguistic variable, the uncertain linguistic variable and the triangular fuzzy linguistic variable. So the research on the MADM problems with the *TFLVs* is very significant. But the related decision making methods based on the *TFLVs* are less. Xu [9] proposed the trapezoid fuzzy linguistic weighted averaging (*TFLWA*) operator to aggregate all the decision making information corresponding to each alternative, and he used the similarity measure to rank the decision alternatives and then the most desirable one is selected. Liang and Chen [10] proposed the trapezoid fuzzy linguistic weighted averaging (*TFLWA*) operator to aggregate the decision making information, and then all the alternatives were ranked by comparing the possibility degree of the *TFLV*.

Based on these, this paper extends the *OWHA* operator [11] and the *UCWHA* operator [12]

to deal with the MADM problems with the trapezoid fuzzy linguistic information, such as the trapezoid fuzzy linguistic weighted harmonic averaging (*TFLWHA*) operator, and the trapezoid fuzzy linguistic ordered weighted harmonic averaging (*TFLOWHA*) operator. The *TFLWHA* operator only focuses on the attribute weight itself, but it ignores the position weight with respect to the attribute value; and the *TFLOWHA* operator focuses on the position weight with respect to the attribute value, but it ignores the weight of the attribute value itself. The two operators are one-sided. So in order to avoid the disadvantage of the two operators, the trapezoid fuzzy linguistic hybrid harmonic averaging (*TFLHHA*) operator is proposed to aggregate the attribute values which take the form of the *TFLVs*. According to the *TFLHHA* operator, the new method is proposed, which can solve MADM problems with the *TFLVs* directly.

To do so, the remainder of this paper is structured as follows: In section 2, this paper reviews the concept and the operational rules of the *TFLVs*, and introduces the comparison method of the *TFLVs*, in which the calculation method of the possibility degree with the *TFLVs* is reviewed. In section 3, three operators are proposed in order to aggregate the *TFLVs*, such as the *TFLWHA* operator, the *TFLOWHA* operator, and the *TFLHHA* operator. In section 4, the decision making steps of the new method is proposed based on the *TFLHHA* operator. In section 5, an illustrative example is given to show the decision making steps, and it verifies the effectiveness of the developed method. The section 6 concludes this paper.

## 2 The Trapezoid Fuzzy Linguistic Variables

### 2.1 The definition of the trapezoid fuzzy linguistic variables

Let  $S = \{s_i \mid i = 1, 2, \dots, t\}$  be a linguistic term set with odd cardinality, any label  $s_i$  represents a possible value of the linguistic variable. Especially,  $s_1$  and  $s_t$  represent the lower and the upper values of the linguistic terms, respectively. For example, a linguistic term set  $S$  could be given as follows:

$S = \{s_1 = \text{extremely poor}, s_2 = \text{very poor}, s_3 = \text{poor}, s_4 = \text{slightly poor}, s_5 = \text{fair}, s_6 = \text{slightly good}, s_7 = \text{good}, s_8 = \text{very good}, s_9 = \text{extremely good}\}$

Usually, in these cases,  $s_i$  and  $s_j$  must satisfy the following additional characteristics [13]:

(1) The set  $S$  is ordered:  $s_i$  is worse than  $s_j$ , if  $i < j$ ;

(2) Maximum operator:  $\max(s_i, s_j) = s_i$ , if  $s_i \geq s_j$ ;

(3) Minimum operator:  $\min(s_i, s_j) = s_j$ , if  $s_i \geq s_j$ .

Some calculation results, however, may not exactly match any linguistic labels in  $S$  in the calculation process. To preserve all the given information, the discrete term set  $S$  is extended to a continuous term set  $\bar{S} = \{s_i \mid s_0 \leq s_i \leq s_q, i \in [0, q]\}$ , where  $s_i$  meets all the characteristics above and  $q (q > t)$  is a sufficient large positive integer. If  $s_i \in S$ , then we call  $s_i$  the original term, otherwise, we call  $s_i$  the virtual term. In general, the decision makers use the original linguistic terms to evaluate the alternatives, and the virtual linguistic terms can only appear in the process of the operation and ranking [13].

**Definition 2.1[8]:** Let  $s_\alpha, s_\beta \in \bar{S}$ , then we defined the distance between  $s_\alpha$  and  $s_\beta$  as:

$$d(s_\alpha, s_\beta) = |\alpha - \beta| \tag{1}$$

**Definition 2.2[8]:** Let  $\tilde{s} = [s_\alpha, s_\beta, s_\gamma, s_\eta] \in \tilde{S}$ , where  $s_\alpha, s_\beta, s_\gamma, s_\eta \in \bar{S}$ , and the subscripts  $\alpha, \beta, \gamma, \eta$  are non-decreasing numbers, and  $s_\beta$  and  $s_\gamma$  indicate the interval in which the membership value is 1, with  $s_\alpha$  and  $s_\eta$  indicating the lower and upper values of  $\tilde{s}$ , respectively, then  $\tilde{s}$  is called the trapezoid fuzzy linguistic variable (*TFLV*), which is characterized by the following membership function (see Figure 1):

$$\mu_{\tilde{s}}(\theta) = \begin{cases} 0 & s_0 \leq s_\theta \leq s_\alpha \\ \frac{d(s_\theta, s_\alpha)}{d(s_\beta, s_\alpha)} & s_\alpha \leq s_\theta \leq s_\beta \\ 1 & s_\beta \leq s_\theta \leq s_\gamma \\ \frac{d(s_\theta, s_\eta)}{d(s_\gamma, s_\eta)} & s_\gamma \leq s_\theta \leq s_\eta \\ 0 & s_\eta \leq s_\theta \leq s_q \end{cases} \tag{2}$$

where  $\tilde{S}$  is the set of all the trapezoid fuzzy linguistic variables. Especially, if any two of  $\alpha, \beta, \gamma, \eta$  are equal, then  $\tilde{s}$  is reduced to a triangular fuzzy linguistic variable; if any three of  $\alpha, \beta, \gamma, \eta$  are equal, then  $\tilde{s}$  is reduced to an uncertain linguistic variable [8].

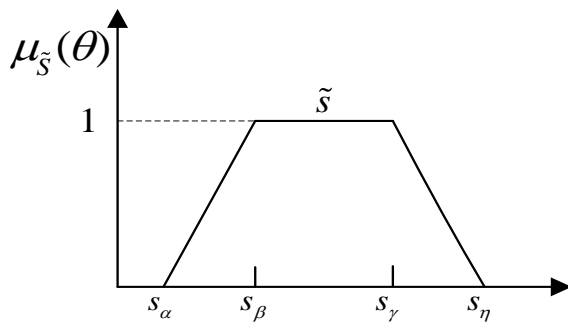


Figure 1 A trapezoid fuzzy linguistic variable  $\tilde{s}$

**2.2 The operational rules and characteristics of the trapezoid fuzzy linguistic variables**

Let  $\tilde{s} = [s_{\alpha}, s_{\beta}, s_{\gamma}, s_{\eta}]$ ,  $\tilde{s}_1 = [s_{\alpha_1}, s_{\beta_1}, s_{\gamma_1}, s_{\eta_1}]$  and  $\tilde{s}_2 = [s_{\alpha_2}, s_{\beta_2}, s_{\gamma_2}, s_{\eta_2}] \in \tilde{\mathcal{S}}$  be any three trapezoid fuzzy linguistic variables, and  $\lambda \in [0,1]$  and  $\lambda_1 \in [0,1]$ , then their operational rules are defined as follows:

- (1)  $\tilde{s}_1 \oplus \tilde{s}_2 = [s_{\alpha_1}, s_{\beta_1}, s_{\gamma_1}, s_{\eta_1}] \oplus [s_{\alpha_2}, s_{\beta_2}, s_{\gamma_2}, s_{\eta_2}] = [s_{\alpha_1+\alpha_2}, s_{\beta_1+\beta_2}, s_{\gamma_1+\gamma_2}, s_{\eta_1+\eta_2}]$ ;
- (2)  $\lambda \tilde{s} = \lambda [s_{\alpha}, s_{\beta}, s_{\gamma}, s_{\eta}] = [s_{\lambda\alpha}, s_{\lambda\beta}, s_{\lambda\gamma}, s_{\lambda\eta}]$
- (3) if  $0 < \alpha \leq \beta \leq \gamma \leq \eta$ , then  $1/\tilde{s} = (\tilde{s})^{-1} = [1/s_{\eta}, 1/s_{\gamma}, 1/s_{\beta}, 1/s_{\alpha}] = [s_{1/\eta}, s_{1/\gamma}, s_{1/\beta}, s_{1/\alpha}]$ .

In addition, the trapezoid fuzzy linguistic variables have the following characteristics:

- (1)  $\tilde{s}_1 \oplus \tilde{s}_2 = \tilde{s}_2 \oplus \tilde{s}_1$ ;
- (2)  $(\lambda \oplus \lambda_1)\tilde{s} = \lambda\tilde{s} \oplus \lambda_1\tilde{s}$ ;
- (3)  $\lambda(\tilde{s} \oplus \tilde{s}_1) = \lambda\tilde{s} \oplus \lambda\tilde{s}_1$ .

**2.3 The comparison method of the trapezoid fuzzy linguistic variables**

**Definition 2.3[10]:** Let  $\tilde{s}_1 = [s_{\alpha_1}, s_{\beta_1}, s_{\gamma_1}, s_{\eta_1}]$  and  $\tilde{s}_2 = [s_{\alpha_2}, s_{\beta_2}, s_{\gamma_2}, s_{\eta_2}]$  be two trapezoid fuzzy linguistic variables, then the possibility degree of  $\tilde{s}_1 \geq \tilde{s}_2$  is defined as follows:

$$p(\tilde{s}_1 \geq \tilde{s}_2) = \min\{\max\{\frac{(\gamma_1 + \eta_1) - (\alpha_2 + \beta_2)}{(\gamma_1 + \eta_1) - (\alpha_1 + \beta_1) + (\gamma_2 + \eta_2) - (\alpha_2 + \beta_2)}, 0\}, 1\}$$

(3)

**Example 1:** Let  $\tilde{s}_1 = [s_2, s_3, s_5, s_6]$  and  $\tilde{s}_2 = [s_4, s_5, s_8, s_9]$  be two trapezoid fuzzy linguistic variables, then the possibility degree of  $\tilde{s}_1 \geq \tilde{s}_2$  is:

$$p(\tilde{s}_1 \geq \tilde{s}_2) = \min\{\max\{\frac{(5+6) - (4+5)}{(5+6) - (2+3) + (8+9) - (4+5)}, 0\}, 1\} = \min\{\max\{0.143, 0\}, 1\} = 0.143$$

The characteristics of the possibility degree  $p(\tilde{s}_1 \geq \tilde{s}_2)$  are shown as follows [10]:

Let  $\tilde{s}_1 = [s_{\alpha_1}, s_{\beta_1}, s_{\gamma_1}, s_{\eta_1}]$ ,  $\tilde{s}_2 = [s_{\alpha_2}, s_{\beta_2}, s_{\gamma_2}, s_{\eta_2}]$ ,  $\tilde{s}_3 = [s_{\alpha_3}, s_{\beta_3}, s_{\gamma_3}, s_{\eta_3}]$  be any three trapezoid fuzzy linguistic variables, then

- (1)  $0 \leq p(\tilde{s}_1 \geq \tilde{s}_2) \leq 1$ ,  $0 \leq p(\tilde{s}_2 \geq \tilde{s}_1) \leq 1$ ;
- (2)  $p(\tilde{s}_1 \geq \tilde{s}_2) + p(\tilde{s}_2 \geq \tilde{s}_1) = 1$ .

Especially, if  $p(\tilde{s}_1 \geq \tilde{s}_2) = p(\tilde{s}_2 \geq \tilde{s}_1)$ , then

$$p(\tilde{s}_1 \geq \tilde{s}_2) = p(\tilde{s}_2 \geq \tilde{s}_1) = \frac{1}{2}$$

- (3) if  $p(\tilde{s}_1 \geq \tilde{s}_2) \geq \frac{1}{2}$ , and  $p(\tilde{s}_2 \geq \tilde{s}_3) \geq \frac{1}{2}$ ,

then  $p(\tilde{s}_1 \geq \tilde{s}_3) \geq \frac{1}{2}$ ;

- (4) if  $p(\tilde{s}_1 \geq \tilde{s}_2) \geq \frac{1}{2}$ , and  $p(\tilde{s}_2 \geq \tilde{s}_3) \geq \frac{1}{2}$ ,

then  $p(\tilde{s}_1 \geq \tilde{s}_2) + p(\tilde{s}_2 \geq \tilde{s}_3) \geq p(\tilde{s}_1 \geq \tilde{s}_3)$

Let  $\tilde{s}_i$  and  $\tilde{s}_j$  be two trapezoid fuzzy linguistic variables, then the steps of the comparison method are shown as follows:

(1) Utilize the formula (3) to compare the size of  $\tilde{s}_i$  and  $\tilde{s}_j$ , and suppose that  $p_{ij} = p(\tilde{s}_i \geq \tilde{s}_j)$ , then we can contribute the possibility degree matrix  $P = (p_{ij})_{n \times n}$ , where  $p_{ij} \geq 0$ ,

$p_{ij} + p_{ji} = 1$ ,  $p_{ii} = \frac{1}{2}$ ,  $i, j = 1, 2, \dots, n$ . We can easily obtain the result that the matrix  $P = (p_{ij})_{n \times n}$  is the complimentary judgment matrix [14].

(2) Sum all the elements of each rows of the possibility degree matrix, and rank the orders of the trapezoid fuzzy linguistic variables based on the values  $p_i$ , where

$p_i = \sum_{j=1}^n p_{ij}$  ( $i = 1, 2, \dots, n$ ). The larger the value of  $p_i$  is, the larger the trapezoid fuzzy linguistic variable  $\tilde{s}_i$  is.

**Example 2:** Let  $\tilde{s}_1 = [s_2, s_3, s_5, s_6]$  and  $\tilde{s}_2 = [s_4, s_5, s_8, s_9]$  be two trapezoid fuzzy linguistic variables, then we can compare the size of  $\tilde{s}_1$  with  $\tilde{s}_2$ :

(1) The possibility degree of  $\tilde{s}_1 \geq \tilde{s}_2$  is:

$$p(\tilde{s}_1 \geq \tilde{s}_2) = \min\{\max\{\frac{(5+6)-(4+5)}{(5+6)-(2+3)+(8+9)-(4+5)}, 0\}, 1\}$$

$$= \min\{\max\{0.143, 0\}, 1\} = 0.143$$

and the possibility degree of  $\tilde{s}_2 \geq \tilde{s}_1$  is:

$$p(\tilde{s}_2 \geq \tilde{s}_1) = \min\{\max\{\frac{(8+9)-(2+3)}{(8+9)-(4+5)+(5+6)-(2+3)}, 0\}, 1\}$$

$$= \min\{\max\{0.857, 0\}, 1\} = 0.857$$

Then we can contribute the possibility degree matrix:

$$P = (p_{ij})_{2 \times 2} = \begin{bmatrix} 0.5 & 0.143 \\ 0.857 & 0.5 \end{bmatrix}$$

$$(2) p_1 = \sum_{j=1}^2 p_{1j} = 0.5 + 0.143 = 0.643,$$

$$p_2 = \sum_{j=1}^2 p_{2j} = 0.875 + 0.5 = 1.375,$$

so  $p_1 < p_2$ .

Then, we can get that:  $\tilde{s}_1 < \tilde{s}_2$  ( $\tilde{s}_1$  is worse than  $\tilde{s}_2$ ).

### 3 Some Harmonic Operators with the Trapezoid Fuzzy Linguistic Variables

**Definition 3.1:** Let  $TFLWHA: \tilde{S}^n \rightarrow \tilde{S}$ , if

$$TFLWHA_w(\tilde{s}_1, \tilde{s}_2, \dots, \tilde{s}_n) = (\sum_{j=1}^n \frac{w_j}{\tilde{s}_j})^{-1} \quad (4)$$

where  $\tilde{S}$  is the set of all trapezoid fuzzy linguistic variables, and  $\tilde{s}_j \in \tilde{S}$  ( $j = 1, 2, \dots, n$ ) is the trapezoid fuzzy linguistic variable.  $w = (w_1, w_2, \dots, w_n)$  is the weight vector, and  $w_i$  is the weight of  $\tilde{s}_i$ , where  $w_i \geq 0$ ,  $i = 1, 2, \dots, n$ ,  $\sum_{i=1}^n w_i = 1$ , then  $TFLWHA$  is called the trapezoid fuzzy linguistic weighted harmonic averaging ( $TFLWHA$ ) operator.

**Example 3:** If  $\tilde{s}_1 = [s_2, s_3, s_5, s_6]$   $\tilde{s}_2 = [s_4, s_5, s_8, s_9]$   $\tilde{s}_3 = [s_5, s_6, s_7, s_9]$  and  $\tilde{s}_4 = [s_3, s_4, s_5, s_7] \in \tilde{S}$  are

four trapezoid fuzzy linguistic variables, and  $w = (0.3, 0.2, 0.1, 0.4)$  is the weight vector, then

$$TFLWHA_w(\tilde{s}_1, \tilde{s}_2, \tilde{s}_3, \tilde{s}_4) = (\sum_{j=1}^4 \frac{w_j}{\tilde{s}_j})^{-1}$$

$$= (\frac{0.3}{[s_2, s_3, s_5, s_6]} \oplus \frac{0.2}{[s_4, s_5, s_8, s_9]} \oplus \frac{0.1}{[s_5, s_6, s_7, s_9]} \oplus \frac{0.4}{[s_3, s_4, s_5, s_7]})^{-1}$$

$$= ([s_{0.05}, s_{0.06}, s_{0.1}, s_{0.15}] \oplus [s_{0.022}, s_{0.025}, s_{0.04}, s_{0.05}] \oplus [s_{0.011}, s_{0.014}, s_{0.017}, s_{0.02}] \oplus [s_{0.057}, s_{0.08}, s_{0.1}, s_{0.133}])^{-1}$$

$$= [s_{0.14}, s_{0.179}, s_{0.257}, s_{0.353}]^{-1} = [s_{2.833}, s_{3.891}, s_{5.587}, s_{7.143}]$$

**Definition 3.2:** Let  $TFLOWHA: \tilde{S}^n \rightarrow \tilde{S}$ , if

$$TFLOWHA_\omega(\tilde{s}_1, \tilde{s}_2, \dots, \tilde{s}_n) = (\sum_{j=1}^n \frac{\omega_j}{\tilde{r}_j})^{-1} \quad (5)$$

where  $\tilde{S}$  is the set of all trapezoid fuzzy linguistic variables, and  $\tilde{s}_j, \tilde{r}_j \in \tilde{S}$  ( $j = 1, 2, \dots, n$ ) are the trapezoid fuzzy linguistic variables.  $\tilde{r}_j$  is the  $j^{th}$  largest of  $\tilde{s}_i$  ( $i = 1, 2, \dots, n$ ), and  $\omega = (\omega_1, \omega_2, \dots, \omega_n)$  is the position weight vector with  $TFLOWHA$ , where

$$\omega_j \geq 0, \quad j = 1, 2, \dots, n, \quad \sum_{j=1}^n \omega_j = 1, \quad \text{then}$$

$TFLOWHA$  is called the trapezoid fuzzy linguistic ordered weighted harmonic averaging ( $TFLOWHA$ ) operator.

The characteristic of the  $TFLOWHA$  operator is: Firstly, The order of the trapezoid fuzzy linguistic variables is ranked, then the position weights are aggregated with them, but there is no relationship between  $\omega_j$  and  $\tilde{s}_j$ , and  $\omega_j$  is only associated with the  $j^{th}$  position in the aggregation process, so  $\omega = (\omega_1, \omega_2, \dots, \omega_n)$  is called the position weight vector.

According to the real situation, the position weight vector  $\omega = (\omega_1, \omega_2, \dots, \omega_n)$  is determined. In this paper, the position weight is determined by the method which proposed in literature [15]. The formula is shown as follows:

$$\omega_{i+1} = \frac{C_{n-1}^i}{2^{n-1}}, i = 0, 1, \dots, n-1 \quad (6)$$

**Example 4:** Let  $\tilde{s}_1 = [s_2, s_3, s_5, s_6]$

and  $\tilde{s}_2 = [s_4, s_5, s_8, s_9]$  be two trapezoid fuzzy linguistic variables, and we already know that  $\tilde{s}_1 < \tilde{s}_2$  (the calculation steps are shown in **Example 1**), then the position weight vector is

$$\omega = (\frac{C_{2-1}^0}{2^{2-1}}, \frac{C_{2-1}^1}{2^{2-1}}) = (0.5, 0.5),$$



$$\begin{aligned}
 TFLOWHA_{\omega}(\tilde{s}_1, \tilde{s}_2) &= \left( \frac{0.5}{[s_2, s_3, s_5, s_6]} \oplus \frac{0.5}{[s_4, s_5, s_8, s_9]} \right)^{-1} \\
 &= \left( \frac{0.5}{[s_2, s_3, s_5, s_6]} \oplus \frac{0.5}{[s_4, s_5, s_8, s_9]} \right)^{-1} \\
 &= ([s_{0.083}, s_{0.1}, s_{0.167}, s_{0.25}] \oplus [s_{0.056}, s_{0.0625}, s_{0.1}, s_{0.125}])^{-1} \\
 &= [s_{0.139}, s_{0.1625}, s_{0.267}, s_{0.375}]^{-1} \\
 &= [s_{2.667}, s_{3.745}, s_{6.154}, s_{7.217}]
 \end{aligned}$$

The *TFLWHA* operator only focuses on the weight of the attribute value itself, but it ignores the position weight with respect to the attribute value; and the *TFLOWHA* operator focuses on the position weight with respect to the attribute value, but it ignores the weight of the attribute value itself. The two operators are one-sided. If the decision makers use these operators to aggregate the decision making information, some information may be lost. So in order to avoid the disadvantage of the two operators, the trapezoid fuzzy linguistic hybrid harmonic averaging (*TFLHHA*) operator is defined as follows:

**Definition 3.3:** Let  $TFLHHA: \tilde{S}^n \rightarrow \tilde{S}$ , if

$$TFLHHA_{\omega, w}(\tilde{s}_1, \tilde{s}_2, \dots, \tilde{s}_n) = \left( \sum_{j=1}^n \frac{\omega_j}{\tilde{r}_j} \right)^{-1} \quad (7)$$

where  $\tilde{S}$  is the set of all trapezoid fuzzy linguistic variables, and  $\tilde{s}_i, \tilde{r}_j \in \tilde{S} (i, j = 1, 2, \dots, n)$  are the trapezoid fuzzy linguistic variables.  $\tilde{r}_j$  is the  $j^{th}$  largest

of  $\tilde{s}_i / nw_i (i = 1, 2, \dots, n)$ , where

$w = (w_1, w_2, \dots, w_n)$  is the weight vector, and  $w_i$  is the

weight of  $\tilde{s}_i, w_i \geq 0 (i = 1, 2, \dots, n), \sum_{i=1}^n w_i = 1$ , and

$n$  is the balancing coefficient.  $\omega = (\omega_1, \omega_2, \dots, \omega_n)$  is the position weight vector with *TFLHHA*, where

$\omega_j \geq 0 (j = 1, 2, \dots, n), \sum_{j=1}^n \omega_j = 1$ , then *TFLHHA*

is called the trapezoid fuzzy linguistic hybrid harmonic averaging (*TFLHHA*) operator.

**Example 5:** Let  $\tilde{s}_1 = [s_2, s_3, s_5, s_6]$  and  $\tilde{s}_2 = [s_4, s_5, s_8, s_9]$  be two trapezoid fuzzy linguistic variables. We already know that the position weight vector is  $\omega = (0.5, 0.5)$  (the calculation steps are shown in **Example 4**), and the weight vector is  $w = (0.3, 0.7)$ , given by the decision makers, then based on the method shown in section 2.2, we can

calculate that:  $\tilde{r}_1 = \tilde{s}_1 / 2w_1 = [s_{3.333}, s_5, s_{8.333}, s_{10}]$ , and

$$\tilde{r}_2 = \tilde{s}_2 / 2w_2 = [s_{2.857}, s_{3.571}, s_{5.714}, s_{6.429}]$$

Then,

$$\begin{aligned}
 TFLHHA_{\omega, w}(\tilde{s}_1, \tilde{s}_2, \dots, \tilde{s}_n) &= \left( \sum_{j=1}^n \frac{\omega_j}{\tilde{r}_j} \right)^{-1} \\
 &= \left( \frac{0.5}{[s_{3.333}, s_5, s_{8.333}, s_{10}]} \oplus \frac{0.5}{[s_{2.857}, s_{3.571}, s_{5.714}, s_{6.429}]} \right)^{-1} \\
 &= ([s_{0.05}, s_{0.06}, s_{0.1}, s_{0.15}] \oplus [s_{0.0778}, s_{0.0875}, s_{0.14}, s_{0.175}])^{-1} \\
 &= (s_{0.1278}, s_{0.1475}, s_{0.24}, s_{0.325})^{-1} \\
 &= (s_{3.077}, s_{4.167}, s_{6.780}, s_{7.826})
 \end{aligned}$$

Especially, if  $w = (1/n, 1/n, \dots, 1/n)$ , then

*TFLHHA* operator is reduced to *TFLOWHA*

operator; if  $\omega = (1/n, 1/n, \dots, 1/n)$ , then *TFLHHA*

operator is reduced to the *TFLWHA* operator.

Obviously, *TFLOWHA* operator and

*TFLWHA* operator are extended from the *TFLHHA*

operator. The *TFLHHA* operator focuses on not only

the importance of the weight of the trapezoid fuzzy

linguistic variables itself, but also the importance of the

position weight of the trapezoid fuzzy linguistic

variables. So this operator is better than the previous

ones.

#### 4 Multiple Attribute Decision Making Method Based on the Trapezoid Fuzzy Linguistic Variables

A multiple attribute decision making problem under the fuzzy linguistic environment is represented as follows:

Let  $X = \{x_1, x_2, \dots, x_n\}$  be the set of the

alternatives, and  $U = \{u_1, u_2, \dots, u_m\}$  be the set of the

attributes. Let  $w = (w_1, w_2, \dots, w_m)^T$  be the weight

vector of the attributes, and  $w_j$  be the weight value of the

$j^{th}$  attribute, where  $w_j \geq 0 (j = 1, 2, \dots, m)$ ,

$\sum_{j=1}^m w_j = 1$ , given by the decision makers directly.

Suppose that  $\tilde{A} = (\tilde{a}_{ij})_{n \times m}$  is the fuzzy linguistic

decision matrix

$$\begin{matrix} u_1 & u_2 & \dots & u_m \end{matrix}$$

$$\tilde{A} = \begin{bmatrix} \tilde{a}_{11} & \tilde{a}_{12} & \cdots & \tilde{a}_{1m} \\ \tilde{a}_{21} & \tilde{a}_{22} & \cdots & \tilde{a}_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \tilde{a}_{n1} & \tilde{a}_{n2} & \cdots & \tilde{a}_{nm} \end{bmatrix} \begin{matrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{matrix}$$

where  $\tilde{a}_{ij} = [a_{ij}^{(\alpha)}, a_{ij}^{(\beta)}, a_{ij}^{(\gamma)}, a_{ij}^{(\eta)}] \in \tilde{\mathcal{S}}$  is the attribute value which takes the form of the trapezoid fuzzy linguistic variables, given by the decision makers, for the alternative  $x_i \in X (i = 1, 2, \dots, n)$  with respect to the attribute  $u_j \in U (j = 1, 2, \dots, m)$ . Let  $\tilde{a}_i = [\tilde{a}_{i1}, \tilde{a}_{i2}, \dots, \tilde{a}_{im}]$  be the vector of the attribute values under the alternative  $x_i (i = 1, 2, \dots, n)$ .

Then the decision making steps are shown as follows

**Step 1:** Construct the weighted linguistic matrix  $\tilde{A}' = (\tilde{a}'_{ij})_{n \times m}$

$$\tilde{A}' = \begin{bmatrix} \tilde{a}'_{11} & \tilde{a}'_{12} & \cdots & \tilde{a}'_{1m} \\ \tilde{a}'_{21} & \tilde{a}'_{22} & \cdots & \tilde{a}'_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \tilde{a}'_{n1} & \tilde{a}'_{n2} & \cdots & \tilde{a}'_{nm} \end{bmatrix} = (\tilde{a}'_{ij})_{n \times m}$$

where  $\tilde{a}'_{ij} = \tilde{a}_{ij} / nw_j$ ,  $w = (w_1, w_2, \dots, w_m)$  is the weight vector of the attributes,  $w_j > 0 (j = 1, 2, \dots, m)$ ,  $\sum_{j=1}^m w_j = 1$ ,  $n$  is the balancing coefficient.

**Step 2:** Utilize the formula (3) to construct the possibility degree matrixes  $P_i = (p_{jk}^{(i)})_{m \times m} = (p_{jk}^{(i)}(\tilde{a}'_{ij} \geq \tilde{a}'_{ik}))_{m \times m}$  with respect to the alternative  $x_i (i = 1, 2, \dots, n)$ , and sum all the elements of each rows of the possibility degree matrix  $P_i$ , then get the ranking vectors  $p^{(i)} = (p_1^{(i)}, p_2^{(i)}, \dots, p_j^{(i)})$ , ( $j = 1, 2, \dots, m$ ), where  $p_j^{(i)} = \sum_{k=1}^m p_{jk}^{(i)}$ . Finally, rank the orders of attribute values  $\tilde{a}'_{ij} (j = 1, 2, \dots, m)$  with respect to the alternative  $x_i$  based on the values  $p_j^{(i)} (j = 1, 2, \dots, m)$ .

**Step 3:** Utilize the formula (6) to calculate the position weight vector  $\omega = (\omega_1, \omega_2, \dots, \omega_m)$  of TFLHHA operator.

**Step 4:** Utilize the formula (7) to calculate the combined attribute values

$$\tilde{z}_i = TFLHHA_{\omega, w}(\tilde{s}_1, \tilde{s}_2, \dots, \tilde{s}_m) = \left( \sum_{j=1}^m \frac{\omega_j}{\tilde{r}_j} \right)^{-1}$$

where  $i = 1, 2, \dots, n$ .

**Step 5:** Utilize the formula (3) to construct the possibility degree matrix  $P = (p_{ij})_{n \times n}$ , based on the combined attribute values  $\tilde{z}_i$  of each alternative, then sum all the elements of each rows of the possibility degree matrix, where  $p_i = \sum_{j=1}^n p_{ij} (i = 1, 2, \dots, n)$ .

Rank all the combined attribute values of each alternative and select the best alternative based on the values  $p_i$ .

### 5 Illustrative Examples

In this section, a decision making problem of evaluating cars for buying (adapted from literature [1, 9]) is used to illustrate the new method.

A decision maker intends to buy a car. Four types of cars  $x_i (i = 1, 2, 3, 4)$  are available. He takes into account four attributes to decide which car he should buy: 1)  $G_1$ : economy, 2)  $G_2$ : comfort, 3)  $G_3$ : design, and 4)  $G_4$ : safety. The decision maker evaluates these four types of cars  $x_i (i = 1, 2, 3, 4)$  under the attributes  $G_j (j = 1, 2, 3, 4)$ , where the weight vector is  $w = (0.3, 0.2, 0.1, 0.4)$  given by the decision makers. He uses the linguistic term set:

$S = \{s_1 = \text{extremely poor}, s_2 = \text{very poor}, s_3 = \text{poor}, s_4 = \text{slightly poor}, s_5 = \text{fair}, s_6 = \text{slightly good}, s_7 = \text{good}, s_8 = \text{very good}, s_9 = \text{extremely good}\}$  and provides the linguistic decision making matrix  $\tilde{A} = (\tilde{a}_{ij})_{4 \times 4}$ :

$$\tilde{A} = \begin{bmatrix} [s_2, s_3, s_5, s_6][s_4, s_5, s_8, s_9][s_5, s_6, s_7, s_9][s_3, s_4, s_5, s_7] \\ [s_3, s_5, s_6, s_7][s_5, s_6, s_7, s_8][s_4, s_5, s_8, s_9][s_4, s_5, s_7, s_8] \\ [s_4, s_6, s_8, s_9][s_4, s_5, s_6, s_7][s_6, s_7, s_8, s_9][s_3, s_4, s_5, s_6] \\ [s_5, s_6, s_7, s_9][s_4, s_7, s_8, s_9][s_3, s_5, s_6, s_7][s_6, s_7, s_8, s_9] \end{bmatrix}$$

**Step 1:** Construct the weighted linguistic matrix  $\tilde{A}' = (\tilde{a}'_{ij})_{4 \times 4}$ , where

$$\tilde{a}'_{ij} = \tilde{a}_{ij} / nw_j (j = 1, 2, 3, 4).$$

$$\tilde{A}' = \begin{bmatrix} [s_{1.67}, s_{2.5}, s_{4.17}, s_5][s_5, s_{6.25}, s_{10}, s_{11.25}] \\ [s_{2.5}, s_{4.17}, s_5, s_{5.83}][s_{6.25}, s_{7.5}, s_{8.75}, s_{10}] \\ [s_{3.33}, s_5, s_{6.67}, s_{7.5}][s_{3.33}, s_{4.17}, s_5, s_{5.83}] \\ [s_{4.17}, s_5, s_{5.83}, s_{7.5}][s_5, s_{8.75}, s_{10}, s_{11.25}] \\ [s_{12.5}, s_{15}, s_{17.5}, s_{22.5}][s_{1.875}, s_{2.5}, s_{3.125}, s_{4.375}] \\ [s_{10}, s_{12.5}, s_{20}, s_{22.5}][s_{2.5}, s_{3.125}, s_{4.375}, s_5] \\ [s_{15}, s_{17.5}, s_{20}, s_{22.5}][s_{1.875}, s_{2.5}, s_{3.125}, s_{3.75}] \\ [s_{7.5}, s_{12.5}, s_{15}, s_{17.5}][s_{3.75}, s_{4.375}, s_5, s_{5.625}] \end{bmatrix}$$

**Step 2:** Utilize the formula (3) to construct the possibility degree matrixes  $P_i = (p_{jk}^{(i)})_{4 \times 4} = (p_{jk}^{(i)}(\tilde{a}'_{ij} \geq \tilde{a}'_{ik}))_{4 \times 4}$  with respect to each alternative  $x_i$  ( $i=1,2,3,4$ ), and sum all the elements of each rows of the possibility degree matrix  $P_i$ , then get the ranking vectors  $p^{(i)} = (p_1^{(i)}, p_2^{(i)}, \dots, p_j^{(i)})$ , ( $j=1,2,3,4$ ), where  $p_j^{(i)} = \sum_{k=1}^4 p_{jk}^{(i)}$ . Finally, rank the orders of attribute values  $\tilde{a}'_{ij}$  ( $j=1,2,3,4$ ) with respect to the alternative  $x_i$  based on the values  $p_j^{(i)}$  ( $j=1,2,3,4$ ).

$$P_1 = \begin{bmatrix} 0.5 & 0 & 0 & 0.59 \\ 1 & 0.5 & 0 & 1 \\ 1 & 1 & 0.5 & 1 \\ 0.41 & 0 & 0 & 0.5 \end{bmatrix}$$

$$p^{(1)} = (p_1^{(1)}, p_2^{(1)}, p_3^{(1)}, p_4^{(1)}) = (1.09, 2.5, 3.5, 0.91)$$

$$\tilde{a}'_{13} > \tilde{a}'_{12} > \tilde{a}'_{11} > \tilde{a}'_{14}$$

$$P_2 = \begin{bmatrix} 0.5 & 0 & 0 & 0.66 \\ 1 & 0.5 & 0 & 1 \\ 1 & 1 & 0.5 & 1 \\ 0.34 & 0 & 0 & 0.5 \end{bmatrix}$$

$$p^{(2)} = (1.16, 2.5, 3.5, 0.84)$$

$$\tilde{a}'_{23} > \tilde{a}'_{22} > \tilde{a}'_{21} > \tilde{a}'_{24}$$

$$P_3 = \begin{bmatrix} 0.5 & 0.73 & 0 & 1 \\ 0.27 & 0.5 & 0 & 1 \\ 1 & 1 & 0.5 & 1 \\ 0 & 0 & 0 & 0.5 \end{bmatrix}$$

$$p^{(3)} = (2.23, 1.77, 3.5, 0.5)$$

$$\tilde{a}'_{33} > \tilde{a}'_{31} > \tilde{a}'_{32} > \tilde{a}'_{34}$$

$$P_4 = \begin{bmatrix} 0.5 & 0 & 0 & 0.78 \\ 1 & 0.5 & 0.0625 & 1 \\ 1 & 0.9375 & 0.5 & 1 \\ 0.22 & 0 & 0 & 0.5 \end{bmatrix}$$

$$p^{(4)} = (1.28, 2.5625, 3.4375, 0.72)$$

$$\tilde{a}'_{43} > \tilde{a}'_{42} > \tilde{a}'_{41} > \tilde{a}'_{44}$$

**Step 3:** utilize the formula (6) to calculate the position vector of *TFLHHA* operator:  $\omega = (0.125, 0.375, 0.375, 0.125)$ .

**Step 4:** utilize the formula (7) to calculate the combined attribute values:

$$\tilde{z}_1 = (s_{2.65}, s_{3.73}, s_{5.73}, s_{7.02}),$$

$$\tilde{z}_2 = (s_{3.67}, s_{5.27}, s_{6.55}, s_{7.55}),$$

$$\tilde{z}_3 = (s_{3.33}, s_{4.5}, s_{5.63}, s_{6.53}),$$

$$\tilde{z}_4 = (s_{4.65}, s_{6.39}, s_{7.39}, s_{8.87})$$

**Step 5:** utilize the formula (3) to construct the possibility degree matrix, based on  $\tilde{z}_i$ :

$$P = \begin{bmatrix} 0.5 & 0.33 & 0.46 & 0.15 \\ 0.67 & 0.5 & 0.66 & 0.29 \\ 0.54 & 0.34 & 0.5 & 0.12 \\ 0.85 & 0.71 & 0.88 & 0.5 \end{bmatrix}$$

then sum all the elements of each row of the possibility degree matrix, we can get  $p_1 = 1.44$   $p_2 = 2.12$   $p_3 = 1.5$   $p_4 = 2.94$ . Based on the values  $p_i$ , rank all combined attribute values of each alternative and select the best alternative, then we can get  $x_4 > x_2 > x_3 > x_1$ , so the best alternative is  $x_4$ .

In order to verify the effective of this method, we utilized the method shown in literature [9] to solve this illustrate example.

Step1: From the linguistic decision making matrix  $\tilde{A} = (\tilde{a}_{ij})_{n \times m}$ , we can get the vector of the ideal point of the attribute values corresponding to the alternative  $x_i$  ( $i=1,2,3,4$ ):  $\tilde{I} = (\tilde{I}_1, \tilde{I}_2, \tilde{I}_3, \tilde{I}_4)$ , and

$$\tilde{I}_1 = (s_5, s_6, s_8, s_9), \tilde{I}_2 = (s_5, s_7, s_8, s_9),$$

$$\tilde{I}_3 = (s_6, s_7, s_8, s_9), \tilde{I}_4 = (s_6, s_7, s_8, s_9)$$

Step2: Utilize the *TFLWA* operator to derive the overall values  $\tilde{z}_i$  ( $i=1,2,3,4$ ) of the alternative  $x_i$  ( $i=1,2,3,4$ ) and  $\tilde{Z}$  of the ideal point  $\tilde{I}$

$$\tilde{z}_1 = (s_{3.1}, s_{4.1}, s_{5.8}, s_{7.3}), \tilde{z}_2 = (s_{3.9}, s_{5.2}, s_{6.8}, s_{7.8}),$$

$$\tilde{z}_3 = (s_{3.8}, s_{5.1}, s_{6.4}, s_{7.4}), \tilde{z}_4 = (s_5, s_{6.5}, s_{7.5}, s_{8.8})$$

$$\tilde{z} = (s_{5.5}, s_{6.7}, s_8, s_9)$$

Step3: We get the similarity degree  $s(\tilde{z}, \tilde{z}_i)$  between  $\tilde{z}$  and  $\tilde{z}_i$  ( $i=1,2,3,4$ ) based on the similarity degree formula

$$s(\tilde{z}, \tilde{z}_1) = 0.876, s(\tilde{z}, \tilde{z}_2) = 0.924,$$

$$s(\tilde{z}, \tilde{z}_3) = 0.910, s(\tilde{z}, \tilde{z}_4) = 0.981$$

Step 4: Rank the order of  $s(\tilde{z}, \tilde{z}_i)$  ( $i=1,2,3,4$ ), then we can get:  $x_4 > x_2 > x_3 > x_1$ .

Analysis:

The order calculated by this method is the same as the order calculated by the method proposed in literature [9], so it is demonstrated that the method proposed in this paper is feasible and effective, and it is also verified that the *TFLHHA* operator is effective. It provided the new idea to solve the MADM problems under the linguistic context, and it provided the new idea of aggregating the trapezoid fuzzy linguistic variables in the MADM problems.

## 6 Conclusions

This paper proposed a new method of the MADM problems based on the *TFLHHA* operator. The new method can deal with the MADM problems where the decision making information takes the form of the *TFLVs* directly, and makes the computation process of the *TFLVs* easily without the loss of the information. This method is easy to use and understand, and it enriched and developed the theory and method of the MADM, This method can solve these MADM problems where the attribute values take the form of the fuzzy linguistic variables, such as fuzzy linguistic variables, the uncertain fuzzy linguistic variables, the triangular fuzzy linguistic variables, the trapezoid fuzzy linguistic variables, and the mixed fuzzy linguistic variables, if we can transform these fuzzy linguistic variables into the trapezoid fuzzy linguistic variables. But this method can only solve the MADM problem under the linguistic context. So it is the limitation of this paper. In the future, we will apply this method to solve the real-life MADM problems in the linguistic context, and we will continued working in the decision making method of the MADM problems with the *TFLV*.

## Acknowledgment

This paper is supported by the Humanities and Social Sciences Research Project of Ministry of Education of China (No.10YJA630073 and No.09YJA630088), the Natural Science Foundation of Shandong Province (No. ZR2011FM036), the Social Science Planning Project Fund of Shandong Province (09BSHJ03), the Soft Science Project Fund of Shandong Province (2009RKA376), and the Doctor Foundation of Shandong Economic University. The authors also would like to express appreciation to the anonymous reviewers for their very helpful comments on improving the paper.

## References

- [1] Herrera, F., Martínez, L.(2000). An Approach for Combining Numerical and Linguistic Information Based on the 2-Tuple Fuzzy Linguistic Representation Model in Decision Making, *International Journal of Uncertainty, Fuzziness and Knowledge -Based Systems*, 8, pp. 539-562.
- [2] Li, D.F., Yang, J.B.(2004). Fuzzy Linear Programming Technique for Multi-attribute Group Decision Making in Fuzzy Environments, *Information Sciences*, 158, pp. 263-275.
- [3] Herrera, F., Martínez L. (2000). A 2-tuple fuzzy linguistic representation model for computing with words, *IEEE Transactions on fuzzy systems*, 8, pp. 746–752.
- [4] Xu, Z.S. (2004).Uncertain Multiple Attribute Decision Making: Methods and Applications, Tsinghua University Press, Beijing.
- [5] Bordogna, G., Fedrizzi, M., Pasi, G. (1997).A Linguistic Modeling of Consensus in Group Decision Making Based on OWA Operators, *IEEE Transactions on Systems, Man, and Cybernetics-Part A* ,27(1), pp.126-132.
- [6] Xu, Z.S. (2004). Uncertain Linguistic Aggregation Operators Based Approach to Multiple Attribute Group Decision Making under Uncertain Linguistic Environment, *Information Sciences*, 168, pp.171-184.
- [7] Wu, Z.B., Chen, Y.H. (2007). The maximizing deviation method for group multiple attribute decision making under linguistic environment, *Fuzzy Sets and Systems*, 158, pp. 1608-1617.
- [8] Xu, Z.S.(2007). Group decision making with triangular fuzzy linguistic variables, Springer Berlin / Heidelberg. pp.17-26.
- [9] Xu, Z.S.(2005). An approach based on similarity measure to multiple attribute decision making with trapezoid fuzzy linguistic variables, *Fuzzy Systems and Knowledge Discovery* 3613, pp. 110-117.
- [10] Liang, X.C., Chen, S.F.(2008). Multiple attribute decision making method based on trapezoid fuzzy linguistic variables, *Journal of Southeast University (English Edition)*, 24(4), pp.478-481.
- [11] Chen, H.Y., Liu, C.L., Sheng Z.H.(2004). Induced Ordered Weighted Harmonic Averaging (IOWHA) Operator and Its Application to Combination Forecasting Method, *Chinese Journal of Management Science*, 12(5), pp. 35-40.
- [12] Liu, J.P., Chen, H.Y.(2007). Uncertain combined weighting harmonic averaging operators and its application, *Operations Research and Management Science*, 16(3), pp.36-40.
- [13] Wei, G.W.(2007). Fuzzy Linguistic Hybrid Geometric Aggregation Operator and Its Application to Group Decision Making, Science paper online. [www.paper.edu.cn/](http://www.paper.edu.cn/).
- [14] Xu, Z.S. (2010). The least variance priority method (LVM) for fuzzy complementary judgment matrix, *System engineering theory& practice*, 21(10), pp.93-96.
- [15] Wang, Y., Xu, Z.S.(2008). A new method of giving OWA weights, *Mathematics in Practice and Theory*, 38(3), pp.51-61.

# Physics Markup Approaches Based on Geometric Algebra Representations

Kuo-pao Yang and Wendy Zhang

Computer Science & Industrial Technology Department, Southeastern Louisiana University, USA

E-mail: {kyang, wzhang}@selu.edu, <http://www2.selu.edu/Academics/Faculty/{kyang, wzhang}>

Frederick Petry

Naval Research Laboratory, Stennis Space Center, USA

<http://www7440.nrlssc.navy.mil>

**Keywords:** geometric algebra, markup languages, MathML, OpenMath, content dictionary

**Received:** January 20, 2012

*This paper presents an approach for a physics markup language using Geometric Algebra which is a unifying language for the mathematics of physics and is useful in an exceptionally wide range of physics problems, particularly those that involve rotations, phases or imaginary numbers. MathML and OpenMath are discussed as potential ways to implement a markup system. Using OpenMath, content dictionaries for Geometric Algebra were developed and used to illustrate the markup of the physics of the rotor which is used in 3-dimensional rotations.*

*Povzetek: Članek predstavi na XMLju temelječ jezik za zapisovanje fizikalnih izrazov.*

## 1 Introduction

In large environmental models, a significant issue is how to effectively integrate the physics used in diverse system components such as found in oceanographic and atmospheric forecasting. A major problem for component integration lies in the implicit assumptions made about the semantics of the physics being used. Usually in such systems the semantics of the physics is only “roughly” described and certainly not in any formal manner. Computational systems lack the ability to use context to understand the semantics of a mathematical denotation. If we wish such meanings to be reliably communicated between such systems, we must mark up the document to provide extra semantic information. A Physics Markup Language (PML) could allow these components to be formally described, tested and used to aid in integration [7], [37].

As physics must be represented mathematically there is a question as to how to consistently deal with various physics concepts. First one must consider the specific mathematical concepts necessary in expressing physical semantics so that they may be handled separately. This separation allows experts, specializing in representing mathematical semantics, to aid in the development of PML by expanding mathematical semantic representations, a pre-requisite in expressing a large body of physical models. A mathematical approach that is able to encompass much of this in a uniform fashion is called Geometric Algebra (GA). GA [35] is a unifying language for the mathematics of physics and is useful in an exceptionally wide range of physics problems, particularly those that involve rotations, phases or

imaginary numbers. Geometric Algebra more compactly and intuitively describes classical mechanics, quantum mechanics, electromagnetic theory and relativity than standard methods do [6], [23]. Our research uses Geometric Algebra to provide a uniform representation of physics concepts.

In order to effectively utilize Geometric Algebra, we need to evaluate which mark-up language to use, as the World Wide Web Consortium (W3C) has proposed a large number of standards for these [36]. Mathematical Markup Language (MathML) [25] is an Extensible Markup Language (XML) application for describing mathematical notation and capturing both its structure and content. MathML deals principally with the presentation of mathematical objects. MathML can be used to encode both mathematical notation and mathematical content. About thirty-eight of the MathML tags describe abstract notational structures, while about one hundred and seventy provide a way of unambiguously specifying the intended meaning of an expression [3]. MathML aims at integrating mathematical formulae into web documents but the semantic contents of mathematical formulae are limited.

OpenMath [30] is a standard aimed at supporting a semantically rich interchange of mathematics among varied computational software tools such as computer algebra systems, theorem provers, and tools for visualizing or editing mathematical text [4]. Open Mathematical Documents (OMDoc) [29] is a semantic markup format for mathematical documents that we use for the Physics Markup Language. OMDoc is used for

mathematical knowledge representation with numerous applications such as creation of customized modules for e-learning, data exchange between different theorem provers, web services, and more. This research focuses on building Content Dictionaries in OpenMath [20] format for Geometric Algebra [21], [38]. Content Dictionaries for Geometric Algebra have not been previously created to interpret semantics of documentations.

## 2 Background

### 2.1 Markup Languages

A markup language is a system for annotating a document for processing, defining, and presenting text syntactically [3]. Hyper Text Markup Language (HTML) is a widely used webpage markup language with predefined structural markers for communicating presentational semantics [24]. These HTML tags markup the document in order to denote the presentation specifications for text and other data. An example of an HTML tag is ‘<head>’ to indicate the heading beginning for a document and ‘</head>’ to indicate the end of the heading.

XML has become as important as HTML for structuring, storing, and transporting information, extensively used in representing arbitrary data structures [13]. XML provides strong support with simplicity, while exclusively regarding the data’s meaning, instead of how the data is displayed. Similarly to HTML, XML is a set of standardized rules for encoding documents with structural markers, known as XML tags. Unlike HTML, XML tags do not have predefined semantics, permitting the author to establish unique and specific XML tags and document structure. Commonly applied together, HTML formats and displays data, while XML stores and transports.

Markup languages based on XML have been developed for a number of specific areas. The Chemical Markup Language (CML) is an approach to supporting interoperable capabilities for a wide variety of chemical concepts such as molecular information, chemical reactions, spectra and analytical data and other information [27], [28]. In the area of biology the Systems Biology Markup Language (SBML) can represent models of biological processes such as metabolic networks, cell-signaling pathways and many others [14], [19]. A Geometry Description Markup Language (GDML) is an XML structured language for describing detector geometries for physics experimental configurations [5]. As it based on pure XML it can be useful for geometry interchange among different applications. In this paper we are concerned with the development of an XML based markup language for the semantics of physics (PML).

Now we discuss MathML and OpenMath as existing approaches that are used to represent mathematics concepts and can be used a basis for PML. MathML, illustrating mathematical notations and capturing their

structures and contents, enables one to display, manipulate, and share mathematical expressions over the web [26], [32]. MathML expressions can be evaluated in computer mathematical systems, rendered in web browsers, edited in word processors, and sent to printers. The XML-based MathML language consists of presentation markups and content markups. The presentation elements, depicting mathematical notations, are used to visually display. The content elements, describing the structures of mathematical expressions, explain what the mathematics means.

The mathematical expression  $(X^2 + 3X - 4)$  shown in Figure 1 is implemented in MathML markup language. The presentation element tags, <row>, <msup>, <mi>, <mn>, and <mo>, help lay out the mathematical expression. The entity reference, “&InvisibleTimes;” indicates it is invisible and “3” and “X” are multiplied. Giving additional information about the meaning of the equation is useful in creating complex mathematical expressions and also in evaluating the markup of computer algebra system. The XSLT stylesheet, “mathml.xsl,” transforms the XML-based implementation into presentation markups and then displays in a Firefox web browser shown in Figure 2.

```
<?xml version="1.0"?>
<?xml-stylesheet type="text/xsl" href=
  "http://www.w3.org/1998/Math/MathML/mathml.xsl"?>
<html xmlns="http://www.w3.org/1999/xhtml">
  <head>A Simple Quadratic Polynomial</head>
  <body>
    <math xmlns=
      "http://www.w3.org/1998/Math/MathML">
      <mrow>
        <msup><mi> X </mi><mn> 2 </mn></msup>
        <mo> + </mo>
        <mrow> <mn> 3 </mn>
          <mo>&InvisibleTimes;
            </mo> <mi> X </mi>
        </mrow>
        <mo> - </mo><mn> 4 </mn>
      </mrow>
    </math>
  </body>
</html>
```

Figure 1: MathML for a Mathematical Expression  $(X^2 + 3X - 4)$ .

OpenMath, a general representation XML-based language for communicating mathematical objects, is about semantic definitions and is used to complement MathML, which determines how expressions are elegantly rendered [10]. OpenMath is an emerging standard for representing mathematical objects with their semantics, allowing them to be exchanged between computer programs, stored in databases, or published on the World Wide Web. The first OpenMath standard [4] encoded in an XML-based markup language was released in 2000. OpenMath consists of the definition of OpenMath Objects, abstract data types for describing the logical structures of mathematical formulae, and the

definition of OpenMath Content Dictionaries, collections of symbol names for mathematical concepts. OpenMath provides XML encodings that meet these requirements to describe the logical structures, and a set of specific Content Dictionaries [9], [20] for some areas of mathematics, in particular covering the K-14 education fragment [22].

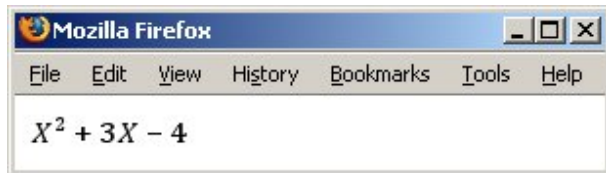


Figure 2: Displaying the MathML Presentation ( $X^2 + 3X - 4$ ) in Mozilla Firefox.

While the original designers were mainly developers of computer algebra systems, it is now attracting interest from other areas of scientific computation and from many publishers of electronic documents with a significant mathematical content [4]. OpenMath is solely concerned with mathematical objects' semantic meaning or content.

OpenMath Content Dictionaries can be embedded and referenced in the content of MathML to define the meaning of mathematical formulae. Strict Content MathML is designed to be an XML encoding of OpenMath Objects. Formal semantics of mathematical expressions in MathML [20] will be fully supported in terms of OpenMath Content Dictionaries. Several applications based on OMDoc were developed in recent years [18], [34], [37].

## 2.2 Geometric Algebra

Geometric Algebra is a consistent computational framework to define geometric primitives and their relationships. This algebraic approach contains geometric operators and permits specification of constructions in a coordinate-free manner. Geometric Algebra gives a geometric extension of the real number system to provide a complete algebraic representation of geometric notations of direction and magnitude. Geometric Algebra provides for a mechanism linking magnitudes and numbers and lends themselves neatly to the representation of physical problems and of reality as we know it. The use of the geometric algebra can provide a unified language for problems in fields such as physics and engineering. GA is often referred to as a unified mathematical language for physics and engineering in the 21st century [15]. Current applications of Geometric Algebra include computer vision, biomechanics and robotics, and distributed data representations [11], [31]. In mathematical physics, a Geometric Algebra is a multilinear algebra described technically as a Clifford algebra over a real vector space equipped with a non-degenerate quadratic form.

The basics concepts leading to the development of geometric algebra occurred during the late 1800's [16] Geometric algebra is a Clifford algebra which has been

used with great success in the modeling of a wide variety of physical phenomena. Clifford algebra is considered a more general algebraic framework than geometric algebra.

However the introduction of the standard Gibbs vector calculus, although having certain limitations, became the major formalism used. One issue addressed by geometric algebra is the limitation of the vector cross product which is only valid in 3 dimensions. The outer product of two vectors,  $a$ ,  $b$ , denoted by  $a \wedge b$ , replaces it and is termed a bivector. A bivector extended by a third vector,  $(a \wedge b) \wedge c$ , is a directed volume element called a trivector. The outer product actually works in all dimensions.

The structure of geometric algebra is based on  $k$ -blades, where  $k$  is called the grade and refers to the dimension of the subspace the blade spans. Vectors are 1-blades bivectors, 2-blades, trivectors, 3-blades and similarly in higher dimensional spaces. A key insight of Clifford was the introduction of a new product, geometric product,  $\otimes$ , combining the inner or dot product and the outer product:

$$a \otimes b = a \cdot b + a \wedge b$$

Since the other products can be expressed in terms of the single geometric product, it can then enable the unification of formalisms across several areas.

## 3 Content Dictionaries for Geometric Algebra

### 3.1 Content Dictionaries in Open Math

Content Dictionaries are used to assign semantics to all symbols used in the OpenMath objects. They define the symbols used to represent concepts arising in a particular area of mathematics. The Content Dictionaries represent the actual common knowledge among OpenMath applications. These provide the "meaning" of objects independently of the application. The application receiving the object may then recognize whether or not, according to the semantics of the symbols defined in the CDs, the object can be transformed to the corresponding internal representation used by the application [1].

A Content Dictionary has been designed to hold two types of information, a header followed by a number of CD Definitions. Each definition is placed inside of the CDDefinition element. It consists of a description, the Commented Mathematical Properties (CMP), and the Formal Mathematical Properties (FMP). A Content Dictionary head consists of the following pieces of information:

1. A CDname gives the name of the Content Dictionary
2. A description of the Content Dictionary
3. A revision date, the date of the last change to the Content Dictionary (Dates should be stored in the ISO-compliant format YYYY-MM-DD, e.g. 1966-02-03, and a review date, a date until

which the content dictionary is guaranteed to remain unchanged)

4. A version number which consists of a major and minor part
5. A status of CD
6. A CD base which, when combined with the CD name, forms a unique identifier for the Content Dictionary. It may or may not refer to an actual location from which it can be retrieved
7. CDURL should be a valid URL where the source file for the Content Dictionary encoding can be found
8. CDComment which can be used in the Content Dictionary header to report the author of the Content Dictionary and to log change information

A CD Definition contains information restricted to a particular symbol definition. This includes a name, a description in natural language, commented and formal properties satisfied by this symbol, and examples of the use of this symbol. A symbol definition consists of the following pieces of information:

1. A mandatory *name* for the symbol, a mandatory *description* of the symbol, which can be as formal or informal as the author likes, and an optional *role*.
2. Zero or more *commented mathematical properties* which are mathematical properties of the symbol expressed in a mechanism other than *OpenMath* and zero or more *formal mathematical properties* which are mathematical properties of the symbol expressed in *OpenMath*. It is common for commented and formal mathematical properties to be introduced in pairs, with the former describing the latter.
3. A Formal Mathematical Property may be given an optional *kind* attribute. An author of a Content Dictionary may use this to indicate whether, for example, the property provides an algorithm for evaluation of the concept it is associated with.
4. Zero or more *examples* which are intended to demonstrate the use of the symbol within an *OpenMath* object.

### 3.2 Content Dictionaries in Open Math

This section gives an example of the inner product using OpenMath CD shown in Figure 3.

#### A. Head of the CD

The CD header contains information pertinent to the whole CD. This includes the name, a description, a date when the CD will next be reviewed, the status of the CD (official, experimental, private, obsolete), and an optional list of CDs on which it depends. The CDName here gives the name of the Content Dictionary as GA-Products.

```
<?xml version="1.0" encoding="UTF-8"?>
<CD xmlns=
  "http://www.openmath.org/OpenMathCD">
<CDComment>
  Author: Joseph B. Collins and Fred
  Petry (2009), Naval Research
  Laboratory. Copyright Notice: This
  is a work of the U.S. Government and
  is not subject to copyright
  protection in the United States.
  Foreign copyrights may apply.
</CDComment>
<CDName>ga_product1</CDName>
<CDBase>http://www.openmath.org/cd
</CDBase>
<CDURL> http://www.openmath.org/
  cd/ga_product1.ocd </CDURL>
<CDReviewDate>2009-07-18</CDReviewDate>
<CDStatus>experimental</CDStatus>
<CDDate>2009-07-18</CDDate>
<CDVersion>1</CDVersion>
<CDRevision>1</CDRevision>
<Description>
  This content dictionary defines the
  fundamental products of
  <a xmlns=
    "http://www.w3.org/1999/xhtml"
    href="http://en.wikipedia.org/wiki/
    Geometric_algebra"> Geometric
    Algebra (GA)
  </a>
  such as inner product, outer
  product, geometric product, and
  scalar product.
  This CD also presents a set of
  axioms for GA associated with the
  products.
</Description>
```

Figure 3: CD Head of GA\_Product.

#### B. CD definition

The CD in the Appendix gives the definition of inner-product in GA. It defines the name as inner-product, role of inner-product as application, and the commented and formal mathematical property of inner-product. For the purposes of use of the inner-product with GAs, we assume a version of a linear algebra CD for which the vector-selector has the capability to select the blade. This means the inner-product can map two blades to a blade. Formal properties are expressed as an XML encoded OpenMath object, whereas commented properties are expressed in natural language. Scalable Vector Graphics (SVG) is a family of specifications of an XML-based file format for describing two-dimensional vector graphics [32]. To better represent the geometric feature of inner-product, instead of using XML text to give examples of the enclosing symbol, a SVG file is used to give graphic features as examples.

#### C. Signature Dictionary

A Small Type System, called STS, has been designed to give semi-formal signatures to OpenMath



symbols [8]. Using the same mechanism, the following example shows how data type in GA systems can be employed to assign types to OpenMath symbols. The following is the STS of inner\_product shown in Figure 4.

D. Display on the Web

The XML-base CD implementation can be transformed by OpenMath XSLT stylesheets into presentation markups and can be displayed in the Firefox web browser. Figure 5 shows the inner\_product Content Dictionary in Firefox with hyper links and graphics.

```
<Signature name="inner_product" >
<OMOBJ xmlns=
"http://www.openmath.org/OpenMath">
<OMA>
<OMS name="mapsto" cd="sts"/>
<OMA>
<OMV name="blade"/>
<OMV name="blade"/>
</OMA>
<OMV name="blade"/>
</OMA>
</OMOBJ>
</Signature>
```

Figure 4: Small Type System of Inner\_Product (ga\_product.sts).

Notice how the elements of each do not directly align. The ‘inner\_product’ CD components were not a straightforward integration from the given GA description. A systematic methodology needed to be constructed to allow GA elements to be transformed into CDs. Once this transformation is complete the GA elements can be provided with a method for standardizing their semantics. By ascertaining an extensive knowledge of GA elements and CD components, a one-to-one mapping was established enabling the required transformations.

The symbol name developed for the inner product was logically formed by simply substituting ‘\_’ for the space. Supplemental information regarding the inner product’s role was given as ‘application’ to further define how the GA element acts. ‘Application’ was decided upon since the inner product is applied upon two n-dimensional vectors. The description for ‘inner\_product’ was obtained through various resources including, but not limited to the papers: [12], [16], [17]. By gathering a comprehensive understanding of the GA element inner\_product from these resources, a concise and basic definition was composed. This enabled us to develop the Commented Mathematical Property (CMP) by expressing the given property for inner product  $A_r \cdot B_s = \langle A_r, B_s \rangle_{(r-s)}$  in plain text, correctly communicating the property’s meaning. The Formal Mathematical Property was constructed by representing the given property, previously expressed within the CMP, in MathML encoding. The image presented in the example for the ‘inner\_product’ was also selected after reviewing several resources, and chosen for its proper communication of the inner product’s semantics in a simple and clear

manner, to even further enhance the understanding of the meaning by the scientists. Lastly, a STS signature file was developed and linked for the ‘inner product’, showing data types within the GA CDs and assigning semi-formal signatures to the OpenMath object, inner\_product. Figure 4 shows the STS of inner\_product.

**inner\_product**

Role: application

Description:

Inner product specific to Geometric Algebra is the generalization of the scalar product, defined in CD linalg1, for arbitrary multivectors. The  $\cdot$  (dot) symbol is used to denote this operator. The inner product is a grade lowering operation.

Commented Mathematical property (CMP):  
For any homogeneous multivectors  $A_r$  of grade  $r$  and  $B_s$  of grade  $s$ , the inner product will lower the grade to  $(r-s)$  if  $r > 0, s > 0$ , and  $(r-s) > 0$ .

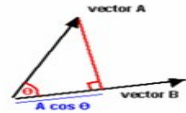
Formal Mathematical property (FMP):

|     |        |        |
|-----|--------|--------|
| xml | prefix | mathml |
|-----|--------|--------|

$$(A_r \cdot B_s) = (A_r \otimes B_s)_{r-s}$$

Example:  
The inner product of two vectors  $a$  and  $b$ , denoted by  $a \cdot b$ , projects  $a$  onto  $b$  resulting in the scalar magnitude of the projection relative to  $b$ 's magnitude.

Scalable Vector Graphics (SVG):



Signatures:  
[sts](#)

Figure 5: Inner\_Product Content Dictionary in Firefox.

The ‘inner\_product’ is expressed as a specialized form of the ‘geometric product’,  $A_r \cdot B_s = \langle A_r, B_s \rangle_{(r-s)}$ , above. Currently OpenMath does not contain the GA element ‘geometric product’, standardizing the semantics. Due to this fact and that CDs must be self-contained, the ‘geometric product’ was defined within the same CD as ‘inner\_product’ and denoted ‘ $\otimes$ ’. This only further exemplifies and supports the desire for standardizing GA semantics in an effort to avoid such notational differences, as seen above in expressing the ‘geometric product’.

Only through a collection of various resources lending to a comprehensive understanding of GA element components and CD components, was a correlation made identifying the required mapping from GA elements to CDs. This mapping, which was previously explored for the ‘inner\_product’ in the example above, becomes the generalized translation approach for all GA elements into OpenMath CDs. This approach was done for several types of GA elements, of which four CDs were produced for the areas of GA Basics, GA Products, GA Spaces, and GA Multivectors. A total of twenty-nine terms were defined within the CDs, all with enabled links and embedded images, verifying that the defined OpenMath mapping between CDs and GA elements were reasonably effective in

standardizing GA semantics, even though human reasoning was necessary.

### 3.3 Representation of Content Dictionaries

This section gives an example of the inner product using OpenMath CD. Content Dictionaries for Geometric Algebra are established and added into OpenMath library. For example, a comprehensive set of basic axioms for GA in terms of the fundamental geometric products such as inner, outer, geometric, and scalar products is implemented in OpenMath format. The developed XML-based Content Dictionary (ga\_product1.ocd) for GA products shown in Figure 6.

```

<CD xmlns=
  "http://www.openmath.org/OpenMathCD">
  <CDName>ga_product1</CDName>
  <CDDefinition>
  <Name>geometric_product</Name>
  <CMP>
  For any multivectors A, B, and C,
  geometric product is associative.
  </CMP>
  <FMP> ...
  <OMS cd="relation1" name="eq"/>
  <OMA><OMS cd="ga_product1"
  name="geometric_product"/>
  <OMA><OMS cd="ga_product1"
  name="geometric_product"/>
  <OMV name="A"/> <OMV name="B"/>
  </OMA>
  <OMV name="C"/>
  </OMA>
  <OMA><OMS cd="ga_product1"
  name="geometric_product"/>
  <OMV name="A"/>
  <OMA><OMS cd="ga_product1"
  name="geometric_product"/>
  <OMV name="B"/> <OMV name="C"/>
  </OMA>
  </OMA>
  </FMP>
  ...
  <Example> ... </Example>
  </CDDefinition>
  <CDDefinition>inner_product ...
  </CDDefinition>
  <CDDefinition>outer_product ...
  </CDDefinition>
  <CDDefinition>scalar_product...
  </CDDefinition>
  </CD>
  
```

Figure 6: Content Dictionary for Geometric Algebra Products (ga\_product1.ocd).

Knowledge of GA is placed inside CDDefinition element of OpenMath. The symbol elements, geometric product for instance, are added to introduce concepts. For the geometric product, one the formal mathematical properties, defining logical laws of the GA theory, is written in prefix notation and is associative:  $(A \otimes B) \otimes C = A \otimes (B \otimes C)$ . The XSLT style sheet (ga\_product1.xsl)

for GA products describes how presentation markups display in web browsers as follows:

```

<xsl:template
  match="om:OMS[@cd='ga_product1' and
  @name='geometric_product']">
  <xsl:call-template name="infix">
    <xsl:with-param name="mo">
      <mo>&#x2297;</mo>
    </xsl:with-param>
    ...
  </xsl:call-template>
</xsl:template>
  
```

To display infix form, the OpenMath symbol, geometric\_product, in Content Dictionary (ga\_product1.ocd) calls the infix stylesheet template. The MathML presentation element tag, <mo>, helps layout  $\otimes$  or x2297 in hexadecimal.

The Content Dictionary for GA products (ga\_product1.ocd), using its stylesheet (ga\_product1.xsl) and OpenMath stylesheets, is transformed into XHTML format (ga\_product1.xhtml) by Apache Ant, a Java-based build tool. This Content Dictionary for GA products displayed in Firefox browser shown in Figure 7. Each Formal Mathematical Property (FMP) has three toggle switch buttons to display XML code, prefix form, and MathML presentation. Each content dictionary is linked to its signature file.

The new stylesheets of GA enable one to write external references and vector graphics in OpenMath Content Dictionary (OCD) files after modifying the original stylesheets of OpenMath. Foreign Objects are containers for non-OpenMath structures according to the OpenMath 2.0 standard. Content Dictionaries for Geometric Algebra should allow us to enable graphics using the OMFOREIGN tag of OpenMath. However Scalable Vector Graphics could not be successfully implemented based on the original stylesheets of OpenMath. To make external reference in XHTML file, it is required to write <a> tags in OCD file, for example:

```

<a xmlns=
  "http://www.w3.org/1999/xhtml"
  href="http://en.wikipedia.org/wiki/Geometric_algebra"> Geometric
  Algebra
  </a>
  
```

This external link, Geometric Algebra, is enabled in web browser after transferring this OCD file into XHTML file.

Developers of Content Dictionaries can write in Scalable Vector Graphics [2], an open XML- based standard that has been under the development of the World Wide Web Consortium since 1999. It is required to write <svg> tags in OCD file. For example, this image (geometric\_product.png) in Portable (Public) Network

Graphic (PNG) format is enclosed in the following SVG code and can be displayed in web browser after transforming into XHTML format.

```
<svg xmlns="http://www.w3.org/2000/svg"
  xmlns:xlink=
  "http://www.w3.org/1999/xlink">
  <image width="100%"
    height="100%" xlink:href=
    "img/geometric_product.png" />
</svg>
```

## 4 Application of GA Markup to Physics

One of the clearest illustrations of GA markups' power is the way with which rotation can be dealt. In order to handle angular momentum and its many applications in dynamics and other topics, the representation of the rotation of vectors is central. The development of the concept of the rotor  $R$  in Geometric Algebra is an approach to this representation. For example, for a bivector  $B = a \wedge b$ , the rotation,  $B'$ , can be expressed as

$$B' = R \otimes B \otimes R^\dagger$$

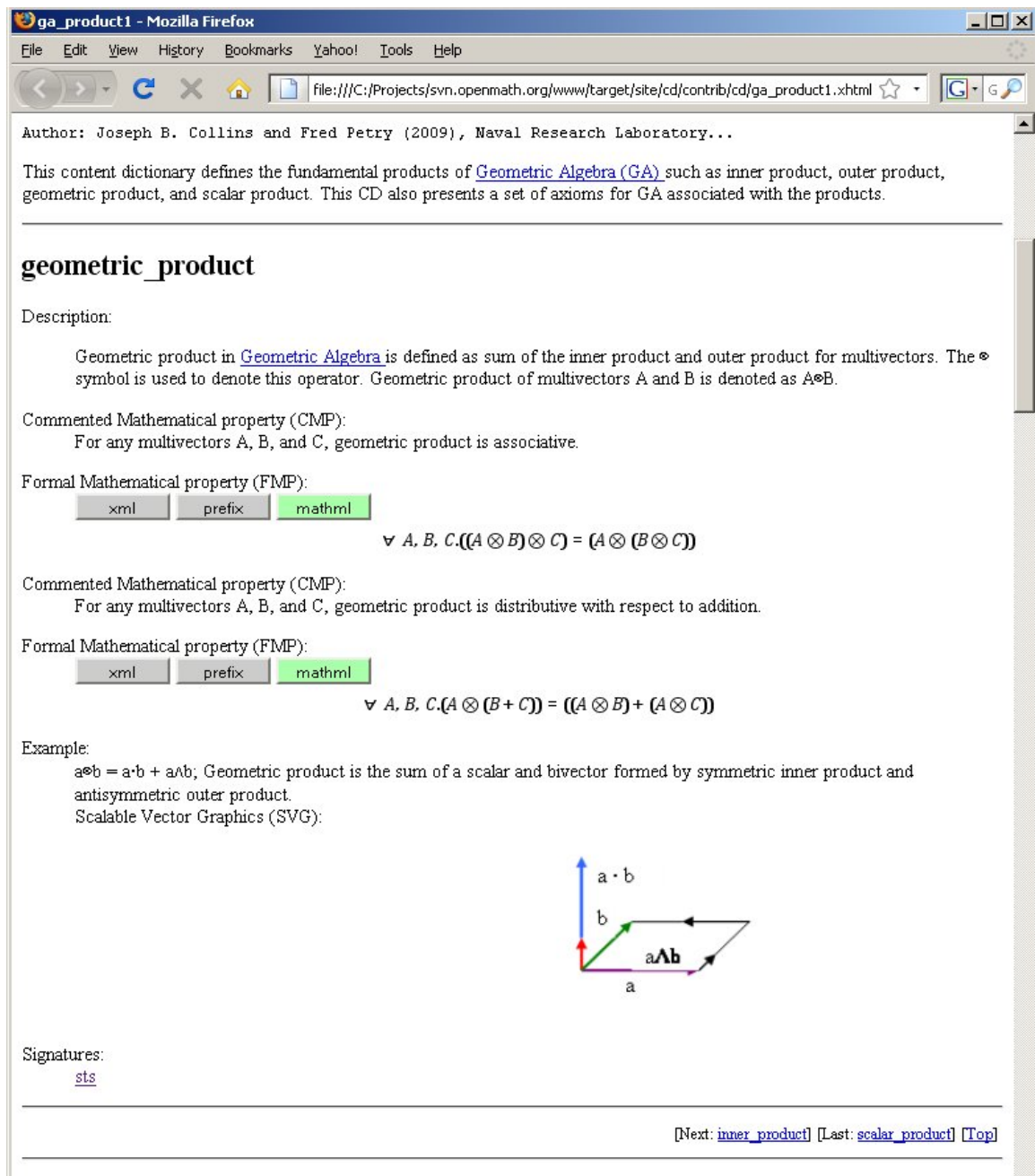


Figure 7: Content Dictionary for Geometric Algebra Products (ga\_product1.xhtml) in Firefox.

where  $\dagger$  is the reversion operation that reverses the order of vectors in a product. The power of the rotor is that this form for the rotation applies to all multivectors. So in this section we will show how we can represent the rotor  $R$  and the reversion operation  $\dagger$  in our markup approach.

### 4.1 Reversion

Reversion is an important operation in geometric algebra that reverses the order of vectors in any product.  $A^\dagger$  denotes the reverse of a multivector  $A$ .

The sign of scalars and vectors is unchanged but bivectors and trivectors change sign. The reverse of a product of vectors is defined by  $(ab\dots c)^\dagger = c \dots ba$ . The reverse can be formed by a series of swaps of anti-commuting vectors, each resulting in a minus sign. Based on the properties of reversion the CD that was developed is shown in Figure 8.

**Reversion**  
**Role:**  
 Operation

**Description:**  
 Reversion aids in reordering factors in a products by a series of swaps. Define the reversion  $\dagger$  as an operation that takes a multivector  $A$  and reverse the order of vectors in any product.  $A^\dagger$  denotes the reverse of a multivector  $A$  and the reverse of a product of vectors is  $(a_1 a_2 \dots a_r)^\dagger = a_r \dots a_2 a_1$

**Commented Mathematical property (CMP):**  
 For any multivectors  $A, B$ , the reversion of the geometric product  $AB$  becomes the geometric product of  $B$  reversion and  $A$  reversion.

**Formal Mathematical property (FMP):**  
 $(\forall A) \wedge (\forall B) \rightarrow (AB)^\dagger = B^\dagger A^\dagger$

**Commented Mathematical property (CMP):**  
 For any multivectors  $A, B$ , the reversion of the sum of  $A$  and  $B$  is the sum of  $A$  reversion and  $B$  reversion.

**Formal Mathematical property (FMP):**  
 $(\forall A) \wedge (\forall B) \rightarrow (A+B)^\dagger = A^\dagger + B^\dagger$

**Example:**  
 The reversion of a bivector  $B = a \wedge b$  is given by  
 $B^\dagger = (a \wedge b)^\dagger = b \wedge a = -a \wedge b = -B$

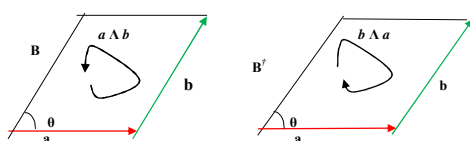


Figure 8: Content Dictionary for Reversion.

### 4.2 Rotors

In a plane  $B$  generated by two unit vectors  $n$  and  $m$ , the rotor is the geometric product of  $n$  and  $m$ :  $R = n \otimes m$ . For any rotor  $R$  and reversion of rotor  $R$ ,  $R^\dagger$ , it satisfies the normalization condition:  $RR^\dagger = R^\dagger R = 1$ . The rotation of the vector  $a$ , denoted as  $a'$ , can be written as  $a' = R \otimes a \otimes R^\dagger$ . In Figure 9 and Figure 10, an example is described for a 3D rotation and illustrated in Scalable Vector Graphics (SVG).

## 5 Conclusion

### 5.1 Summary

We have developed four Content Dictionaries with twenty-nine definitions for OpenMath relevant to basic GA terms based on the first chapter (page 1 – 19) of Hestenes and Sobczyk [16]. CDs with the fundamental symbols and operations of GA, fundamental vectors, fundamental spaces of GA, and GA products were created. In these CDs, the signature files were created, and embedded graphics and hyperlinks were also enabled. This forms a foundation toward providing the necessary mathematical semantics needed for a Physics Markup Language.

We created limited CDs to lay down the foundation for the PML and we took some basic physics examples such as rotational physics issues, and showed how we can represent the physics in the marked-up GA.

The fundamental Content Dictionaries for Geometric Algebra has been established in OpenMath format. The XML-based implementation is transformed by XSLT stylesheets into presentation markup and then displayed in the web browsers. After modifying the original OpenMath stylesheets, these content dictionaries are easily readable and understandable by providing more features such as external references and graphics.

### 5.2 Future Work

We are planning to develop more Content Dictionaries for Geometric Algebra to fully support PML and its applications. Moreover, we plan to create our own OpenMath environment with expansions towards unifying Geometric Algebra and higher-level math in OpenMath. This research will be expanded to general mathematics and computer science topics to enhance the k-16 education. Further research will be employed to continue the extension and development of Geometric Algebra, ultimately unifying mathematics and science applications.

### Acknowledgement

We would like to thank the Naval Research Laboratory’s Base Program, Program Element No. 0602435N for sponsoring this research. Also we wish to acknowledge Dr Joe Collins for his inspiration for this research.

**rotor**

Role:  
operation

Description:  
Rotor R describes the rotation directly in terms of the planes and angle.

Commented Mathematical property (CMP):  
In a plane B generated by 2 unit vectors n and m, the rotor is the geometric product of n and m.

Formal Mathematical property (FMP):  
xml prefix mathml  
$$\forall m, n. R = (n \otimes m) = ((n \cdot m) + (n \wedge m)) = (\cos(x) + (n \wedge m))$$

Commented Mathematical property (CMP):  
For any rotor R and reversion of rotor R,  $R^\dagger$ , it satisfies the normalization condition:  $R \otimes R^\dagger = R^\dagger \otimes R = 1$

Formal Mathematical property (FMP):  
xml prefix mathml  
$$\forall R. (R \otimes R^\dagger) = (R^\dagger \otimes R) = 1$$

Commented Mathematical property (CMP):  
For any rotor R and reversion of rotor R,  $R^\dagger$ , the rotation of the vector a, denoted as  $a'$  can be written as  $a' = R \otimes a \otimes R^\dagger$

Formal Mathematical property (FMP):  
xml prefix mathml  
$$\forall R. a' = (R \otimes (a \otimes R^\dagger))$$

Commented Mathematical property (CMP):  
Define a bivector in a  $\wedge$  b plane by  $B = a \wedge b$ , for any rotor R and reversion of rotor R,  $R^\dagger$ , the rotation of the bivector B, denotes as  $B'$ , can be written as  $B' = R \otimes B \otimes R^\dagger$

Formal Mathematical property (FMP):  
xml prefix mathml  
$$\forall R. B' = (R \otimes (B \otimes R^\dagger))$$

Example:  
A rotation in 3D Example: The vector a is rotated to  $a' = R \otimes a \otimes R^\dagger$   
Scalable Vector Graphics (SVG):

Signatures:  
[sts](#)

Figure 9: Content Dictionary for Rotor (ga\_rotor1.xhtml)

## References

- [1] Buswell, S., Caprotti, O., Carlisle, D., Dewar, M., Gaëtano, M. and Kohlhase, M. (2004) "OpenMath v2.0.-OpenMath Standard": <http://www.openmath.org/standard/om20-2004-06-30/omstd20.pdf>.
- [2] Bruhn, R. and Burton, P. (2004) "Displaying mathematics in a web browser using MathML and SVG", *Proceedings of Mid-South College Computing Conference*, pp. 97 – 106.
- [3] Caprotti, O. and Carlisle, D. (1999) "OpenMath and MathML: Semantic Mark Up for Mathematics", *ACM Crossroads*, vol. 6(2), pp. 11-14.
- [4] Carlisle, D. (2000) "OpenMath, MathML and XSL", *ACM Special Interest Group on Symbolic and Algebraic Manipulation (SIGSAM)*, vol. 34 (2), pp. 6 – 11.
- [5] Chytracek, R., McCormick, J., Pokorski, W., and Santin, G. (2006) "Geometry Description Markup Languages for Physics Simulation and Analysis Applications", *IEEE Trans. On Nuclear Physics*, vol.53 (5), pp. 2892-2896.
- [6] Collins, J. (2008) "Mathematical Type for Physical Variables", *Lecture Notes in Artificial Intelligence vol. 5144: Intelligent Computer Mathematics*, pp. 370-381, Springer-Verlag, Berlin.
- [7] Collins, J. (2008) "Thematical and Scientific Markup as an Approach to Model Specification", *Proceedings of the Society for Modeling and Simulation International [SCS]: Grand Challenges in Modeling & Simulation (GCMS 08)*, Edinburgh, Scotland.
- [8] Davenport, J. (2000) "Interactive A Small OpenMath Type System", *ACM Special Interest*

- Group on Symbolic and Algebraic Manipulation (SIGSAM)*, vol. 34 (2), pp. 16 – 21.
- [9] Davenport, J. (2000) “On Writing OpenMath Content Dictionaries”, *ACM Special Interest Group on Symbolic and Algebraic Manipulation (SIGSAM)*, vol. 34 (2), pp. 12 – 15.
- [10] Dewar, M. (2000) “OpenMath: An Overview”, *ACM Special Interest Group on Symbolic and Algebraic Manipulation (SIGSAM)*, vol. 34 (2), pp. 2 – 5.
- [11] Dorst, I., Doran, C. and Lasenby, J. (2002). *Applications of Geometric Algebra in Computer Science and Engineering*. Springer-Birhauser, Boston MA.
- [12] Dorst, L., Fontijne, D., Mann, S. (2007) *Geometric Algebra for Computer Science: An Object-Oriented Approach to Geometry*, Morgan Kaufmann.
- [13] Evjen, B., Sharkey, K., Thangarathinam, T., Vernet, A., and Ferguson, S. (2007) *Professional XML*, Wiley Indianapolis, IN.
- [14] Finney, A., Hucka, M., Bornstein, B.J., Keating, S.M., Shapiro, B.E., Matthews, J., Kovitz, B.L., Schilstra, M.J., Funahashi, A., Doyle, J.C., and Kitano, H. (2006). “Software Infrastructure for Effective Communication and Reuse of Computational Models.” In *Systems Modeling in Cell Biology: From Concepts to Nuts and Bolts*. MIT Press, pp. 369–378.
- [15] Hestenes, D. (2003) “Reforming the Mathematical Language of Physics”, *American Journal of Physics*, vol. 71, pp. 104-106.
- [16] Hestenes, D. and Sobczyk, G. (1987) *Clifford Algebra to Geometric Calculus A Unified Language for Mathematics and Physics*. Kluwer Academic Publishers, Dordrecht.
- [17] Hestenes, D. (1999) *New Foundations for Classical Mechanics*, 2<sup>nd</sup> edition, Kluwer Academic Publishers.
- [18] Hilf, E., Kohlhase, M., and Stamerjohanns, H. (2006) “Capturing the Content of Physics Systems, Observables, and Experiments” *Mathematical Knowledge Management 2006*, LNAI 4108, pp. 165–178, Springer-Verlag Berlin Heidelberg.
- [19] Hucka, M., Finney, A., Sauro, H.M., Bolouri, H., Doyle, J.C., Kitano, H. (2003) “The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models”, *Bioinformatics*, vol. 19 (4), pp. 524–31.
- [20] Kohlhase, M. (2000) “OMDoc: An Infrastructure for OPENMATH Content Dictionary Information”, *ACM Special Interest Group on Symbolic and Algebraic Manipulation (SIGSAM)*, vol. 34 (2), pp. 43 – 48.
- [21] Kohlhase, M. (2006) *An Open Markup Format for Mathematical Documents OMDoc 1.2.*, LNAI 4180, Springer-Verlag Berlin Heidelberg..
- [22] Kohlhase, M. (2008). “Semantic Knowledge Management for Education.”, *Proceedings of the IEEE*, vol. 96 (6), pp. 970 - 989.

```

<CD xmlns=
  "http://www.openmath.org/OpenMathCD">
<CDName>rotor</CDName>
<CDDefinition>
<Name>rotor</Name>
<CMP>
  In a plane B generated by 2 unit
  vectors n and m, the rotor is the
  geometric product of n and m.
</CMP>
<FMP><OMOBJ xmlns=
  "http://www.openmath.org/OpenMath"
  version="2.0"
  cdbase="http://www.openmath.org/cd">
<OMBIND>
<OMS cd="quant1" name="forall"/>
  <OMBVAR>
    <OMV name="m"/><OMV name="n"/>
  </OMBVAR>
  <OMA>
    <OMS cd="relation1" name="eq"/>
    <OMV name="R"/>
    <OMA>
      <OMS cd="relation1" name="eq"/>
      <OMA>
        <OMS cd="ga_product1"
          name="geometric_product"/>
        <OMV name="n"/><OMV name="m"/>
      </OMA>
      <OMA>
        <OMS cd="relation1" name="eq"/>
        <OMA><OMS cd="ga_rotor1"
          name="plus"/>
        <OMA><OMS cd="ga_product1"
          name="inner_product"/>
        <OMV name="n"/>
        <OMV name="m"/>
      </OMA>
        <OMA><OMS cd="ga_product1"
          name="outer_product"/>
        <OMV name="n"/>
        <OMV name="m"/>
      </OMA>
      </OMA>
    </OMA><OMS cd="ga_rotor1"
      name="plus"/>
    <OMA><OMS cd="transcl"
      name="cos"/>
    <OMV name="x"/>
  </OMA>
  <OMA><OMS cd="ga_product1"
    name="outer_product"/>
  <OMV name="n"/>
  <OMV name="m"/>
  </OMA>
</OMA>
</OMBIND>
</OMOBJ>
</FMP>

```

Figure 10: Content Dictionary for Rotor in XML.

- [23] Lasenby, J., Lasenby, A., and Doran, C. (2000) “A unified mathematical language for physics and engineering in the 21st century” *Phil. Trans. R. Soc. Lond. A*, vol. 358, pp. 21-39.
- [24] Lloyd, I. (2008) *The Ultimate HTML Reference*, Sitepoint, Melbourne Australia.
- [25] MathML, (2009) *Mathematical Markup Language*, DOI= <http://www.w3.org/Math>.
- [26] Miner, R. (2005) “The Importance of MathML to Mathematics Communication”. *Notices of the AMS*, vol. 52(5), pp. 532-538.
- [27] Murray-Rust, P., and Rzepa, H. (1999) “Chemical Markup, XML, and the Worldwide Web. 1. Basic Principles”, *J. Chem. Inf. Comput. Sci.*, vol. 39 (6), pp. 928–942,
- [28] Murray-Rust, P., Rzepa, H., and Wright, M. (2001) “Development of Chemical Markup Language (CML) as a System for Handling Complex Chemical Content”, *New J. Chem.*, pp. 618–634
- [29] OpenMath, (2009) *OpenMath Content Dictionaries*, DOI=<http://www.openmath.org>.
- [30] Overview of OpenMath [Online]. Available:<http://www.openmath.org/overview/index.html>
- [31] Patyk-Lonska, A., Czachor, M. and Aerts, D. (2011) “Distributed Representations based on Geometric Algebra: The Continuous Model”, *Informatica*, vol. 35, pp. 407-417.
- [32] Sandhu P. (2003) *The MathML Handbook*: Charles River Media, Boston MA.
- [33] Scalable Vector Graphics (2011), <http://www.w3.org/Graphics/SVG>.
- [34] Shrestha, R., Watts, M. Zhang, W. and Yang, K. (2009). “Representing Clifford Algebra Into OMDoc”, *Proc. of CCSC-SC 2009*, pp. 262-268.
- [35] Suter, J. (2003) “Geometric Algebra Primer”, [[http://www.jaapsuter.com/paper/ga\\_primer.pdf](http://www.jaapsuter.com/paper/ga_primer.pdf)]
- [36] W3C Working Draft 24 (2009) “Mathematical Markup Language (MathML) Version 3.0”, [Online].: <http://www.w3.org/TR/MathML3/>
- [37] Yang, K.P., Zhang, W., and Petry, F., (2009) “Building Content Dictionaries for Geometric Algebra in OMDoc Format”, *Proceedings of the 47th ACM Southeast Conference*, Clemson, South Carolina, Article No. 45.
- [38] Yang, K.P., Petry, F., and Collins, J. (2010) “Building Content Dictionaries for Geometric Algebra in OpenMath Format,” *Jour of Computing Sciences in Colleges*, vol. 25(4), pp. 22-29.

```

<a xmlns="http://www.w3.org/1999/xhtml"
href="http://en.wikipedia.org/wiki/Geometric_algebra">
  Geometric Algebra
</a>
is the generalization of the scalar product, defined in
CD linalg1, for arbitrary multivectors. The
&#x2219;(dot) symbol is used to denote this operator.
The inner product is a grade lowering operation.
</Description>
<CMP>
  For any homogeneous multivectors Ar of grade r and
  Bs of grade s, the inner product will lower the grade to
  (r-s) if r>0, s>0, and (r-s) > 0.
</CMP>
<FMP>
<OMOBJ
  xmlns="http://www.openmath.org/OpenMath"
  version="2.0" cdbase="http://www.openmath.org/cd">
<OMA>
  <OMS cd="relation1" name="eq"/>
<OMA>
  <OMS cd="ga_product1" name="inner_product"/>
<OMA>
  <OMS cd="linalg1" name="vector_selector"/>
  <OMV name="r"/>
  <OMV name="A"/>
</OMA>
<OMA>
  <OMS cd="linalg1" name="vector_selector"/>
  <OMV name="s"/>
  <OMV name="B"/>
</OMA>
</OMA>
<OMA>
  <OMS cd="linalg1" name="vector_selector"/>
<OMA>
  <OMS cd="arith1" name="minus"/>
  <OMV name="r"/>
  <OMV name="s"/>
</OMA>
<OMA>
  <OMS cd="ga_product1"
  name="geometric_product"/>
<OMA>
  <OMS cd="linalg1" name="vector_selector"/>
  <OMV name="r"/>
  <OMV name="A"/>
</OMA>
<OMA>
  <OMS cd="linalg1" name="vector_selector"/>
  <OMV name="s"/>
  <OMV name="B"/>
</OMA>
</OMOBJ>
</FMP>

```

## Appendix

### CD Definition of Inner\_Product

```

<!-- Inner Product -->
<CDDefinition>
<Name>inner_product</Name>
<Role>application</Role>
<Description>
  Inner product specific to

```

<Example>

The inner product of two vectors  $a$  and  $b$ , denoted by  $a \cdot b$ , projects  $a$  onto  $b$  resulting in the scalar magnitude of the projection relative to  $b$ 's magnitude.

```
<svg xmlns="http://www.w3.org/2000/svg"
      xmlns:xlink="http://www.w3.org/1999/xlink">
  <image width="100%" height="100%"
        xlink:href="img/inner_product.png"/>
</svg>
```

</Example>

</CDDefinition>



# Local Graph Embedding Based on Maximum Margin Criterion (LGE/MMC) for Face Recognition

Minghua Wan and Shan Gai  
 School of Information Engineering, Nanchang Hangkong University  
 Nanchang 330063, China  
 E-mail: wmh36@sina.com, gaishan@yahoo.com

Jie Shao  
 School of computer Science, Shangqiu Institute of Technology  
 Shangqiu, 476000, China  
 E-mail: sj012328@163.com

**Keywords:** locally linear embedding, dimensional reduction, face recognition, maximum margin criterion, local graph embedding

**Received:** November 14, 2010

*Locally linear embedding (LLE) is an efficient dimensional reduction algorithm for nonlinear data, and the low dimensional data can maintain topological relations in the original space after the processing. But this algorithm main application is not very good in the data dimensional reduction, the visualization and learning effects of data classification question and so on. In ordered to solve the above question, this paper proposes an efficient dimensional reduction and data classification method--local graph embedding method based on maximum margin criterion (LGE/MMC) for dimensional reduction, which is applied in face recognition. This goal of algorithm is preserved under nearest neighbour premise, where MMC criterion is used to construct the intrinsic graph and the penalty graph. In the intrinsic graph, the nonlinear structure is discovered in the high dimensional data space by the locally symmetric of linear restructuring, which is caused the similar sample as far as possible to gather in together. At the same time, the different class sample is far away as far as possible in the penalty graph. LGE/MMC seeks to minimize the difference, rather than the ratio, between the locality preserving between-class scatter and locality preserving within-class scatter. The results of face recognition experiments on ORL, YALE and AR face databases demonstrate the effectivity of the proposed method.*

*Povzetek: Članek opisuje algoritem za zmanjšanje števila dimenzij podatkov, ki se uporablja pri prepoznavanju obrazov.*

## 1 Introduction

Face recognition has been active areas of research because of their potential applications in human-computer interfaces, image and computer vision. Linear dimensionality reduction seeks to find a meaningful low dimensional subspace in a high-dimensional input space. The subspace can provide a compact representation of the input data when the structure of data embedded is linear in the input space. Principal components analysis (PCA) [1] maintains the global Euclidean structure of the data in the high-dimensional space and preserves the total variance by maximizing the trace of the feature covariance matrix. Linear discriminant analysis (LDA) [2] preserves discriminative information between data of different classes and finds the optimal set of projection vectors by maximizing the ratio between the interclass and intraclass scatters.

PCA, LDA, and their variants [5, 6] are not able to reveal the underlying non-linear [3, 4] structure of the face data. Recently, many manifold learning-based algorithms with locality preserving abilities have been presented. Among them, isometric feature mapping

(ISOMAP) [7], locally linear embedding (LLE) [8, 9], Laplacian eigenmap (LE) [10, 11] and local tangent space alignment (LTSA) [12] are widely used. He et al. [13, 14] proposed locality preserving projections (LPP), which is a linear subspace learning method derived from Laplacian Eigenmap. LPP can find an embedding space that preserves local information, and it is an unsupervised method. Many modified LPP algorithms have been put forward to consider the discriminant information of recognition task in recent years [15-18].

LLE is another representative local linear manifold learning method. Based on the assumption of the local linearity, LLE first constitutes local coordinates with the least constructed cost and then maps them to a global one. Some supervised versions of LLE [19-22] are introduced to deal with data sets labelled with class information and some other supervised LLE algorithms combined with LDA are becoming popular. Zhang et al. presented a unified framework of LLE and LDA [23,24]. Recently, He et al. [25] proposed another linear dimensionality reduction technique neighbourhood

preserving embedding(NPE), which is the linearization of the locally linear embedding(LLE) algorithm and aims at finding a low-dimensional embedding that optimally preserves the local neighbourhood reconstruction relationships on the original data manifold. Some extension methods of NPE [26,27] are introduced to feature extraction. Experiments have proven that LLE and NPE are effective method for visualization.

Some other discriminant manifold learning algorithms, local discriminant embedding (LDE) [28], marginal Fisher analysis (MFA) [29] and neighbourhood preserving discriminant embedding (NPDE) [30] are proposed where their combine the Fisher criterion [31] with manifold criterion. They can be unified under the Fisher graph framework. However, they are different on their objective functions in terms of different graph embedding types and derivations. LDE utilizes LPP to form the intra-class and inter-class graph pair; MFA models the intra-class graph to characterize the intra-class compactness and the inter-class graph to characterize the inter-class separability with binary graph coefficient; and NPDE utilizes NPE to form the intra-class and inter-class graph pair to model the within- and between-neighbourhood scatters.

Above manifold learning algorithms can all be interpreted as the implementations of the linear graph embedding framework (LGE) [32] with different weight matrices or some variations. However, some limitations are exposed when LGE is applied to pattern recognition. One limitation is that some LGE such as LPP, LLE and NPE neglect the class information, which will impair the recognition accuracy. Another limitation lies in that some LGE such as LDE, MFA and DLPP involve inverse matrix of discriminant criterion, which will impair the recognition accuracy. So, in this paper we present local graph embedding method based on maximum margin criterion [33] (LGE/MMC) for dimensional reduction. Therefore, much computational time would be saved for feature extraction which is not necessary to convert the image matrix into high-dimensional image vector and avoids inverse matrix.

The rest of this paper is organized as follows: We review the ideas of linear methods in section 2. In Section 3, we propose the idea of LGE/MMC algorithm in detail. In section 4, we introduce the connections between LLE, NPE and LGE/MMC. Experiments are presented to demonstrate the effectiveness of LGE/MMC on face recognition in section 5. Finally, we give concluding remarks and a discussion of future work in Section 6.

## 2 Outline of Linear Methods

Let us consider a set of  $N$  sample  $X = \{x_1, x_2, \dots, x_N\}$ ,  $x_i \in R^D$  taking values in an  $n$ -dimensional image space. Let us also consider a linear transformation mapping the original  $n$ -dimensional space into a  $d$ -dimensional feature space  $Y = \{y_1, y_2, \dots, y_N\}$ , where  $y_i \in R^d$

and  $n > d$ . The new feature vectors  $y_i \in R^d$  are defined by the following linear transformation:

$$y_i = U^T x_i, \quad i = 1, \dots, N \quad (1)$$

where  $U \in R^{n \times d}$  is a transformation matrix. In this section, we briefly review how the LDA, LPP and UDP algorithms realize subspace learning.

### 2.1 Linear discriminant analysis (LDA)

LDA [2] is a supervised learning algorithm. Let  $c$  denote the total class number and  $c_i$  denote the number of training samples in the  $i$ -th class. Let  $x_i^j$ , denote the  $j$ -th sample in  $i$ -th class,  $\bar{x}$  be the mean of all the training samples,  $\bar{x}_i$  be the mean of the  $i$ -th class. The between-class and within-class scatter matrices can be evaluated by:

$$S_b = \sum_{i=1}^c l_i (\bar{x}_i - \bar{x})(\bar{x}_i - \bar{x})^T \quad (2)$$

$$S_w = \sum_{i=1}^c \sum_{j=1}^{c_i} (x_i^j - \bar{x}_i)(x_i^j - \bar{x}_i)^T \quad (3)$$

LDA aims to find an optimal projection  $U$  such that the ratios of the between-class scatter to within-class scatter is maximized, i.e.

$$U = \arg \max_U \frac{|U^T S_b U|}{|U^T S_w U|} \quad (4)$$

where  $\{U_i | i = 1, 2, \dots, d\}$  is the set of generalized eigenvectors of  $S_b$  and  $S_w$  corresponding to the  $d$  largest generalized eigenvalues  $\{\lambda_i | i = 1, 2, \dots, d\}$ , i.e.

$$S_b U_i = \lambda_i S_w U_i, i = 1, 2, \dots, d. \quad (5)$$

### 2.2 Linear preserving projection (LPP)

The similarity matrix  $S$  of LPP [13,14] can be Gaussian weight or uniform weight of Euclidean distance using  $k$ -neighbourhood or  $\varepsilon$ -neighbourhood, defined as

$$S_{ij} = \begin{cases} 1, & \|x_i - x_j\|^2 < \varepsilon \\ 0, & \text{otherwise} \end{cases} \quad (6)$$

Hence, the objective function of LPP is defined as :

$$\min \sum_{i,j} \|y_i - y_j\| S_{ij} \quad (7)$$

where  $\|\bullet\|$  means the  $L_2$  norm. After some matrix analysis steps, the minimization problem becomes

$$\begin{aligned} & \arg \min_U U^T X L X^T U \\ & \text{s.t. } U^T X D X^T U = 1 \end{aligned} \quad (8)$$

where  $X = [X_1, X_2, \dots, X_N]$  is the training space of size  $n \times N$ , and  $D$  is a diagonal matrix whose entries

are column or row sums of  $S$ .  $L = D - S$  is the Laplacian matrix.

The optimal  $d$  projection vectors that minimizes the objective function can be computed by the minimum eigenvalues solutions to the generalized eigenvalues problem

$$XLX^T U_i = \lambda_i XDX^T U_i \quad (9)$$

### 2.3 Maximum margin criterion (MMC)

The MMC is based on the difference of between-class scatter matrix and within-class scatter matrix, which is defined as follows:

$$J_s(w) = \text{tr}(U^T (S_b - \alpha S_w) U) \quad (10)$$

where the parameter  $\alpha$  is a nonnegative constant which balances the relative merits of maximizing the between-class scatter to the minimization of the within-class scatter. The between-class scatter matrix  $S_b$  and within-class scatter matrix  $S_w$  can be denoted as

$$S_b = \frac{1}{n} \sum_{i=1}^c n_i (\mathbf{f}_i - \mathbf{f}_0)(\mathbf{f}_i - \mathbf{f}_0)^T \quad (11)$$

$$S_w = \frac{1}{n} \sum_{i=1}^c \sum_{j=1}^{n_i} (\mathbf{x}_j^i - \mathbf{f}_i)(\mathbf{x}_j^i - \mathbf{f}_i)^T \quad (12)$$

where  $n_i$  is the number of training samples in class  $i$ . In class  $i$ , the  $j^{\text{th}}$  training sample is denoted by  $\mathbf{x}_j^i$ , the mean vectors of training samples in class  $i$  is denoted by  $\mathbf{f}_i$  and the mean vector of all training samples is  $\mathbf{f}_0$ . Let  $S_t = S_b + S_w$  and  $S_t$  denotes the total scatter matrix. As we know,  $S_b$ ,  $S_w$  and  $S_t$  are all positive semi-definite.

## 3 Local Graph Embedding Based on Maximum Margin Criterion

### 3.1 The idea of LGE/MMC

When only a small number of training samples is available the within-class scatter matrix used by many feature extraction techniques (LDA, LPP etc) is singular, which represents a major obstacle for most techniques as they require an inversion of this singular matrix. Motivated by the idea of MMC, LGE/MMC seeks to minimize the difference, rather than the ratio, between the locality preserving between-class scatter and locality preserving within-class scatter. Then the singularity is avoided. LGE/MMC is theoretically elegant and can derive its discriminant vectors from both the range of the locality preserving between-class scatter and the range space of locality preserving within-class scatter. To gain more discriminative power, it is desirable to minimize the locality preserving between-class scatter and maximize the locality preserving within-class scatter simultaneously.

### 3.2 Locality preserving within-class scatter

To begin with, we propose to minimize the local scatter compactness of each data point by linear coefficients that reconstruct the data point from other points. The technique of local representation is the same as LLE [8, 9]. LLE regards each data point and its nearest neighbors as the locality. The algorithm can be described in three steps.

The first step of LLE is to select  $K_c$ -nearest neighbors of each data points  $x_i$  using Euclidean distances.

The second step of LLE is to calculate the reconstructing weight matrix  $W = [w_{ij}]_{N \times N}$ , which reconstructs each point  $x_i$  from its  $K_c$ -nearest neighbours. We can obtain the coefficient matrix  $W$  by minimizing the reconstruction error:

$$\min J_L(W) = \sum_{i=1}^N \left\| x_i - \sum_{j=1}^{K_c} w_{ij} x_j \right\|^2 \quad (13)$$

where  $w_{ij} = 0$  if  $x_i$  and  $x_j$  are not neighbors, and the rows of  $W$  sum to 1:  $\sum_{j=1}^{K_c} w_{ij} = 1$ .

The reconstruction error can be converted to this form:

$$\begin{aligned} \xi &= \left\| x_i - \sum_{j=1}^N w_{ij} x_j \right\|^2 = \left\| \sum_{j=1}^N w_{ij} (x_i - x_j) \right\|^2 \\ &= \sum_{j=1}^N w_{ij} (x_i - x_j) \sum_{t=1}^N w_{it} (x_i - x_t) = \sum_{j=1}^N \sum_{t=1}^N w_{ij} w_{it} G_{jt}^i \end{aligned} \quad (14)$$

where  $G_{jt}^i = (x_i - x_j)^T (x_i - x_t)$ , called the local Gram matrix. By solving the least-squares problem with the constraint  $\sum_{j=1}^{K_c} w_{ij} = 1$ , the optimal coefficients are given:

$$w_{ij} = \frac{\sum_{t=1}^{K_c} G_{jt}^{-1}}{\sum_{p=1}^{K_c} \sum_{q=1}^{K_c} G_{pq}^{-1}} \quad (15)$$

After repeating the first step and the second step are performed on all the  $N$  data points, we can calculate the reconstruction weights to construct a weight matrix  $W = [w_{ij}]_{N \times N}$ .

The third step of LLE is to reconstruct represented  $y_i$  by the weight matrix  $W$ . To maintain the intrinsic geometrical feature of the data after the embedding process, the reconstruction error function must be minimized:

$$\min J_L(Y) = \sum_{i=1}^N \left\| y_i - \sum_{j=1}^N w_{ij} y_j \right\|^2 \quad (16)$$

where  $y_i$  is the output of  $x_i$ ,  $y_j$  is a neighbor of  $y_i$ .

Considering the map in Eq. (1), the objective function reduces to

$$\begin{aligned} J_L(U) &= \sum_{i=1}^N \left\| y_i - \sum_{j=1}^{K_c} w_{ij}^j y_j \right\|^2 \\ &= \sum_{i=1}^N \text{tr} \left\{ \left( y_i - \sum_{j=1}^{K_c} w_{ij}^j y_j \right) \left( y_i - \sum_{j=1}^{K_c} w_{ij}^j y_j \right)^T \right\} \\ &= \text{tr} \left\{ \sum_{i=1}^N \left( y_i - \sum_{j=1}^{K_c} w_{ij}^j y_j \right) \left( y_i - \sum_{j=1}^{K_c} w_{ij}^j y_j \right)^T \right\} \quad (17) \\ &= \text{tr} \left\{ Y(I-W^T)(I-W^T)^T Y^T \right\} \\ &= \text{tr} \left\{ Y(I-W)^T (I-W) Y^T \right\} \\ &= \text{tr} \left\{ U^T X M X^T U \right\} \end{aligned}$$

where  $M = (I - W)^T (I - W)$ .

### 3.3 Locality preserving between-class scatters

To begin with, for the first aspect of our consideration, we propose to maximize the sum of pair wise squared distances between outputs if they have different labels. So, maximize locality preserving between-class scatter of samples is considered :

$$\max J_G(Y) = \sum_{i=1}^N \sum_{j=1}^N \|y_i - y_j\|^2 \quad (18)$$

Considering the map Eq.(1), the objective function reduces to

$$\begin{aligned} J_G(Y) &= \sum_i \sum_j \|y_i - y_j\|^2 W_{ij}^p \\ &= \sum_i \sum_j \|U^T x_i - U^T x_j\|^2 W_{ij}^p \\ &= 2U^T X(D^p - W^p)X^T U \\ &= 2U^T X L^p X^T U \quad (19) \end{aligned}$$

The variance of between-class points is deemed as local information. We construct the similarity matrix  $W_{ij}^p$  as follows:

$$W_{ij}^p = \begin{cases} 1, & \text{if } x_i \text{ is in the } K_p \text{ nearest} \\ & \text{from different classes of } x_j \\ 0, & \text{otherwise} \end{cases} \quad (20)$$

### 3.4 Criterion of LGE/MMC

At last, when the locality preserving between-class scatter and the locality preserving within-class scatter have been constructed, an intuitive motivation is to find a common projection that minimizes between-class scatter  $J_L(W)$  and maximizes within-class scatter  $J_G(U)$  at the same time. Actually, we can obtain such a projection by the following multi-object optimized problem, that is:

$$\begin{cases} \min \text{tr} \{U^T X M X^T U\} \\ \max \text{tr} \{U^T X L^p X^T U\} \end{cases} \quad (21)$$

$$\text{s.t. } U^T X X^T U = I$$

The solution to the constrained multi-object optimized problem is to find a subspace which minimize the locality preserving between-class scatter and maximize the locality preserving within-class scatter simultaneously. Motivated by the idea of MMC, LGE/MMC seeks to minimize the difference, rather than the ratio, between the locality preserving between-class scatter and the locality preserving within-class scatter. So it can be changed into the following constrained problem:

$$\begin{aligned} \min \text{tr} \{U^T X (M - \mu L^p) X^T U\} \\ \text{s.t. } U^T X X^T U = I \end{aligned} \quad (22)$$

where  $\mu$  is an adjustable parameter to balance between-class scatter and within-class scatter.

Eq. (22) can be solved by Lagrange multiplier method:

$$\begin{aligned} L(U, \lambda) &= \{U^T X (M - \mu L^p) X^T U - \lambda (U^T X X^T U - I)\} \\ &= 0 \end{aligned} \quad (23)$$

where  $\lambda$  is the Lagrange multiplier. Thus we get:

$$X (M - \mu L^p) X^T U = \lambda X X^T U \quad (24)$$

where  $U_i$  is generalized eigenvector correspondingly to generalized eigenvalue  $\lambda_i$ .

## 4 Connection between LLE, NPE and LGE/MMC

In this Section, LGE/MMC seems to be formally similar to LLE and NPE. However, LGE/MMC is also obviously different from them. In order to investigate the similarity and the difference, we discuss the connections between LLE, NPE and LGE/MMC.

### 4.1 Connection between LLE and NPE

LLE and NPE aim to discover the local structure of the data manifold. LLE is defined only on the training samples, and there are no natural maps of the testing sample. Instead, NPE is defined on both the training and test samples. NPE is a linear approximation to LLE.

In NPE, the matrix  $XX^T$  is symmetric and semi-positive definite. In order to remove an arbitrary scaling factor in the projection, we impose a constraint as follows:

$$YY^T = I \Rightarrow U^T XX^T U = I \quad (25)$$

Finally, the minimization problem reduces to finding  $U$ :

$$\min_{U^T XX^T U = I} \text{tr} \{ U^T X M X^T U \} \quad (26)$$

The transformation matrix  $U$  that minimizes the objective function is given by the minimum eigenvalue solution to the following generalized eigenvector problem:

$$X M X^T U_i = \lambda_i X X^T U_i \quad (27)$$

#### 4.1 Connection between NPE and LGE/MMC

As a result, LGE/MMC is formulated as the following constrained minimization problem:

$$\min_{U^T XX^T U = I} \text{tr} \{ U^T X (M - \mu L^p) X^T U \} \quad (28)$$

Thus we have:

$$X \hat{M} X^T U_i = \lambda_i X X^T U_i \quad (29)$$

where  $\hat{M} = M - \tilde{M}$ ,  $\tilde{M} = \mu L^p$ . It is easy to see that NPE is a special case of LGE/MMC (i.e. when  $\mu = 0$ ).

#### 4.2 Connection between LLE, NPE and LGE/MMC

From above discussed, NPE and LGE/MMC yield mappings that are defined not only on the training data points but also on novel testing points. The essence of NPE is the linear approximation to LLE. As we know, the graph construction of LLE and NPE fails to use the global discriminative information. However, we can see from  $\hat{M}$  first that LGE/MMC preserves the locality characteristic since  $\tilde{M}$  still exists and second that it adds the discriminant information through  $\tilde{M}$ . From what has been discussed above, it can be concluded that LGE/MMC builds a new graph with different edge weight assignment method, integrating both local information and discriminant information. Thus, by integrating the discriminant into the objective function, LGE/MMC will be more robust than LLE and NPE.

### 5 Comparisons of Computation complexity and space complexity

In Table 1, we compare the computational and the memory space complexities of the six methods. Here  $m$  and  $n$  is the number of the rows and the columns of the image matrix.  $L$ ,  $M$  and  $N$  are the number of the projection vectors, the testing and the training samples, respectively.

Table 1: The computational and the memory space complexities of the six methods.

| Method Complexity |                         |                |               |
|-------------------|-------------------------|----------------|---------------|
|                   | Time (training)         | Time (testing) | Memory        |
| PCA               | $O(m^2 n^2 L)$          | $O(MNL)$       | $O(m^2 n^2)$  |
| LDA               | $O(m^2 n^2 L)$          | $O(MNL)$       | $O(m^2 n^2)$  |
| MMC               | $O(m^2 n^2 L)$          | $O(MNL)$       | $O(m^2 n^2)$  |
| LLE               | $O(m^2 n^2 L + mnN^2)$  | $O(MNL)$       | $O(m^2 n^2)$  |
| LLE+LDA           | $O(2m^2 n^2 L + mnN^2)$ | $O(2MNL)$      | $O(2m^2 n^2)$ |
| LGE/MMC           | $O(m^2 n^2 L + 2mnN^2)$ | $O(MNL)$       | $O(m^2 n^2)$  |

In Table 1, for the PCA, LDA and MMC, since we need to perform  $O(MN)$  tests when using the nearest neighbour rule for classification and for each test it has the time complexity of  $O(L)$ , the testing time is  $O(MNL)$ . The memory cost is determined by the size of the matrices of the associated eigen equations, which is  $O(m^2 n^2)$ . The training time complexity depends on both the size of the matrices in the eigen equations and the number of the projection vectors that are required to be computed, which is  $O(m^2 n^2 L)$ . For the LLE method, an extra time cost to construct the similarity matrix, i.e.,  $O(mnN^2)$ , will be taken into account. So, LLE+LDA has the time complexity of  $O(2m^2 n^2 L + mnN^2)$ , the testing time is  $O(2MNL)$  and Memory is  $O(2m^2 n^2)$ . The proposed method LGE/MMC has the time complexity of  $O(m^2 n^2 L + 2mnN^2)$ , the testing time is  $O(MNL)$  and Memory is  $O(m^2 n^2)$ . So the proposed method is much than other methods in testing time.

### 6 Experiments and results

To evaluate the proposed LGE/MMC algorithm, we systematically compare it with the PCA [1], LDA [2], LLE [8-9], MMC [33] and LLE+LDA [23-24] algorithm in three face databases: ORL, YALE and AR. When the projection matrix was computed from the training part, all the images including the training part and the test part were projected to feature space. Euclidean distance and nearest neighborhood classifier are used in all the experiments. The experiments were carried out on the same PC (CPU: P4 2.8 GHz, RAM: 1024 MB).

#### 6.1 Database

The ORL face database [34] contains images from 40 individuals, each providing 10 different images where the pose, face expression and sample size vary. The facial expressions and facial details (glasses or no glasses) also vary. The images were taken with a

tolerance for some tilting and rotation of the face of up to 20 degrees. Moreover, there is also some variation in this scale of up to about 10 percent. All images normalized to a resolution of 56×46. We test the recognition performances of the six methods: PCA, LDA, LLE, MMC, LLE+LDA and LGE/MMC. In the experiments,  $l$  images ( $l$  varies from 2 to 6) are randomly selected from the image gallery of each individual to form the training sample set. The remaining  $10-l$  images are used for testing. For each  $l$ , we independently run 50 times. In the PCA phase of LDA, LLE, MMC, LLE+LDA and LGE/MMC, we keep 95 percent image energy.

The YALE face database [35] contains 165 gray scale images of 15 individuals, each individual has 11 images. The images demonstrate variations in lighting condition, facial expression (normal, happy, sad, sleepy, surprised, and wink). In this experiment, each image in Yale database was manually cropped and resized to 50×40. In the PCA phase of LDA, LLE, MMC, LLE+LDA and LGE/MMC, we keep 95 percent image energy. In the experiments,  $l$  images ( $l$  varies from 2 to 6) are randomly selected from the image gallery of each individual to form the training sample set. The remaining  $11-l$  images are used for testing. For each  $l$ , we independently run 50 times.

The AR face database [36] contains over 4,000 color face images of 126 people (70 men and 56 women), including frontal views of faces with different facial expressions, lighting conditions, and occlusions. The pictures of 120 individuals (65 men and 55 women) were taken in two sessions (separated by two weeks) and each section contains 13 colour images. The face portion of each image is manually cropped and then normalized to 50×40 pixels. These images vary as follows: 1. neutral expression 2. smiling 3. angry 4. screaming 5. left light on 6. right light on 7. all sides light on 8. wearing sun glasses 9. wearing sun glasses and left light on 10. wearing sun glasses and right light on. In this experiment,  $l$  images ( $l$  varies from 2 to 6) are randomly selected from the image gallery of each individual to form the training sample set. The remaining  $20-l$  images are used for testing. For each  $l$ , we independently run 10 times. In the PCA phase of LDA, LLE, MMC, LLE+LDA and LGE/MMC, the number of principle components is set as 150. The dimension steps are set to be 5 in final low-dimensional subspaces obtained by the seven methods.

Fig.1, Fig.2 and Fig.3 show the sample images from the three databases.



Fig.1 Images of one person on the ORL database



Figure 2: Images of one person on the YALE database.



Figure 3: Images of one subject of the AR database. The first line and the second line images were taken in different time (separated by two weeks).

### 6.2 Experimental results and analysis

Except PCA and LDA, the local methods involved in the experiments are manifold learning based approaches, where  $K_c$  and  $K_p$  nearest neighbourhood search are contained. Thus how to select  $K_c$  and  $K_p$  are an important problem in feature extraction. If the value of  $K_c$  and  $K_p$  are too small, it is very difficult to preserve the topologic structure in low-dimensional space. On the contrary, if the value of  $K_c$  and  $K_p$  are too big, it is very difficult to depict the assumption of local linearity in high dimensional space. So it will affect the dimensionality manifold reduction result by the value of  $K_c$  and  $K_p$ . In the first experiment, we investigate the performance of the LGE/MMC algorithm over the reduced dimensions versus the corresponding varied the value of  $K_c$  and  $K_p$ . To find how  $K_c$  and  $K_p$  affect the recognition performance, we change  $k_c = l - 1$  and  $K_p$  are from 1 to 20 with step 1. Fig.4 displays the average recognition rates with varied the value of  $K_p$  by carrying out LGE/MMC when only two images per class were randomly selected for training on the YALE face databases. LGE/MMC obtains the best average recognition rate is 94.79% when  $K_p=4$ . This indicates that the locality and the globality are with the same importance. In the next experiment, the value of adjustable parameter  $K_p$  is taken to be 4.

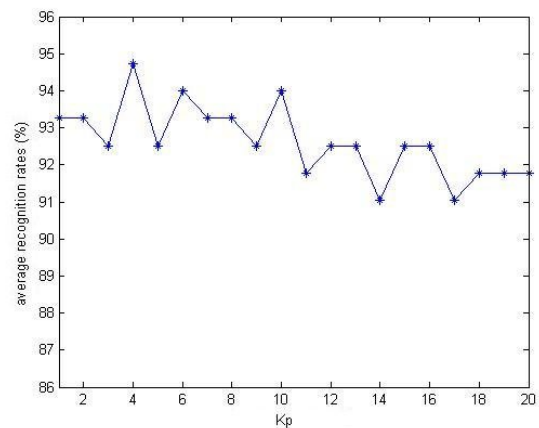


Figure 4: The average recognition rates (%) of LGE/MMC versus the corresponding varied the value of

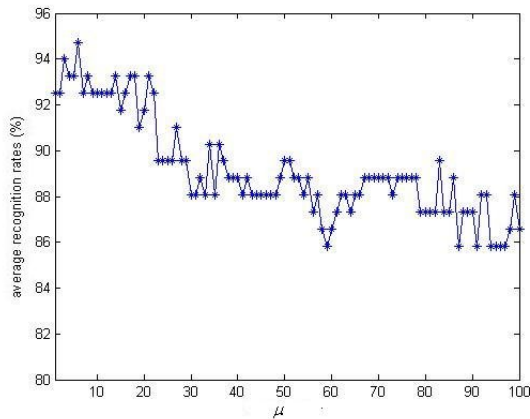


Figure 5: The maximal average recognition rates (%) of LGE/MMC versus the corresponding varied the value of  $\mu$  when only two images per class were randomly selected for training on the YALE face database.

$K_p$  when only two images per class were randomly selected for training on the YALE face databases.

In the second experiment, we also test the impact of  $\mu$  on the performance when only two images per class were randomly selected for training on the YALE face database, which can be found in Fig. 5. We varied  $\mu$

from 1 to 100 with step 1. Fig. 5 displays the maximal average recognition rates with varied parameter  $\mu$  by carrying out LGE/MMC. From Fig. 5, it can be found that the effectiveness of the LGE/MMC algorithm is sensitive to the value of the parameter  $\mu$ . LGE/MMC obtains the best average recognition rate is 94.79% when  $\mu = 6$ . This indicates that the locality and the globality are with the same importance. In the next experiment, the value of adjustable parameter  $\mu$  is taken to be 6.

In the third experiment, we randomly select  $l$  images ( $l$  varies from 2 to 6) of each individual for training, and the remaining ones are used for testing. We compare the performances of different algorithms. The average recognition rates obtained by different algorithms as well as the corresponding dimensionality of reduced subspace (the numbers in parentheses) on the ORL, YALE and AR face databases are given in the Table 2, Table 3 and Table 4, respectively. Fig.6. is the average recognition rates (%) of LGE/MMC versus the corresponding varied dimensions when only six images per class were randomly selected for training on the ORL, YALE and AR face databases. We change the number of eigenvectors from 2 to 50 with step 2 on the ORL, YALE face databases and the number of eigenvectors from 5 to 150 with step 5 on the AR face database, respectively.

Table 2: The average recognition accuracy(%)of different algorithms on the ORL face database and the corresponding standard deviations and dimensions.

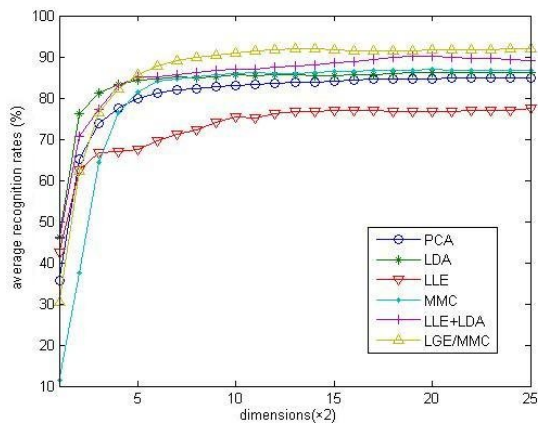
| Methods \ $l$ | 2                    | 3                    | 4                    | 5                    | 6                    |
|---------------|----------------------|----------------------|----------------------|----------------------|----------------------|
| PCA           | 72.25 ± 2.61<br>(22) | 81.62 ± 0.65<br>(22) | 83.98 ± 1.00<br>(22) | 85.41 ± 1.30<br>(22) | 85.41 ± 1.98<br>(22) |
| LDA           | 76.35 ± 1.07<br>(28) | 83.40 ± 1.69<br>(28) | 84.13 ± 2.03<br>(28) | 86.07 ± 1.16<br>(28) | 88.60 ± 0.78<br>(28) |
| LLE           | 69.78 ± 0.82<br>(32) | 72.66 ± 0.73<br>(30) | 76.32 ± 1.14<br>(36) | 78.85 ± 1.15<br>(20) | 88.44 ± 1.55<br>(28) |
| MMC           | 75.01 ± 1.77<br>(26) | 83.96 ± 0.48<br>(28) | 84.37 ± 2.48<br>(28) | 87.47 ± 0.85<br>(28) | 90.49 ± 1.17<br>(26) |
| LLE+LDA       | 73.36 ± 1.24<br>(18) | 85.84 ± 1.23<br>(20) | 89.30 ± 0.83<br>(22) | 88.61 ± 1.79<br>(20) | 94.53 ± 1.82<br>(10) |
| LGE/MMC       | 75.53 ± 1.88<br>(30) | 86.00 ± 1.41<br>(36) | 91.65 ± 0.33<br>(36) | 94.17 ± 1.45<br>(36) | 96.53 ± 0.72<br>(36) |

Table 3: The average recognition accuracy(%) of different algorithms on the YALE face database and the corresponding standard deviations and dimensions.

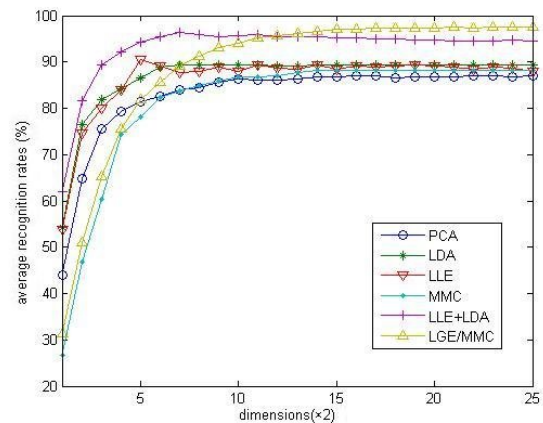
| Methods \ $l$ | 2                    | 3                    | 4                    | 5                    | 6                    |
|---------------|----------------------|----------------------|----------------------|----------------------|----------------------|
| PCA           | 77.29 ± 1.20<br>(18) | 80.20 ± 1.27<br>(20) | 83.99 ± 1.38<br>(18) | 84.72 ± 1.24<br>(20) | 86.21 ± 0.80<br>(22) |
| LDA           | 80.45 ± 1.48<br>(8)  | 84.55 ± 1.06<br>(8)  | 87.85 ± 0.45<br>(8)  | 87.20 ± 1.64<br>(8)  | 88.54 ± 0.82<br>(8)  |
| LLE           | 83.55 ± 1.38<br>(14) | 83.58 ± 0.59<br>(6)  | 85.90 ± 0.75<br>(6)  | 88.37 ± 1.63<br>(8)  | 88.97 ± 1.56<br>(8)  |
| MMC           | 80.58 ± 0.71<br>(12) | 82.84 ± 0.88<br>(6)  | 85.87 ± 1.12<br>(6)  | 86.83 ± 0.37<br>(6)  | 87.79 ± 0.50<br>(6)  |
| LLE+LDA       | 88.33 ± 1.21<br>(22) | 92.45 ± 0.75<br>(18) | 92.84 ± 1.01<br>(12) | 95.34 ± 1.21<br>(10) | 95.21 ± 1.24<br>(10) |
| LGE/MMC       | 94.38 ± 0.41<br>(36) | 94.54 ± 0.83<br>(16) | 93.84 ± 1.16<br>(38) | 95.47 ± 0.11<br>(38) | 96.29 ± 1.26<br>(38) |

Table 4: The average recognition accuracy(%) of different algorithms on the AR face database and the corresponding standard deviations and dimensions.

| Methods \ $l$ | 2                    | 3                    | 4                    | 5                    | 6                     |
|---------------|----------------------|----------------------|----------------------|----------------------|-----------------------|
| PCA           | 66.68 ± 1.11<br>(85) | 70.21 ± 1.62<br>(85) | 77.60 ± 1.27<br>(85) | 79.03 ± 0.81<br>(80) | 81.54 ± 1.04<br>(85)  |
| LDA           | 71.50 ± 0.71<br>(70) | 75.58 ± 0.76<br>(70) | 82.53 ± 0.81<br>(70) | 87.12 ± 0.33<br>(70) | 87.58 ± 0.75<br>(70)  |
| LLE           | 70.40 ± 0.89<br>(85) | 74.31 ± 1.16<br>(75) | 83.35 ± 1.49<br>(70) | 84.78 ± 0.76<br>(80) | 86.98 ± 0.40<br>(85)  |
| MMC           | 69.39 ± 1.15<br>(80) | 75.71 ± 0.59<br>(80) | 82.98 ± 0.62<br>(85) | 85.27 ± 0.94<br>(80) | 87.86 ± 0.82<br>(80)  |
| LLE+LDA       | 69.38 ± 1.20<br>(90) | 79.23 ± 0.89<br>(80) | 88.17 ± 1.36<br>(80) | 89.09 ± 1.49<br>(85) | 92.420 ± 1.08<br>(80) |
| LGE/MMC       | 71.99 ± 0.99<br>(75) | 81.00 ± 0.78<br>(50) | 90.46 ± 0.89<br>(85) | 91.39 ± 0.91<br>(45) | 95.60 ± 0.80<br>(50)  |



(a) ORL face database ( $l = 4$ )



(b) Yale face database ( $l = 6$ )



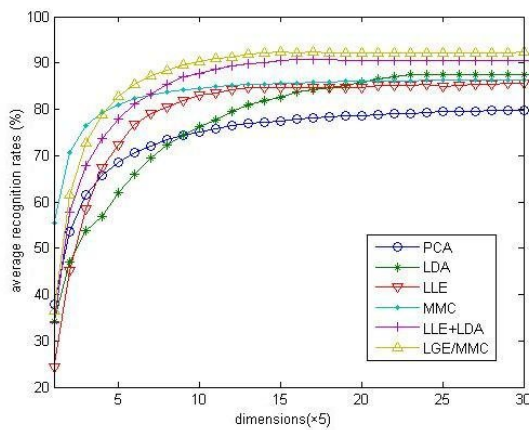
(c)AR face database ( $l = 5$ )

Figure 6: The average recognition rates (%) of LGE/MMC versus the corresponding varied dimensions on the ORL, YALE and AR face databases.

The above experiments showed that the maximal average recognition rates of all methods increases with the increase in training sample size in Table 2, Table 3 and Table 4 respectively. The proposed LGE/MMC algorithm consistently outperforms better than other methods in all experiments in three face databases. From Fig.6 we can find that with the increase number of eigenvectors on three face databases, the average recognition rates also improved.

## 7 Conclusions

In pattern recognition, feature extraction techniques are widely employed to reduce the dimensionality of data and enhance the discriminatory information. In this paper, we proposed a new method for feature extraction and recognition, namely local graph embedding method based on maximum margin criterion (LGE/MMC) for dimensional reduction. The results of face recognition experiments on ORL, YALE and AR face databases demonstrate the effectivity of the proposed method. In the future, we will make more tests on other types of data and decide the optimal parameter  $\mu$ ,  $K_c$  and  $K_p$ . For future work, we will extend LGE/MMC to supervised and semi-supervised cases.

## Acknowledgements

This work is partially supported by the National Science Foundation of China under grant no. 60632050, 90820306, 60873151, 60973098, 61162002, 61005008, 61005005, 60963002, the National Science Foundation of Jiangxi Provincial under grant no. 20114BAB201034, China's Aviation Science no. 20115556007 and Youth Foundation of Jiangxi Provincial Department of Education no. GJJ12459.

## References

[1] M. Turk, A.P. Pentland (1991) Face recognition using eigenfaces. In: Proceedings of IEEE

Computer Society Conference on Computer Vision and Pattern Recognition, Maui, Hawaii, June 1991, pp. 586–591

[2] P.N. Belhumeur, J.P. Hespanha, D.J. Kriegman (1997) Eigenfaces vs. Fisherfaces: recognition using class specific linear projection. *IEEE Trans. Pattern Anal. Mach. Intelligence* 19(7): 711–720

[3] M. Kirby, L. Sirovich (1990) Application of the KL procedure for the characterization of human faces. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 12(1): 103–108

[4] [4] J.M. Lee (1997) *Riemannian Manifolds: An Introduction to Curvature*. Springer, Berlin

[5] B. Scholkopf, A. Smola, K.R. Muller (1998) Nonlinear component analysis as a kernel eigenvalue problem. *Neural Comput* 10 (5): 1299–1319

[6] S. Mika, G. Ratsch, J. Weston, B. Scholkopf, A. Smola, K.-R. Muller (2003) Constructing descriptive and discriminative nonlinear features: Rayleigh coefficients in kernel feature spaces. *IEEE Trans. Pattern Anal. Mach. Intell.* 25(5): 623–628.

[7] J.B. Tenenbaum, V. de Silva, J.C. (2000) Langford A global geometric framework for nonlinear dimensionality reduction. *Science* 290(5500): 2319–2323.

[8] S.T. Roweis, L.K. Saul (2000) Nonlinear dimensionality reduction by locally linear embedding. *Science* 290(5500): 2323–2326.

[9] L.K. Saul, S.T. Roweis (2003) Think globally, fit locally: unsupervised learning of low dimensional manifolds. *J. Mach. Learn. Res* 4: 119–155.

[10] M. Belkin and P. Niyogi, T. G. Dietterich, S. Becker, and Z. Ghahramani (2000) Laplacian eigenmaps and spectral techniques for embedding and clustering. *Advances in Neural Information Processing Systems* 14: 873–878

[11] M. Belkin, P. Niyogi (2003) Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Comput* 15(6): 1373–1396.

[12] Z. Zhang, H. Zha (2004) Principal manifolds and nonlinear dimensionality reduction via tangent space alignment. *SIAM Journal of Scientific Computing* 26(1): 313–338

[13] X. He and P. Niyogi (2003) Locality Preserving Projections. In: *Proceedings of the 17th Annual Conference on Neural Information Processing Systems*, Vancouver and Whistler, Canada, December 2003, pp. 153–160

[14] X. He, S. Yan, Y. Hu, P. Niyogi, and H.-J. Zhang (2005) Face Recognition Using Laplacianfaces. *IEEE Trans. Pattern Analysis and Machine Intelligence* 27(3): 328–340

[15] H. Hu (2008) Orthogonal neighborhood preserving discriminant analysis for face recognition. *Pattern Recognition* 41: 2045–2054.

[16] W. Yu, X. Teng, C. Liu (2006) Face recognition using discriminant locality preserving projections. *Image Vision Computing* 24: 239–248

[17] L. Yang, W. Gong, X. Gu, W. Li, Y. Liang (2008) Null space discriminant locality preserving

- projections for face recognition. *Neurocomputing* 71: 3644–3649
- [18] G. F. Lu, Z. Lin, Z. Jin (2010) Face recognition using discriminant locality preserving projections based on maximum margin criterion. *Pattern Recognition* 43: 3572-2579
- [19] D. de Ridder, R.P.W. Duin, Locally linear embedding for classification, Technical Report PH-2002-01, Pattern Recognition Group, Department of Imaging Science and Technology, Delft University of Technology, Delft, The Netherlands, 2002.
- [20] D. de Ridder, O. Kouropteva, O. Okun, M. Pietikainen, R.P.W. Duin, Supervised locally linear embedding, artificial neural networks and neural information processing, in: *ICANN/ICONIP 2003 Proceedings, Lecture Notes in Computer Science*, vol. 2714, Springer, Berlin, 2003, pp. 333–341.
- [21] X. Bai, B. Yin, Q. Shi, Y. Sun, Face recognition based on supervised locally linear embedding method, *J. Inf. Comput. Sci.* (4) (2005) 641–646.
- [22] O. Kouropteva, O. Okun, M. Pietikainen, Supervised locally linear embedding algorithm for pattern recognition, *IbPRIA 2003, Lecture Notes in Computer Science*, vol. 2652, Springer, Berlin, 2003, pp. 386–394.
- [23] J. Zhang, H. Shen, Z.-H. Zhou, Unified Locally Linear Embedding and Linear Discriminant Analysis Algorithm for Face Recognition, *Lecture Notes in Computer Science*, Springer, Berlin, 2004.
- [24] J. Zhang, H. Shen, Z.-H. Zhou, Ensemble-based discriminant manifold learning for face recognition, *ICNC 2006, Part I, Lecture Notes in Computer Science*, vol. 4221, Springer, Berlin, 2006, pp. 29–38.
- [25] X.F. He, D. Cai, S.C. Yan, H.J. Zhang (2005) Neighborhood preserving embedding. In: *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, Beijing, china, October 2005, pp. 1208–1213.
- [26] X.H. Zeng, S.W. Luo (2007) A supervised subspace learning algorithm: supervised neighborhood preserving embedding. In: *Proceedings of 3rd International Conference on Advanced Data Mining and Applications*, Harbin, China, August 2007, pp 81-88
- [27] Y. Wang, Y. Wu (2010) Complete neighborhood preserving embedding for face recognition. *Pattern Recognition* 43:1008-1015
- [28] C. Hwann-Tzong, C. Huang-Wei, L. Tyng-Luh, Local discriminant embedding and its variants, in: *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 2* (IEEE Computer Society, 2005).
- [29] S.C. Yan, D. Xu, B.Y. Zhang, H.J. Zhang, Q. Yang, S. Lin, Graph embedding and extensions: A general framework for dimensionality reduction, *IEEE Trans. Pattern Anal. Mach. Intell.* 29 (2007) 40-51.
- [30] P.Y. Han, A.T.B. Jin, F.S. Abas, Neighbourhood preserving discriminant embedding in face recognition, *J. Vis. Commun. Image Represent.* 20 (2009) 532-542.
- [31] M.A. Turk, A.P. Pentland, Face recognition using eigenfaces, *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (91CH2983-5)*, (1991).
- [32] S. Yan, D. Xu, B. Zhang, and H.-J. Zhang, Graph Embedding: A General Framework for Dimensionality Reduction, *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 2, pp. 830-837, 2005, San Diego.
- [33] H. Li, T. Jiang, K. Zhang, Efficient and robust feature extraction by maximum margin criterion, *IEEE Transactions on Neural Networks* 17 (1) (2006) 1157–1165.
- [34] AT&T Laboratories Cambridge, The ORL database of faces, [\\_http://www.uk.research.att.com/facedatabase.html](http://www.uk.research.att.com/facedatabase.html).
- [35] Yale database, 1997. Available from: <http://cvc.yale.edu/projects/yalefaces/yalefaces.html>.
- [36] A. Martinez, R. Benavente, The AR face database, CVC.

# Ensembles for Predicting Structured Outputs

Dragi Kocev

Department of Knowledge Technologies, Jožef Stefan Institute, Jamova cesta 39, Ljubljana, Slovenia

E-mail: Dragi.Kocev@ijs.si Web: <http://kt.ijs.si/DragiKocev>

## Thesis Summary

**Keywords:** ensemble methods, predictive clustering trees, predicting structured outputs

**Received:** February 7, 2012

*This article presents a summary of the doctoral dissertation of the author on the topic of building ensembles for predicting structured outputs.*

*Povzetek: Članek predstavlja povzetek doktorske disertacije avtorja, ki obravnava temo gradnje ansamblov za napovedovanje strukturiranih vrednosti.*

## 1 Introduction

In many real life problems of predictive modelling, the output is structured, meaning that there can be dependencies between classes or some internal relations between the classes (e.g., classes are organized into a tree-shaped hierarchy or a directed acyclic graph). These types of problems occur in domains such as life sciences (gene function prediction, drug discovery), ecology (analysis of remotely sensed data, habitat modelling), multimedia (annotation and retrieval of images and videos) and the semantic web (categorization and analysis of text and web). Having in mind the needs of the application domains and the increasing quantities of structured data, the task of “mining complex knowledge from complex data” was listed as one of the ten most challenging problems in data mining [5].

## 2 Methods and evaluation

In the thesis [2], we address the task of learning models for predicting structured outputs that take as input a tuple of attribute values and produce as output a structured object. In contrast to classification and regression, where the output is a single scalar value, in our case the output is a data structure, such as a tuple or a directed acyclic graph. We consider both global and local prediction of structured outputs, the first based on a single model that predicts the entire output structure and the latter based on a collection of models, each predicting a component of the output structure.

A variety of methods, specialized for predicting a given type of structured output, have been proposed [1]. However, many of them are computationally demanding and not suited for dealing with large datasets (especially large outputs). In the thesis, we propose to use predictive clustering trees (PCTs) [3] for efficient and accurate prediction of structured outputs. PCTs offer a unifying approach to deal-

ing with different types of structured outputs. We extend PCTs in the direction of ensemble methods [4] to further increase their predictive performance.

In particular, we take the notion of an ensemble, i.e., a collection of predictive models whose predictions are combined, and apply it in the context of predicting structured outputs. We develop methods for learning different types of ensembles of PCTs for global and local prediction of different types of structured outputs. The different types of ensembles include bagging, random forests, random subspaces and bagging of subspaces. The types of outputs considered correspond to the different predictive modeling tasks, i.e., multi-target regression, multi-target classification, and hierarchical multi-label classification. Each of the combinations can be applied both in the context of global prediction (producing a single ensemble) or local prediction (producing a collection of ensembles).

Computational complexity analyses of the methods show that the global ensembles are the most efficient, especially random forests. The analyses also indicate that the proposed approaches are scalable to datasets which can be large along any of the following dimensions: number of attributes, number of examples, and size of the target. This is confirmed also by the empirical evaluation of the proposed methods on a large number of datasets.

## 3 Conclusion

The thesis makes several contributions to the area of machine learning and the respective application areas. First, we propose ensemble learning methods for predicting structured outputs that use PCTs as base predictive models. The proposed methods are general in terms of the type of the structured output: they support the tasks of predicting multiple continuous targets, predicting multiple discrete targets, and hierarchical multi-label classification.

Second, we perform an extensive empirical evaluation of

the proposed methods over a variety of benchmark datasets. We construct ensembles of up to 1000 predictive models and select ensembles of 50 global predictive models as optimal in terms of predictive performance and efficiency.

Third, we compare the performance of ensembles of global models and single global models, as well as ensembles of local models. Both global and local ensembles perform better than the single model counterparts in terms of predictive power. Global and local ensembles perform equally well, with global ensembles being more efficient and producing smaller models, as well as needing fewer trees in the ensemble to achieve the maximal performance.

Fourth, we apply the proposed methods in three practically relevant domains. (1) We constructed models that assess vegetation condition from remotely sensed data and generated maps of the state of Victoria, Australia [6]. (2) On the task of hierarchical annotation of medical X-ray images, the ensembles of PCTs provided the best annotation results reported so far in the literature [8]. (3) Extensive experimental evaluation over several tasks of gene function prediction in three organisms showed that bagging of PCTs is superior to or competitive with state-of-the-art methods [7].

In the thesis, we also present some preliminary results that further explore the proposed paradigm of ensembles for structured prediction. We first discuss structured prediction for different types of structured outputs. Next, we propose a method for feature ranking in the context of structured outputs, based on random forests. Finally, we suggest a novel ensemble learning method that is based on the beam search strategy and can control directly the diversity in the ensemble.

## References

- [1] G. Bakır, T. Hofmann, B. Schölkopf, A. Smola, B. Taskar, S. Vishwanathan (2007) *Predicting structured data*, The MIT Press.
- [2] D. Kocev (2007) *Ensembles for predicting structured outputs*, PhD Thesis, IPS Jožef Stefan, Ljubljana, Slovenia.
- [3] H. Blockeel (1998) *Top-down induction of first order logical decision trees*, PhD Thesis, Katholieke Universiteit Leuven, Belgium.
- [4] G. Seni, J. Elder (2010) *Ensemble methods in data mining: Improving accuracy through combining predictions*, Morgan & Claypool Publishers.
- [5] Q. Yang, X. Wu (2006) 10 Challenging Problems in Data Mining Research, *International Journal of Information Technology & Decision Making*, 5(4):597–604.
- [6] D. Kocev, S. Džeroski, M. White, G. Newell, P. Griffoen (2009) Using single- and multi-target regression trees and ensembles to model a compound index of vegetation condition, *Ecological Modelling*, 220(8):1159–1168.
- [7] L. Schietgat, C. Vens, J. Struyf, H. Blockeel, D. Kocev, S. Džeroski (2010) Predicting gene function using hierarchical multi-label decision tree ensembles, *BMC Bioinformatics*, 11(2):1–14.
- [8] I. Dimitrovski, D. Kocev, S. Loskovska, S. Džeroski (2011) Hierarchical annotation of medical images, *Pattern Recognition*, 44(10-11):2436–2449.

# Rapid Ontology Development Model Based on Rule Management Approach in Business Applications

Dejan Lavbič  
University of Ljubljana, Faculty of Computer and Information Science,  
Tržaška 25, 1000 Ljubljana, Slovenia  
E-mail: Dejan.Lavbic@fri.uni-lj.si

## Thesis Summary

**Keywords:** ontologies, semantic web, methodologies for ontology development, business rules, ontology evaluation, rapid ontology development, ROD.

**Received:** October 12, 2011

*In this paper rapid ontology development with emphasis on facilitating development with constant evaluation of steps in the process of ontology development is presented. The review of related work pointed out that existing methodologies for ontology development are complex and high level of technical knowledge is required. Business users and developers usually don't pose such knowledge therefore ontology completeness indicator was introduced in this approach. The role of this indicator is to guide developer throughout the development process and constantly aid user with recommendations to progress to next step and improve the quality of ontology. While evaluating the ontology, several aspects are considered; from description, partition, consistency, redundancy and to anomaly. The approach was verified on Financial Instruments and Trading Strategies (FITS) ontology and compared to other approaches.*

*Povzetek: V članku je predstavljen hiter razvoj ontologij.*

## 1 Introduction

The Semantic Web vision is the idea of having data on the Web defined and linked in a way that it can be used by machines not just for display purposes (e.g. user), but for automation, integration and reuse of data across various applications. Next generation of the Web is expected to provide automated services based on machine processable semantics of data, reasoning techniques and heuristics that make use of these data. The applications of ontologies are mainly restricted to academia while successful employment in business environments is rare.

The simplicity of using approaches for ontology construction and accompanying tool support is an important issue which needs a lot of attention and further work.

## 2 Related work

Current approaches [1-3] in ontology development are technically very demanding and require long learning curve and are therefore inappropriate for business users. In majority of existing approaches an additional role of knowledge engineer is required for mediation between actual knowledge that business users possess and ontology engineers who encode knowledge in one of selected formalisms. Introduction of several abstraction layers as suggested in systems for business rules manipulation and MDA approach has turned out to be very effective in development of ontologies and using it

in business applications [4]. Besides simplifying the process of ontology construction we also have to focus on very important aspect of ontology completeness. Several researches [5],[6] have discussed error free ontologies and identified frequent errors and anomalies in ontology development, which is advised to be included in the development process and therefore aiding users at prevention and elimination of repeated design errors.

## 3 Rapid Ontology Development

The main purpose of this thesis was to define innovative model for rapid ontology development (ROD) that is suitable for users without extensive technical knowledge and knowledge of ontology design [7]. The proposed process includes pre-development, development and post-development activities, from business vocabulary acquisition to employment of developed ontology as a functional component in information system (see Figure 1).

ROD process also includes a constant evaluation of developed ontologies which is conducted at every step of the process and gives user recommendations on how to improve the quality of developing ontology. This functionality is implemented using ontology completeness indicator (OC) that is used for following the steps of ROD process in simplified manner (see Figure 2).

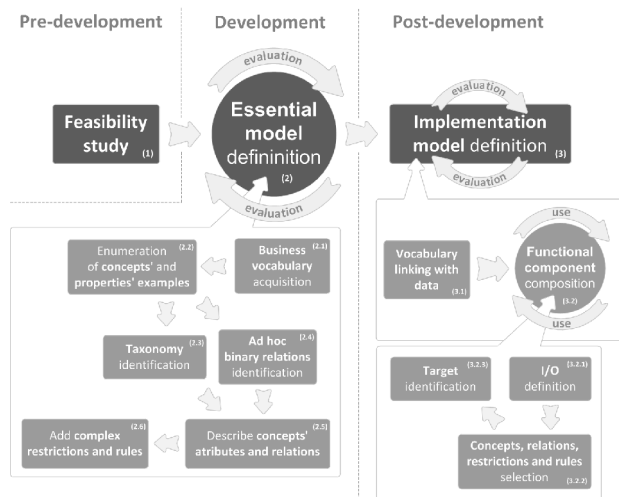


Figure 1: Process of Rapid Ontology Development.

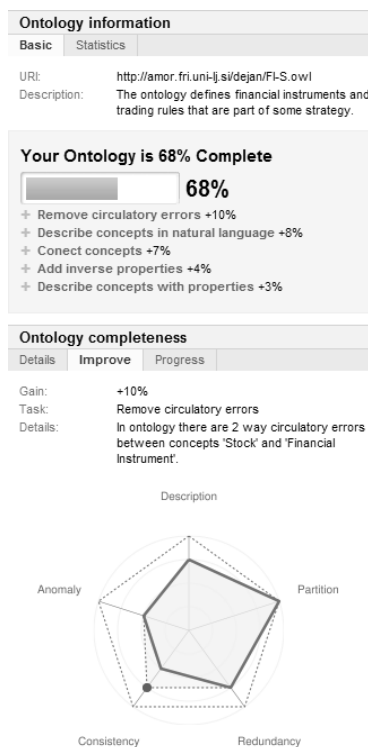


Figure 2: Display of ontology completeness results and improvement recommendations.

Constant evaluation of developing ontology is also performed with dynamic adaption of weights in calculation which in turn aids user with recommendations on how to improve ontology. Along with the development of ROD model several possibilities of using ontology as a functional component was investigated. Ontology can be used as whole or just partly by using only schematic part. To improve integration with existing data sources, interfaces for direct linking of semi-structured data in ontology were developed. This was accomplished with a generic approach of regular expressions that enables us to

connect to any semi-structured source of data, e.g. document, web page, data base, CSV file etc. For the evaluation purposes of proposed approach FITS ontology for trading with financial instruments was developed. Its generic design enables users very straightforward reuse. In this experiment ROD approach turned out to be very effective. Less iterations were required to develop a working version of ontology and the required confirmation level of ontology quality was also achieved earlier.

## 4 Conclusions

The thesis represents integral model for rapid ontology development that enables users without extensive technical knowledge development of ontology. By doing this they have an ability to employ the advantages of semantic web in building semantically enabled application that can very intuitively reuse data from several (also semi-structured) data sources.

## References

- [1] G. Schreiber, H. Akkermans, A. Anjewierden, R. de Hoog, N. Shadbolt, W. van de Velde, and B. Wielinga, *Knowledge Engineering and Management - The CommonKADS Methodology*, London, England: The MIT Press: Cambridge, Massachusetts, 1999.
- [2] Y. Sure, "Methodology, Tools & Case Studies for Ontology based Knowledge Management," University of Karlsruhe, Institute AIFB, 2003.
- [3] J. Davies, R. Studer, and P. Warren, *Semantic Web Technologies - trends and research in ontology-based systems*, Chichester, England: John Wiley & Sons, 2006.
- [4] A. Smaizys and O. Vasilecas, "Business Rules based agile ERP systems development," *Informatica*, vol. 20, 2009, pp. 439-460.
- [5] M. Fahad and M.A. Quadir, "Ontological errors - Inconsistency, Incompleteness and Redundancy," *International Conference on Enterprise Information Systems (ICEIS) 2008*, Barcelona, Spain: 2008.
- [6] R. Porzel and R. Malaka, "A Task-based Approach for Ontology Evaluation," 2004.
- [7] D. Lavbič and M. Krisper, "Facilitating Ontology development with continuous evaluation," *Informatica*, vol. 21, 2010, pp. 533-552.

## 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 900 staff, has 700 researchers, about 250 of whom are postgraduates, around 500 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 Slovenia (or S<sup>o</sup>nia). The capital today is considered a crossroad between East, West and Mediter-

anean 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](mailto:matjaz.gams@ijs.si)  
Public relations: Polona Strnad

**INFORMATICA**  
**AN INTERNATIONAL JOURNAL OF COMPUTING AND INFORMATICS**  
**INVITATION, COOPERATION**

**Submissions and Refereeing**

Please submit a manuscript at: <http://www.informatica.si/Editors/PaperUpload.asp>. 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 L<sup>A</sup>T<sub>E</sub>X 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](mailto: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 eighteen 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 intelligent systems in the European computer science, informatics and cognitive 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: .....                                  | Office Address and Telephone (optional): ..... |
| Title and Profession (optional): .....       | .....  |
| .....  | E-mail Address (optional): .....               |
| Home Address and Telephone (optional): ..... | .....  |
| .....  | Signature and Date: .....                      |



## **Informatica WWW:**

**<http://www.informatica.si/>**

### **Referees from 2008 on:**

Ajith Abraham, Siby Abraham, Renato Accornero, Raheel Ahmad, Cutting Alfredo, Hameed Al-Qaheri, Gonzalo Alvarez, Wolfram Amme, Nicolas Anciaux, Rajan Arora, Costin Badica, Zoltán Balogh, Andrea Baruzzo, Borut Batagelj, Norman Beaulieu, Paolo Bellavista, Steven Bishop, Marko Bohanec, Zbigniew Bonikowski, Borko Bosković, Marco Botta, Pavel Brazdil, Johan Brichau, Andrej Brodnik, Ivan Bruha, Maurice Bruynooghe, Wray Buntine, Dumitru Dan Burdescu, Yunlong Cai, Juan Carlos Cano, Tianyu Cao, Norman Carver, Marc Cavazza, Jianwen Chen, L.M. Cheng, Chou Cheng-Fu, Girija Chetty, G. Chiola, Yu-Chiun Chiou, Ivan Chorbev, Shauvik Roy Choudhary, Sherman S.M. Chow, Lawrence Chung, Mojca Ciglarič, Jean-Noël Colin, Vittorio Cortellessa, Jinsong Cui, Alfredo Cuzzocrea, Darko Čerepnalkoski, Gunetti Daniele, Grégoire Danoy, Manoranjan Dash, Paul Debevec, Fathi Debili, Carl James Debono, Joze Dedic, Abdelkader Dekdouk, Bart Demoen, Sareewan Dendamrongvit, Tingquan Deng, Anna Derezinska, Gaël Dias, Ivica Dimitrovski, Jana Dittmann, Simon Dobrišek, Quansheng Dou, Jeroen Doumen, Erik Dovgan, Branko Dragovich, Dejan Dragic, Jozo Dujmovic, Umut Riza Ertürk, CHEN Fei, Ling Feng, YiXiong Feng, Bogdan Filipič, Iztok Fister, Andres Flores, Vladimir Fomichov, Stefano Forli, Massimo Franceschet, Alberto Freitas, Jessica Fridrich, Scott Friedman, Chong Fu, Gabriel Fung, David Galindo, Andrea Gambarara, Matjaž Gams, Maria Ganzha, Juan Garbajosa, Rosella Gennari, David S. Goodsell, Jaydeep Gore, Miha Grčar, Daniel Grosse, Zhi-Hong Guan, Donatella Gubiani, Bidyut Gupta, Marjan Gusev, Zhu Haiping, Kathryn Hempstalk, Gareth Howells, Juha Hyvärinen, Dino Ienco, Natarajan Jaisankar, Domagoj Jakobovic, Imad Jawhar, Yue Jia, Ivan Jureta, Dani Juričić, Zdravko Kačič, Slobodan Kalajdziski, Yannis Kalantidis, Boštjan Kaluža, Dimitris Kanellopoulos, Rishi Kapoor, Andreas Kassler, Daniel S. Katz, Samee U. Khan, Mustafa Khattak, Elham Sahebkar Khorasani, Ivan Kitanovski, Tomaž Klobučar, Ján Kollár, Peter Korošec, Valery Korzhik, Agnes Koschmider, Jure Kovač, Andrej Krajnc, Miroslav Kubat, Matjaz Kukar, Anthony Kulis, Chi-Sung Lai, Niels Landwehr, Andreas Lang, Mohamed Layouni, Gregor Leban, Alex Lee, Yung-Chuan Lee, John Leggett, Aleš Leonardis, Guohui Li, Guo-Zheng Li, Jen Li, Xiang Li, Xue Li, Yinsheng Li, Yuanping Li, Shiguo Lian, Lejian Liao, Ja-Chen Lin, Huan Liu, Jun Liu, Xin Liu, Suzana Loskovska, Zhiguo Lu, Hongen Lu, Mitja Luštrek, Inga V. Lyustig, Luiza de Macedo, Matt Mahoney, Domen Marinčič, Dirk Marwede, Maja Matijasevic, Andrew C. McPherson, Andrew McPherson, Zuqiang Meng, France Mihelič, Nasro Min-Allah, Vojislav Misić, Vojislav Mišić, Mihai L. Mocanu, Angelo Montanari, Jesper Mosegaard, Martin Možina, Marta Mrak, Yi Mu, Josef Mula, Phivos Mylonas, Marco Di Natale, Pavol Navrat, Nadia Nedjah, R. Nejabat, Wilfred Ng, Zhicheng Ni, Fred Niederman, Omar Nouali, Franc Novak, Petteri Nurmi, Denis Obrul, Barbara Oliboni, Matjaž Pančur, Wei Pang, Gregor Papa, Marcin Paprzycki, Marek Paralič, Byung-Kwon Park, Torben Bach Pedersen, Gert Schmeltz Pedersen, Zhiyong Peng, Ruggero G. Pensa, Dana Petcu, Marko Petkovšek, Rok Piltaver, Vid Podpečan, Macario Polo, Victor Pomponiu, Elvira Popescu, Božidar Potočnik, S. R. M. Prasanna, Kresimir Pripuzic, Gabriele Puppis, HaiFeng Qian, Lin Qiao, Jean-Jacques Quisquater, Vladislav Rajković, Dejan Rakovic, Jean Ramaekers, Jan Ramon, Robert Ravnik, Wilfried Reimche, Blagoj Ristevski, Juan Antonio Rodriguez-Aguilar, Pankaj Rohatgi, Wilhelm Rossak, Eng. Sattar Sadkhan, Sattar B. Sadkhan, Khalid Saeed, Motoshi Saeki, Evangelos Sakkopoulos, M. H. Samadzadeh, MariaLuisa Sapino, Piervito Scaglioso, Walter Schempp, Barabara Koroušić Seljak, Mehrdad Senobari, Subramaniam Shamala, Zhongzhi Shi, LIAN Shiguo, Heung-Yeung Shum, Tian Song, Andrea Soppera, Alessandro Sorniotti, Liana Stanescu, Martin Steinebach, Damjan Strnad, Xinghua Sun, Marko Robnik Šikonja, Jurij Šilc, Igor Škrjanc, Hotaka Takizawa, Carolyn Talcott, Camillo J. Taylor, Drago Torkar, Christos Tranoris, Denis Trček, Katarina Trojancanec, Mike Tschierschke, Filip De Turck, Aleš Ude, Wim Vanhoof, Alessia Visconti, Vuk Vojisavljevic, Petar Vračar, Valentino Vranić, Chih-Hung Wang, Huaqing Wang, Hao Wang, Hui Wang, YunHong Wang, Anita Wasilewska, Sigrid Wenzel, Woldemar Wolynski, Jennifer Wong, Allan Wong, Stefan Wrobel, Konrad Wrona, Bin Wu, Xindong Wu, Li Xiang, Yan Xiang, Di Xiao, Fei Xie, Yuandong Yang, Chen Yong-Sheng, Jane Jia You, Ge Yu, Borut Zalik, Aleš Zamuda, Mansour Zand, Zheng Zhao, Dong Zheng, Jinhua Zheng, Albrecht Zimmermann, Blaž Zupan, Meng Zuqiang

# *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 2012 (Volume 36) 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: ABO grafika d.o.o., Ob železnici 16, 1000 Ljubljana.

Orders may be placed by email ([drago.torkar@ijs.si](mailto: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.*

# *Informatica*

**An International Journal of Computing and Informatics**

|   |  |            |
|---|--|------------|
| Editors' Introduction to the Special Issue on IPTV and Multimedia Services  | E. Mikóczy, I. Vidal, D. Kanellopoulos | <b>1</b>   |
| IPTV Evolution Towards NGN and Hybrid Scenarios   | E. Mikóczy, I. Vidal, D. Kanellopoulos | <b>3</b>   |
| IPTV Services Personalization Using Context-Awareness   | S. Song, H. Moustafa, H. Afifi         | <b>13</b>  |
| Privacy Aware Recommender Service using Multi-agent Middleware- an IPTV Network Scenario  | A.M. Elmisery, D. Botvich              | <b>21</b>  |
| An RTSP Proxy for Implementing the IPTV Media Function Using a Streaming Server   | Z.S. Shibeshi, A. Terzoli, K. Bradshaw | <b>37</b>  |
| Secure Key Exchange Scheme for IPTV Broadcasting  | R.S. Pippal, S. Tapaswi, C.D. Jaidhar  | <b>47</b>  |
| <hr/> <i>End of Special Issue / Start of normal papers</i>  |  |            |
| 'The Frozen Accident' as an Evolutionary Adaptation: A Rate Distortion Theory Perspective on the Dynamics and Symmetries of Genetic Coding Mechanisms | J.F. Glazebrook, R. Wallace            | <b>53</b>  |
| Times Limited Accountable Anonymous Online Submission Control System from Single-Verifier $k$ -times Group Signature                                  | X. Zhao, F. Zhang                      | <b>75</b>  |
| Multiple Attribute Decision Making Method Based on the Trapezoid Fuzzy Linguistic Hybrid Harmonic Averaging Operator                                  | P. Liu, Y. Su                          | <b>83</b>  |
| Physics Markup Approaches Based on Geometric Algebra Representations  | K-p. Yang, W. Zhang, F. Petry          | <b>91</b>  |
| Local Graph Embedding Based on Maximum Margin Criterion (LGE/MMC) for Face Recognition  | M. Wan, S. Gai, J. Shao                | <b>103</b> |
| Ensembles for Predicting Structured Outputs   | D. Kocev                               | <b>113</b> |
| Rapid Ontology Development Model Based on Rule Management Approach in Business Applications   | D. Lavbič                              | <b>115</b> |

