2007

Proceedings of the Sixth International Workshop for Applied PKC (IWAP2007)

Dongguang Li

Proceedings of the Sixth International Workshop for Applied PKC (IWAP2007)

Edited by Dongguang Li

ISBN: 0-7298-0644-6

Published by School of Computer and Information Science, Edith Cowan University
Perth, Western Australia

The conference was organized by School of Computer and Information Science at the Mount Lawley Campus of Edith Cowan University.
3rd - 4th December, 2007

Preface

IWAP2007 will be the sixth of a series of successful international workshops with focus on research and engineering issues of the applied aspects of public key cryptosystems. The inaugural IWAP event was held in Korea in 2001, and was subsequently held in 2002, 2004, 2005 and 2006 respectively in Taipei, Japan, Singapore and China. The IWAP2003 was cancelled due to the SARS breakout. Theoreticians and practitioners interested in the applied issues of PKC were encouraged to participate and contribute to the continuous success of the IWAP workshop series. The host of the IWAP2007 is Edith Cowan University. It is my pleasure to have the opportunity to chair the IWAP conference in Perth, Australia in 2007.

Security is well recognized as a most important issue in e-commerce applications and national security systems while Public Key Cryptosystem (PKC) is widely accepted to be a key mechanism in secure application systems. As such, infrastructures that facilitate the management and deployment of public key cryptosystems have received much attention from the security community. Authorities and regulators have spent a lot of effort standardizing PKC-related standards and enacting legislations for recognizing PKC in business transactions. However, in reality, due to systems issues and engineering considerations, the adoption of PKC has not been as pervasive as the security community anticipated while high sensitivity application systems remain as vulnerable as they used to be.

The theme of this workshop is to provide a forum for discussing the systems and engineering aspects of security systems that make use of PKC as a basic security mechanism. It's observed that, over the past few years, there have been a growing number of critical application areas (such as the new generation travel documents by ICAO and payment cards by EMV) rely on the presence of some well-designed, well-engineered PKC. While there are existing venues for promoting theoretical aspects of PKC, IWAP 2007 aims to provide a platform for researchers to exchange ideas on applied aspects of PKC, and to stimulate further researchers to innovating and/or important applications of PKC as well as systems and engineering aspects for PKC deployment in a large complex environment.

There are 21 selected papers included in the proceedings. 13 of them are from Australian universities such as ECU, ANU, Deakin and James Cook University. Others are from 8 different countries including USA, Germany, China, Korea, Japan, Ukraine, Iran, and India. Although the conference is small in size it is really an international one. The conference will be opened by Prof Tony Watson, General Co-Chair and Pro Vice Chancellor, and Executive Dean of the Faculty of Computing, Health and Sciences, Edith Cowan University. Participants from ten countries will deliver many outstanding presentations over two intense days. The papers in the Proceedings are ordered according to the original program sessions and their corresponding themes.

All the papers included in the proceedings have been peer reviewed in full by at least two independent reviewers selected from the international program committee.
Mention must be made to Associate Professor Jim Cross, Associate Dean of the Faculty of Computing, Health and Sciences, Edith Cowan University, whom have worked tirelessly with me in the preparation of this conference, starting almost one year ago.

IWAP2007 would not have been possible without the dedicated support of the International Steering Committee:

Kwang-Jo Kim (School of Engineering, Information and Communications Univ, Korea)
Kwok-Yan LAM (Tsinghua University, Singapore)
Kouichi Sakurai (Kyushu University, Japan)
Craig Valli (Edith Cowan University, Australia)
Jialin Cao (Shanghai University of Electric Power, China)

I would like to thank all of the members of the International Programme Committee for carrying out the paper reviews with care and competence:

Dongguang Li Chair (Edith Cowan University, Australia)
Xinmin Geng Co-Chair (Shanghai University of Electric Power, China)
Weiguo Pan Co-Chair (Shanghai University of Electric Power, China)
Lisa McCormack Secretary (Edith Cowan University, Australia)
David Veal (Edith Cowan University, Australia)
Heping Tu (Shanghai University of Electric Power, China)
Shaoguang Li (The Jackson Laboratory, USA)
Clifton Smith (Edith Cowan University, Australia)
Paul Maj (Edith Cowan University, Australia)
Lipo Wang (Nanyang Technological University, Singapore)
Jitian Xiao (Edith Cowan University, Australia)
Alfred Tan (Edith Cowan University, Australia)
Huaizhong Li (Wenzhou University, China)
Cui yongrui (Dalian University of Science and Technology, China)
Yingxu Wang (University of Calgary, Canada)
Yi Zhuang, (Nanjing University of Aeronautics & Astronautics, China)
Mingchu Li, (Dalian University of Technology, China)
Haibin Zhu, (Nipissing University, Canada)
Huaisheng Wang (Shanghai University of Electric Power, China)
Yoshifumi Ueshige (Institute of Systems & IT /KYUSHU, Japan)

My thanks also go to the members of the Organizing Committee:

Jim Cross Chair (Edith Cowan University, Australia)
Craig Valli Co-Chair (Edith Cowan University, Australia)
Hao Zhang Co-Chair (Shanghai University of Electric Power, China)
Liz John (Edith Cowan University, Australia)
Rebecca Treloar-Cook (Edith Cowan University, Australia)
Lisa McCormack .(Edith Cowan University, Australia)
Heping Tu (Shanghai University of Electric Power, China)
Fei Han (Shanghai University of Electric Power, China)
Bryan Garnett-Law (Edith Cowan University, Australia)

In particularly, I want to thank our invited keynote speakers:

Dr Michiharu Kudo
Manager, Security & Privacy, I&I, IBM Research, Tokyo Research Laboratory, Japan
Prof Chan Yeob Yeun
Information Communication University, Korea
Prof Clifton Smith

Foundation Director of the Australian Institute of Security and Applied Technology, Edith Cowan University, Australia
Prof Saeid Nahavandi
Alfred Deakin Professor at Deakin University, Australia

I wish all the participants of IWAP2007 a fruitful conference and a wonderful stay in Perth.

A/Prof Dongguang Li

General Chair of IWAP2007
School of Computer and Information Science
Faculty of Computing, Health and Science
Edith Cowan University
Australia

30 November 2007

Table of Contents

Title	Author(s)	Page
Preface		2
Table of Contents		5
Trusted Computing Infrastructure and PKI	Michiharu Kudo	6
New Novel Approaches for Securing VoIP Applications	Chan Yeob Yeun, Kyusuk Han, and Kwangjo Kim	11
Survey of RSA Implementations and Attacks	Srinivasa Rao Subramanya Rao	21
An Efficient Homomorphic Coercion Resistant Voting Scheme Using Binary Search Tree	Vinodu George and M P Sebastian	30
Compressed Nested Certificates Provide More Efficient PKI	A Jancic and LM Batten	40
Fast Arithmetic In Jacobian Of Hyperelliptic Curves Of Genus 2 Over GF(p)	V. Kovtun and J. Pelzl	51
Australian firearm identification system based on the ballistics images of projectile specimens	Dongguang Li	60
Firearm Identification with Hierarchical Neural Networks by analyzing the firing pin Images retrieved from cartridge cases	Dongguang Li	71
A General Model for Oblivious Transfer	Hossein Ghodosi	80
A Non-Interactive Multiparty Computation Protocol	Hossein Ghodosi and Rahim Zaare-Nahandi	89
China's Information Security Policy Inadequate To Protect Its Netizen's Personal Rights	Fenyu Zeng and Pan Hua	97
Sequencing Clusters of Spatial Join Operations Using Weighted Match	Jitian Xiao	107
Mobile Agent for Electrical Power Infrastructure Protection	Michael W. David	117
The Evaluation of Security Systems: Testing Biometric and Intelligent Imaging Systems	Clifton L Smith	126
Triangulation Based Static Wide Angle Laser Scanning For Obstacle Detection	K. Sahba, K. E. Alameh and C. L. Smith	138
Improving Security for Vision Impaired ATM Access via Dynamic Patterns	D. Veal	149
State Model Diagrams and Home Security System Control	D. Veal & S. P. Maj	157
Study on Whole Lifecycle Protection of Digital Contents	Mei Xue, Xinmin Geng	166
Data Mining and Genetic Algorithm Application In Bioinformatics With Microarray	C. Y. Jiao and D. G. Li	177
Three-Dimensional Cellular Automation LFSR Algorithm	Yong Wang , Xinmin Geng, Yu Wang	188
An Agent-Oriented Programming Based on OOP	Yong Wang , Xinmin Geng, Yu Wang	195

Trusted Computing Infrastructure and PKI

Michiharu Kudo

Tokyo Research Laboratory
IBM, Japan
E-mail: kudo@jp.ibm.com

ABSTRACT

The Trusted Computing Group (TCG) has been standardizing a hardware-rooted "health check" of computing platforms (called Remote Attestation). The trusted computing infrastructure defined by the TCG relies on PKI technologies with several interesting extensions. Both the trusted computing and PKI technologies are required to establish a more secure and trusted computing infrastructure. The TCG has standardized a broad range of specifications from hardware to software, though there still remain many unresolved problems and open issues. This paper reviews trusted computing standards and describes where trusted computing infrastructure and PKI come into play.

INTRODUCTION

Currently, security incidents are frequently reported, such as information leakage of sensitive personal data, swindles based on phishing sites, password leakage by key-loggers, zero-day attacks, and so on. Why do these incidents happen so repeatedly? A typical root cause is the installation of malicious software which is not recognized by users as being malicious. For example, an ignorant user may invoke an executable file attached in email sent from a malicious Web site, which installs a malicious key-logger into the operating system, and which then stealthily transmits any typed passwords to a remote site.
Why is the Public Key Infrastructure unable to prevent this from happening? The PKI technology provides reliable ways of authenticating users and servers and of verifying digital signatures. For example, Java supports a code-signing scheme so that the users can verify the integrity and authenticity of a Java application before executing it, and determine with confidence whether or not the program is safe. However, user-based involvement like this often fails because typical users do not have sufficient knowledge about the trust and security of IT artifacts. Beyond PKI, various kinds of security software are available these days, ranging from intrusion detection systems and personal firewalls to virus detection tools that also protect a user's computing platform from malicious attacks by such malware as BotNet installers. However, security tools based on software-rooted trust are sometimes not effective against smart viruses and self-evolving malware that can sneak through the protective mechanisms provided by the security tools. These problems are becoming more and more critical since end-user devices such as client PCs, PDAs, and smart phones are playing important roles in the handling and storage of sensitive business data.
The Trusted Computing Group (TCG) [10] focuses on these problems using a completely different and innovative approach that did not previously exist in the PC industry. The TCG's trust model is hardware-rooted, based on the existence of a specific security chip embedded in the computing platform. By using hardware-rooted trust, the TCG technology makes it possible to check the health of a computing platform (using a technique called Remote Attestation) and can alert the user whenever the platform is contaminated by unknown software. The trusted computing infrastructure relies on PKI technology with several extensions. Both the trusted computing and PKI technologies are needed to establish a more secure and trusted computing infrastructure. The TCG has standardized a broad range of specifications including hardware, software, and usage scenarios. However, there still remain many unresolved problems and open issues. In this paper, I review the TCG activities and related specifications along with the technical challenges.

TRUSTED COMPUTING GROUP

The Trusted Computing Group (TCG) is a not-for-profit organization which was established in 2003 to develop, define, and promote open standards for hardware-enabled trusted computing and security technologies [11]. The TCG Promoters and board members include AMD, Hewlett-Packard, IBM, Intel Corporation, Microsoft, Sony Corporation and Sun Microsystems. The primary goal of the TCG is to help users protect their information assets from compromise due to external software attack and physical theft. TCG adopted the specifications of Trusted Computing Platform Alliance (TCPA) which was formed in 1999. As of October 2007, 139 companies from hardware manufacturers to solution vendors are participating in the TCG.
One of the unique technologies of the TCG is the Trusted Platform Module (TPM), a security chip embedded in the hardware platform, e.g. the motherboard of a PC, to support hardware-based platform security. Using the TPM, a trust chain is built up from the TPM security chip, as a Root of Trust, to the upper software stack such as the BIOS, the OS, and the applications, thus providing a secure computing platform that can protect IT assets from various kinds of software attacks.
In the TCG, "Trust" is defined as the expectation that a device will behave in a particular manner for a specific purpose. A trusted platform should provide at least three basic features: protected capabilities, integrity measurement, and integrity reporting. The crucial protected capabilities in the TCG are a set of commands with exclusive permission to access shielded locations, memory, registers, etc., where sensitive data can be guaranteed to be handled safely.

Working Group and Scope

The TCG is not only standardizing the Trusted Platform Module (security chip) but also specifications for a broad range of platforms such as personal computers, mobile phones, and personal digital assistants where reliable computing environments are needed. These specifications include the TPM software stack and its interface to be called from application programs, network protocols to exchange platform configurations, and a reference architecture. In the TCG, there are several working groups (WGs) under the Technical Committee, such as the TPM WG, the TPM Software Stack WG, the Mobile Phone WG, and the Infrastructure WG.

Fundamental Features of TPM

The TPM implements the protected capabilities and shielded locations (called Platform Configuration Registers or PCRs) used to store and protect the integrity measurements. The TPM-protected capabilities include additional security functions such as cryptographic key management and random number generation. Attestation is the process of vouching for the accuracy of certain information so that external entities can attest to the contents of shielded locations, the proper execution of protected capabilities, and the integrity of the Roots of Trust. A platform can attest to its own descriptions of platform characteristics that affect the integrity (trustworthiness) of the platform. All forms of attestation require reliable evidence of the integrity of the attesting entity. Integrity Measurement as defined in the TCG is the process of obtaining metrics of platform characteristics that can prove the integrity (trustworthiness) of a platform and of putting digests of those metrics into the PCRs. The starting point of measurement is called the Root of Trust for Measurement (RTM). A static root of trust for measurement begins measuring from a well-known starting state such as the power on self-test state. Attestation and integrity measurement are described in detail in the next section.

Remote Attestation

The Remote Attestation concept supported by the TCG consists of two trust anchors. One is a new Platform Root of Trust consisting of the security chip TPM and write-protected initial boot-up code, the Core Root of Trust Measurement (CRTM). The other anchor is a PKI. The remote attestation becomes possible only after successfully extending the Trust Chain that begins from these Trust Anchors to a certain level of the software stack of the target system. This section describes the basic scheme of Remote Attestation.
The distributor of the binary image generates a hash value for the components, and signs them to prove the authenticity of their origin using a Reference Integrity Measurement Manifest (RIMM). In a platform which supports Transitive Trust, the integrity of each component is measured at the time of its execution, and recorded in one of the TPM's Platform Configuration Register (PCR). The platform also retains this event in

its Stored Measurement Log (SML). The integrity of the measured component is listed in the SML and the integrity of the SML is protected by a PCR in the TPM chip. The platform has an Attestation Identity Key (AIK) with certification. The integrity information stored in the PCR is sent to a verifier with a signature created by the AIK, as an Integrity Report. The AIK is a special key used solely for TCG attestation. The AIK credential is issued by a CA after justification by the Endorsement Key (EK) Credential of the TPM and the Platform Credential and this proves that the AIK belongs to a genuine TPM. The Remote Verifier verifies the signature and compares each of the component hash values with the RIMM, and finally checks the proper state of the target platform.

PKI AND TRUSTED COMPUTING INFRASTRUCTURE

The PKI provides the fundamental security infrastructure for user authentication, server authentication, and digital signature verification. The server authentication using SSL is one of the technologies used to validate the authenticity of a server platform. For example, users who connect to a server via SSL can be certain that the server's identity is authenticated by some Certificate Authority (CA) or a Registration Authority (RA), if the server's certificate is valid when checked against the CA policy. Note that the server authentication does not necessarily mean that the server is correctly configured or that the server will behave as the users expect. There are several possible problems: First, the issued server certificate could have been copied to other servers so that it is not possible to identify a specific server platform instance. Second, the server certificate has nothing to do with the integrity of the platform configuration. For these reasons, the PKI by itself cannot be used as a platform attestation.
The TCG's trusted computing infrastructure requires PKI as an indispensable component of its infrastructure. We can explain how PKI is incorporated in the trusted computing infrastructure in relation to the asset management lifecycle. In the provisioning phase, a TPM chip manufactured by a TPM vendor is assembled with the PC H/W components. An EK and an EK credential are created by the TPM vendor and a platform manufacturer may issue a Platform Credential to guarantee that the platform hardware is authentic and correctly implemented as defined in the TCG specifications. In the deployment phase, the owner of the platform takes ownership (requiring initialization of the platform and the TPM). An AIK and AIK credential are issued by the Privacy CA (described later). For retirement (or redeployment), the keys and credentials are erased, migrated, or backed-up according to the asset's disposition. The data formats of the AIK, EK, and platform credentials are X.509 certificates. For the EK credential, a new TPM specification field consists of the Family, Level, and Revision values.

Privacy Certificate Authority

The Endorsement Key (EK) is an RSA public key pair which is tightly coupled with each TPM chip and embedding platform. Therefore it is possible to recognize each platform using the Endorsement Key's unique values if the Endorsement Key pair is used in the remote attestation process. This linkability property poses a privacy threat, because the identity of a platform could be easily discovered by unknown or unauthorized entities. For this reason, the TCG introduced platform identity aliases, known as Attestation Identity Keys (AIKs) that can be associated with the information relating to a specific use or domain. First a new Attestation Identity Key pair is created by the TPM and then the Privacy Certificate Authority (Privacy CA) issues an AIK credential as requested (after verifying the EK credential and the platform credential that are included in the request). Since the AIK credential contains no information about the EK credential, the remote attestation process does not reveal the identity of the platform. The Privacy CA is trusted not to reveal sensitive information including the public Endorsement Key or Personally Identifiable Information. It is also trusted not to misrepresent the trust properties of platforms for which AIK credentials are being issued.

Direct Anonymous Attestation

The ideas for the Privacy CA are included in the TCG specifications, but there are several concerns when it is deployed in the real world. The first concern is that the platform needs to get an Attestation Identity Key credential from the Privacy CA whenever the user of the platform wants privacy from relying parties. In other words, a platform might request AIK credentials from the Privacy CA thousands of times to insure the

maximum privacy of the platform. If that happened, then the Privacy CA could become a bottleneck for the remote attestation process. Second, the Privacy CA should be a trusted third party, and not controlled by any of the parties who have a vested interest in the integrity test result. To address these concerns, Direct Anonymous Attestation (DAA) has been proposed [1] and included in the TPM specification version 1.2. The key idea of DAA is not to provide a certificate to the relying party who is seeking to verify the integrity but to use a zero-knowledge cryptographic proof to show that the user has a corresponding AIK signature key. The DAA assumes the existence of a trusted third party called the DAA Issuer, who provides a signature on a DAA key for the requesting platform. One of the advantages of the DAA approach is that the platform does not need to trust the DAA Issuer that is protecting the privacy information of the AIK, not even the DAA Issuer can link the AIK to a particular instance of an EK. In addition, even when seeking maximum privacy, the platform no longer needs to connect to the Privacy CA after it receives the signature on the DAA Key. Therefore the DAA approach minimizes the scalability problem.

Subject Key Attestation Evidence

The TCG defines Subject Key Attestation Evidence (SKAE) that stores the TPM-relevant information in an extension of the standard X.509 certificate [12]. The SKAE extension makes possible more secure verification of the public key used by the system (e.g. a key for server authentication), to make sure that the key was generated and managed by a TPM chip. With the SKAE extension, the X.509 certificate can store the information about the TPM key structure and the information on how to access the authority of the AIK credential. One typical use scenario is that the public-key pair for server authentication would be generated and managed by a TPM chip and the corresponding server certificate would include the SKAE extension. The client would be assured that the possibility of tempering with the server's public key was very low.

Available Implementations

OpenTC [9], a trusted infrastructure research project in the European Union, has developed a Privacy CA implementation that covers AIK lifecycle management [8]. The DAA tool is also available from the IBM alphaWorks site [7].

OPEN ISSUES

There are still many issues to address in performing remote attestation of platforms in the trusted computing infrastructure. The following is a selected (non-exhaustive) list of problems related to PKI:

1. Most TPM chip manufactures do not issue EK Credentials.
2. No commercial Privacy CA yet exists (and thus no AIK Credentials are available)
3. No platform vendors yet issue Platform Credentials

As far as the author knows, only the German TPM chip manufacturer Infineon issues the EK Credentials. VeriSign certifies Infineon's Trusted Platform Module (TPM) CA using its Trusted Computing Root CA [13]. Since EK credentials are critical to implement TCG remote attestation, every TPM chip manufacturer should issue EK credentials. Second, since there is no commercial Privacy CA available at this moment, it is impossible to get Attestation Identity Key credentials. In addition, no platform vendor is issuing Platform Credentials for the platforms they manufacture. These three obstacles need to be removed to allow for remote attestation in the client PC industry segment.

Related Work

Several papers propose practical schemes for Remote Attestation. Munetoh [3] proposed a novel way of checking the integrity on Linux that broadens the applicability of remote attestation to a Commercial Off-the-Shelf (COTS) operating environment. This approach uses several patterns of integrity measurement which enable efficient verification of the COTS components. Sailer et al. [5] proposed the Integrity Management Architecture (IMA) to enable generic integrity measurement of software modules for each object module. The measurement scheme is called binary attestation, in which a set of the measured hash

values for each object module is compared with a set of pre-calculated correct hash values, resulting in binary values of correct or incorrect. Sadeghi and Stüble [4] and Haldar et al. [2] proposed another approach called property-based attestation that gives not only binary results, but also semantic results such as fully compliant with company policy or SOX compliant. Yoshihama et. al. [6] proposed a fine-grained attestation mechanism on top of WS-Security technologies, as well as infrastructure support for attestation validation.

CONCLUSION

This paper reviews the TCG standardization activities and some of the technical specifications. TCG's trusted platform infrastructure has opened a new path to securing computing platforms by using the platform identity, integrity measurements of the platform, and secure storage for sensitive data. There are still many open issues, but in the near future the combination of hardware-rooted platform attestation and PKI-based authentication technologies will create a more secure and trusted computing infrastructure for future IT environments.

Reference

[1] Camenisch J. (2004) Better Privacy for Trusted Computing Platforms. In *Proceedings of 9th European Symposium on Research in Computer Security* (ESORICS 2004)
[2] Haldar V., Chandra D., and Franz M. (2004) Semantic remote attestation - a virtual machine directed approach to trusted computing. In *3rd Virtual Machine Research and Technology Symposium*, May
[3] Munetoh S. (2006) Practical Integrity Measurement and Remote Verification for Linux Platform, *Workshop on Advances in Trusted Computing*, 2006.
[4] Sadeghi A. and Stüble C. (2004) Property-based attestation for computing platforms: Caring about properties, not mechanisms. In *2004 Workshop on New Security Paradigms*, pages 67–77.
[5] Sailer R., Zhang X., Jaeger T., and van Doorn L., (2004) Design and Implementation of a TCG-based Integrity Measurement Architecture, *13th Usenix Security Symposium*, California, August
[6] Yoshihama S., Ebringer T., Nakamura M., Munetoh S., Mishina T., and Maruyama H., (2007) WS-Attestation: Enabling Trusted Computing on Web Services, in *Test and Analysis of Web Services*, Springer
[7] IBM Direct Anonymous Attestation Tools, http://www.alphaworks.ibm.com/tech/daa
[8] OpenTC PKI, http://opentc.iaik.tugraz.at/index.php?item=pca/pcaover
[9] Open Trusted Computing, URL: http://www.opentc.net/
[10] Trusted Computing Group, http://www.trustedcomputinggroup.org/
[11] TCG Infrastructure Workgroup (2007), Specification Architecture Overview Specification Revision 1.4, https://www.trustedcomputinggroup.org/groups/TCG_1_4_Architecture_Overview.pdf
[12] TCG Infrastructure Workgroup (2005), TCG Specification Architecture Overview, *Subject Key Attestation Evidence Extension*, Specification Version 1.0, Revision 7, 16 June 2005
[13] VeriSign and Infineon Collaborate to Increase Usability of Trusted Computing Solutions, http://www.verisign.com/verisign-inc/news-and-events/news-archive/us-news-2005/page_036159.htm

New Novel Approaches for Securing VoIP Applications

Chan Yeob Yeun, Kyusuk Han, and Kwangjo Kim

Information and Communications University (ICU)
119,Munjiro,Yuseonggu,Daejeon, 305-732, Korea
{cyeun, hankyusuk, kkj}@icu.ac.kr

Abstract. SIP message authentication and SRTP key agreement are the important issue in the SIP-based VoIP service. Several secure solutions such as HTTP Digest Authentication, SSL/TLS, and S/MIME, are used for the SIP message authentication and key agreement. When the VoIP is used in the wireless environments, the efficiency of security service is one of the important matters in question. For such efficiency, WPKI is a better substitution than the traditional PKI, while it still requires the effort on the certificate management. Therefore, we would like to propose efficient ID-based cryptosystem for the VoIP in the wireless environments. In this paper, we present the overview of WPKI and the application of ID-based cryptosystem for the SIP message authentication as well as the authenticated one-way key agreement for SRTP. Our novel design reduces delaying for the key generation and provides the explicit mutual authentication.

1 Introduction

Voice over internet protocol (VoIP) is becoming more common and widely used everywhere, where the various security shortcomings are frequently incurring: Session initiation protocol (SIP) [16] message forgery during SIP transaction and eavesdropping Secure real-time transport protocol (SRTP) [3] packet are critical security problems in the SIP based VoIP services.

Currently, HTTP digest authentication between VoIP user and servers, SSL/TLS among servers, and S/MIME for the message authentication are the solutions for the security of VoIP services.

There are several approaches that consist of the SIP message authentication are shown in the VoIP systems as follows. At first, RFC 4474 [14] defines the VoIP server of user side signs the SIP message, when users send their SIP messages to the VoIP server, the server sign the messages. Users do not provide the security of SIP message. However there are too much overhead in the server with the large number of SIP transactions.

The second approach is signing by users themselves. In this case, users ought to posses the enough computational power with the certificate management. In addition, Kong *et al.* [8] proposed the scheme that users create their own public key pairs and the servers share the information of the public key.

Since the construction of traditional PKI [1] has too much communication overhead in the wireless environments, we would like to consider the WPKI (wireless PKI) [11, 17] as the more efficient way to utilize PKI. However, even WPKI gives the significant efficiency, the overhead from the certificate management still remains.

Therefore, we consider the certificate-less environments with the employment of ID-based cryptography. In 2006, Ring *et al.* [15], proposed the authentication and key agreement schemes for the VoIP employing ID-based cryptography. Their design is based on two-pass key agreement protocol with signatures and it takes relatively much time for verifying the signature in ID-based cryptography that may o Θ ccur the delay in key generation in their design.

In this paper, we present the overview of WPKI and the application of ID-based cryptosystem for the SIP message authentication as well as the authenticated one-way key agreement for SRTP. Our novel design reduces delaying for the key generation and provides the explicit mutual authentication.

2 Related Works

2.1 VoIP security

For the authentication, SIP presently uses *HTTP digest authentication* [7], which does not provide message integrity, end-to-end security, and has lack of scalability to multi-domain because of the shared user password based model.

Secure/Multipurpose Internet Mail Extensions (S/MIME) [2] is a protocol that adds digital signatures and encryption to Internet MIME (Multipurpose Internet Mail Extensions) messages described in RFC 1521 [5]. SIP allows sections of the messages to be encrypted using S/MIME, however S/MIME is dependent upon a Certificate Authority (CA) and accompanying Public Key Infrastructure (PKI), and therefore limited by the adoption of such a system. Also, it is possible that S/MIME is likely to be too heavy for resource constrained handsets.

The model in the RFC 4474 [14] defines the server signs user address binding and contact address with own domain certificate. Please see Figure 1 for more detail. In this model, users do not have to keep their own certificate and allow user's message authentication in the outside of the user domain. In this case, the public key is not used by every user so that the delegation of signature generation is required for the practical solution.

However, the message signing in the environment with the large number of user will be the server's overhead. When a great number of transactions happen, the server might be vulnerable against DoS attack. Furthermore, the computational power of mobile devices is continually being improved.

Kong *et al.* [8] proposed the model that users sign their own SIP messages with their public keys. Users self-generate public key pairs and register them to their registered VoIP server, and sign the SIP message with the private key. In the mobile environments, generating public key pairs and registering them to servers will be the computational overhead.

They showed their model is efficient because of the overhead from the message signing is distributed to each user. However, their model still has the overhead from the public key registration to all servers. Since the public key pairs are self-generated by each user, the cost to register the public key pairs to all servers should not be ignored.

Generic public key cryptosystem requires the verification of the public key in the certificate, and the communication with the trusted third party (TTP), whom the servers role in [8].

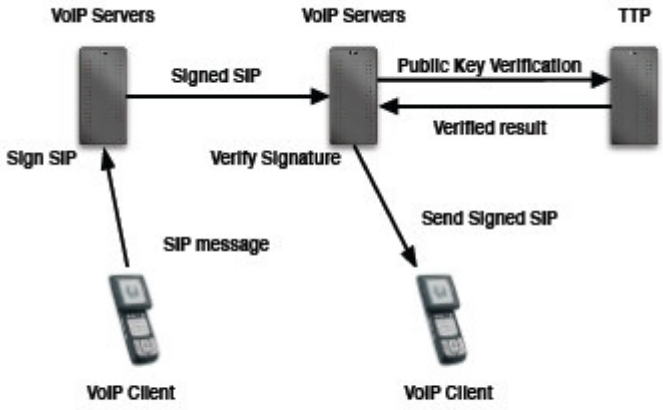


Fig. 1. Server signs SIP message (in RFC 4474)

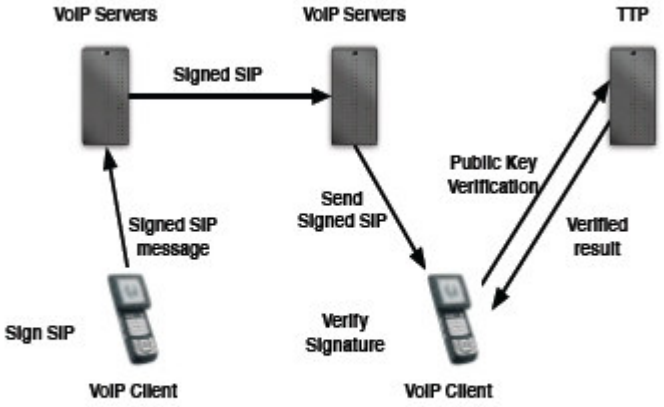


Fig. 2. User signs own SIP message (Generic PKI)

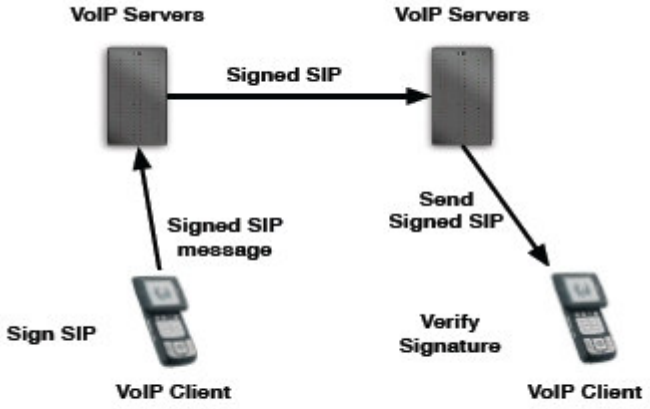


Fig. 3. User signs own SIP message (ID-based cryptography)

Also, each user has to manage other user's public key.
In this paper, we propose an e˙cient and practical secure VoIP service with applying ID-based cryptography. We address the combination of the signature scheme in [6] and the key agreement scheme in [10]. With such combination, we achieve the one way authenticated key agreement for the SRTP as well as the message integrity and authentication for the SIP. Using ID-based cryptosystem, user has the benefit from the removing of public key verification.
Also we would like to discuss Ring *et al.* [15]'s design with our proposed protocol.

2.2 WPKI

Generally, WPKI is known to be more efficient than traditional PKI. [9] showed the implementation result of RSA and ECDSA in the mobile phone. From their results, the performance of ECDSA showed about 5 times faster than RSA. The key size was 163 bits for ECDSA, while 1024 bits required for RSA to achieve the same security level. Thus the following section describe briefly about wireless PKI. Please refer [17] for more details.

TCP/IP and PKI are computationally intensive solutions, also incur a large communication overhead, which are undesirable in wireless environments. Nevertheless, the basic elements of PKI and certificate remain the same. Also, it is trivial that most VoIP applications will be employed in the wireless environments, which brings the requirements of more efficient way.

The wireless PKI (WPKI) can be used for the same applications as those with PKI. However, the characteristics of the wireless environment can give rise to the development of a whole new set of revolutionary applications including banking, payments, ticketing and receipt, stock trading, gambling and public administration. Compared to a PKI, WPKI applications have to work in an environment with less powerful CPUs, less memory, restricted power consumption, smaller displays, and diverse input devices [11–13]. Figure 4 shows the WPKI architecture [9]. The enhancements of WPKI are described as follows.

WPKI Protocols. The traditional method used to handle PKI service requests relies on the ASN.1 Basic Encoding Rules (BER) and Distinguished Encoding Rules (DER). BER/DER requires more processing resources than a WAP device should effectively have to handle. WPKI protocols are implemented using WML 2.0 and WMLSCrypt. WML 2.0 and SignText function in WMLSCrypt provide for significant savings when encoding and submitting PKI service requests as compared to the methods used in traditional PKI.

WPKI Certificate Format. The WPKI certificate format specification sought to reduce the amount storage required for a public key certificate. One of the mechanisms was to define a new certificate format for server side certificates, which significantly reduces the size as compared to a standard X.509 certificate. Another significant reduction in the WPKI certificate can be attributed to Elliptic Curve Cryptography (ECC). With ECC, the saving in the overall size of the certificate is typically more than 100 bytes due to the smaller keys needed for ECC vs. other signature schemes. WPKI has also limited the size of some of the data fields of the IETF PKIX certificate format. Because the WPKI certificate format is sub-profile of the PKIX certificate format, it is possible to maintain interoperability between standard PKIs.

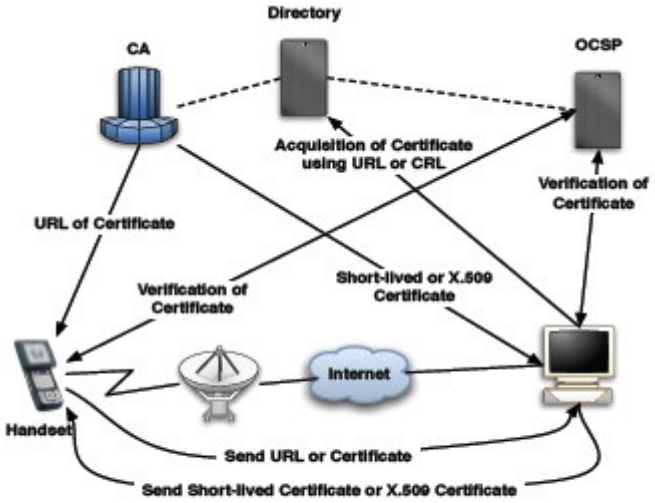


Fig. 4. Wireless PKI Architecture [9]

WPKI Cryptographic Algorithms and Keys. While traditional signature schemes are optionally supported by the WAP security standards, they are viewed as impractical to implement in the wireless environment from a performance and resource viewpoint. Traditional signature schemes demand much more processing, memory, and storage resources in the WAP device when compared to the resource requirements of more efficient cryptographic-ECC.
ECC techniques are recognized as the most optimized, and therefore the best suited for supporting security in the wireless environment. The keys for elliptic curve are typically of the order of six times smaller than equivalent keys in other signature schemes, for example 164 bits vs. 1024 bits. This creates great efficiencies in key storage, certificate size, memory usage and digital signature processing. ECC is fully supported by the WAP security standards and has been widely accepted by WAP device manufacturers. However, one must carefully choose good Elliptic Curves otherwise it might be prone to various attacks.

WPKI is an extension of, and includes most of the technologies and concepts that are present in traditional PKI. WPKI must be optimized using more efficient cryptography such as ECC but one must carefully select the good curves in order to prevent known attacks. One of the issues with getting WPKI widely accepted is the management of certificates via CA's.

By using WPKI with VoIP applications, we are able to provide swift key agreement protocol with signing and verifying.

2.3 ID-based signature scheme

In this section, we describe the signature scheme used for our model, which is based on the scheme 1 in [6].

At first, we define $h : \{0, 1\}^* \times V \rightarrow (Z/lZ)^x$, $H : \{0, 1\}^* \rightarrow G^*$, where $G^* := G \setminus \{0\}$.

ID-based signature scheme consists of 4 algorithms, *Setup*, *Extract*, *Sign*, and *Verify*, and 3 entities, the trusted authority (TA), the signer, and the verifier.

Setup: TA select a random integer $t \in (Z/lZ)^x$, computes $QTA = tP$, where t remains secret. And then, TA publishes QTA.

Extract: The signers request own private keys $SID = tH(ID)$ to TA, where ID is signers' identities.

Sign: To sign the SIP message m The signer selects arbitrary length $P1 \in G^*$ and a random integer $k \in (Z/lZ)^x$, and computes followings;

1. $r = e(P1, P)^k$
2. $v = h(m, r)$
3. $u = vS_{ID} + k_{P1}$

Verify: The verifier receives the message m and the signature (u, v), computes followings;

1. $r = e(u, P) \vee e(H(ID), QTA)^v$
2. Accepts if and only if $v = h(m, r)$

2.4 ID-based Key Agreement Scheme

Assume two entities A, and B who exchange the key, where A requests the key exchange. Key agreement methods are defined as the following forms. *Non-interactive* is the method that A pre-shares the key for each entity. A encrypts the self-generated session key using the pre-shared key with B and sends the encrypted session key to B. However, there is claim that the session key is controlled by A and at least one communication is required, which is no more *non-interactive*.

The other way is *Two-pass* method, which A and B mutually exchange key generating information. [15] is based on the two-pass key agreement protocol.

Another way is *One-way* method, which A sends key generating information and encrypted message using the session key to B at the same time. In this model, the communication is required only once. We conclude the one-way method is practical considering the cost and security.

In 2006, Ring *et al.* [15] showed the two-pass key agreement model, which is shown in Figure 5.

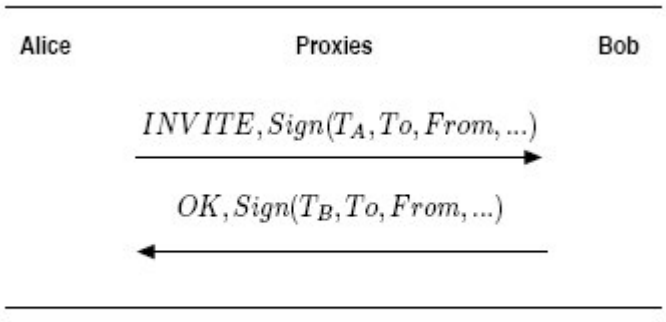


Fig. 5. Ring *et al.*'s Key Agreement Model for SIP [15]

To reduce the delay from computing the session key used for SRTP encryption, we propose the one-way key agreement model. The example is shown in Figure 6.

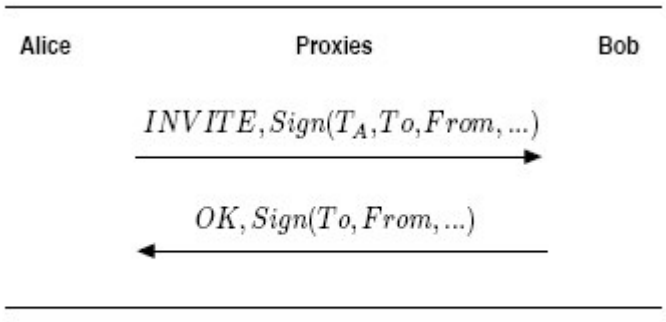


Fig. 6. Our proposed Key Agreement Model for SIP

Figure 7 shows the comparison of our one-way key agreement and two-pass key agreement [15] employing in VoIP.

As shown in Figure 7, Alice can pre-compute the session key when she send the **INVITE** message to Bob. When Alice and Bob agreed to the session key and send SRTP transaction, they can reduce the delay, which is shown in two-pass model. In two-pass key agreement model, Alice can compute the session key after Bob responds with **OK** message. In practical VoIP application, employing our model, the delay is reduced.

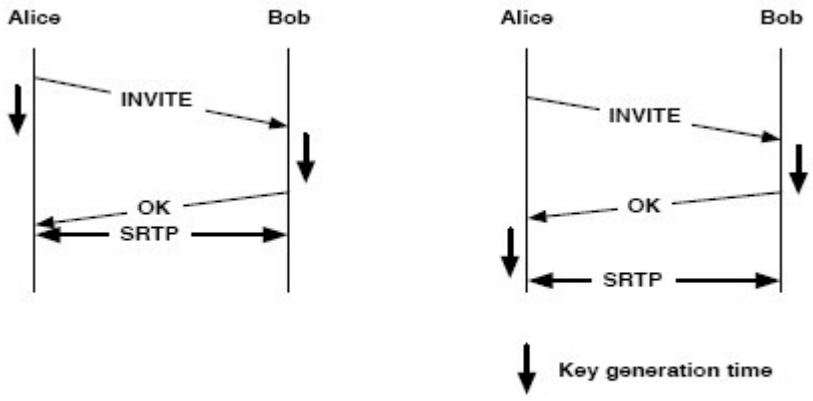


Fig. 7. Comparison of one-way and two-pass key agreement

For the one-way key agreement protocol, we apply the scheme 1 in [10], which is one-way method. The protocol is as following.

We assume two entities A and B in the protocol. SA denote the private key of A, which is $sH(IDA)$, where s is the master secret of KGC. $H(IDA)$ is A's public key, where H is the hash function, $H : \{0, 1\}^* \rightarrow G1$, $G1$ is additive cyclic group. IDA is the identity of A. Please refer to [4] for the bilinear maps from elliptic curve pairings.

Parameter Distribution: A selects a random integer $r \in Z^*_q$ and computes $XA = rH(ID_A)$.
A sends X_A to B via the public channel.

Established Key: A and B computes followings;
– A: $k_{AB} = e(S_A,H(ID_B))r \ \Theta \ e(S_A,H(ID_B))$,
– B: $k_{BA} = e(X_A, S_B) \ \Theta \ e(H(ID_A), S_B)$.
Θ denotes XOR operation.

3 Proposed Scheme

We assume a sender A, a receiver B, and a server in a certain VoIP service. The sender and the receiver generate messages for VoIP service, as clients, while the server provides VoIP service.

To generate SIP message, the sender (denote A) generates followings;
– $r = e(P_1, P)^k$
– $t = H^*(r) \bullet H(ID_A)$
– $v = h(m, t)$
– $u = vd_A + kP_1$
Here, $h : \{0, 1\}^* \times G_1 \rightarrow (Z/lZ) \times$, $H : \{0, 1\}^* \rightarrow G_1$, and others follow [6]. To generate t, r should be transformed from elliptic curve to finite fields. H^* is a *map-to-point* hash function, which $H^* : G2 \rightarrow \{0, 1\}$. To compute with $H(ID_A)$, the transformation is necessary.
$e : G_1 \times G_1 \rightarrow G_2$. G_1 is cyclic additive group; generated by P with order q. $G2$ is cyclic multiplicative group with the same prime order q. d_A denotes A's private key, $d_A = sH(ID_A)$. m is the SIP message, which includes the sender's address, the receiver's address, message generated time, Session Description Protocol (SDP) and other necessary information.

Then, A sends $(u,\ v) \in (G,\ (Z/lZ)^{x})$ to the receiver B.

After receiving $(u,\ v)$, B generates the following.
$t = H^*(r) \bullet H(ID_A) = H^*\ (e(u,\ P) \bullet e(H(ID_A),\text{-}sP\ )^v) \bullet H(ID_A)$.

After that, A and B generate the session key simultaneously.
– $A : k_{AB} = e(d_A,H(ID_B))^{H^*(r)}\ \Theta\ e(d_A,H(ID_B))$.
– $B : k_{BA} = e(t,\ d_B)\ \Theta\ e(H(ID_A),\ d_B)$.
The correctness of $k_{AB} = k_{BA}$ follows,
$k_{AB} = e(d_A,H(ID_B))^{H^*(r)}\ \Theta\ e(d_A,H(ID_B))$
$= e(H(ID_A),H(ID_B))^{H^*(r)s}\ \Theta\ e(H(ID_A),H(ID_B))^s$
$= e(td_B)\ \Theta\ e(H(ID_A)),\ d_B)$
$= k_{BA}$

Therefore, r can be used for both SIP message signature and the key generation, which reduces the additional communication only for the key generation.

Θ is the additive operation in G_2. When the hash function $H\prime : G2 \rightarrow \{0,\ 1\}$ is used, Θ can be XOR operation in $k_{BA} = H'(e(t,\ d_B))\ \Theta\ H'(e(H(ID_A),\ d_B)$.

4 Security Analysis

We describe the security analysis for our secure VoIP design as follows.

4.1 Security in SIP message authentication

The security in SIP message authentication is the same as the security in [6]. When the attack is succeeded, the Diffie-Hellman problem is solved. However, the DH problem is known as the mathematical hard problem. It is also secure against Man-in-the-middle attack due to the explicit digital signature scheme is applied. Therefore, we also achieve the authenticated one-way key agreement.

4.2 Security in SRTP key generation

For the key generation protocol in [10] is followings.
– **Known-key security** The session key in each session should be independent. When the session key is leaked, it should not threat the other session keys.
– **Unknown key share** When A and B exchange the session key, The other entity C is not exchanging the key.
– **Key control** No entity should not use the previous parameter for the session key.
– **Sender's key-compromise impersonation** When the private key of A is leaked, the attacker can impersonate A, but not other entities.
– **Sender's forward security** When A's private key is leaked, the security of previous session has guaranteed.
– **Random number compromise security** The leakage of the certain parameters selected by A doesn't affect to the leakage of A's private key or session key.

Known-key security To generate r, where $r = e(P_1,\ P)^k$, the sender randomly choose P_1 and k in each session. The leakage of P_1 or k doesn't affect the previous session.

Unknown key-share To generate to key the receiver B verifies the signature of the sender A first. Also, the sender self-generates the session key without any information from the receiver. Therefore, any other entities except A and B cannot exchange the key. To succeed the attack, the adversary should be able to generate the signature of A or know the private key of B.
Key control since the key generating parameter t is selected by A, and the process is done in one-way, B cannot control the session key, also it is difficult for A to pre-compute the random integer r and the generator P_1 to control t.

Random number compromise The random integer r is easily known from (u, v). However it is difficult know A and B's private keys or session key from public parameters P, sP, and r. To attack the session key, The knowledge of A or B's private key is necessary. The success of attack with P, sP and r is the same as the success of attack on the signature.

Attacks on sender When A's private key is leaked, the adversary can impersonate A, since r is known to A, while it is not possible to impersonate other entity. However, sender's forward security is not guaranteed unlike [10], since r is sent with the signature.

4.3 Efficiency

Signature generation requires one exponentiation operation in G_2, two hash operations, two multiplications in G_1. Verification requires one exponentiation operation in G_2, two pairing operations, and one multiplication operation. When the several messages are sent by the same identity, the sender pre-compute $e(H(ID),-sP)$ to reduce one pairing operation. For the key generation, one pairs operation of the sender, one multiplication over elliptic curve, one exponentiation operation, and two pair operation of the receiver.
When we apply to SIP message, two exponentiation, three multiplications, two pairs in the sender side, three pairs and two exponentiation operation in the receiver side.
Using one-way key agreement with signature, we can reduce the delay using two-pass key agreement.

5 Conclusion

In this paper, we suggested new approaches by using WPKI and proposed the efficient and practical method for the SIP message authentication with signature and authenticated one-way key agreement for SIP-based VoIP service with ID-based cryptosystem. In conclusion, our new approaches can reduce the cost for the public key management and additional process for the key generation with re-using the parameter for the signature verification.

References

1. Public-key infrastructure (x.509) pkix. http://www.ietf.org/html.charters/pkix-charter.html.
2. S/mime mail security (smime). http://www.ietf.org/html.charters/smime-charter.html.
3. M. Baugher, D. McGrew, M. Naslund, E. Carrara, and K. Norrman. The secure real-time transport protocol (srtp), March 2004.
4. D. Boneh and M. Franklin. Identity-based encryption from the weil pairing. In *SIAM Journal on Computing*, volume 32, pages 585{615, 2003.
5. N. Borenstein and N. Freed. Mime (multipurpose internet mail extensions) part one: Mechanisms for specifying and describing the format of internet message bodies. RFC 1521, September 1993.

6. Florian Hess. E_cient identity based signature schemes based on pairings. *9th Annual International Workshop, SAC 2002*, 2595/2003:310{324, Jan 2003.
7. J.Frank, P Hallam-Baker, J Hostetler, S Lawrence, P Leach, A Loutonen, and L Stewart. Http authentication: Basic and digest access authentication. RFC 2617, 1999.
8. Lei Kong, Vijay Arvind Balasubramaniyan, and Mustaque Ahamad. A lightweight scheme for securely and reliably locating sip users. *The 1st IEEE Workshop on VoIP Management and Security (VoIP MaSe 2006)*, 2006.
9. Yong Lee, Jeail Lee, and JooSeok Song. Design and implementation of wireless pki technology suitable for mobile phone in mobile-commerce. *Comput. Commun.*, 30(4):893{ 903, 2007.
10. Takeshi Okamoto, Raylin Tso, and Eiji Okamoto. One-way and two-party authenticated id-based key agreement protocols using pairing. *Modelling Decisions for Artificial Intelligence, Second International Conference, MDAI 2005, Tsukuba, Japan, July 25-27, 2005. Proceedings*, 3558/2005:122{133, 2005.
11. OMA. Wireless application protocol - wireless public key infrastructure. WAP-217-WPKI, April 2001.
12. OMA. Wireless application protocol architecture specification. WAP-210-WAPArch, July 2001.
13. OMA. Wireless transport layer security. WAP-261-WTLS, April 2001.
14. J. Peterson and C. Jennings. Enhancements for authenticated identity management in the session initiation protocol (sip). RFC4474, August 2006.
15. J Ring, KR Choo, E Foo, and M Looi. A new authentication mechanism and key agreement protocol for sip using identity-based cryptography. Proceedings AusCERT Asia Pacific Information Technology Security Conference 2006.
16. J. Rosenberg, H. Schulzrinne, G. Camarillo, A. Johnston, J. Peterson, R. Sparks, M. Handley, and E. Schooler. Rfc 3261, "sip: Session initiation protocol", June 2002.
17. Chan Yeob Yeun and Tim Farnham. Secure m-commerce with wpki. In *In Proceedings of IWAP'01*, pages 171{183, Daejeon, Korea, October 2001.

Survey of RSA Implementations and Attacks

Srinivasa Rao Subramanya Rao

Mathematical Sciences Institute
Australian National University, Australia,
E-mail: subrmnya@maths.anu.edu.au

Abstract: We survey a few RSA implementations or variants and give some insights into the security aspects of these implementations. We also survey a few non-mathematical attacks and motivate the effectiveness of these attacks against RSA implementations.

Keywords: RSA Variants, Multiprime RSA, Batch RSA, Rebalanced RSA, Twin RSA, Dual RSA, Side Channel attacks, Fault attacks

INTRODUCTION:

RSA is one of the most widely deployed public key cryptosystems in the world today. Factors that have contributed to this wide deployment are (i) RSA is based on factorization, a widely studied problem and (ii) the inherent simplicity of the system. This popularity has resulted in RSA being adapted to many different environments, thus giving rise to many different implementations or variants proposed in the literature.

This paper surveys some RSA variants. RSA variants can be broadly classified into (i) Variants that speed up the encryption or/and decryption operations (ii) Variants that reduce storage requirements and (iii) Variants that provide a very high level of security.

Variants that speed up encryption or decryption operations include Batch RSA, Multi-prime RSA, Rebalanced RSA, RSA-Small-e, RSA-Small-d and Generalized-Rebalanced RSA. Variants that facilitate reduced storage include Twin RSA, while implementations such as Unbalanced RSA aim for higher security levels. This paper compares and contrasts different RSA Variants, along with some application scenarios where the different variants could be usefully employed. For instance, variants such as Rebalanced RSA or RSA-Small-d may be useful in obtaining efficient RSA digital signatures in constrained environments or heavily loaded environments. Variants that facilitate negligible difference between encryption and decryption times may be desirable in real time duplex data exchange.

Attacks on various RSA implementations may be classified into (i) Mathematical attacks and (ii) Non-Mathematical attacks. Mathematical attacks are those attacks in which attempts are made either to factorize the RSA modulus or to obtain the decryption exponent without factoring the RSA modulus under certain circumstances. Non-mathematical attacks are those attacks in which the attacker makes use of the physical characteristics of the cryptosystem under investigation. This paper surveys some non-mathematical attacks such as side-channel attacks. These attacks are relevant in scenarios where the attacker has access to the physical equipment that facilitates the cryptosystem.

The rest of the paper is organised as follows: Section-2 reviews standard RSA before different variants of RSA are surveyed in Section-3. Section-4 surveys some non-mathematical attacks and then motivates this against some RSA implementations.

REVIEW OF RSA

Standard RSA (Stinson 2002, Sec 5.3) consists of three important ingredients -- (i) key generation (ii) encryption and (iii) decryption.

Key generation: Two large random primes p and q are chosen and N=pq is computed. N is called the RSA modulus. The euler totient function φ is defined as $\varphi(N) = (p-1)(q-1)$. Some small value of e is chosen such that $\gcd(\varphi(N), e) = 1$. The public key is <N, e>. An integer d is computed as $d=e^{-1} \bmod \varphi(N)$. The private key is <pq, d>.

Encryption: A message M in Z_n is encrypted to yield the ciphertext C using $C=M^e \bmod N$.

Decryption: The ciphertext C is decrypted using $C^d \bmod N = M$.

The mathematical security of RSA depends on the difficulty of factorization and the difficulty of computing the e^{th} roots modulo some composite integer whose factorization is unknown. There is a very well known attack (Boneh 2000) which places a restriction on the values that the decryption exponent can take. Specifically if $d < N^{0.292}$, the RSA instance can be considered unsafe. There are no well known attacks for small values of e for Standard RSA. This means that RSA can be used with small public exponents but not with small private exponents. This means that decryption costs in Standard RSA can be expensive compared to the encryption costs.

VARIANTS OF RSA

Time Efficient Variants: Since decryption in Standard RSA is an expensive operation, it is useful to optimize this operation. Decryption using CRT is the first step in this direction. Since the private keys p and q are known to the decryptor, the following can be computed: $M_p \equiv C^{dp} \bmod p$ and $M_q \equiv C^{dq} \bmod q$ where $dp = d \bmod p-1$ and $dq = d \bmod q-1$. The CRT could be used to compute the unique solution modulo N which is $M = C^d \bmod N$. Subsequent optimizations require other techniques and are outlined below.

Multiprime RSA: In this variant, the modulus N is a product of three or more primes $p_1, p_2, \ldots . p_r$. The public exponent e and private exponent d are generated as in standard RSA with $\varphi(N) = (p_1-1)(p_2-1) \ldots .(p_r-1)$. The encryption algorithm in Multiprime RSA would be the same as that of Standard RSA. The decryption procedure when used with CRT would be a direct extension of Standard RSA decryption with CRT. It turns out that the theoretical speed up factor of Multiprime RSA over Standard RSA is about $r^2/4$ (Boneh 2002). However, for a given RSA modulus size, r cannot be indiscriminately increased as this would result in smaller prime factors which can then be tackled by factorization algorithms where running times depend on the size of the prime factors. For modulus size of 1024 bits, r=3 primes could be used (Lenstra 2001, Sec 4.1). There is a trade-off between efficient decryption, due to large r on one hand and security of the system on the other (Boneh 2000, Sec 2). The security of Multiprime RSA, from the perspective of mathematical attacks has been studied by Hinek (Hinek 2006, Hinek 2007). It can be shown that if r primes are used, (1-1/r) of the most significant bits of N and φ(N) would contain the same values (Hinek 2007, page 75).

A specialised form of Multiprime RSA is Multipower RSA where the RSA modulus N is of the form $N=p^2q$ or more generally $N=p^{r-1}q$ where p and q would be of size n/r bits and n is the size of N. This form is called Multipower RSA (Takagi 1998). A major difference between Multiprime and

Multipower variants is in the decryption process. In Multiprime RSA a mathematical technique called Hensel lifting is used to perform mod p^{r-1} computations more efficiently compared to a full exponentiation modulo p^{r-1} (Boneh 2002, Sec 3.2). It turns out that the theoretical speed up of Multipower RSA over Standard RSA is about $r^3/8$ (Boneh 2002). Multipower RSA is subject to a trade-off between the magnitude of r and security just as in Multiprime RSA. For instance, one could use a value of utmost r = 3 when the modulus is of size 1024 bits.

Batch RSA: Batch RSA (Fiat 1996) is a RSA variant where for the cost of a single RSA decryption plus some additional arithmetic, it is possible to decrypt two or more RSA ciphertexts. For instance, if C_1 is the ciphertext obtained by encrypting M_1 using the public exponent e_1=3 and if C_2 is the ciphertext obtained by encrypting M_2 using the public exponent e_2=5, then by computing $A = (C_1^5 C_2^3)^{1/15}$, one can obtain the two decryptions by using

$$C_1^{1/3} = A^{10} / (C_1^3 C_2^2) \quad \text{and} \quad C_2^{1/5} = A^6 / (C_1^2 C_2)$$

The above scheme can be generalized to batch r ciphertexts. Clearly, the RSA modulus should be the same to meaningfully batch decryptions. It can be shown that the public exponents need to be distinct. Also, batching can become computationally expensive if the public exponents are not small values, thus negating the advantages of batching.

There is no trade-off here between security and the value of r. However, as r increases, there is an increase in
wait times for decryption
memory requirements to store the r ciphertexts
number of RSA certificates issued by the certification authorities adding to the memory requirements as well as the costs to procure the certificates.
Thus there is a trade-off between the size of r on one hand and time, space and certificate costs on the other.

Rebalanced RSA: Multiprime and Batch RSA were attempts to speed up the decryption operations by optimizing the associated mathematical operations. An alternate to this would be to have small values for the decryption exponent d. However, the $N^{0.292}$ attack mandates that the size of d cannot be too small. To overcome this problem, dp = d mod p-1 and dq = d mod q-1 are chosen such that both dp and dq are small, while at the same time d is not too small. The public exponent would then be computed as $e = d^{-1} \bmod \varphi(N)$. The decryption procedure is the same as that of Standard RSA. This variant is called the Rebalanced RSA (Sun 2005b) as it enables in rebalancing a property of Standard RSA, which is to have faster public exponent operations and slower private exponent operations. In other words, faster private exponent operations and slower public exponent operations are enabled by suitably shifting some of the work load from the decryptor to the encryptor. It can be shown that if the size of dp (and dq) is r and if the size of the modulus N is n, then the theoretical speed up of Rebalanced RSA over Standard RSA is about n/(2r). This would mean to say that if the size of dp and dq are made arbitrarily small, the efficiency of Rebalanced RSA would increase. However, dp and dq cannot be made too small because an attacker can factor N in time $O(\min(dp^{1/2}, dq^{1/2}))$ (Boneh 1999). Thus there is a trade-off between the size of dp (and dq) on one hand and security (due to factorization attacks) on the other.

In (Sun 2005a), the authors discuss RSA-Small-e and RSA-Small-d in the context of RSA variants. RSA-Small-e and RSA-Small-d focus on reducing the values of the public exponent and the private exponent respectively. The values are intended to be much smaller than $\varphi(N)$.

RSA-Small-e: Even though the size of the public exponent is small compared to the size of the private exponent it would be convenient to specify the size of the public exponent. Still better would be to fix a value for the public exponent such as e=3 or $e=2^{16}+1$ and then choose two distinct large primes p and q such that gcd((p-1)(q-1),e)=1 and then computing N=pq. The private exponent is then computed as $d=e^{-1} \bmod \varphi(N)$. Since the size of d will be almost the same as that of φ(N), the decryption costs would match Standard RSA decryption costs, while the encryption costs would be more favourable compared to Standard RSA.

RSA-Small-d: This variant is a mirror image of RSA-Small-e. Since e and d can be used interchangeably, it is a simple matter to generate keys as in RSA-Small-e and swap the values of e and d. However, the $N^{0.292}$ attack needs to be taken into account. While RSA-Small-d and Rebalanced RSA both intend to achieve reduced private key computation times, the key generation process is different in the two variants.

Generalized Rebalanced RSA: While the previous variants surveyed in this article optimized either the decryption or encryption operations, a new set of variants proposed by Sun and Yang (Sun 2005a) and Galbraith (Galbraith 2005) intend to achieve a balance between the two, thus enabling variants whose private key operations compares with known fastest methods while at the same time the public key operations perform better than corresponding public key operations that go with known fastest methods for private key operations. In (Sun et al 2007) the authors provide a method of constructing the private exponent d, given dp and dq, which can be applied in the case of Rebalanced RSA as well.

Time efficient variants such as Batch RSA, Multiprime RSA, Multipower RSA and Rebalanced RSA can be used in scenarios where private key operations are a major bottleneck. For instance every SSL handshake using RSA requires a RSA private key operation. A server using SSL for secure communication can quickly become loaded with multiple private key operations when there are multiple handshake requests and thus can become very slow especially in the light of Standard RSA's expensive private key operations. This can become very annoying to web browsers transacting with SSL servers. Another instance where time efficient RSA variants are desirable is in the field of mobile computing. There are a growing number of applications in this segment where signature generation and authentication may be of importance. Slow private key operations can become a major bottleneck in these applications.

The Generalized-Rebalanced variant can be used in applications where it is desired that the encryption and decryptions costs be equal. This may be required for instance in two-way authentication protocols. Balanced encryption and decryption costs also ensure that all communicating entities are treated fairly and equally. This might not be feasible when Standard RSA is used.

Storage Efficient Variants:

It is common for storage space to be expensive in constrained environments. Thus any attempts at storage reduction in RSA variants would be useful.

Dual RSA: Dual RSA proposed by Sun (Sun et al 2007) is a scenario where two RSA instances share the same public and private exponents. If (e,N_1) and (e,N_2) are the public keys in the two instances, (d, p_1, q_1) and (d, p_2, q_2) would be the corresponding private keys. These two instances can be considered as one instance of the Dual RSA enabling the private and public exponents to be stored just once.

Dual RSA can be utilized in the generation of blind signatures and in the resolution of the reblocking problem of RSA Signatures (Sun et al, Sec IV).

Twin RSA: In this variant, two RSA moduli N and (N+d) are stored for almost the price of one. If d is a small integer, it would suffice to store N and d, thus avoiding the need to store (N+d) separately. The key generation process should ensure that N and (N+d) satisfy

$$N = pq \quad \text{and} \quad (N+d) = rs$$

where N and (N+d) would be of the same size and p, q, r and s would be distinct prime numbers. After the key generation process the public and private exponents are instantiated as in Standard RSA.

High-Security Variant:

Unbalanced RSA: This variant proposed by Shamir (Shamir 1995), intends to provide a very high level of security for the cost of normal security. In this variant, the modulus N is the product of primes p and q of completely different sizes and thus the name 'Unbalanced RSA'. In the paper (in 1995 512 bits were getting to be inadequate), Shamir proposes increasing the size of p to about 500 bits and q to 4500 bits. (Thus N would be of size 5000 bits). The idea is to make use of the different rates of progress between factorization algorithms whose running times depend on the size of the factors and factorization algorithms whose running times depend only on the size of the number to be factored. Though Shamir (Shamir 1995) argues that RSA is generally used to exchange session keys and that the session key would typically be lesser than the smaller prime factor resulting in reduced decryption times, this may not hold when the plain text is greater than the smaller prime factor.

The various variants are summarized in the following table:

RSA Variant	Time-Efficient	Space-Efficient	Comments
Multiprime RSA	Yes	No	Multiple prime factors make up the RSA modulus Faster decryption when used with CRT Trade-off between number of primes and security
Batch RSA	Yes	No	Multiple public exponents, but single modulus Faster decryption. can be used in loaded server environments large latency when batch queue is long. not easily interoperable with Standard RSA more memory required to store multiple public exponents
Rebalanced RSA	Yes	No	Faster decryption but slower encryption compared to standard RSA Not easily interoperable with standard RSA
RSA-Small-e	Yes	No	Faster encryption and comparable decryption costs relative to Standard RSA Interoperable with Standard RSA
RSA-Small-d	Yes	No	Faster decryption and comparable encryption costs relative to Standard RSA

			$N^{0.292}$ attack should be taken care of
Generalized Rebalanced RSA	Yes	No	Faster encryption and decryption can be used in real time duplex data exchange applications
Dual RSA	No	Yes	Can be used to generate signatures without the signer knowing the contents of the message(blind signatures) Insecure when public exponent $e < N^{0.25}$
Twin RSA	No	Yes	
Unbalanced RSA	No	No	High security cannot be guaranteed for all applications

Other proposed variants of RSA include Dual-RSA-Small-e, Dual-RSA-Small-d, Dual-Generalized Rebalanced RSA, Twin-Small-e, Twin-Small-d, Multiple-RSA and Big-Brother RSA. The last two variants are generalizations of Twin RSA. Dual-RSA-Small-e may not be secure because when $e < N^{0.25}$, Dual RSA is insecure (Sun 2007).

There are also recommendations to select the public (or the private) exponent in RSA to be a sparse integer (Banks 2002, Lim 1996).

NON-MATHEMATICAL ATTACKS:

There have been extensive studies in the area of mathematical attacks on RSA. In (Koblitz 2007), the authors write 'Throughout the history of public key cryptography almost all of the effective attacks on the most popular systems have succeeded not by inverting the one-way function, but rather by finding a weakness in the protocol'. Thus it would not suffice if cryptographers focus on the mathematical structure alone for security related studies.

Non-mathematical attacks (also known as implementation attacks in the literature) may be broadly classified as
Side-Channel attacks
Protocol attacks
Physical Fault attacks

Side-Channel attacks: In addition to having plain text and ciphertext as input and output, a cryptosystem also requires time and power to run. Side Channel attacks make use of these physical characteristics of the implementation. Measurement of the physical characteristics such as the time or power consumed by the physical cryptographic device can lead to security vulnerabilities. These vulnerabilities can become more exaggerated in constrained environments (smart card cryptography, cryptography over mobile networks etc.) due to
an increased risk of exposure of cryptographic devices to the attacker (no high wall security in constrained environments)
easier measurement of time, power consumed in the absence of noise (noise is a characteristic of generic, non constrained environments)
branching instructions, software optimizations and other such patterns in the cryptographic implementations.
These physical characteristics of the computation are directly related to the secret key(s) that is(are) used by the cryptographic device and thus is a measure of the secret key itself.

The first side-channel attack was proposed by Kocher (Kocher 1996). The attack targeted RSA implementations which used the repeated square and multiply algorithm for private exponentiation operations. The attack necessarily intends to measure the time taken for each iteration through the square and multiply loop. The attack proposed in (Kocher 1996) was followed by another attack related to power measurements (Kocher et al 1999). Though the authors in (Kocher et al 1999) describe the power analysis attacks mainly on DES, the attacks can be adapted to RSA implementations as well.

Protocol Attacks: Protocol attacks try to exploit weaknesses in the protocol that are responsible for the desired cryptographic objective. A famous example of this attack is due to Bleichenbacher (Bleichenbacher 1998), who proposed a chosen ciphertext attack against protocols based on PKCS #1. This attack makes use of error messages that the decryptor might generate when the plain text message obtained after decrypting a chosen ciphertext is not of a particular format.

Fault Attacks: Fault attacks (Boneh et al 1997) make use of computational errors that can occur in systems running the cryptographic algorithms. These errors can occur either as a result of hardware problems or as a result of an attacker gaining access to the cryptographic device and inducing a fault into the device. For instance, any RSA variant in which the CRT is used for efficient signature generation, can be subject to a fault-attack if an error occurs during one partial signature generation (Joye 1996). If m is a message to be signed, d is the decryption exponent and p and q are the two prime factors of the RSA modulus N, the signature generator would compute $S_p = m^d \bmod p$ and $S_q = m^d \bmod q$ and the signature S would then be encrypted as

$$S = ((S_p \cdot q(q^{-1} \bmod p) + (S_q \cdot p(p^{-1} \bmod q)) \bmod N$$

However, if an error occurs during the computation of say S_p, then S_p would be faulty and then S would be faulty as well. If the faulty S is denoted as S_f, then the attacker can compute $\gcd((S_f^e - m) \bmod N, N) = q$ and thus obtain the factorization of N. It turns out that Garner's algorithm, when used to do the CRT recombination has an advantage in countering certain fault attacks (Sun 2003).

Impact of Non-mathematical attacks on RSA Implementations:

Some RSA variants may be used in constrained environments where there is a risk of exposing the physical cryptographic device to a Side Channel attack or a Fault attack. This is more so because most RSA variants used in constrained environments would be optimized, for instance by using CRT for private key operations. As outlined above, under certain circumstances, when the CRT is used for optimized private key operations, the implementation might be subject to Fault attacks. As an other example, when certain parameter sizes are small as can happen in certain RSA variants, Side-Channel attacks are plausible. In (Lim 1996), the authors describe a RSA variant where the private exponent d is sparse with an intention to reduce the number of multiplications required for private exponent operations. There are no known mathematical attacks on this variant, but clearly a Timing attack is highly feasible.

It is not always necessary for the success of Timing attacks to be limited to constrained environments such as smart card cryptography. Timing attacks can apply to more generic cryptosystems as well. In fact, a timing attack on RSA embedded in Open SSL is described in (Brumbly 2003). A power analysis based attack on RSA implementations with CRT based decryption (where Garner's Algorithm is used for CRT recombination) is described in (Novak 2002).

We have already seen that Standard RSA when used with CRT is vulnerable to fault attacks. This can be extended to Multiprime RSA as well. In Multiprime RSA with r primes, (r-1) different

faulty signatures will enable the factorization of the r-prime RSA modulus if each of the faulty (r-1) signatures is caused by a different prime (Yang et al 2005).

It is not always necessary for a non-mathematical attack to be completely successful to break a system. In fact under certain circumstances, when a fraction of the bits of the private exponent d is obtained, the idea of partial-key exposure attack can be made use of. It was shown in (Boneh et al 1998) that given a fraction (about one quarter) of the bits in d, all of d can be reconstructed provided $e < N^{1/2}$. Thus, if a non-mathematical attack like Kocher's timing analysis is used against RSA variants such as RSA-Small-e, it would suffice to obtain 25% of the bits of d. Thus we have a notion of a non-mathematical attack plus a mathematical attack.

As an other example of a non-mathematical attack plus a mathematical attack, we consider a fault attack on Multiprime RSA. Suppose that the number of primes is 3 and that RSA is used for signature generation. If p, q, r are the three prime factors, S_p, S_q and S_r are computed as

$$S_p \leftarrow m^d \bmod p,\ S_q \leftarrow m^d \bmod q \text{ and } S_r \leftarrow m^d \bmod r$$

and then the signature S on message m is computed using the CRT.
If S_p is faulty, p could be recovered because

$$\gcd((S^e - m) \bmod N, N) = qr$$

In (Hinek 2006, Hinek 2007), the author presents a method to factor the r-prime RSA, given certain parts of the most significant or the least significant bits in the unknown primes. Using this attack, it turns out that in the 3-prime RSA, after a fault attack yields one of the primes as shown above, a complete factorization of N is possible in this case, if on the average, $(1/6)^{th}$ of the least significant or the most significant of the other two prime factors are known.

In (Brumbley 2003), the authors state 'Many crypto libraries ignore the timing attack and have no defenses implemented to prevent it'. However, as we have just seen, non-mathematical attacks cannot be overlooked and crypto-libraries should not confine themselves to mathematical attacks alone.

CONCLUSIONS:

In this paper, we surveyed a few RSA variants and then highlighted the need for studies of non-mathematical attacks against RSA implementations. Though no attack has come anywhere close to breaking RSA, a knowledge of different possible attacks is useful especially in the light of implementation attacks that may be possible against different RSA variants.

References:

Banks W D and Shparlinski I E. (2002) On the Number of Sparse RSA Exponents, available at http://citeseer .ist.psu.edu/306202.html.
Bleichenbacher D. (1998) Choosen Ciphertext attacks against protocols based on the RSA encryption standard PKCS #1, CRYPTO '98, LNCS Volume 1462, pp 1-12, Springer-Verlag.
Boneh D, DeMillo R and Lipton R. (1997) On the importance of checking cryptographic protocols for faults, EUROCRYPT 97, LNCS Volume 1233, pp 37-51, Springer-Verlag.
Boneh D, Durfee G and Frenkel Y. (1998) An attack on RSA given a fraction of the private key bits, ASIACRYPT '98, LNCS Volume 1514 pp 25-34, Springer-Verlag.
Boneh D. (1999) Twenty years of Attacks on the RSA Cryptosystem, Notices of the American Mathematical Society, February 1999, Volume 46, Number 2.
Boneh D and Durfee G. (2000) Cryptanalysis of RSA with private key d less than N0.292, IEEE Transaction on Information Theory 46(4): July 2000.

Boneh D and Shacham H. (2002) Fast variants of RSA, Cryptobytes, RSA Laboratories Volume 5, No 1 Winter/Spring 2002.
Brumbly D and Boneh D. (2003) Remote Timing Attacks are Practical, proceedings of 12th Usenix Security Symposium, Washington DC 4-8 August 2003, pp 1-14.
Fiat A. (1996) Batch RSA, Proceedings of Crypto '98, Volume 435 LNCS pp175-185, Springer-Verlag.
Galbraith S D, Heneghan C and McKee J F. (2005) Tunable Balancing of RSA, ACISP 2005, LNCS 3574, pp 280-292, 2005.
Hinek J. (2006) On the security of Multiprime RSA, available at www.cacr.math.waterloo.ca/techreports/2006/cacr2006-16.pdf on Oct 31st 2007.
Hinek J. (2007) On the security of some variants of RSA, PhD Thesis, University of Waterloo, Canada.
Joye M and Quisquater J J. (1996) Attacks on Systems using Chinese Remaindering, Tech Report CG-1996/9, UCL Cryptogroup, Louvain-la-Neuve, March 1997.
Koblitz N and Menezes A. (2007) Another look at provable security, Journal of Cryptology, Volume 20, Issue 1, January 2007.
Kocher P. (1996) Timing attack on implementations of Diffie-Hellman, RSA, DSS and other systems, CRYPTO '96, Volume 1109 of LNCS, pp 104-113 Springer-Verlag 1996.
Kocher P, Jaffe J and Jun B. (1999) Differential Power Analysis, Advances in Cryptology, CRYPTO '99, Volume 1666 of LNCS, pp 388-397, Springer.
Lenstra A K. (2001) Unbelievable Security – Matching AES Security using Public Key Systems, Advances in Cryptology, ASIACRYPT 2001: 7th International Conference on the Theory and Application of Cryptology and Information Security, Gold Coast, Australia, Dec 9-13 2001, proceedings(LNCS).
Lim C H and Lee P J. (1996) Sparse RSA Secret Keys and Their Generation, 3rd Annual Workshop on Selected Areas in Cryptography(SAC), August 1996, Ontario, pp 117-131.
Novak R. (2002) SPA-based Adaptive Choosen-Ciphertext Attack on RSA implementation, PKC 2002, LNCS 2247, pp 252-262.
Shamir A. (1995) RSA for paranoids, Cryptobytes, RSA Laboratories, Volume 1, No 3 Autumn 1995.
Stinson D. (2002) Cryptography Theory and Practice, Chapman and Hall/CRC.
Sun H M and Yang C T. (2003) Permanent Fault attack on the Parameters of RST with CRT, ACISP 2003, LNCS 2727 pp 285-296, Springer.
Sun H M and Yang C T. (2005a) RSA with balanced short exponents and its applications to entity authentication, PKC 2005, LNCS 3386 pp 199-215, Springer.
Sun H M, Hinek J and Wu M E. (2005b) On the Design of Rebalanced RSA-CRT, available at www.cacr.math.waterloo.ca/techreports/2005/cacr2005-35.pdf on Oct 31st 2007.
Sun H M, Wu M E, Ting W C and Hinek M J. (2007) Dual RSA and its security analysis, IEEE Transaction on Information Theory 53(8), August 2007.
Takagi T. (1998) Fast RSA-type Cryptosystems Modulo pkq, proceedings of Crypto'98, LNCS Volume 1462, pp 318-326, Springer-Verlag 1998.
Yang Y, Abid Z, Wang W, Zhang Z and Yang C. (2005) Efficient Multiprime RSA Immune against Hardware Fault Attack, IEEE International Symposium on Circuits and Systems, ISCAS 2005.

An Efficient Homomorphic Coercion Resistant Voting Scheme Using Binary Search Tree

Vinodu George[1] and M P Sebastian[2]

[1]Assistant Professor,
Department of Computer Science and Engineering
LBS College of Engineering, Kasaragod
Kerala, India –671542
E-mail: vinodu.george@gmail.com

[2] Professor and Head
Department of Computer Science and Engineering
National Institute of Technology
Calicut, Kerala
India – 673601
E-mail: sebasmp@nitc.ac.in

Abstract: The paper presents a voting scheme that coalesces many of the advantageous features of an efficient e-voting scheme like receipt-freeness, uncoercibility and write-in ballot, without requiring untappable channels. Some of the previous schemes in the literature provide most of these features with a penalty of increased running time. The proposed scheme utilizes the advantages of binary search tree for reducing the running time in the order of *O(n log n)* and also satisfies the desirable features of existing write-in and coercion resistant voting schemes, such as fair degree of efficiency, and protection against any kind of adversarial behavior with lowest running time and is applicable to any kind of real time elections.

Keywords. *Electronic Voting, Receipt-Freeness, Write-in Ballots, Coercion-Resistance, Homomorphic Encryption, Paillier Cryptosystem, Mix Networks*.

INTRODUCTION

Electronic voting is a rising social application of cryptographic protocols. It promises the possibility of a convenient, efficient and secure facility for recording and tallying votes. Lot of literature on electronic voting has been developed over the last two decades. The uses of insecure Internet, incorrect implementations, and the resulting security breaches have caused a lot of rework in this area. Several of these schemes were for a secure electronic voting (Report,2002), (Sampigethaya, Poovendran 2006) .

One of the main requirements (Smith,2005) of voting protocol is the privacy of voter. Receipt freeness, a stronger notion of privacy that restricts the voter to show how he voted. The first receipt free protocol appeared in Benaloh. and D. Tuinstra (1994). Since then, several schemes (Benaloh. and D. Tuinstra 1994, Okamoto 1996) were proposed in order to meet the condition of receipt-freeness. Coercion resistant protocols, introduced by Juels Catalano and Jakobsson (Juels et al 2002), gives more freedom to the adversary and restricts any party to force another party to vote in a particular way. A coercion resistant and receipt free voting protocol prevents vote buying and vote selling.

Cryptographic ballots do not naturally support write-in (Acquisti 2004) votes. Generally, when Alice wants to write in a name, other than from the predetermined list of candidate, she "write-in" the pseudo-candidate option, and follows a separate process to specify her candidate. Kiayias and Yung (2004) proposed, a vector ballot approach, which provides write-in ballot.

In this paper we propose a simple and practical voting scheme that is both receipt-free and coercion resistant. Also it has the write-in property for cast the vote to a pseudo-candidate. Since it is using the

Paillier crypto system (Paillier 1999), which is a homomorphic encryption scheme, the tallying of vote can also be done easily.

Previous Work

The approaches to electronic voting scheme can be broadly classified into four types.

Early investigations of e-voting considered a simple voting approach. Boardroom voting protocols are examples (DeMillo et al 1984, Yao 1982). Failure of a single voter causes an election failure in these classes. Simple voting schemes like Voter Verified Paper Audit Trial (VVPAT) and vote by mail are inherently insecure and there is no guarantee for the security or privacy during ballot casting.
Receipt based voting schemes (Chaum 1981, Benaloh 1987, Cramer 1997, Baudron 2001) enable the voter to make sure that the vote cast by a person is counted properly. By the introduction of the receipt, threats like vote buying and selling become more serious. Receipts can be used as a solid proof for buyers and sellers.

Receipt free voting protocol was introduced by (Benaloh 1994). Receipt freeness restricts the voter to prove the attacker how he voted. The voter is unable to provide any receipt as a proof for their choice of candidate selection. As a result the attacker is unable threaten the voter for his choice of candidate. In effect the property of receipt freeness reduces vote buying and selling. The scheme in Sako (1995) proposes a multi-authority scheme that uses re-encryption mixnets to mask candidate choices, and homomorphic encryption scheme is used for the calculation of the final tally. The modeling of their scheme was clarified and refined by Michels and Horster (1996). Hirt and Sako(1995) followed Sako and Kilian (1995) and presented an efficient and correct voting scheme.

Voting schemes by David Chaum (2002), Neff (2001, 2003) and Kiayias and Yung (2002, 2004) satisfy receipt freeness and allow write-in ballots. The protocol in David Chaum (2002) has the physical constraints of visual cryptography and voting stations. The scheme in Neff (2003) is an efficient voting scheme based on shuffle mixnet protocol. This scheme is also has some physical constraint such as “code book” which confirms the correspondence between the election codes and the voter’s preferences. Protocol proposed in Kiayias and Yung (2004) incorporates write-in choice into a homomorphic encryption scheme using “vector ballot” approach. This ballot has a fixed representation, which distinguishes between predetermined candidate and write-in one ballot. It has the potential to achieve more efficient tallying for e-voting and is universally verifiable at the same time. But this scheme lacks coercion resistance property. So this is vulnerable to coercion.

The notion of coercion resistance was proposed by Juels et al (2002). This concept captures the fullest possible range of adversarial behavior in a real-world, Internet-based voting scheme. According to this scheme the coercers have more privileges. Voters access an anonymous channel during voting stage. Anonymous channels can be realized in a practical way by use of mixnets, while untappable channels require largely unrealistic physical assumptions. A drawback of this scheme is that overhead for tallying authorities is quadratic in the number of voters. Processing V votes by N voters requires $O(NV)$ steps to perform the entire cross checks, i.e. at least N^2 steps. Thus the scheme is only practical for small elections. Also this scheme does not openly discuss the possibility of write-in ballots, allowing voters to choose their own ballots.

Acquisti (2004) proposed a new voting scheme, which guarantees receipt-freeness, uncoercibility and write-in ballot without any physical assumptions and by addressing some of the problems in Juels et al (2002). In addition, this protocol allows flexible ballot formats to be used, including write-in ballots without the specific procedural constraints or physical assumptions needed as in David Chaum (2002) and Neff (2003). Unlike David Chaum (2002), Neff (2003), and Kiayias and Yung (2004), this protocol can at least neutralize forced abstention and randomization attacks”(Juels et al 2002). So this scheme, offers write-in ballot with coercion resistance. But the drawback in Acquisti (2004) is more serious with respect to Juels et al (2002). For validation and tallying of votes, it takes all possible ballots for each credential. So the time complexity is more than what is given in Juels et al (2002). So this scheme also is practicable only for small elections.

Our Contribution

We proposes a protocol to the literature mainly by presenting a scheme which guarantees all the desirable properties proposed previously and by addressing the major issues of Kiayias and Yung (2004), Juels et al (2005) and Acquisti (2004). The problem with Kiayias and Yung (2004) is lack of coercion resistance and with Juels et al (2005) is the time needed for the checking of credential validity, and the scheme is practically applicable only for small elections. The same problem is more serious in Acquisti (2004). Even though it accomplishes write in ballot and coercion resistance, the time complexity for the same is $O(n^3)$. It is because it checks all possible ballots for each credential.

In this paper we present a scheme which is similar to Acquisti (2004), but it takes only $O(n\log n)$ steps to perform the entire cross checks and elimination of duplications. Also the elimination of duplicate votes and validation of credential can be combined as a single step. So further reduction of running time is possible. Since this scheme uses homomorphic encryption, tallying time can also be reduced. Thus our scheme will have all the features of Kiayias and Yung (2004), Juels et al (2005) and Acquisti (2004) with complexity of $O(n\log n)$. So it applies to any type of election. This protocol also allows multiple casting, so the coerced voter will get the chance to cast the intended vote.

Organization

Rest of the paper is organized as follows. In next section , we describe our voting model and required steps in the voting process. We discuss the necessity for coercion resistant voting schemes also in that section. In the section after that we present the actual scheme, and the detailed analysis of the scheme. Conclusions are given in the last section . Proof for the complexity analysis is given in Appendix A. Details of Paillier cryptosystem is in Appendix B.

VOTING MODEL

An election system consists of several sets of entities

Authorities: Denoted by $A = \{A_1, A_2, \ldots\ldots An_A\}$, authorities are responsible for jointly issuing keying material, i.e. credential (Brands 2002) to voters
Validator: Denoted by D is responsible for the construction of binary search tree and validation of the votes.
Talliers: The set of n_T Talliers denoted by $T = \{T_1, T_2, \ldots\ldots Tn_T\}$. Talliers are responsible for processing the ballots and jointly counting the votes and publishing a final tally.
Voters: The set of n_V voters, denoted by $V = \{V_1, V_2, \ldots\ldots V_{nV}\}$, are the entities participating in a given election.

Voting Life Cycle

A simplified life cycle of an electronic voting scheme may be divided into four main stages (Report 2002).

Setup. This stage involves the initialization of the technical part of the election system as well as the initialization of the organizational structure.
Voting. In this stage the votes are cast. To do so, the voters will need some form of authentication to validate the (digital) ballot form completed by them. Furthermore, some (cryptographic) technique needs to be applied to ensure ballot secrecy.
Validation. This stage includes validation of votes cast by voters. If any of the votes are found to be invalid, they are discarded.
Tabulation. In this stage the Talliers will compute the election result from the valid votes, after which the election officials may announce it.

Why Coercion Resistance

Coercion resistance (Juels et al 2002) is a stronger property for the voting scheme. Intuitively, an election protocol is coercion-resistant if a voter *V* cannot prove to a potential coercer *C* that he voted in a particular way. We assume that *V* cooperates with *C* in an interactive fashion. Receipt-freeness is a weaker property, for which we assume that *V* and *C* cannot interact during the protocol: to break receipt-freeness, *V* later provides an evidence (the receipt) of how he voted. Coercion resistance implies receipt-freeness, which implies privacy, the basic property of voting protocols being anonymous.

Coercion-resistance (Juels et al 2002) guarantees that the coercer cannot become convinced of how a voter votes, even if the voter cooperates with him. Of course the voter can tell a coercer how he voted, but coercion resistance asserts that he is unable to prove it, so the coercer has no reason to believe him. Intuitively, coercion resistance is a stronger property than privacy, since if it is possible for a coercer to detect the value of a voter's vote without the voter's cooperation, then it must also be possible with the voter's co-operation (Delaune et al 2006).
So we need a (Cramer et al 1997)voting scheme, which offers not only receipt-freeness, but also defense against randomization, forced-abstention(Okamoto 1997), and simulation attacks (Okamoto 1997), (Hirt and Sako 2000). Here comes the stronger notion of privacy, coercion resistance, which will provide shield against the fullest possible range of adversarial behavior in a real-world Internet-based voting scheme.

VOTING SCHEME

The Authority includes of independently functioning servers that manage registration, binary search tree and tallying of all the cast votes through the bulletin board. During registration stage each responsible authority creates shares of voting credential for each eligible voter, which used by the voter during voting stage to authenticate the ballot. Each authority posts theses credentials shares, encrypted with authority's public key and also with voter's public key on a bulletin board on the place allotted for each voter. Authority also publishes the candidate slate on the bulletin board. Validator multiplies the shares of encrypted credential and creates the binary search tree with the key as encrypted credential and the resulting tree is also published on the bulletin board.
Each voter multiplies the decrypted shares of he has selected from the bulletin board using the homomorphic property of Paillier cryptosystem. The voter sends the encrypted credential along with the encrypted candidate choice and the proof to the bulletin board. At the end of voting stage, the authority mixes all ciphertexts posted by eligible voters and stores the verified ballots in the binary search tree.

Voting Protocol

Setup: The Paillier public/private key pairs (PA, SA, VA, VA_j) and (PT, ST, VT, VT_j) for Authority and Tallier respectively are generated in a suitably reliable manner, and *PA* and *PT* are published along with all system parameters. The set of candidates is defined as Candidates=$\{1,B,B^2,....B^{nc-1}\}$ where *B* is an integer with the property $B > n_v$ (number of voters) and *nc* is the number of candidates.

Registration: The Authority A_j generates the string σ_{ij} for voter V_i that serves as the credential of the voter V_i. More details about the creation of credential are given in (Brands 2002). For each σ_{ij} it creates, A_j encrypts σ_{ij} using *PA* and appropriate secret randomization, and then encrypts the resulting ciphertext with the public key of V_i, and publishes it on bulletin board on a row publicly reserved for the shares of credential of voter V_i. It is possible for A_j to furnish V_i with a proof that $(E^{PA}(\sigma_{ij}))$ is a ciphertext on σ_{ij}.

$S_i = E^{PVi}(E^{PA}(\sigma_{ij}))$
E^{PVi} represents RSA encryption under V_i's public key.

The Authority A_j sends the encrypted credential for each voter V_i to a common trusted Validator in a random order so that the source of information about A_j will be unknown. So the Validator will have only the

complete set of encrypted credential and mapping between voter and credential will not be possible for him. Since the Authority A_j sends the encrypted credential share in a random order, the Validator will not be able to find which Authority is responsible for each of the credential. Validator calculates the respective credential for each voter by multiplying the credential shares.

$$\prod_{j=1,..m} (E^{PA}(\sigma_{ij})) = E^{PA}_{j=1,...m}(\Sigma\, \sigma_{ij}) \equiv E^{PA}(\sigma_i)$$

Validator mixes the encrypted credentials in order to eliminate the relation between the voters and credential and then creates a binary search tree with the encrypted credential as the key of the tree node.

Tree Node Structure:

Encrypted credential	Time stamp	Encrypted Ballot	Write-in ballot bit	Left Node	Right Node

Voting: Voter receives the credential share from the space allotted for the voter and verifies the designated verifier proof of S_i. Upon successful verification, he multiplies together the shares $E^{PA}(\sigma_{ij})$. For each encrypted share of credential he receives, a voter V_i verifies the designated verifier proof of S_i. Upon successful verification, he multiplies together the shares $E^{PA}(\sigma_{ij})$. If the voter selects the candidate choice B^c from the bulletin board published by the Authority, then the write-in ballot bit will not be set and he encrypts it with the public key of Tallier, $E^{PT}(B^c)$. Otherwise he will set the bit and cast his vote. Along with the encrypted credential, voter will attach the non-interactive zero knowledge proof for the correctness of the credential. So that validator can verify the authenticity of the credential along with the ballot. The voter V_i casts his vote as

$$(E^{PK}(E^{PT}(B^c)),\ E^{PK}(E^{PA}(\sigma_i))).$$

The first is a ciphertext on the candidate choice of the voter; the latter is a ciphertext on the credential of the voter. The voter wraps the resulting ciphertext with the RSA public key, *PK* of Validator. The encryption using Paillier (1999) public key *PT* of Tallier should be semantically secure (Goldwasser and Micali 1984) for preventing any kind of passive attacks (See appendix B for more details about Paillier cryptosystem).

Tallying: To tally the ballots posted to bulletin board, the Tallier performs the following steps:

Mixing of the ballots: To eliminate the relation between the voter and his vote, a mixing of ballots is carried out.

Eliminating duplicates and checking credentials: Search the binary tree for the encrypted credential of the voter. The hit indicates a valid credential and otherwise an invalid credential which is discarded. Check for already stored vote for the valid credential and if not, store the corresponding encrypted ballot and the time stamp of the vote. Otherwise identify the last vote cast by comparing the time stamp of the ballot and store the corresponding vote. Also before tallying the time stamp of each ballot is removed in order to avoid the uniquely identifiable ballots.

Tallying: Tallier multiplies all ciphertexts of ballots in tree nodes whose write-in ballot bit is not set and decrypts the sum. The Tallier decrypts all encrypted ballots in the tree nodes whose bit is set since write-in ballots need to be read individually and by combining the output of these two steps, the Tallier makes the final result. There can be a threshold (Damgard and Jurik 2003) decryption by sharing the secret key among the Talliers. i.e. using the shares of the secret key ST_j and the verification keys VT and VT_j. Each Tallier runs the decryption algorithm and produces a partial decryption of $E^{PT}(B^c)$, providing a proof of validity for the partial decryption. The combiner can then produce the decryption of ciphertext $E^{PT}(B^c)$, if enough partial decryptions (t or more) are valid.

Properties and Security Issues

Robustness/Collusion: Each authority sends the encrypted credential shares to each voter to the Bulletin Board along with the non-interactive zero knowledge proof for the authenticity of the credential. Hence the voter can identify the malfunctioning of any of the authorities. Since the credential is generated in a distributed manner all the authorities have to collude to cheat the election process. Our assumption is that there will be at least one honest authority. Since the outputs of each step are stored in the bulletin board, any malfunctioning of the Validator can be monitored even though Validator and Tallier collude.
Universal verifiability: The list of cast ballots is published in the bulletin board and hence the voter can verify whether his vote is taken into account. The transactions of Validator are stored in the bulletin board and also the outputs of mixing and tallying stages. Thus all the communications during protocol execution are stored in the bulletin board and hence the scheme is universally verifiable.
Correctness: Even though the voter is allowed for multi-casting, only one choice of candidate is stored in the binary search tree. Duplicates are getting eliminated and hence one credential is counted only once during tallying. Moreover, the Validator can't add or eliminate any vote since all functions are publicly monitored.
Coercion Resistance and Receipt Freeness: Coercion resistance is that property which prevents the voter from proving to an attacker that he voted in a particular way. In this scheme it gains this property by allowing the attacker to cast the vote by providing a fake credential with non-interactive zero knowledge proof for the credential validity. During voting stage the voter is not getting any indication about the validity of credentials, and credentials are validated during tallying phase only. Before the validation of credentials a mixing of ballots are performed and hence the relation between the voter and the ballot has been removed. See Jules et al (2002) for a rigorous proof.

Since this scheme supports multiple casting for the same credential, the coerced voter can cast his right choice during any time of voting. But only the last cast ballot is counted for the election result. Hence the scheme is free from "forced abstention attack"(Juels et al 2002). Moreover, the coercer can't distinguish the candidate choice of the voter, which added to the final tally. Also, the voter can't be uniquely identified with the time stamp as it is removed from the ballot during Tallying stage. As the voter can furnish fake credential to the adversary, the coercer is unable to identify the candidate choice of the voter and hence the scheme is free from "randomization attack" by Schoenmakers as given in Juels et al (2002).

Complexity Analysis

The registration phase will have running time of O(No. of Voters X No. of Authorities), as the Validator has to multiply the credential share of each Authority in order to get the credential of each voter. After issuing the credential, the Validator creates the binary search tree with a running time of O(No. of Voters X *log*(No. of Voters)). Validator takes the credential from the mix net, and hence the binary tree will be a randomly built binary search tree. The height of a randomly built binary search tree is *logn* (Knuth 1973), (Cormen et al 2000) (see appendix A for proof which is taken from Cormen et al 2000). The validation and duplication elimination can be done by traversing the binary search tree, and it takes the running time of O(No. of Voters X *log*(No. of Voters)). Tallying can be done by traversing through the tree. Hence the total running time will be O(No. of Voters X *log*(No. of Voters)). By the addition of write-in property, will not cause any additional running time, since the tallying can be completed by a single traverse through the binary tree. Thus, this scheme is applicable to any real time election.

CONCLUSION

We have presented an electronic voting scheme that achieves privacy, uncoercibility, and receipt freeness with write-in property. Our protocol contributes to the literature by presenting a scheme, which guarantees receipt freeness and uncoercibility with lowest running time compared to other major coercion resistant receipt free voting schemes in the literature. The scheme can be used to cast different types of ballots that include yes/no, multi-candidate, 1 out of t choices, as well as write-in ballots in any type of election. In addition, our protocol allows for flexible ballot formats to be used in receipt-freeness and universal

verifiability elections. So our scheme will serve as a simple and secure means for any kind of real time electronic voting.

REFERENCES

[1]. Aggelos Kiayias and Moti Yung. Self-tallying elections and perfect ballot secrecy. PKC '02, pages 141–158. Springer-Verlag, 2002. LNCS no. 2274.
[2] Aggelos Kiayias and Moti Yung. The vector-ballot e-voting approach, 2004. Mimeo, University of Connecticut and Columbia University.
[3]. Alessandro Acquisti. Receipt-free homomorphic elections and write-in ballots. Cryptology ePrint Archive, Report 2004/105, 2004. http://eprint.iacr.org/
[4]. C. Andrew Neff. A verifiable secret shuffle and its application to e-voting. Proc. 8th ACM conference on Computer and Communications Security, pages 116–125. ACM Press, 2001.
[5]. Andrew Neff. Detecting malicious poll site voting clients, 2003. http://www.votehere.net/.
[6]. Ari Juels, Dario Catalano, and Markus Jakobsson. Coercion resistant electronic elections. ACM Workshop On Privacy; The Electronic Society 2005 (WPES'05), pages 61–70, November 2005.
[7]. Ari Juels, Dario Catalano, and Markus Jakobsson. Coercion-resistant electronic elections. Cryptology ePrint Archive, Report 2002/165, 2002. http://eprint.iacr.org/.
[8]. O. Baudron, P.A. Fouque, D. Pointcheval, G. Poupard, and J. Stern Practical Multi-Candidate Election System. Proc. 20th ACM Symposium on Principles of Distributed Computing (PODC '01), pages 274-283. ACM Press 2001.
[9] J Benaloh. and D. Tuinstra. Receipt-free secret-ballot elections (extended abstract). Proc. 26th Annual Symposium on Theory of Computing (STOC'94), pages 544–553. ACM Press, 1994.
[10]. Cramer R, Gennaro R, Schoenmakers B. A secure and optimally efficient multi-authority election scheme. Proc. of EUROCRYPT'97, pages 103-118. Springer-Verlag, 1997. LNCS no.1233.
[11]. David Chaum. Untraceable electronic mail, return addresses, and digital pseudonyms. Communications of the ACM, 24(2):84–88, 1981.
[12]. David Chaum. Secret-ballot receipts and transparent integrity. Draft, 2002. www.vreceipt. com/article.pdf.
[13]. DeMillo R., Lynch N and Merritt M. C. Cryptographic Protocols. Proc. 14th ACM Symp. On Theory of Computing, San Francisco, pages 383-400. CA (May 1982).
[14] Donald E. Knuth. Sorting and Searching, volume 3 of The Art of Computer Programming. Addison-Wesley, 1973
[15]. S. Goldwasser and S. Micali. Probabilistic Encryption. Journal of Computer and System Sciences, 28:270-299, 1984.
[16]. M. Hirt and K. Sako. Efficient receipt-free voting based on homomorphic encryption. B. Preneel, editor, EUROCRYPT '00, pages 539–556. 2000. LNCS no. 1807.
[17]. Ivan B. Damgard and Mads J. Jurik. A Length-Flexible Threshold Cryptosystem with Applications. BRICS Report Series RS-03-16, ISSN 0909-0878 March 2003.
[18]. Josh C. Benaloh. Verifiable secret-ballot elections. PhD Thesis, Yale University, Department of Computer Science, 1987. Number 561.
[19]. Krishna Sampigethaya, Radha Poovendran. A framework and taxonomy for comparison of electronic voting schemes. Computers & Security, 25:137-153. 2006.
[20]. M. Michels and P. Horster. Some remarks on a receipt-free and universally verifiable mix-type voting scheme. K. Kim and T. Matsumoto, editors, ASIACRYPT '96. Springer-Verlag, 1996. LNCS Vol 1163.
[21].T.Okamoto. An electronic voting scheme. IFIP World Conference on IT Tools, Canberra, Australia, pages 21–30. 1996.
[22]. T. Okamoto. Receipt-free electronic voting schemes for large scale elections. B. Christianson et al., editor, Security Protocols Workshop, pages 25–35. Springer-Verlag, 1997. LNCS no 361.
[23]. Pascal Paillier. Public-key cryptosystems based on composite degree residuosity classes. J. Stern, editor, EUROCRYPT '99, pages 223–238. Springer-Verlag, 1999.LNCS no. 1592.
[24]. Report on Review of Cryptographic Protocols and Security Techniques for Electronic Voting CyberVote January 28, 2002.

[25]. K. Sako and J. Kilian. Receipt-free mix-type voting scheme - a practical solution to the implementation of a voting booth. In L. Guillou and J.-J. Quisquater, editors, EUROCRYPT '95, pages 393–403. Springer-Verlag, 1995. LNCS no. 921.
[26]. Stefan Brands. A Technical Overview of Digital Credentials, Technical Report, February 2002.
[27]. Stephanie , Steve Kremer, Mark Ryan. Coercion-Resistance and Receipt-Freeness in Electronic voting. Proc. 19th IEEE Computer Security Foundations Workshop 2006.
[28] Thomas H. Cormen, Charles E. Leiserson and Ronald L. Rivest. Introduction to Algorithms. Prentice Hall, 2000
[29]. Warren D. Smith. New cryptographic voting scheme with best known theoretical properties. Workshop on Frontiers in Electronic Elections (FEE 2005), Milan, Italy, September 2005.
[30]. Yao, A. Protocols for Secure Computations. Proc. 23rd IEEE Symp.on Foundations of Computer Science, Chicago,IL, pages 160-164. 1982.

APPENDIX

A. Height of a Randomly Built Binary Tree

All the basic operations on a binary search tree(Knuth 1973, Cormen et al 2000) run in O(h) time, where h is the height of the tree. The height of a binary search tree varies, however, as items are inserted and deleted. In order to analyze the behavior of binary search trees in practice, it is reasonable to make statistical assumptions about the distribution of keys and the sequence of insertions and deletions.

Unfortunately, little is known about the average height of a binary search tree when both insertion and deletion are used to create it. When the tree is created by insertion alone, the analysis becomes more tractable. Let us therefore define a "randomly built binary tree" on n distinct keys as one that arises from inserting the keys in random order into an initially empty tree, where each of the $n!$ permutations of the input keys is equally likely. The goal of this appendix is to show that the expected height of a randomly built binary search tree on n keys is $O(\lg n)$.
We begin by investigating the structure of binary search trees that are built by insertion alone.

Lemma 1
Let T be the tree that results from inserting n distinct keys $k_1, k_2, \ldots, k_n$ (in order) into an initially empty binary search tree. Then k_i is an ancestor of k_j in T, for $1 \le i < j \le n$, if and only if
$K_i = \min \{k_l : 1 \le l \le i \text{ and } k_l > k_j\}$
or
$k_i = \max \{k_l : 1 \le l \le i \text{ and } k_l > k_j\}$

Proof : Suppose that k_i is an ancestor of k_j. Consider the tree T_i that results after the keys $k_1, k_2, \ldots, k_i$ have been inserted. The path in T_i from the root to k_i is the same as the path in T from the root to k_i. Thus, if k_j were inserted into T_i, it would become either the left or the right child of k_i. Consequently, k_i is either the smallest key among $k_1, k_2, \ldots, k_i$ that is larger than k_j or the largest key among $k_1, k_2, \ldots, k_i$ that is smaller than k_i.
Suppose that k_i is the smallest key among $k_1, k_2, \ldots, k_i$ that is larger than k_j. Comparing k_j to any of the keys on the path in T from the root to k_i yields the same results as comparing k_i to the keys. Hence, when k_j is inserted, it follows a path through k_i and is inserted as a descendant of k_i.

Corollary 1
Let T be the tree that results from inserting n distinct keys $k_1, k_2, \ldots, k_n$ (in order) into an initially empty binary search tree. For a given key k_j, where $1 \le j \le n$, define

$G_j = \{k_i : 1 \le i < j \text{ and } k_l > k_i > k_j \text{ for all } l < i \text{ such that } k_l > k_j\}$
and
$L_j = \{k_i : 1 \le i < j \text{ and } k_l < k_i < k_j \text{ for all } l < i \text{ such that } k_l < k_j\}$

Then the keys on the path from the root to k_j are exactly the keys in $G_j \cup L_j$, and the depth in T of any key k_j is

$d(k_j, T) = |G_j| + |L_j|$

Lemma 2

Let $k_1, k_2, \ldots, k_n$ be a random permutation of n distinct numbers, and let $|S|$ be the random variable that is the cardinality of the set

$S = \{k_i: 1 \le i \le n \text{ and } k_l > k_i \text{ for all } l < i\}$. (1)

Then Pr $\{|S| \ge (\beta + 1) H_n\} \le 1/n^2$, where H_n is the n^{th} harmonic number and $\beta \approx 4.32$ satisfies the equation (In β - 1) $\beta = 2$.

Proof: We can view the cardinality of the set S as being determined by n Bernoulli trials, where a success occurs in the i^{th} trial when k_i is smaller than the elements $k_1, k_2, \ldots, k_{i-1}$. Success in the i^{th} trial occurs with probability $1/i$. The trials are independent, since the probability that k_i is the minimum of $k_1, k_2, \ldots, k_i$ is independent of the relative ordering of $k_1, k_2, \ldots, k_{i-1}$.

We have the result that for a sequence of n Bernoulli trials, where in the i^{th} trial for i =,2,…n, success occurs with probability p_i and failure occurs with probability $q_i = 1 - p_i$. Let X be the random variable describing the total number of success, and let $\mu = [X]$. Then for $r > \mu$,

$\Pr\{X - \mu \ge r\} \le (\mu e/r)^r$

We can use this result to bound the probability that $|S| \ge (\beta + 1) H_n$.

The expectation of $|S|$ is $\mu = H_n \ge \ln n$. Since $\beta > 1$, the above result yields

$$\begin{aligned} \Pr\{|S| \ge (\beta + 1) H_n\} &= \Pr\{|S| - \mu \ge \beta H_n\} \\ &\le (eH_n / \beta H_n)^{\beta Hn} \\ &= e^{(1-\ln \beta)\, \beta Hn} \\ &\le e^{-(\ln \beta - 1)\, \beta \ln n} \\ &= n^{-(\ln \beta - 1)\, \beta} \\ &= 1/n^2 \end{aligned}$$

which follows from the definition of β.

We now have the tools to bound the height of a randomly built binary search tree.

Theorem The average height of a randomly built binary search tree on n distinct keys is O(lg n).

Proof: Let $k_1, k_2, \ldots, k_n$ be a random permutation on the n keys, and let T be the binary search tree that results from inserting the keys in order into an initially empty tree. We first consider the probability that the depth $d(k_j, T)$ of a given key k_j is at least t, for an arbitrary value t. By the characterization of $d(k_j, T)$ in Corollary 1, if the depth of k_j is at least t, then the cardinality of one of the two sets G_j and L_j must be at least $t/2$. Thus,

$$\Pr\{ d(k_j, T) \ge t\} \le \Pr\{| G_j | \ge t/2\} + \Pr\{| L_j | \ge t/2 \} \quad (2)$$

Let us examine $\Pr\{| G_j | \ge t/2\}$ first. We have

$$\begin{aligned} &\Pr\{| G_j | \ge t/2\} \\ &\quad = \Pr\{|\{ k_i : 1 \le i < j \text{ and } k_l > k_i > k_j \text{ for all } l < i \}| \ge t/2\} \\ &\quad \le \Pr\{|\{ k_i : 1 \le n \text{ and } k_l > k_i \text{ for all } l < i \}| \ge t/2\} \\ &\quad = \Pr\{| S | \ge t/2\} \end{aligned}$$

where S is defined as in equation (1). To justify this argument, note that the probability does not decrease if we extend the range of i from $i < j$ to $i \le n$, since more elements are added to the set. Likewise, the probability does not decrease if we remove the condition that $k_i > k_j$ since we are substituting a random permutation on possibly fewer than n elements (those k_i that are greater than k_j) for a random permutation on n elements. .

Using a symmetric argument, we can prove that

$$\Pr\{| L_j | \ge t/2\} \le \Pr\{| S | \ge t/2\},$$

and thus, by inequality (2), we obtain

$$\Pr\{ d(k_j, T) \ge t\} \le 2 \Pr\{| S | \ge t/2\}.$$

If we choose $t = 2(\beta + 1)H_n$, where H_n is the n^{th} harmonic number and $\beta \approx 4.32$ satisfies (In β - 1) $\beta = 2$, we can apply Lemma 2 to conclude that

$$\Pr\{ d(k_j, T) \geq 2(\beta + 1)H_n\} \leq \Pr\{|S| \geq (\beta + 1) H_n\} \leq 2/n^2$$

Since there are at most n nodes in a randomly built binary search tree, the probability that any node's depth is at least $2(\beta + I)H_n$ is therefore, by Boole's inequality, at most $n(2/n^2) = 2/n$. Thus, at least $1 - 2/n$ of the time, the height of a randomly built binary search tree is less than $2(\beta + 1)H_n$, and at most 2/n of the time, it is at most n. The expected height is therefore at most $(2(\beta + 1)H_n)(1 - 2/n) + n(2/n) = O(\lg n)$.

B. Paillier Cryptosystem

Various cryptosystems based on randomized encryption schemes E(M) which encrypt a message M by raising a basis g to the power M and suitably randomizing. Their security is based on the intractability of various "residuosity" problems. As an important consequence of this encryption technique, those schemes have homomorphic properties. Only a 'trapdoor' allows the owner of a private key to decrypt the cipher text without knowing the randomizing component. The trapdoor discrete logarithm mechanism that we use in our voting scheme is Paillier's(1999). It can be described in the following way:

Key Generation: Let N be an RSA modulus $N = pq$, where p and q are prime integers. Let g be an integer of order a multiple of N modulo N^2. The public key is $PK = (N, g)$ and the secret key is $SK = \lambda(N)$ where $\lambda(N)$ is defined as $\lambda(N) = lcm((p-1)(q-1))$.

Encryption: To encrypt a message $M \varepsilon Z_N$, randomly choose x in Z^*_N and compute the cipher text $c = g^M x^N \bmod N^2$.

Decryption: To decrypt c, compute $M = L(c^{\lambda(N)} \bmod N^2)/L(g^{\lambda(N)} \bmod N^2) \bmod N$ where the L-function takes in input elements from the set $S_N = \{u < N^2 \mid u = 1 \bmod N\}$ and computes $L(u) = u-1/N$.

Proof of knowledge of an encrypted message: Let N be a k-bit RSA modulus. Given $c = g^m r^N \bmod N^2$, under the assumption that the decryptions of the c lie in an interval $[0,2^l]$, the prover P convinces the verifier V that the c's encrypt the same message m.

1. P picks at random $\rho \varepsilon [0,2^k]$ and $s \varepsilon Z^*_N$. Then he computes $u = g^\rho s^N \bmod N^2$ and commits to the u.
2. V chooses at random a challenge e in $[0,A]$ and sends it to P.
3. P computes $z = \rho + me$ and $v = sr^e \bmod N$ and sends z. V checks that $z \varepsilon [0,2^k]$ and that $g^z v^N = uc^e \bmod N$.

Compressed Nested Certificates Provide More Efficient PKI

A Jancic[1] and LM Batten[2]

[1]School of Engineering and Information Technology
Deakin University
E-mail: anaj@deakin.edu.au

[2]School of Engineering and Information Technology
Deakin University
E-mail: lmbatten@deakin.edu.au

Abstract. Certificate verification in PKI is a complex and time consuming process. In the classical PKI methodology, in order to obtain a public key and to accept a certificate as valid, a verifier needs to extract a certificate path from the PKI and to verify the certificates on this path recursively. Levi proposed a nested certificate model with the aim to simplify and speed up certificate verification. Such a nested certificate-based PKI significantly improves certificate verification, but it also requires a large increase in the number of issued certificates, which makes this model impractical for real life deployment. In order to solve this drawback of nested PKI, while retaining its speed in certificate verification, we propose in this paper the innovative concept of a *compressed nested certificate*, which is a significantly modified version of the nested certificate model. Compressed nested certificate PKI deploys compressed nested certificates which speed up and simplify certificate verification while keeping certificate load to a minimum, thus providing implementers the option of integrating it into the existing PKI model or building it separately as an independent model.

1 INTRODUCTION

Public Key Infrastructure, or PKI, is designed to provide an authentic public key distribution across a large range of applications through digital certificates that include a combination of personal data about the certificate holder and the certificate, as well as the certificate's public key and digital signature. Various PKIs have been proposed in the literature, such as Privacy Enhanced Mail (PEM), Secure Electronic Transaction (SET), Simple Public Key Infrastructure (SPKI) and Domain Name System Security Extensions (DNSSEC). Most of them are based on the third edition of the ISO/ITU-T X.509 (1998) certificate standard. Although the X.509 standard does not enforce any topology for a standard PKI, X.509-based PKIs are generally hierarchical and centralised (Adams & Lloyd 2003). That is why we will suppose that the PKI analysed in this paper has a hierarchical structure. Three important characteristics of hierarchical X.509-based PKI topology are: a tree with 3 or more levels; strict distinction between CAs and end-users (i.e. only CAs issue certificates); forming optional networks via cross certificates. However, one of the major limitations of hierarchical PKI models is that their relatively long certificate paths make certificate validation complex. In the classical PKI methodology, in order to obtain a public key and to accept a certificate as valid, a verifier needs to extract a certificate path from the PKI, and to verify the certificates on this path recursively.

Levi (1999) proposed the concept of nested certificates with the aim to simplify and to speed up the certificate verification process. Nested certificates are a special kind of certificate issued for other certificates. A PKI that deploys nested certificates, known as a nested certificate-based PKI (NPKI), speeds up cryptographic digital certificate verification and reduces the number of certificate revocation controls. In order to verify a certificate with its certificate path, a verifier is required to perform a cryptographic verification for the first certificate only while other certificates are verified just by using fast hash computations. The nested certificate-based PKI model efficiently improves verification, but it also generates large numbers of certificates in a system. This large certificate load is the major problem that makes NPKI impractical for real life deployment. In this paper, in order to solve the problem of high certificate load in NPKI, while retaining its computational efficiency, we introduce the concept of a compressed nested certificate.

A compressed nested certificate is an advanced version of the nested certificate that is issued for several certificates simultaneously in order to speed up verification and to minimize the number of certificates generated. A PKI that deploys compressed nested certificates, which we call a Compressed Nested PKI (CNPKI), also has the advantage of reducing the number of certificate revocation controls since at most two certificate revocation checks are sufficient. Thus, our CNPKI model has the advantages of fast and simple verification with, as we shall show, virtually no increase in the number of certificates over classical hierarchical PKI (for large systems), thus providing a superior solution to both classical and nested PKI.

In section 2, we introduce the certificate validation problem. In section 3, we describe nested certificates. Sections 4, 5, 6 and 7 introduce and analyse our new model as well as presenting revocation and security scenarios. We conclude in section 8.

2 PKI AND THE CERTIFICATE VALIDATION PROBLEM

In every PKI model (with the exception of PGP) certificates are issued by trusted Certification Authorities (CAs). A verifier always verifies first a digital signature of the CA over the CA's certificate before he verifies the user's certificate. In this way the trustworthiness of the user's certificate is established (Henderson et al. 2002).. Most often there are too many end-users and the work-load is too large to be handled by a single CA; thus more than one CA is deployed within a network, which leads to the most common PKI model, that with a hierarchical structure. In this case, each user receives her/his certificate with its certificate path starting with the Root CA's certificate and finishing with the end-user's certificate (Cooper et al. 2005).. A verifier only wants to find the public key of the target entity, and to check correctness of binding between a public key and the identity of the certificate holder. However, he needs to verify all certificates on the certificate path one by one in order to check the validity of his certificate.

Ordinary verification of a single certificate is a complex and time consuming process (Lloyd 2002), which consists of:
• Calculation and checking the certificate's signature value;
• Checking the validity period and the validity of the certificate policy;
• Checking intended key usage;
• Checking the revocation status of the certificate.
Cryptographic verification of the certificate's signature value and checking of the revocation status of the certificate are the two most time-consuming and expensive parts of the verification. In order to verify the certificate path, the verifier needs to check the revocation status of each single intermediate certificate and to store the public keys of the certificates of all CAs involved within his security domain (Lloyd 2002).

One way of improving certificate path verification time is to make the CA responsible for verifying the public keys of the end users via certificate paths and for issuing direct classical certificates for them. A similar approach is used within the ICE-TEL (Chadwick et al. 1997) model. However, direct certification can rarely be used for hierarchical PKIs where the topology and trust relationships must be preserved due to pre-established relationships between CAs at different levels. Levi in (1999), proposes NPKI to extract efficiently verifiable certificate paths. NPKI is based on a *nested certification* concept, where both classical and nested certificates are used together. The verifier, in order to verify a certificate using its *nested certificate path* (described in section 3), performs only one cryptographic computation during the certificate verification process, whereas in the classical PKI model he must perform two or more cryptographic computations depending on the path length.

However, as pointed out earlier, the NPKI model is impractical for large scale deployment, since it greatly increases the certificate load in the system. This increase is not evenly distributed among CAs in the hierarchy. Higher level CAs produce more certificates than lower-level CAs, with the Root CA supporting the largest increase. This is the major drawback of the NPKI model. A large additional increase in the number of certificates issued by the Root CA evolves down to additional certificate issuing costs at lower levels. A comparative analysis of the number of certificates issued is given in section 5.

In this paper we solve the drawback of Levi's NPKI model by introducing Compressed Nested PKI in which certificate verification time is identical to that in the NPKI model while the total number of certificates issued in a system (and by the Root CA) is very close to the total number of certificates issued. Thus, our model provides a fast certificate verification process while preserving the trust structure and topology of the original model allowing implementers the option of partial integration with existing PKI or building it as an independent model.

3 NESTED CERTIFICATES

In tackling the problem of time spent verifying the public key of each certificate in a chain of certificates, Levi (1999) introduced the nested certificate model with the aim of reducing the number of public key cryptographic computations during a verification process. In the classical PKI model, the verifier only wants to find a public key of the certificate user, and to check correctness of binding between the public key and the identity of the certificate holder. However, he needs to cryptographically verify all certificates on the certification path in order to verify the certificate of the certificate user. Cryptographic verification of a certificate is done by:
Obtaining the public key from the CA and cryptographically un-signing the signature to obtain the hash of the certificate data;
Calculating the hash of the same certificate data directly; and
Comparing these two values.
This is a time inefficient process.

Levi's model is based on a nested certification [7] concept. A *nested certificate* is simply defined as "a certificate for another certificate" (Figure 1).

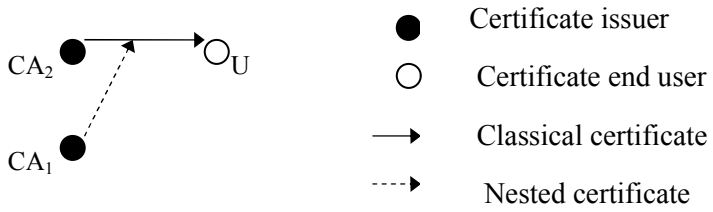
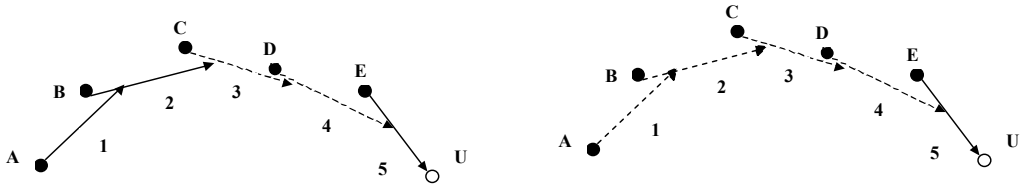


Fig. 1. The certificate relationships

A classical certificate assures that the binding between the public key and the user's identity is correct. A nested certificate, on the other hand, certifies another certificate. The latter is implemented to verify the legitimacy of the digital signature of another certificate and the integrity of data within that other certificate. Levi's NPKI model is based on this nested certification concept. A certificate that has been certified by a nested certificate is called a *subject certificate* (Levi et al. 2004).. A certification path in this model may contain both nested and classical certificates, but it should always end with a classical certificate. Also within any nested certificate path, only a first certificate needs to be verified cryptographically while all other certificates can be verified as subject certificates.

In order to verify a subject certificate a verifier needs to perform the following two steps:
To recalculate a hash of the subject certificate content and to check whether it is the same as the one stored within the nested certificate;

To compare the signature over the content of the subject certificate with the subject certificate's signature stored within the nested certificate. These two values should be the same.
As can be seen, the subject certificate verification does not use public key cryptographic computation. It performs only one fast hash function computation. Therefore, since subject certificate verification is more efficient than public key cryptographic protocol computation, the use of nested certificates reduces the overall verification time along a certificate chain. Levi shows in (1999) that the certificate path verification time within his model is reduced by a factor of 8 to a factor of 3000 depending on the size and complexity of the structure.



a. A certificate path with classical and nested certificates

b. A nested certificate path

Fig. 2. Examples of certificate paths with nested certificates

A *nested certificate path* is a certificate path in which all certificates in a path are nested certificates except a last certificate which must always be a classical certificate (Levi et al. 2004). Figure 2b shows a *nested certificate path*, whereas Figure 2a is a certificate path that contains nested certificates in it but is not a nested certificate path. In Figure 2a certificates 1, 2 and 3 are verified cryptographically, whereas certificates 4 and 5 are verified as subject certificates of the nested certificate. That means that the verifier would need to perform three public key cryptographic computations for the certificate with this certification path. A nested certificate path is given by 3, 4 and 5.

4 COMPRESSED NESTED CERTIFICATES – OUR MODEL

A nested certificate by definition does not provide any identity assurance of an entity. The purpose of a nested certificate is to assure legitimacy and integrity of a subject certificate. In other words it assures that the subject certificate content has been signed by the claimed CA, and that it has not been maliciously modified. This means that a nested certificate does not need to verify a single subject certificate. Therefore, our proposed new certificate, which we call a *compressed nested certificate (CNC)*, allows simultaneous verification of a number of different subject certificates. We say that a nested certificate which verifies data of two or more subject certificates has the *one-to-many property.* Classical and nested certificates have the *one-to-one property*.

Figure 3 shows a Compressed NPKI where each compressed nested certificate verifies two subject certificates.
In order to provide identity assurance and integrity of subject certificates, a compressed nested certificate needs to contain both hash values and digital signatures over the subject certificate's contents along with some unique identifier of the subject certificate, such as the certificate serial number.

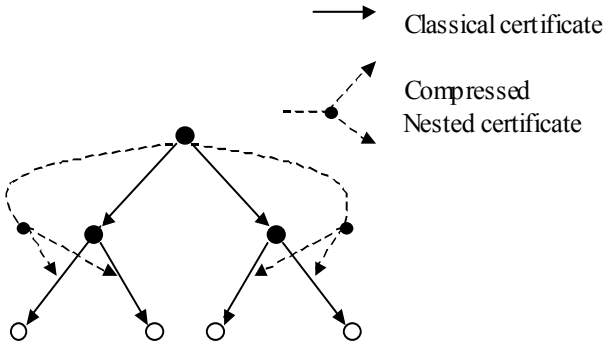


Fig. 3. An example of a CNPKI

A compressed nested certificate is a certificate for other certificates; or in other words it is a node-to-arc arc (as described by Figure 3). Compressed nested certificates can be used together with classical certificates in the corresponding CNPKI model. In the CNPKI, a Compressed Nested CA behaves like a classical CA. The aim of the compressed nested certificate is to improve the verification speed of the end-user's certificate

with a minimal increase in the certificate load. Therefore, we stipulate that compressed nested certificate paths be issued only for end-user certificates and not for classical certificates that belong to other CAs. Now, the basic compressed nested certificate issuance rule is the following:

Let CA be a fixed upper certification authority and CA^c be the set of certificate authorities that have been certified by the CA. In order to issue a compressed nested certificate, CA needs to check the validity of the public keys of each member of CA^c and also the validity of the certificates issued by members of CA^c. After that the CA can safely issue compressed nested certificates for certificates issued by authorities in CA^c.

In the initial step, a compressed nested certificate path propagates from end nodes towards the Root CA. The bottom level CAs issue only end-user certificates. Compressed nested certificates are issued by the lowest level CAs who certify other CAs. Also each CA in the CNPKI has a choice whether it wants to issue a compressed nested certificate or a classical certificate for the lower level CAs. In this way, CAs may choose to issue end-user certificates with compressed nested certificate paths only to some certificates in the system.

Each compressed nested certificate is created for a number of subject certificates. End-users usually apply for their certificates in a random manner. Therefore, a drawback of this new model would be additional accumulated time spent on collection of those subject certificates. If each compressed nested certificate is supposed to verify up to *n* subject certificates, then this may not be achieved with a sufficient rapidity to satisfy end-users. Therefore, we propose to incorporate a time constraint as well as a subject certificate constraint on the system. The time interval may vary in any particular CNPKI application, but should be determined using both nested certificate creation time and nested certificate verification time. CNC creation takes approximately the same time as for creation of a classical certificate. In fact, in the hardware, processing speeds of up to 600 1024-bit RSA signatures per second are possible (HSM 2006). This number is insignificant for larger verifiers that need to manage several millions of certificates per day. Thus, two constraints that need to be identified within our CNPKI model are:
A waiting-time constraint *T* and,
A maximum number *n* of subject certificates verified with each nested certificate.
A CNC is then validated when either the waiting time constraint *T* is achieved or when *n* certificates have accumulated, whichever occurs first.
The number *n* should not be too large as otherwise the compressed nested certificate data structure becomes unwieldy; on the other hand, a larger value of *n* results in a larger reduction in certificate repository storage load, which is an especially important factor for root CAs.

4.1 Implementation of Constraints

The method of deployment of the constraints within CNPKI depends on the requirements of the security environment. Where the order and consistency of the system is top priority, it is appropriate to have the RCA control them and fix them across the system. This would be the case for Web PKIs where certificate information is usually already en-coded and fixed within the software. The drawback in this situation is that only a small number of certificates might accumulate in some parts of the CNPKI structure before time *T* is reached. Alternatives are to allow each CA to control *n* or *T,* or both, independently of the RCA. This is useful for changeable environments where flexibility and adaptability of the system is most important. This alternate approach is especially useful for security environments with changeable conditions, particularly where human factors are involved. For example, suppose that a large organization deploys a PKI structure across separate branches in different geographic locations each with its own CAs. It might happen that those branches have different strategic purposes within the organization, and that consequently security requirements with different certification policies for PKI are deployed.

This leads us to define two methods for deployment of constraints within the CNPKI:
1. *Fixed* – where the RCA initially sets up the values for those two constraints and publishes them within the certification policy (following X.509).
2. *Flexible*- where each CA is responsible to deploy those two values independently.

In implementation, e*ach CA within the CNPKI might be allowed to decide itself the number of certificates each CNC* is allowed to certify, and how long it will wait for subject certificates to arrive for a certification. On the other hand, a general policy might be established with guidelines for deciding them. In either case, the algorithm below describes the actions of a CA in generating a nested certificate.

Algorithm for nested certificate generation:
Summary: Given a nested certificate time interval T and a maximum number of subject certificates certified by each nested certificate n, a CA should do the following:
Set an internal clock at $t=0.00$ and activate it at the moment when the first subject certificate SC_1 arrives

Collect all incoming subject certificates generated by the lower level CAs until either the number of them reaches n or the internal clock reaches the time T (i.e. $t=T$)
For the i-th subject certificate SC_i, where $i \leq k$ (where $k \leq n$ is the number of subject certificates waiting to be certified by the nested certificate) the CA does the following:
Assign the hash value of the content of the SC_i to the CNC,
Assign the signature value of the content of the SC_i to the CNC,
Assign the serial number of the SC_i to the CNC.
Create other certificate fields within the CNC.

5 PERFORMANCE ANALYSIS

In order to compare performance efficiency of the *CNPKI* against that of classical PKI and NPKI, a numerical analysis of verification speed and load on servers is given below. The number of paths is specific to the topology of PKI. Since it is almost impossible to formulate the average path lengths and the average number of certificates issued by CAs for irregular graph shaped PKIs, we use a balanced tree topology, prevalent in practice, in the analysis. The topology analysed in this section is that of an l-level, m-ary balanced tree where each non-end node (or CA) issues m classical certificates for their child nodes (CAs or end-users), and there are l non-leaf node levels (l is also the length of each end-user's certificate path). Performance measurements took place on a P3 workstation with a speed of 2.1 GHz and a RAM of 256 Mbytes, running the Windows XP operating system. Data from Figure 4 was obtained from (Levi et al. 2004). It shows an improvement in certificate verification speed of the CNPKI model over the classical PKI. This improvement in verification speed, referred to as the speed-up factor, is measured as the time spent on cryptographic verification of a certificate with a compressed nested certificate path divided by the time spent on cryptographic verification of a certificate with a classical certificate path of the same length and with the same algorithms used. The certificate verification speeds have small variations for different types of hash and digital signature algorithms deployed. Therefore, measurements were performed with different types of digital signature algorithms used with various lengths of keys (RSA and DSA algorithms with 512 bit, 1024 bit and 2024 bit keys are tested) and also with different hash algorithms applied (MD5 and SHA1 were used) and average values for certificate paths of various lengths have been used for further measurements. An improvement of certificate verification speed within CNPKI over certificate verification in the classical PKI model is shown in Figure 4.

The graph shows that the certificate verification speed-up factor based on those two models increases almost linearly as the certificate verification path length increases, which means that CNPKI is more efficient for certificates with medium to long certificate paths. Thus, if a certificate has certificate path with length l, the verification of a certificate with a compressed nested certificate path is l times faster then with the classical certificate path. According to our comparison between Levi's NPKI and our model, the certificate verification speed of a certificate with a compressed nested certificate path is almost identical to the verification speed of a certificate with a nested certificate path.

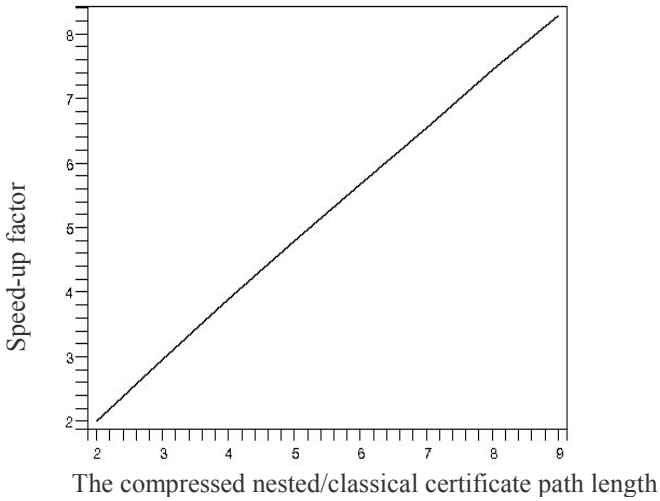
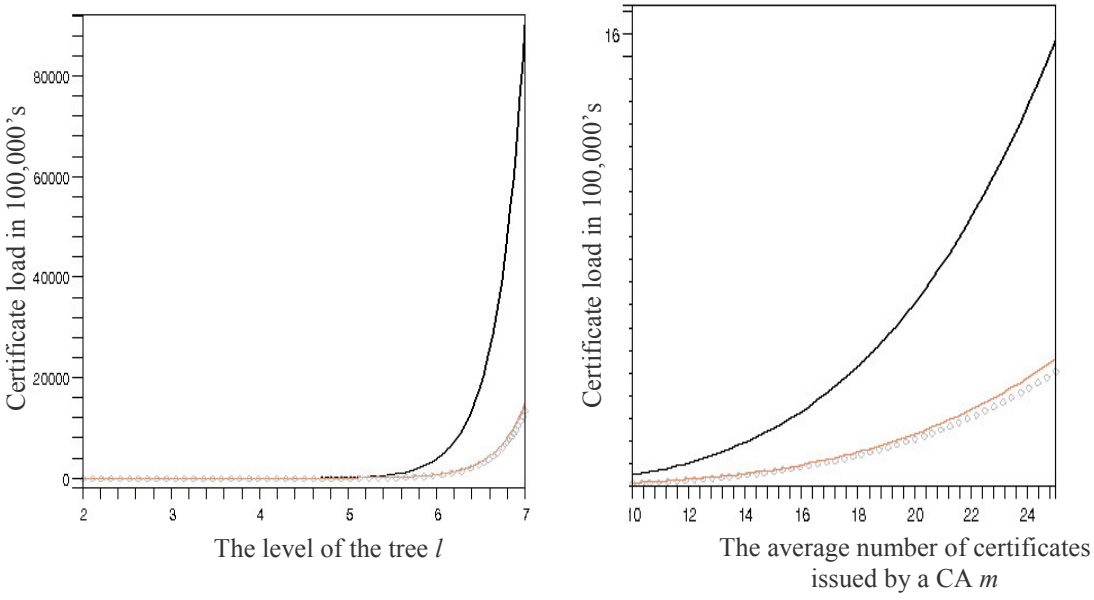


Fig 4. The speed-up factor or quotient of increase in certificate verification speed of compressed nested certificate paths compared to classical certificate paths for different levels *l* of the tree

In the next two figures we show an improvement in certificate load within the CNPKI model over the NPKI model. Figures 5 and 6 represent changes in certificate load for the whole system and for the Root CA respectively. Computations are performed with respect to three different factors: level of the tree *l*; average number of certificates *m* issued by CAs; average number of subject certificates *k* issued by each compressed nested certificate.



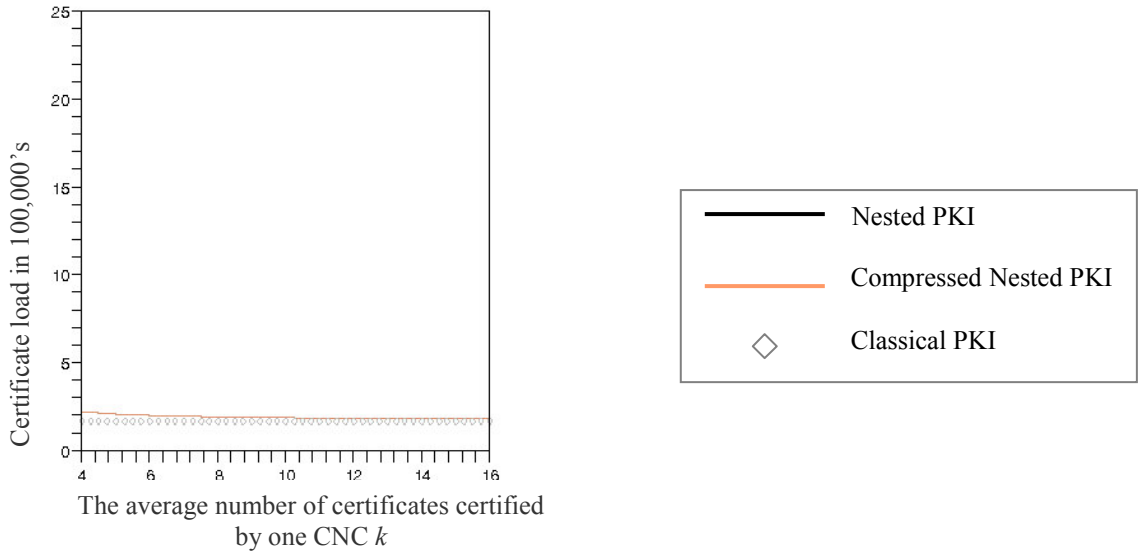


Fig 5. The total number of certificates in a system issued within classical PKI, NPKI and CNPKI with respect to *l, m* and *k*

Data from Figure 5 shows that the number of certificates within the CNPKI converges towards the number of certificates issued within the classical PKI, whereas the number of certificates within NPKI becomes significantly bigger when one of the factors increases.

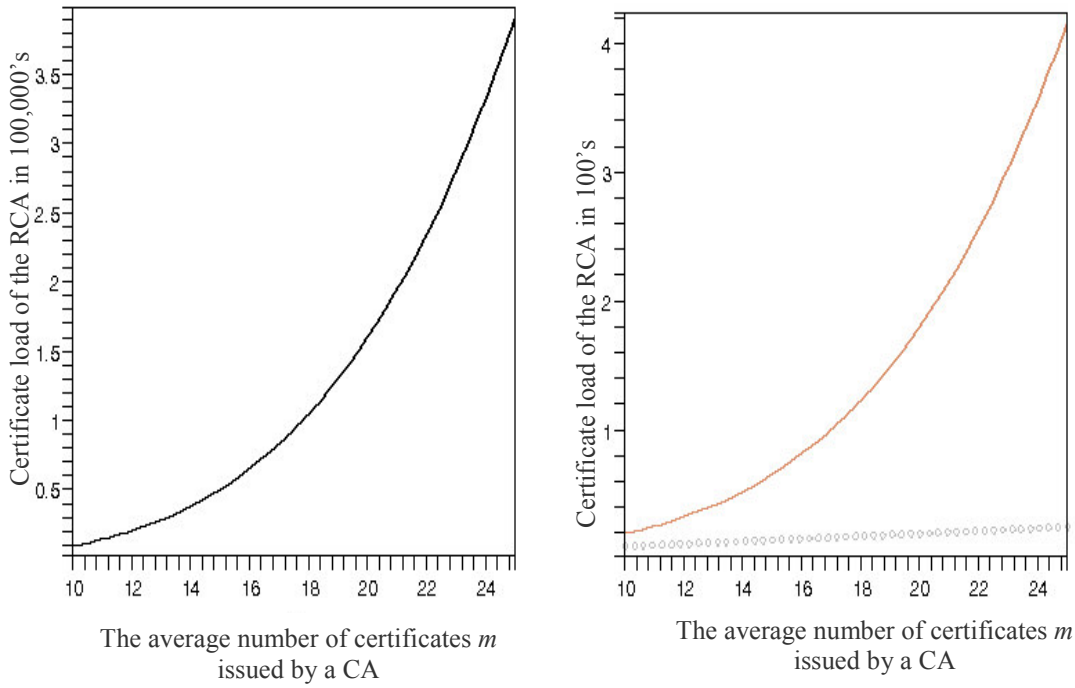


Fig 6. The number of certificates issued by the Root CA within classical PKI, NPKI and CNPKI with respect to the average number of certificates *m* issued by a CA

Measurements provided above show that a difference in the certificate load of NPKI and CNPKI is significantly larger for the Root CA than for the whole system. This is best shown by the first graph in Figure 6 where the Root CA within NPKI issues more than 200,000,000 certificates, whereas the number of certificates issued by the Root CA within CNPKI is less than 700 for a tree of level 6 (i.e. if we suppose that m=20 and the average number of certified certificates by one CNC is k=10). In order to represent an increase

in the number of certificates in a model, Levi et al. in (2004) defined the *nested certificate overhead*, or *NCO*. The *NCO* is the quotient of the total number of certificates, defined as *NTotal*, in his model and the total number of certificates issued in classical PKI, defined as *CTotal,* represented as follows:

$$NCO = \frac{NTotal}{CTotal} = \frac{(l-1)(m-1)m^{l-1}}{m^l - 1} + 1. \quad (1)$$

This notion can also be defined in our model as *compressed nested certificate overhead*, or *CNCO*, to represent the quotient of the total number of certificates in the CNPKI model, defined as *CNTotal*, and the total number of certificates in classical PKI, as follows:

$$CNCO = \frac{CNTotal}{CTotal} = \frac{m^{l-1}(m-1)(1-k^l)}{k^l(1-k)(m^l-1)} + 1. \quad (2)$$

The ratio between the number of certificates issued by the Root CA in our model and in Levi's model is given by:

$$(m + \left\lceil \frac{m^l}{n^{l-1}} \right\rceil) / (m + m^l) \rightarrow \frac{1}{n^{l-1}},$$

as m increases, which means that for larger *n* and longer certificate paths, the difference between the number of certificates issued within NPKI and CNPKI becomes significantly large (eg. for the number of subject certificates certified by one CNC $n10$ and length of a certificate path $l>3$, the difference between the number of certificates issued by the Root CA of CNPKI and the Root CA of NPKI is in the thousands). Thus, with CNPKI we achieve fast certificate verification time with an insignificant increase in the number of certificates issued by upper level authorities. Therefore, CNPKI appears to be most advantageous for highly used public key infrastructures where keeping certificate verification time to a minimum is a priority, such as, for instance, the Domain Name System Security (Eastlake 1999).

The next section will describe how our model deals with two important issues related to certificate verification: certificate revocation, and certificate expiry and renewal.

6 CERTIFICATE REVOCATIONS AND RENEWAL

Classical digital certificates are issued with a limited lifespan, but very often they need to be revoked before the expiration time. However, digital certificates are hard-coded data and they can not be undone (Stallings 2003). Therefore, CAs need to issue a separate, signed revocation statement, invalidating unexpired certificates. Certificates are normally revoked for one of two reasons:

1. The user's private key (or its CA's private key) has been compromised;
2. The relationship under which the private key was issued has changed.

There is a number of different certificate revocation mechanisms proposed in the literature. The three most commonly deployed revocation models are Certificate Revocation Lists (CRLs), Online Certificate Status Protocol (OCSP) and Certificate Revocation Trees (CRTs).
In classical PKI the verifier should verify a path of certificates in order to trust the end-user's certificate. Therefore, the verifier should obtain and check the revocation status of all certificates on the path. Thus, the difficulty of certificate revocation control is proportional to the number of CAs involved on the path.

A large number of compressed nested certificates (or nested certificates) are issued within the CNPKI and the NPKI model, and it seems that this would cause an increased certificate revocation burden in addition to the revocation of classical certificates in CNPKI. However, CNPKI requires only two revocation checking controls for a certificate path of any given length. The responsibility of the upper level CA in the

compressed nested certificate path creation process is to check the trustworthiness of the lower level CA and its compressed nested certificate before it issues a compressed nested certificate for it. Therefore, an intermediate compressed nested certificate from the path does not need to be revoked. In fact, it can be used in a path as a valid certificate even though its certifying pair of keys has been revoked or compromised. However, the verifier needs to check the revocation status for the first Root CA's compressed nested certificate in order not to start the revocation process with the bogus certificate. The revocation status for the end-user's certificate should be checked independently of the PKI model it is used in.

All classical certificates have a limited lifespan in order to limit the chance of misusing a compromised private key, or the privileges that a certificate holder has with the certificate. The aim of the compressed nested certificate is to verify a number of classical certificates at one time. Consequently, a compressed nested certificate can exist on several compressed nested certificate paths and its expiration time depends only on the expiration time of its subject certificates. Therefore, when the classical certificate is revoked or expired at the end of one of those paths, a compressed nested certificate on the path should not be removed unless there is no other valid classical certificate verified by this certificate. It should remain in the system as long as there is at least one valid subject certificate to verify. Thus, a compressed nested certificate automatically expires when the last certificate at the end of one of its paths become invalid. Therefore, the expiration time period of the compressed nested certificate within the path is equal to the latest expiration time period defined within all its end-users' classical certificates, located at the end of all its paths. All new compressed nested certificates need to be reissued when the new classical certificate has been reissued. If classical certificates have been temporarily revoked and are reissued later on, their original compressed nested certificate path can be re-used.

7 SECURITY ASSURANCE OF CNPKI

An important precondition of our model is that it is at least as secure as classical PKI. In other words, CNPKI must be resistant to those attacks common for all PKI models, such as:
Extensive crypto-analytic attack;
Forging a certificate with an expired key;
Impersonation or identity fraud attack;
Bribe an employee of a CA.

There are two possible kinds of crypto-analytic attacks on PKI: an attack on a certificate's private key and an attack on a hash function used in a certificate. The risk of the second attack on hash functions used by compressed nested certificates is larger then that for classical certificates, since a larger data space has been hashed in these certificates. The risk of a crypto-analytic attack on the certificate users' private keys is identical to the risk within the classical PKI. The solution to this kind of attack is usage of longer keys that need to be regularly changed. However, compressed nested certificates issued by intermediate CAs do not need to be revoked if the CAs' keys have expired, as long as they are issued before their expiration time. One factor that makes CNPKI more vulnerable is that the Root CA is regularly using a private key. In the classical PKI, a CA's private keys are strictly protected and activated rarely, and under strict supervision.

The risk of forging a certificate with a CA's expired key is lower within CNPKI than within the classical PKI model, since an upper level authority within CNPKI needs to check the validity of a compressed nested certificate issued by a lower level CA before it issues a compressed nested certificate, whereas this is not the case for the classical PKI model.

It might seem that CNPKI is more vulnerable to an impersonation attack than classical PKI, since compressed nested certificates carry some private information (eg. subject certificate hash and signature values) about other users' certificates. However, data within any digital certificate is publicly available, so this is, in fact, not an issue. Additionally, users usually send their certificates with data signed with their private key, and they are the only entities that know their private key. Therefore, the CNPKI model invades the privacy of entities (eg. end-users and CAs) involved in a system to only a very slight extent.

The risk of a bribing an employee of a CA is slightly lower within CNPKI than within classical PKI, since CAs within CNPKI are more dynamically involved in the end-users' certification process than they are in classical PKI and there is greater control of the work of lower level CAs by upper level authorities.

8 CONCLUSION

In this paper we have introduced an innovative compressed nested PKI model with compressed nested certificates as an alternative, more efficient solution to PKI that simplifies many important components of the verification. A major practical advantage of our model is that for large PKIs with longer certificate paths, where fast and simple certificate verification is a priority, it provides both speed and simplicity. It also preserves the trust structure and the topology of hierarchical PKI. CNPKI can be easily integrated with minimal cost into existing X.509 certificate schemes. Compressed nested certificates can be built by using currently deployed X.509 version3 digital certificate data structures with some minimal editing to the X.509 standard. CNPKI can be deployed either as an independent model or as an extension to classical PKIs where entities decide which certificate path model they want to use.

9 REFERENCES

Adams, C. and Lloyd, S, 2003, *Understanding PKI- Concepts, Standards, and Deployment Considerations*, Addison-Wesley, 2nd Edition, Boston.

Chadwick, D. W. , Young, A. J. & Cicovic, N. K. May/June 1997, 'Merging and Extending the PGP and PEM Trust Models – The ICE-TEL Trust Model', *IEEE Network*, vol. 11, no. 3, pp.16-24.

Eastlake, D. 1999, *Domain Name System Security Extensions*, RFC 2535, http://www.rfc-editor.org/rfc/rfc2535.txt

Henderson, M., Coulter, R., Dawson E. & Okamoto, E. 2002 'Modelling Trust Structure for Public Key Infrastructures' *ACISP 2002*, Springer-Verlag Berlin Heidelberg 2002, LNCS 2384, pp.56-70.

ISO/IEC 9594-8: 1998 Information Technology-Open Systems Interconnection – The Directory: Authentication Framework 1998, ITU-T Recommendation X.509.

Levi, A. 1999, *Design and Performance Evaluation of the Nested Certification Scheme and its Application in Public Key Infrastructures,* PhD Thesis, Department of Computer Engineering, Bogazici University.

Levi, A., Caglayan, M. U. & Koc C. K. February 2004, 'Use of Nested Certificates for Efficient, Dynamic, and Trust Preserving Public Key Infrastructure', *ACM Transactions on Information and System Security*, vol. 7, no. 1, pp. 21-59.

Lloyd, S. 2002, *Understanding Certification Path Construction,* PKI Forum: Understanding Certification Path Construction Key Infrastructure: Certification Path Building. White Paper, http://www.pkiforum.org/pdfs/Understanding_Path_construction-DS2.pdf

Cooper, M., Dzambasow, Y., Hesse, P., Joseph, S. & Nicholas, R. September 2005, *Internet X.509 Public Key Infrastructure: Certification Path Building*, RFC 4158.

The Hardware Security Module: HSM 2006, SafeNet's Family of HSMs, http://www.safenet-inc.com/products/pki/index.asp

Stallings, W. 2003, *Cryptography and Network Security Principles and Practise*, 3rd ed. Prentice-Hall, Englewood Cliffs, NJ (Chapter 15).

Fast Arithmetic In Jacobian Of Hyperelliptic Curves Of Genus 2 Over GF(p)

V. Kovtun[1] and J. Pelzl[2]

[1]Senior researcher, Kharkiv Air Force University, Ukraine
E-mail: vladislav.kovtun@gmail.com

[2]Chief Technology Officer, Escrypt GmbH, Germany
E-mail: jpelzl@escrypt.com

Abstract

In this paper, we suggest a new fast transformation for a divisor addition for hyperelliptic curves. The transformation targets the Jacobian of genus-2 curves over odd characteristic fields in projective representation. Compared to previously published results, the modification reduces the computational complexity and makes hyperelliptic curves more attractive for applications.

Introduction

Cryptographic systems have become an integral element in a wide spectrum of modern information and telecommunication systems. Nowadays, bulks of circulating information have to be processed by efficient ITS clients. With the extensive growth of computational power and mathematical methods for cryptanalysis, the requirements to existing and perspective cryptosystems are challenging. In other words, the main requirement when developing a cryptosystem is the capability to assure a required security level, taking recent development in cryptanalysis into account.
It is widely agreed that new cryptosystems such as elliptic curve cryptosystems (ECC) and hyperelliptic curve cryptosystems (HECC) will very likely dominate applications on constraint devices in the future. This work is another step towards an efficient HECC.

The challenge to develop practical systems is dominated by the difficulty of its efficient implementation on modern hardware such as, e.g. embedded systems. The efficiency of a cryptosystem is directly related to the public key primitives that form its basis.
Important primitives use transformations in rings and fields. More recently, transformations of the group of points on an elliptic curve (EC) have been widely used and are to be found in various international and government standards (ISO 2002, IEEE 2000, DSTU 2002). Such transformations allow for building a cryptosystem with a high security level at quite low computational costs.
Further development of cryptographic transformations using algebraic curves implies applications of even more complex curves –hyperelliptic curves (HEC) – which increases the computational complexity (hereinafter complexity) of such systems. Therefore, papers dealing with the decrease of complexity of prospective cryptosystems are of extreme importance.
Cryptosystems based on HEC are dominated by operations over reduced divisors, in particularly by scalar multiplications of reduced divisors (Koblitz 1989) with its corresponding basic operations of addition and doubling divisors.
The significant decrease in the complexity of calculations based on divisor transformations in the Jacobian of a HEC can be achieved on account of the decrease in complexity of a divisor scalar multiplication algorithm.

The earliest publications dedicated to the arithmetic in the Jacobian of HEC are due to (Cantor 1987, Lange 2002a) and are rather of theoretical interest.
Recently, multiple scientists done intense research in the effective implementation of the arithmetic in the Jacobian of HEC (Cantor 1987, Koblitz 1989, Spallek 1994, Harley 2000, Kriger 2001, Miyamoto et al. 2002, Lange 2001, 2002a, 2002b, 2002c, Takahashi 2002, Suguzaki 2002). In these contributions, the attention is particularly paid to the reduction of field operations by applying HEC with fixed genus. This allows various modifications of the arithmetic in the Jacobian of HEC. Papers (Koblitz 1989, Kriger 2001) deal with methods of addition and doubling of divisors in the Jacobian of genus 2 HEC. The first practical implementation of these methods was described in (Harley 2000). In (Wollinger 2004), generalized results (Harley 2000) for curves over the even characteristic fields are shown. The development of addition and doubling methods is described in (IEEE 2000, Cantor 1987, Harley 2000).
The time required for a scalar multiplication in the Jacobian of genus 2 HEC as the scalar multiplication in the EC curve is main part of discrete logarithm based cryptosystems. Under this circumstance, the challenge of improving the performance of a cryptosystem and, in particular, a divisor scalar multiplication in the Jacobian of HEC becomes of special urgency.
Curves of genus 2 are of our main interest when decreasing the arithmetic complexity in the Jacobian of HEC. Thus, we will take genus-2 curves into our further consideration.
As it was proved in (Lange 2002b, Hankerson et al. 2000), the divisor addition and doubling operations in the Jacobian of HEC perform a very complex field operation – the field inversion. According to publications Hankerson et al. (2000), Brown et al. (2000), for $\mathbf{GF}(p)$, the complexity of a field inversion I depends on the actual platform and lies in the interval ($80M$, $90M$) Brown et al. (2000), where M is the complexity of field multiplication.
At publication (Miyamoto et al. 2002) proposes for the first time an approach to implement the arithmetic in a Jacobian of HEC of genus 2 without using any field inversion in intermediate computations. The further development of the proposed approach was shown in papers (Lange 2002b, Lange 2002c), where the results were improved and extended to a wider class of HEC over even characteristic fields (Lange 2002b, Lange 2002c). As a prototype for practicable methods under discussion, the results of (Lange 2002b, Lange 2002c) are considered.

Mathematical Background

In this paper, we are considering genus-2 HEC: $v^2 + h(u)v = f(u)$ over field $\mathbf{GF}(p)$ with odd characteristic, where $h(x) = 0$, $f(x) = x^5 + f_3x^3 + f_2x^2 + f_1x + f_0$, $f_i \in \mathbf{GF}(p)$.
The divisor representation in Mumford form is given as $[u, v]$, $u(x) = x^2 + u_1x + u_0$, $v(x) = v_1x + v_0$, $\deg v < \deg u \le 2$ and will be denominated as affine representation. The representation that does not involve field inversion will be denominated as projective. In our case, the divisor in projective representation is given by $[u, v], u(x) = x^2 + U_1/Z\,x + U_0/Z$, $v(x) = V_1/Z\,x + V_0/Z$, and is represented as $[U_1, U_0, V_1, V_0, Z]$ (Miyamoto et al. 2002, Lange 2002b). Where the divisor in weighted representation $[u, v], u(x) = x^2 + U_1/Z_1^2\,x + U_0/Z_1^2$, $v(x) = V_1/Z_1^3Z_2\,x + V_0/Z_1^3Z_2$, is represented as $[U_1, U_0, V_1, V_0, Z_1, Z_2]$ (Lange 2002b).
The aim of this paper is to develop a modified method of the arithmetic in the Jacobian of genus 2 HEC in the projective representation with the purpose to increase the performance of the scalar multiplication in a HECC.
Under the accepted model (Lange 2002a, Harley 2000), a typical addition of divisors is given by the addition of the divisors $[u_1(x), v_1(x)]$ and $[u_2(x), v_2(x)]$, where the resultant $r(u_1(x), u_2(x))$ is not

equal to zero, and under doubling of divisor $[u_1(x), v_1(x)]$, where the resultant $r(u_1(x), h(x)+2v_1(x))$ is not equal to zero.
The proposed modification that provides a reduced complexity is based on Harley's method (Harley 2000) and its modification (Lange 2002a, Wollinger 2004). In this context, we suggest to use the projective representation of divisors in the method being described.
In the algorithms of addition and doubling (Harley 2000, Wollinger 2004), the most difficult parts in terms of computational complexity are operations in the polynomial function ring: division, inversion, reduction, and multiplication.

Proposed Efficient Arithmetic In Jacobians Of HEC Over $\mathbf{GF}(p)$

To decrease these operations, we propose to modify the addition and doubling algorithms as follows:
pass directly from operations in the polynomial function ring to the field operations using HEC with fixed and small genus (i.e., 2) (Spallek 1994, Harley 2000);
simplify the arithmetic in the polynomial function ring by polynomial normalizations;
- normalize and minimize the Hamming weight of HEC parameters $h(x)$ and $f(x)$. These parameters set up a special kind of HEC (Miyamoto et al. 2002, Wollinger 2004);
- normalize polynomial function $u(x)$, steps 3 and 4 of algorithms 1, 2; steps 4 and 5 of algorithms 3, 4 (Miyamoto et al. 2002, Wollinger 2004);
- simultaneously invert several field elements with the Montgomery trick (Lange 2002b, Wollinger 2004);
- multiply polynomial functions of different powers with the Karatsuba method, step 5 of algorithms 1, 2; step 6 of algorithms 3, 4 (Wollinger 2004);
- reduce polynomial functions by the Karatsuba method; step 3 of algorithms 1, 2; step 4 of algorithms 3, 4 (Harley 2000);
- exclude multiplicative field inversion by using the projective representation of divisors; step 2 of algorithms 1, 2, 3, 4 (Lange 2002b, Miyamoto at al 2002).

Using the above proposed modifications, we obtain arithmetic algorithms on HEC defined by the equation $v^2 + h(u)v = f(u)$ over field $\mathbf{GF}(p)$ with odd characteristic in projective representation, where $f(x) = x^5 + f_3x^3 + f_2x^2 + f_1x + f_0$, $f_i \in \mathbf{GF}(p)$, $h(x) = 0$. The proposed addition algorithm is described in Algorithm 1.
Particularly, with the algorithm of addition, there often arises a situation where one of the input divisors is affine (Z is equal to 1), and the other one is projective. The result of addition is obtained in the projective representation. For Algorithm 1, this input data allows for its simplification to Algorithm 2 which comes with a decreased number of field operations and, hence, is of decreased complexity.

Algorithm 1. Addition of divisors
Input: $\mathrm{div}(U_{11}, U_{10}, V_{11}, V_{10}, Z_1)$, $\mathrm{div}(U_{21}, U_{20}, V_{21}, V_{20}, Z_2)$
Output: $\mathrm{div}(U_1', U_0', V_1', V_2', Z') = \mathrm{div}(U_{11}, U_{10}, V_{11}, V_{10}, Z_1) + \mathrm{div}(U_{21}, U_{20}, V_{21}, V_{20}, Z_2)$

	Operation	Cost
1	Compute resultant r: $Z = Z_1 \cdot Z_2$, $\widetilde{U}_{21} = Z_1 \cdot U_{21}$, $\widetilde{U}_{20} = Z_1 \cdot U_{20}$, $\widetilde{V}_{21} = Z_1 \cdot V_{21}$, $\widetilde{V}_{20} = Z_1 \cdot V_{20}$, $y_1 = U_{11} \cdot Z_2 - \widetilde{U}_{21}$, $y_2 = \widetilde{U}_{20} - U_{10} \cdot Z_2$, $y_3 = U_{11} \cdot y_1 + y_2 \cdot Z_1$, $r = y_2 \cdot y_3 + y_1^2 \cdot U_{10}$.	1S, 11M
2	Compute almost inversion $inv = r/u_2 \bmod u_1$: $inv_1 = y_1$, $inv_0 = y_3$.	

3	Compute $s=(v_1-v_2)inv \bmod u_1$: $w_0=V_{10}\cdot Z_2-\widetilde{V}_{20}$, $w_1=V_{11}\cdot Z_2-\widetilde{V}_{21}$, $w_2=inv_0\cdot w_0$, $w_3=inv_1\cdot w_1$, $s_1=(inv_0+Z_1\cdot inv_1)\cdot(w_0+w_1)-w_2-w_3\cdot(Z_1+U_{11})$, $s_0=w_2-U_{10}\cdot w_3$. If $s_1=0$ then consider special case.	8M
4	Precomputation: $R=r\cdot Z$, $s_2=s_0\cdot Z$, $s_3=s_1\cdot Z$, $\widetilde{R}=R\cdot s_3$, $w_0=s_1\cdot s_0$, $w_1=s_1\cdot s_3$, $w_2=s_0\cdot s_3$, $w_3=w_1\cdot\widetilde{U}_{21}$, $w_4=R\cdot s_1$	9M
5	Compute $l=su_2$: $l_0=w_0\cdot\widetilde{U}_{20}$, $l_2=w_3+w_2$, $l_1=(w_1+w_0)\cdot(\widetilde{U}_{21}+\widetilde{U}_{20})-l_0-w_3$	2M
6	Compute $u'=(s(l+2v_1)-k)u_1^{-1}$, $k=(f-v_1^2)/u_1$: $\widetilde{U}_1'=2w_2-s_3\cdot s_1y_1-R^2$, $\widetilde{U}_0'=s_2^2+s_1\cdot y_1\cdot(s_1\cdot\widetilde{U}_{11}-2s_2)+y_2\cdot w_1+2w_4\cdot\widetilde{V}_{21}+R\cdot r\cdot(y_1+2\widetilde{U}_{21})$	2S, 8M
7	Correction: $U_0'=\widetilde{U}_0'\cdot\widetilde{R}$, $U_1'=\widetilde{U}_1'\cdot\widetilde{R}$, $Z'=s_3^2\cdot\widetilde{R}$	1S, 3M
8	Compute $v'\equiv-(s_1l+v_2)\bmod u'$: $V_1'=\widetilde{U}_1'\cdot(l_2-\widetilde{U}_1')+s_3^2\cdot(\widetilde{U}_0'-w_4\widetilde{V}_{21}-l_1)$, $V_0'=\widetilde{U}_0'\cdot(l_2-\widetilde{U}_1')-s_3^2\cdot(l_0+w_4\cdot\widetilde{V}_{20})$	5M
		4S, 46M

Algorithm 2. Mixed addition of divisors
Input: $\mathrm{div}(U_{11},U_{10},V_{11},V_{10},1)$, $\mathrm{div}(U_{21},U_{20},V_{21},V_{20},Z_2)$
Output: $\mathrm{div}(U_1',U_0',V_1',V_2',Z')=\mathrm{div}(U_{11},U_{10},V_{11},V_{10},1)+\mathrm{div}(U_{21},U_{20},V_{21},V_{20},Z_2)$

	Operation	Cost
1	Compute resultant r for u_1 and u_2: $\widetilde{U}_{11}=Z_2\cdot U_{11}$, $y_1=\widetilde{U}_{11}-U_{21}$, $y_2=U_{20}-U_{10}\cdot Z_2$, $y_3=U_{11}\cdot y_1+y_2$, $r=y_2\cdot y_3+y_1^2\cdot U_{10}$	1S, 5M
2	Compute almost inversion $inv=r/u_2 \bmod u_1$: $inv_1=y_1$, $inv_0=y_3$	
3	Compute $s=(v_1-v_2)inv \bmod u_1$: $w_0=V_{10}\cdot Z_2-V_{20}$, $w_1=V_{11}\cdot Z_2-V_{21}$, $w_2=inv_0\cdot w_0$, $w_3=inv_1\cdot w_1$, $s_0=w_2-U_{10}\cdot w_3$, $s_1=(inv_0+inv_1)\cdot(w_0+w_1)-w_2-w_3\cdot(U_{11}+1)$. If $s_1=0$ then consider special case	7M
4	Precomputation: $R=r\cdot Z_2$, $s_2=s_0\cdot Z_2$, $s_3=s_1\cdot Z_2$, $\widetilde{R}=R\cdot s_3$, $w_0=s_1\cdot s_0$, $w_1=s_1\cdot s_3$, $w_2=s_0\cdot s_3$, $w_3=w_1\cdot U_{21}$, $w_4=R\cdot s_1$.	9M
5	Compute $l=su_2$: $l_0=w_0\cdot U_{20}$, $l_2=w_3+w_2$, $l_1=(w_1+w_0)\cdot(U_{21}+U_{20})-l_0-w_3$	2M
6	Compute $u'=(s(l+2v_1)-k)u_1^{-1}$, $k=(f-v_1^2)/u_1$: $\widetilde{U}_1'=2w_2-s_3\cdot s_1y_1-R^2$, $\widetilde{U}_0'=s_2^2+s_1\cdot y_1\cdot(s_1\cdot\widetilde{U}_{11}-2s_2)+y_2\cdot w_1+2w_4\cdot V_{21}+R\cdot r\cdot(y_1+2U_{21})$	2S, 8M
7	Correction: $U_0'=\widetilde{U}_0'\cdot\widetilde{R}$, $U_1'=\widetilde{U}_1'\cdot\widetilde{R}$, $Z'=s_3^2\cdot\widetilde{R}$	1S, 3M
8	Compute $v'\equiv-(h+s_1l+v_2)\bmod u'$: $V_1'=\widetilde{U}_1'\cdot(l_2-\widetilde{U}_1')+s_3^2\cdot(\widetilde{U}_0'-w_4V_{21}-l_1)$, $V_0'=\widetilde{U}_0'\cdot(l_2-\widetilde{U}_1')-s_3^2\cdot(l_0+w_4\cdot V_{20})$	5M
		4S, 39M

Similarly, with the algorithm of doubling, there often arises a situation when the input divisor is affine (Z is equal to 1). The result of doubling is produced in projective representation. For Algorithm 3, this input data allows for its simplification to Algorithm 4 which incorporates a decreased number of field operations and which is of decreased complexity.

Algorithm 3. Doubling of divisor
Input: $\mathrm{div}(U_1,U_0,V_1,V_0,Z)$

Output: $\mathrm{div}(U_1', U_0', V_1', V_2', Z') = 2\,\mathrm{div}(U_1, U_0, V_1, V_0, Z)$

	Operation	Cost
1	Compute resultant r for u and $2v$ (where $\tilde{v} \equiv (2v) \bmod u$): $Z_2 = Z^2$, $\widetilde{V}_1 = 2V_1$, $\widetilde{V}_0 = 2V_0$, $w_0 = V_1^2$, $w_1 = U_1^2$, $w_2 = \widetilde{V}_1^2 = 4w_0$, $w_3 = \widetilde{V}_0 \cdot Z - U_1 \cdot \widetilde{V}_1$, $r = \widetilde{V}_0 \cdot w_3 + w_2 \cdot U_0$.	3S, 4M
2	Compute almost inversion $inv \equiv r/\tilde{v} \bmod u$: $inv_1 = -\widetilde{V}_1$, $inv_0 = w_3$.	
3	Compute $k \equiv \left[\left(f - v^2\right)/u\right] \bmod u$: $w_3 = f_3 \cdot Z + w_1$, $k_1 = 2w_1 + w_3 - Z \cdot 2U_0$, $k_0 = U_1 \cdot (4ZU_0 - w_3) + Z \cdot (f_2 \cdot Z - w_0)$.	5M
4	Compute $s = k \cdot inv \bmod u$: $w_0 = k_0 \cdot inv_0$, $w_1 = k_1 \cdot inv_1$, $s_0 = w_0 - ZU_0 \cdot w_1$, $s_3 = (inv_0 + inv_1) \cdot (k_0 + k_1) - w_0 - w_1 \cdot (1 + U_1)$, $s_1 = s_3 \cdot Z$. If $s_1 = 0$ then consider special case.	6M
5	Precomputation: $R = r \cdot Z_2$, $\widetilde{R} = R \cdot s_1$, $w_0 = s_1 \cdot s_3$, $w_1 = s_0 \cdot s_3$, $w_3 = w_1 \cdot Z$, $w_4 = R \cdot s_3$.	6M
6	Compute $l = su$, $l = l_2 x^2 + l_1 x + l_0$: $l_0 = U_0 \cdot w_1$, $l_2 = U_1 \cdot w_0$, $l_1 = (w_1 + w_0) \cdot (U_1 + U_0) - l_0 - l_2$.	3M
7	Compute $u' = \left\lfloor l^2 + \frac{1}{s} l 2v - \frac{1}{s^2}\left(f - v^2\right)\right\rfloor / u^2$: $\widetilde{U}_0' = s_0^2 + 2w_4 \cdot V_1 + R \cdot r \cdot 2U_1$, $\widetilde{U}_1' = 2w_3 - R^2$.	2S, 2M
8	Correction: $U_0' = \widetilde{U}_0' \cdot \widetilde{R}$, $U_1' = \widetilde{U}_1' \cdot \widetilde{R}$, $Z' = s_1^2 \cdot \widetilde{R}$.	1S, 3M
9	Compute $v' \equiv -(s_1 l + v_2) \bmod u'$: $V_1' = \widetilde{U}_1' \cdot \left(l_2 - \widetilde{U}_1' + w_3\right) + s_1^2 \cdot \left(\widetilde{U}_0' - w_4 V_1 - l_1\right)$, $V_0' = \widetilde{U}_0' \cdot \left(l_2 - \widetilde{U}_1' + w_3\right) - s_1^2 \cdot \left(l_0 + w_4 \cdot V_0\right)$.	5M
		6S, 35M

Algorithm 4. Mixed doubling of divisor
Input: $\mathrm{div}(U_1, U_0, V_1, V_0, 1)$
Output: $\mathrm{div}(U_1', U_0', V_1', V_2', Z') = 2\,\mathrm{div}(U_1, U_0, V_1, V_0, 1)$

	Operation	Cost
1	Compute resultant r for u and $2v$ (where $\tilde{v} \equiv (2v) \bmod u$): $\widetilde{V}_1 = 2V_1$, $\widetilde{V}_0 = 2V_0$, $w_0 = V_1^2$, $w_1 = U_1^2$, $w_2 = \widetilde{V}_1^2 = 4w_0$, $w_3 = \widetilde{V}_0 - U_1 \cdot \widetilde{V}_1$, $r = \widetilde{V}_0 \cdot w_3 + w_2 \cdot U_0$.	2S, 3M
2	Compute almost inversion $inv \equiv r/\tilde{v} \bmod u$: $inv_1 = -\widetilde{V}_1$, $inv_0 = w_3$.	
3	Compute $k \equiv \left[\left(f - v^2\right)/u\right] \bmod u$: $w_3 = f_3 + w_1$, $w_4 = 2U_0$, $k_1 = 2w_1 + w_3 - w_4$, $k_0 = U_1 \cdot (2w_4 - w_3) + f_2 - w_0$.	1M
4	Compute $s = k \cdot inv \bmod u$: $w_0 = k_0 \cdot inv_0$, $w_1 = k_1 \cdot inv_1$, $s_0 = w_0 - U_0 \cdot w_1$, $s_1 = (inv_0 + inv_1) \cdot (k_0 + k_1) - w_0 - w_1 \cdot (1 + U_1)$. If $s_1 = 0$ then consider special case.	5M
5	Precomputation: $\widetilde{R} = r \cdot s_1$, $w_0 = s_1^2$, $w_1 = s_0 \cdot s_1$.	1S, 2M
6	Compute $l = su$: $l_0 = U_0 \cdot w_1$, $l_2 = U_1 \cdot w_0$, $l_1 = (w_1 + w_0) \cdot (U_1 + U_0) - l_0 - l_2$.	3M
7	Compute $u' = \left\lfloor l^2 + \frac{1}{s} l 2v - \frac{1}{s^2}\left(f - v^2\right)\right\rfloor / u^2$: $\widetilde{U}_0' = s_0^2 + 2\widetilde{R} \cdot V_1 + r^2 \cdot 2U_1$, $\widetilde{U}_1' = 2w_1 - r^2$.	2S, 2M
8	Correction: $U_0' = \widetilde{U}_0' \cdot \widetilde{R}$, $U_1' = \widetilde{U}_1' \cdot \widetilde{R}$, $Z' = w_0 \cdot \widetilde{R}$.	3M
9	Compute $v' \equiv -(s_1 l + v_2) \bmod u'$: $V_1' = \widetilde{U}_1' \cdot \left(l_2 - \widetilde{U}_1' + w_1\right) + w_0 \cdot \left(\widetilde{U}_0' - \widetilde{R} V_1 - l_1\right)$, $V_0' = \widetilde{U}_0' \cdot \left(l_2 - \widetilde{U}_1' + w_1\right) - w_0 \cdot \left(l_0 + \widetilde{R} \cdot V_0\right)$.	5M
		5S, 24M

Computational Complexity

Table 1 summarizes estimates of the computational complexity for the known arithmetic (Miyamoto et al. 2002, Lange 2002a, Lange 2002b, Wollinger 2004) and the proposed Algorithm 1 to Algorithm 4 in the Jacobian of genus-2 HEC for common cases. The computational complexity is evaluated in terms of field operations.

Table 1. Summarized estimates of computational complexity of arithmetic in the Jacobian of HEC

Parameters of genus 2 HEC	Algorithm: Addition			Mixed addition			Doubling			Mixed doubling		
	$()^{-1}$	^2	*	$()^{-1}$	^2	*	$()^{-1}$	^2	*	$()^{-1}$	^2	*
Odd characteristic fields												
Affine coordinates												
$f_4 = 0$ (Lange 2002a)	1	3	22				1	5	22			
Projective coordinates $[U_1, U_0, V_1, V_0, Z]$												
$\deg(h) = 2$, $h_i \in \mathbf{F}_2$ (Lange 2002b)		4	47		3	40		6	40		5	25
$\deg(h) = 2$, $h_i \in \mathbf{F}_2$ (Kovtun, Wollinger 2007)		4	46		4	39		6	39		5	25
$h(x) = 0$, $f_4 = 0$ [proposed]		4	46		4	39		6	35		5	24
Weighted coordinates $[U_1, U_0, V_1, V_0, Z_1, Z_2, Z_1^2, Z_2^2]$												
$h(x) = 0$, $f_4 = 0$ (Lange 2002c)		7	47		5	36		7	34		5	21
Even characteristic fields												
Affine coordinates												
$f_4 = 0$ (Lange 2002a)	1	3	21				1	5	20			
$h_2 = 0$, $f_4 = 0$ (Lange 2002a)	1	3	21				1	5	17			
$h(x) = x$, $f_4 = 0$, $f_3 = f_2 = 0$ (Wollinger 2004)							1	6	9			
Projective coordinates $[U_1, U_0, V_1, V_0, Z]$												
$h(x) = x$, $f_4 = 0$, $f_3 = f_2 = 0$ (Wollinger 2004)		5	45		5	38		6	31		5	18
$h(x) = x$, $f_4 = 0$, $f_3 = f_2 = 0$ (Kovtun, Wollinger 2004)		4	44		4	37		6	30		4	17
Weighted coordinates $[U_1, U_0, V_1, V_0, Z_1, Z_2, Z_1^2, Z_2^2, Z_1 Z_2, Z_1^3 Z_2]$												
$f_4 = 0$, $h_2 \neq 0$ (Lange 200c)		4	46		5	35		6	35		5	24
$f_4 = 0$, $h_2 = 0$ (Lange 200c)		6	44		6	34		6	29		6	19

From Table 1, one can see that for $\mathbf{GF}(p)$ the proposed algorithms show a decreased complexity in comparison to (Lange 2002b, Wollinger 2004).

Conclusions

We developed a modification of the arithmetic in the Jacobian of genus-2 HEC in projective coordinates. The new formulas have a lower complexity in comparison to the existing algorithms (Lange 2002a, Wollinger 2004). The modification is characterized as follows:
- a decreased number of recomputed values due to pre-computations and reordering in comparison to (Harley 2000, Lange 2002b, Wollinger 2004);

- an increased number of pre-computed values due to reordering of field operations in comparison with (Harley 2000, Lange 2002b, Wollinger 2004);
- an increased number of pre-computed values due to the use of dependencies among resulting polynomial functions.
- a decreased Hamming weight of HEC parameters in analogy to (Lange 2002b).

The proposed modifications of the arithmetic in the Jacobian of genus 2 HEC allows for a 2% decrease in complexity compared to the best previous results (Lange 2002a, Lange 2002b, Kovtun, Wollinger 2007). From (Hankerson et al. 2000) it is known that the complexity of a simple scalar multiplication is, on average:

$$DSM = \tfrac{1}{2}t \cdot DA + t \cdot DD,$$

where t is the bitlength of the scalar, DSM is the complexity of the divisor scalar multiplication, DA is the complexity of the divisor addition, DD is the complexity of a divisor doubling. Thus, using the proposed algorithms, we decrease the complexity of the scalar multiplication by $1\tfrac{1}{2}t$ field multiplications in comparison with (Kovtun, Wollinger 2007) and by $7t$ filed multiplications in comparison with (Lange 2002a).
For the implementation formulas from (Lange 2002a, Kovtun, Wollinger 2007) and proposed are used Microsoft Visual C++ 2005 without assembler. For the performance benchmark is used workstation with AMD AthlonXP (Barton core) 2500+ MHz CPU and Microsoft Windows XP OS. Jacobian arithmetic based on the own fast finite field library tuned to the latest x86 processors family.

Table 2. Experimental timings for the simple scalar multiplication of weight 2 divisor [ms]

Results / Base field	$\mathbf{GF}(p_{80})$	$\mathbf{GF}(p_{84})$	$\mathbf{GF}(p_{161})$
$\deg(h)=2,\ h_i \in \mathbf{F}_2$ (Lange 2002b)	11352.6	12209.4	66288
$\deg(h)=2,\ h_i \in \mathbf{F}_2$ (Kovtun, Wollinger 2007)	11156.24	11998.22	65268.8
$h(x)=0,\ f_4=0$ [proposed]	10463	11252.66	61313.6

When using the proposed formulas, we can reduce the time for a scalar multiplication by approximately 6.2%.

References

Brown, M., Hankerson, D., Lopez, D., Menezes, A. (2000) Software implementation of the NIST elliptic curves over prime fields. Research Report CORR 2000–55. Department of

Combinatorics and Optimization, University of Waterloo. –Canada: Waterloo, Ontario, 21p.
Cantor, D.G. (1987) Computing in the Jacobian of hyperelliptic curve. Math. Comp., No 48. pp. 95-101.

Hankerson, D., Lopez, J., Menezes, A. (2000) Software implementation of elliptic curve cryptography over binary fields fields / In Cetin K. Koc and C. Paar editors. Workshop and embedded systems. –CHES'99. –LNCS 1717. –Berlin: Springer–Verlag. pp.1–24.

Harley, R. (2000) Fast arithmetic on genius 2 curves. Available at: http://cristal.infra.fr/~harley/hyper.
Institute of Electrical and Electronics Engineers. (2000) IEEE P1363–2000: Standard Specifications for Public Key Cryptography. 206p.

International Organization for Standardization. (2002) ISO/IEC FCD 15946-2: Information technology - Security techniques - Cryptographic techniques based on elliptic curves - Part 2: Digital signatures, Final Committee.

Koblitz, N. (1989) Hyperelliptic cryptosystems. Journal of cryptology, No 1. pp.139–150.
Kovtun, V., Wollinger, T. (2007) Fast explicit formulae for genus 2 hyperelliptic curves using projective coordinates. In pro. 4th International conference on Information Technology (ITNG'2007). pp. 893–897.

Kriger, U. (2001) Anwendung hyperellipischer kurven in der kryptographie. Master's thesis, Universitat Gesamthochschule Essen.

Lange, T. (2001) Efficient arithmetic on hyperelliptic curves. PhD thesis, Universitat Gesamthochschule Essen.

Lange, T. (2002a) Efficient arithmetic on genius 2 hyperelliptic curves over finite fields via explicit formulae. Cryptology ePrint Archive. Report 2002/121. Available http://eprint.iacr.org.

Lange T. (2002b) Inversion-free arithmetic on genius 2 hyperelliptic curves. Cryptology ePrint Archive. Report 2002/147. Available http://eprint.iacr.org

Lange, T. (2002c) Weighted coordinates on genius 2 hyperelliptic curves. Cryptology ePrint Archive. Report 2002/153. Available http://eprint.iacr.org.

Miyamoto, Y., Doi, H., Matsuo, K., Chao, J., Tsujii, S. (2002) A fast addition algorithm of genius two hyperelliptic curve. In the 2002 Symposium on cryptography and information security. – SCIS 2002, IEICE Japan, pp.497–502. In Japanese.

Spallek, A.M. (1994) Kurven vom geschlecht 2 und ihre anwendung in public-key-kryptosystemen. PhD thesis, Universitat Gesamthochschule Essen.

State Standards of Ukraine (Derzhavni Standarty Ukrainy, DSTU) (2002) DSTU 4145-2002. Information technologies. Cryptographic information security. Digital signature, what based on elliptic curves. Generation and verification. –K.: Derzhstandart Ukrainy, 40p.
Suguzaki, H., Matsuo, K., Chao, J., Tsujii, S.(2002) An extension of Harley algorithm addition algorithm for hyperelliptic curves over finite fields of characteristic two. Technical report ISEC2002-9 (2002-5), IEICE Japan, pp. 49–56.

Takahashi, M. (2002) Improving Harley algorithms for jacobians of genius 2 hyperelliptic curves. In Proc. of SCIS2002, IEICE Japan. in Japanese.

Wollinger, T. (2004) Software and hardware implementation of hyperelliptic curve cryptosystems. PhD dissertation. Bochum, Germany, May 2004.

Australian firearm identification system based on the ballistics images of projectile specimens

Dongguang Li

School of Computer and Information Science, Faculty of Computing, Health and Science Edith Cowan University
2 Bradford Street, Mount Lawley, WA 6050
Perth, Australia
d.li@ecu.edu.au

Abstract. Characteristic markings on the cartridge case and projectile of a fired bullet are created when it is fired. Over thirty different features within these marks can be distinguished, which in combination produce a "fingerprint" for a firearm. By analyzing features within such a set of firearm fingerprints, it will be possible to identify not only the type and model of a firearm, but also each every individual weapon as effectively as human fingerprint identification. A new analytic system based on fast Fourier transform (FFT) for identifying the projectile specimens by the line-scan imaging technique is proposed in this paper. Experimental results show that the proposed system can be used for firearm identification efficiently and precisely through digitizing and analyzing the fired projectiles specimens.

1 Introduction

The analysis of marks on bullet casings and projectiles provides a precise tool for identifying the firearm from which a bullet is discharged [1] [2]. Characteristic markings on the cartridge case and projectile of a bullet are produced when a gun is fired. Over thirty different features within these marks can be distinguished, which in combination produce a "fingerprint" for identification of a firearm [3]. This forensic technique is the vital element for legal evidence, in cases where the use of firearms is involve.

Projectile bullets fired through the barrel of a gun will exhibit extremely fine striation markings, some of which are derived from minute irregularities in barrel produced during the manufacturing process. The examination of these striations on land marks and groove marks of the projectile is difficult using conventional optical microscopy. However, digital imaging techniques have the potential to detect the presence of striations on ballistics specimens for identification.

Given a means of automatically analyzing features within such a firearm "fingerprint", identifying not only the type and model of a firearm, but also each individual weapon as effectively as human fingerprint identification can be achieved. Due to the skill required and intensive nature of ballistics identification, law enforcement agencies around the word have expressed considerable interest in the application of ballistics imaging identification systems to both greatly reduce the time for identification and to introduce reliability (or repeatability) to the process.

Several ballistics identification systems are available either in a commercial form or in a beta-test state. A Canadian company, Walsh Automation, has developed a commercial system called "Bulletproof", which can acquire and store images of projectiles and cartridge cases, and automatically search the image database for particular striations on projectiles. However the impressed markings or striations on projectiles must be matched by user. This inherent limitation of the system with respect to projectiles has prohibited its use. The Edith Cowan University of Australia, in conjunction with the Australia Police, has developed a prototype database called FIREBALL [4]. It has the capability of storing and retrieving images of cartridge cases heads, and of interactively obtaining position metrics for the firing-pin impression, ejector mark, and extractor

mark. The limitation of the system is that the position and shape of the impression images must be traced manually by users. For the time being, we still have unsolved problems on projectiles imaging, storing and analyzing although the system has been put in use for 4 years. The efficiency and accuracy of FireBall system must be improved and increased.

The research papers on the automatic identification of cartridge cases and projectiles are hardly to be found. L. P. Xin [5] proposed a cartridge cases based identification system for firearm authentication. His work was focused on the cartridge cases of center-firing mechanism. And he also provided a decision strategy by which the high recognition rate would be achieved interactively. C. Kou et al. [6] described a neural network based model for the identification of the chambering marks on cartridge cases. But no experimental results were given in their paper.

In this paper, a new analytic system based on fast Fourier transform (FFT) for identifying the projectile specimens captured by the line-scan imaging technique is proposed. The system gives an approach for projectiles capturing, storing and analyzing automatically and makes a significant contribution towards the efficient and precise identification of projectiles. Firstly, in Section 2, the line-scan imaging technique for projectiles capturing is described. Secondly, the analytic approach based on FFT for identifying the projectile characteristics and the experimental results are presented in Section 3. Finally, Section 4 gives a short conclusion.

2 Line-scan Imaging Technique for Projectile Capturing

2.1 Line-scan Imaging

The traditional optical microscopy for imaging the cylindrical shapes of ballistics specimen is inherently unsuitable for the expected high contrast imaging. It is difficult to maintain image quality using oblique lighting on a cylindrical surface, from low magnification microscopy, as the specimen is translated and rotated [7].

However, we can obtain the surface information from a cylindrical shaped surface using a line-scan imaging technique by scanning consecutive columns of picture information and storing the data in a frame buffer to produce a 2D image of the surface of the cylindrical specimen.

The periphery camera was the precursor imaging device to the line-scan camera, which consists of a slit camera with moving film in order to ‘unwrap’ cylindrical objects by rotating them on a turntable [8]. Relative motion between the line array of sensors in the line-scan camera and the surface being inspected is the feature of the line-scan technique. This relative motion is achieved by rotating the cylindrical ballistics specimen relative to the stationary line array sensor [7].

With the line-scan technique, because the cylindrical ballistics specimen is rotated about an axis of rotation relative to a stationary line array of sensor, all points on the imaging line of the sample are in focus. Hence, all points on the rotating surface will be captured on the collated image during one full rotation of the cylindrical ballistics specimen [7].

The line-scan imaging analysis system for projectiles in our study is shown in Fig. 1. The stepper motor rotates with 2400 steps/360 degrees, namely 0.15 degree each step. We use CCD camera instead of the traditional camera used in [7] [8]. The graphic capturing card installed in PC has a resolution of 240×320 pixels/inch. A ring light source is adopted, which can provide uniform lighting conditions [9].

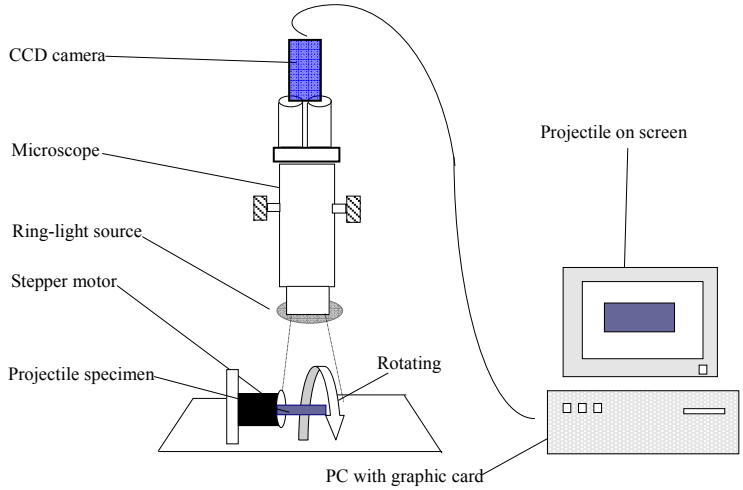


Fig. 1. The line-scan imaging and analyzing system

Being quite different from the method used in [7] [8], the procedure in our line-scan imaging approach is as follows:

1) With the stepper motor's every step
2) CCD camera captures the current image of projectile specimen and
3) Sends the image to Graphic card in PC;
4) The middle column of pixels in this image is extracted and saved consecutively in an array in the buffer on PC, and
5) step1) and 2) are repeated until the whole surface of the projectile specimen is scanned;
6) The array in the buffer is used to produce a 2-D line-scanned image for the whole surface of the projectile.

The resolution of the line-scan image is dependent on the rotational degree per step of the stepper motor, the resolution of CCD camera, the resolution of graphic capturing card, and the columns captured at each step in step 3). Therefore, the resolution of the line-scanned image of projectile specimen could be manipulated by adjusting the length of each step of the stepper motor and the number of columns captured in each step to meet forensic investigation requirements.

2.2 Projectile specimens and their line-scanned images

The projectile specimens in our study, provided by Western Australia Police Department, are in four classes and belong to four different guns. They are:
1) Browning, semiautomatic pistol, caliber 9mm.
2) Norinco, semiautomatic pistol, caliber 9mm.
3) and 4) Long Rifle, semiautomatic pistol, caliber 22.

All the projectile specimens in our study are recorded using the line-scan imaging technique discussed in Section 2.1 under the same conditions (such as the light conditions, the stepping angle of stepper motor, etc.). The stepping angle of the stepper motor is adjusted to produce a image by just one full rotation (360 degrees), so that all the land marks and groove marks of projectile specimen are captured and displayed in the line-scanned image. Line-scanned images of four classes of projectile specimens in our study are shown Fig. 2.

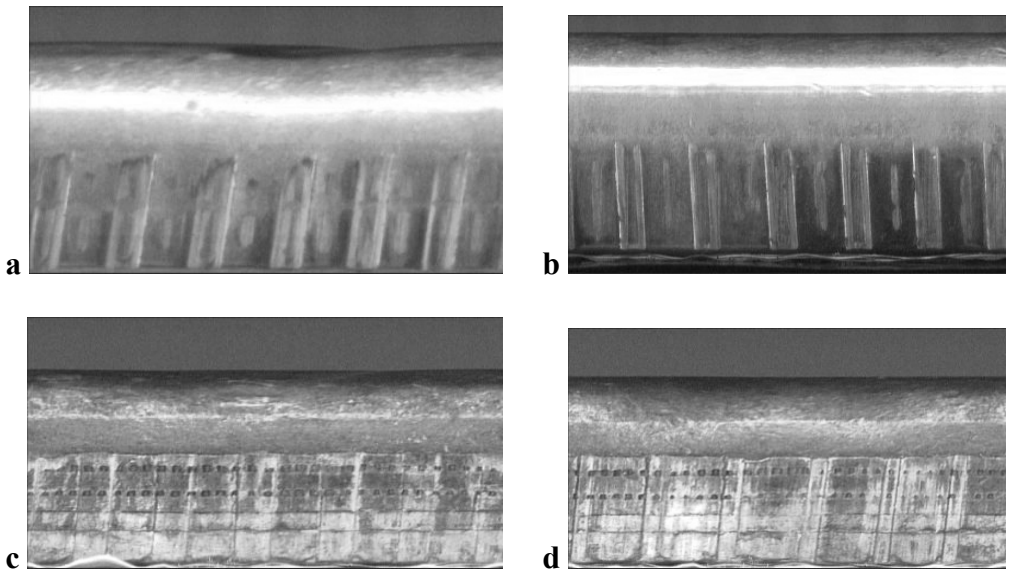


Fig. 2. Four classes of line-scanned images of projectiles in our study (with code: a, 101; b, 201; c, 301; d, 401)

2.3 Image Pre-processing for FFT analysis

In a practical application, the quality of the line-scanned image of a projectile specimen can be affected and noised by many factors such as the lighting conditions, the materials of the specimen, the original texture on the surface of specimen, and the deformed shapes. All these can bring strong noise into the line-scanned image or damage the shape of the line-scanned image, and would result in many difficulties to extract and to verify the important features used for identifying the individual specimen, such as the contours, edges, the directions and the width (or distance) of land marks and groove marks. In order to remove or decrease the affection mentioned above, the following image pre-processing operations are applied to the line-scanned images obtained in Section 2.2.

2.3.1 Contrast Enhancement:

One of the general functions in image preprocessing is the contrast enhancement transformation [10]. Low-contrast images can result from poor lighting condition, lack of dynamic range of the imaging sensor, and a wrong setting of a lens aperture during image acquisition. The idea behind

contrast enhancement is to increase the dynamic range of the gray levels in the image being processed. In our study, for the reason of the strong reflection from the metal surface, the images obtained are often blurred in a certain extent. The land marks or groove marks may be hidden within. Therefore, the contrast enhancement transformation is used upon the images obtained in Section 2.2. For the line-scanned images, only the regions that include the land marks and groove marks are useful for analyzing and identifying the characteristics of the projectile specimens. Hence, we only select the regions in images that are necessary and useful to our study. The images (the effective regions in original images) shown in Fig. 3 are transformed versions corresponding to the images in Fig. 2 by the region selecting and the contrast enhancement transformation.

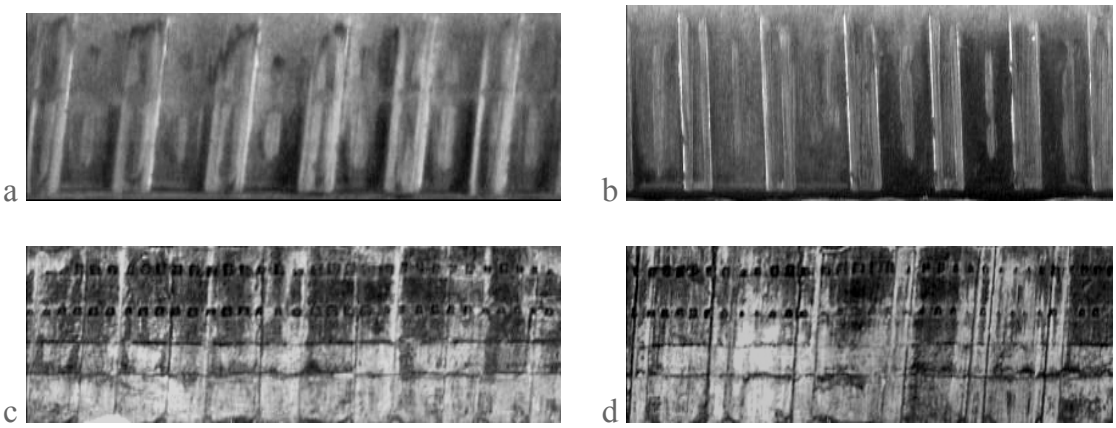


Fig. 3. Contrast enhancement results (a, b with size 400×110, and c, d with size 400×100)

2.3.2 Feature Extraction:

Feature extracting plays an important role in identification system. We pick up the first derivatives [10] as the images features. For a digital image, the most popular powerful masks used to approximate the gradient of f at coordinate (i, j) are *Sobel* operators in vertical and horizontal directions (shown in Fig. 4 with 3×3 window). In our study, we adopt the *Sobel* operators to extract the contours and edges of the land and groove marks on line-scanned images of projectile specimens. For the reason that the directions of the land and the groove marks of the projectile specimens are all or almost along 90 degree in the line-scanned images, we only adopt the vertical direction mask (Fig. 4) for extracting the features of the line-scanned images. By observing the Fig. 5 in which there are lots of noises and disconnection on the land and groove marks, the conventional spatial techniques are not suitable for the nature of locally. Hence, a FFT-based analysis for projectiles is introduced in.

-	-	-
0	0	0
-	-	-

-	0	-
-	0	-
-	0	-

Fig. 4. *Sobel* masks in vertical and horizontal directions

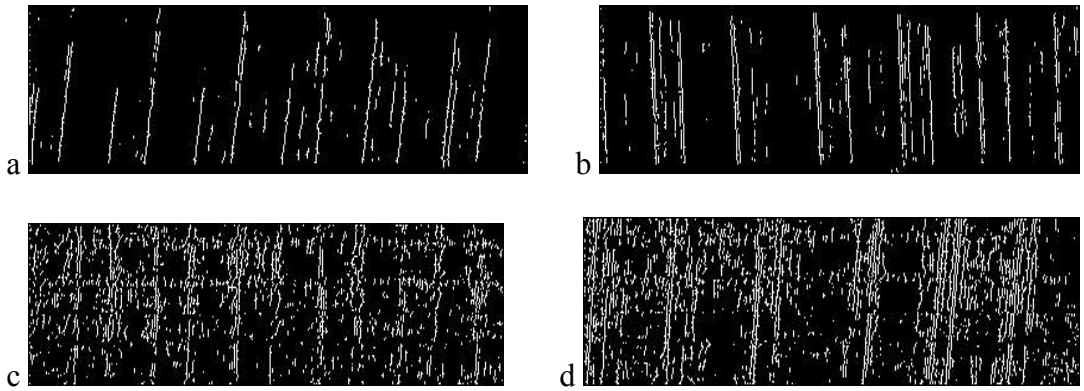


Fig. 5. The contours and edges extracting using *Sobel* operator in vertical direction

3 FFT-based Analysis

3.1 FFT and Spectrum Analysis

The Fourier transform of a two variables, continuous function, $f(x,y)$, is defined by the equation [10]

$$F(u,v)=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)e^{-j2\pi(ux+vy)}dxdy \qquad (1)$$

where $j=\sqrt{-1}$.

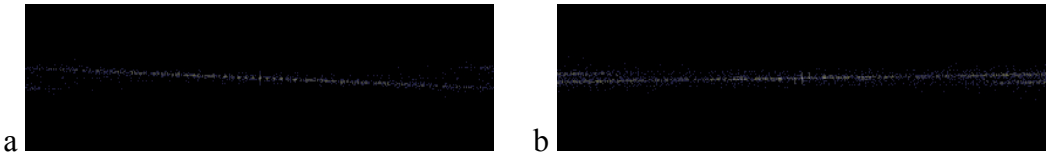
The Fourier transform of a two variables, discrete function (image), $f(x,y)$, of size $M\times N$, is given by the equation

$$F(u,v)=\frac{1}{MN}\sum_{x=0}^{M-1}\sum_{y=0}^{N-1} f(x,y)e^{-j2\pi(ux/M+vy/N)} \qquad (2)$$

where $j=\sqrt{-1}$, for all $u=0,1,2,\cdots,M-1$, $v=0,1,2,\cdots,N-1$. We define the Fourier spectrum by the equation

$$|F(u,v)|=\left[R^2(x,y)+I^2(x,y)\right]^{1/2} \qquad (3)$$

where $R(x,y)$ and $I(x,y)$ are the real and imaginary parts of $F(u,v)$, respectively.



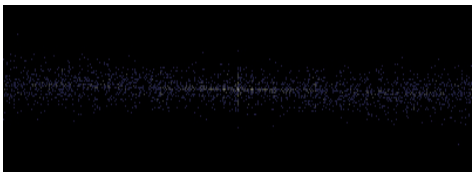
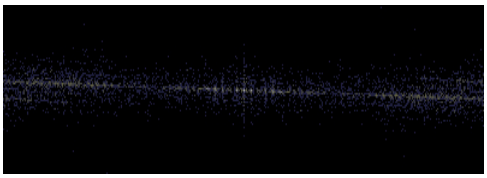
c  d 

Fig. 6. Fourier transformation results of the images in **Fig. 5**

The Fourier spectrum is ideally suited for describing the directionality of periodic or almost periodic 2-D patterns in an image. These global texture patterns, although easily distinguishable as concentrations of high-energy burst in the spectrum, generally are quite difficult to detect with spatial methods because of the local nature of these techniques.

Here, we consider a set of features of the Fourier spectrum that are used for analyzing and description the line-scanned images of projectiles:

(1) Prominent peaks in the spectrum give the principal direction of the texture patterns.

(2) The location of the peaks in the frequency plane gives the fundamental spatial period of the patterns.

(3) Some statistical features of the spectrum.

Detection and interpretation of the spectrum features just mentioned often are simplified by expressing the spectrum in polar coordinates to yield a function $S(r,\theta)$, where S is the spectrum function, and r and θ are the variables in this coordinate system. For each direction θ, $S(r,\theta)$ is a 1-D function $S_\theta(r)$. Similarly, for each frequency r, $S_r(\theta)$ is a 1-D function. Analyzing $S_\theta(r)$ for a fixed value of θ yields the behavior of the spectrum (such as the presence of peaks) along a radial direction from the origin, whereas analyzing $S_r(\theta)$ for a fixed value of r yields the behavior along a circle centered on the origin. A more global description is obtained by integrating (summing for discrete variables) these functions [10]:

$$S(r) = \sum_{\theta=0}^{\pi} S_\theta(r) \qquad (4)$$

and

$$S(\theta) = \sum_{r=1}^{R_0} S_r(\theta) \qquad (5)$$

where R_0 is the radius of a circle centered at origin.

The results of Equations (4) and (5) constitute a pair of values $[S(r), S(\theta)]$ for each pair of coordinates (r,θ). By varying these coordinates, we can generate two 1-D functions, $S(r)$ and $S(\theta)$, that constitute a spectral-energy description of texture for an entire image or region under consideration. Furthermore, descriptors of these functions themselves can be computed in order to characterize their behavior quantitatively.

3.2 FFT-based Analysis, Identification and Experimental Results

Some characteristics and descriptors of the line-scanned images for identification of projectiles using the radius spectrum and angular spectrum are discussed in details in this section.

We know that the slowest varying frequency component ($u = v = 0$) corresponds to the average gray level of an image. As we move away from the origin of the transform, the low frequencies correspond to the slowly varying components of an image. In a line-scanned image of projectile specimen, for example, these might correspond to the land and groove marks which are large in scale and regular in shape. As we move further away from the origin, the higher frequencies begin to correspond to faster and faster gray level changes in the image. These are the small or irregular marks and other components of an image characterized by abrupt changes in gray level, such as noises. So we focused our attention on the analysis of low frequencies in the radius and angle spectrum of line-scanned images.

The plots of radius and angle spectrum corresponding to images in Fig. 6 a, b are shown in Fig. 7 a, b, c and d, respectively. The results of FFT clear exhibit directional 'energy' distributions of the surface texture between class one and two. Comparing Fig. 7 a to b, the plots of radius spectrum, the former contains six clear peaks in the range of low frequencies (r <20), the later has only three peaks in the same range and is smooth in shape, this indicates that 'energy' of the class one specimen is distributed in several permanent positions, and also reveals that the class one specimens has a coarse surface texture and the wide land and groove marks, while the surface texture of class two is fine and the wide of land and groove marks is thin.

The angular spectrums (Fig. 7 c and d) display a great distinctness in position of prominent peaks between class one and two. Further study reveals that the angular spectrum can clearly indicate the angular position of periodic grooves or scratches on the surface with respect to the measurement coordinate. It can be seen from the angular spectrum there is a maximum peak at about 81 degree in Fig. 7 c. This is indicative of scratches (the land or groove marks) oriented in the direction 81 degree on the surface of projectiles , while the maximum peak in Fig. 7 d at about 95 degree. In addition, the former plot has a second prominent peak (corresponding to small or shallow marks on the projectile's surface) at about 100 degree. However, it is noted that the second peak of Fig. 7 d is at about 85 degree.

The characteristics of projectile specimen surface textures can also be revealed by examining quantitative differences of spectrums using a set of features. To compare and analyze the spectrums differences between the class one and two easily, a set of features is used, and the quantitative results are shown in Table 1 (where, r_1 and a_2, Max; r_2 and a_3, Mean; r_3 and a_4, Std; r_4 and a_5, Max**:** Median; and a_1, Position of maximum peak).

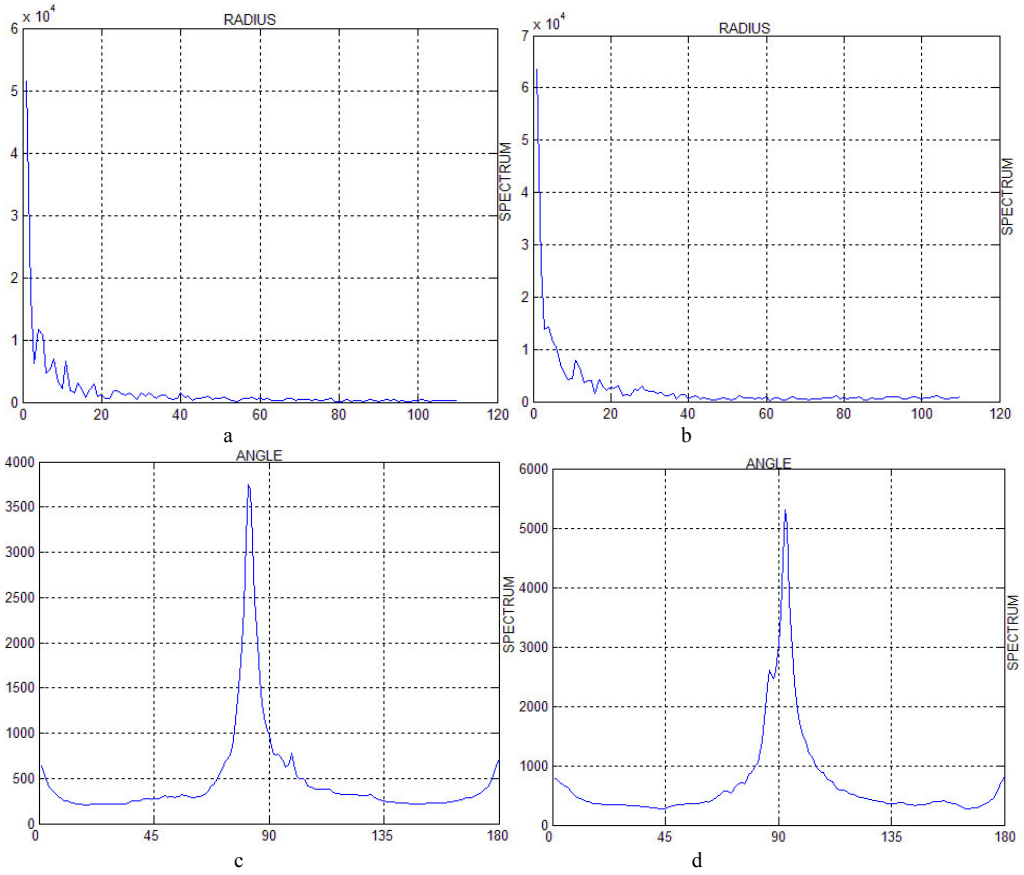


Fig. 7. Radial spectrum (a, b) and Angular spectrum (c, d) of the images (a, b) in **Fig. 6**

It can be observed from Table 1 that the Max, Mean, and Std of the class one are relatively smaller than the class two, while the relative variation for radio between Max and Mean is greater. And the difference between prominent peaks (corresponding to the orientations of land and groove marks) of class one and two is 14 degrees. This provides evidence that FFT spectrum analysis, in the form of quantification, can reveal characteristic and directional surface textures of projectile specimen.

Table 1. Radial spectrum and angular spectrum statistics results of **Fig. 6** a and b

Class	Code	Radial spectrum				Angular spectrum				
		r_1	r_2	r_3	r_4	a_1	a_2	a_3	a_4	a_5
1	101	51517	1728	5372	29.81	81	3753	475.2	546.2	7.898
2	201	63646	2538	6809	25.07	95	5308	697.9	794.7	7.600

All experimental results based on the projectiles in our study are listed in Table 2.

Table 2. Radial spectrum and angular spectrum statistics results based on the specimens in our study

Class	Code	Radial spectrum				Angular spectrum				
		r_1	r_2	r_3	r_4	a_1	a_2	a_3	a_4	a_5
1	101	51517	1728	5372	29.81	81	3753	475.2	546.2	7.898
	102	51162	1591	5187	32.16	81	3709	437.6	545.6	8.487
	103	51200	1520	5087	33.68	81	3583	418.2	509.8	8.571
	104	51556	1699	5348	30.34	81	3589	467.3	514.1	7.685
	105	62715	1962	6299	31.96	81	4219	539.5	617.8	7.827

	201	63646	2538	6809	25.07	95	5308	697.9	794.7	7.600
2	202	64381	2738	7038	23.51	95	5257	752.9	777.7	6.990
	203	64059	2545	6707	25.16	95	5193	700.0	794.0	7.419
	301	63959	2899	6942	22.06	86	2514	724.7	451.9	3.469
3	302	64448	2478	6889	26.01	86	2714	719.4	445.5	3.774
	303	64288	2743	7090	23.43	86	2517	685.8	439.9	3.669
	304	63694	3011	6999	21.23	86	2750	752.7	512.8	3.657
	401	76059	4040	8554	18.27	79	4965	1010	787.8	4.916
4	402	76406	5026	8982	15.20	79	4972	1256	835.6	3.959
	403	75607	3735	8035	20.23	79	4897	933.9	753.3	5.249
	404	76796	3786	8498	20.28	79	4135	946.3	738.6	4.371

By observing Table 2 and recalling that the calibers of type one and two are same, and so are the type three and four, we can easily identify the projectiles using any one of features listed in Table 2. For example, all the values of r_4 for type one are greater than 28.0, while for type two, no one is greater than 26.0. Same result can be obtained using other features in Table 2. The characteristics we used in spectrum analysis can be formed as a set of features vectors for building an artificial intelligent (AI) system for the automatic firearm identification based on the spent projectiles.

4 Conclusion

In this paper, a new analytic system for firearm identification based on the projectile specimens automatically is proposed. We not only present an approach for capturing and storing the surface image of the spent projectiles at high resolution using line-scan imaging technique for the projectiles database, but also presented a novel and effective FFT-based analysis technique for analyzing and identifying the projectiles. This system can make a significant contribution towards the efficient and precise analysis of firearm identification based on ballistics projectiles. The study demonstrates that different types of land and groove marks generated by different guns have distinctive surface textures, and these textures can be measured and identified effectively by spectral analysis. It is clear that spectral analysis is an effective method to study line-scanned images of projectile specimen. The method can surmount difficulties with descriptions in the normal spatial domain in identifying texture features formed by land and groove marks on the surface of projectiles.

Reference

1. C.L. Smith, and J.M. Cross, (1995): Optical Imaging Techniques for Ballistics Specimens to Identify Firearms. Proceedings of the 29th Annual 1995 International Carnahan Conference on Security Technology, pp. 275-289, Oct. 1995, England.

2. R. Saferstein (ED), (1988) Forensic Science Handbook: Volume 2. Englewood Cliffs: Prentice Hall, 1988.

3. G.Burrard, (1951): Identification of Firearms and Forensic Ballistics. London: Herbert Jenkins,1951.

4. C.L. Smith, J.M. Cross, and G.J. Variyan, (1995): FIREBALL: An Interactive Database for the Forensic Ballistic Identification of Firearms. Research Report, Australian Institute of Security and Applied Technology, Edith Cowan University, Western Australia, 1995.

5. Le-Ping Xin, (2000): A Cartridge Identification System for Firearm Authentication, Signal Processing Proceedings, 2000. WCCC_ICSP 2000. 5th International Conference on Volume: 2, P1405-1408.

6. Chenyuan Kou, Cheng-Tan Tung and H. C. FU, (1994): FISOFM: Firearms Identification based on SOFM Model of Neural Network, Security Technology, 1994. Proceedings. Institute of Electrical and Electronics Engineers 28th Annual 1994 International Carnahan Conference on , 12-14 Oct. Pages: 120-125.

7. C.L. Smith, Robinson, M. and Evans, P.: (2000): Line-scan Imaging for the positive identification of ballistics, 2000. IEEE International Carnahan Conference on Security Technology, 269-275. 2000.

8. Kingslake, R., Optics in Photography. SPIE Optical Engineering Press. Bellingham, Washington, USA, 1992.

9. Jun Kong, D. G. Li., A. C. Watson: A Firearm Identification System Based on Neural Network, AI 2003, Lecture Notes in Artificial Intelligence, 315-326, 2003, Springer.

10. Rafael C. Gonzalez, Richard E. Woods: Digital Image Processing, Second Edition, Beijing: Publishing House of Electronics Industry, 2002, 7, 519-566.

Firearm Identification with Hierarchical Neural Networks by analyzing the firing pin Images retrieved from cartridge cases

Dongguang Li

School of Computer and Information Science
Edith Cowan University
2 Bradford Street, Mount Lawley 6050
Perth, Western Australia
d.li@ecu.edu.au

Abstract
When a gun is fired, characteristic markings on the cartridge and projectile of a bullet are produced. Over thirty different features can be distinguished from observing these marks, which in combination produce a "fingerprint" for identification of a firearm. In this paper, through the use of hierarchial neural networks a firearm identification system based on cartridge case images is proposed. We focus on the cartridge case identification of rim-fire mechanism. Experiments show that the model proposed has high performance and robustness by integrating two levels Self-Organizing Feature Map (SOFM) neural networks and the decision-making strategy. This model will also make a significant contribution towards the further processing, such as the more efficient and precise identification of cartridge cases by combination with more characteristics on cartridge cases images.
Keywords: Firearm identification; Neural networks; Image processing.

Introduction

A precise tool for identifying the firearm from which a bullet is discharged **[1] [2]** is the analysis of marks on bullet casings and projectiles. When a gun is fired, characteristic markings on the cartridge and projectile of a bullet are produced. Over thirty different features within these marks can be distinguished, which in combination produce a "fingerprint" for identification of a firearm **[3]**. In cases where the use of firearms is involve, this forensic technique is the vital element for legal evidence. It will be possible to identify not only the type and model of a firearm, but also each individual weapon as effectively as human fingerprint identification can be achieved; given this means of automatically analyzing features within such a firearm fingerprint.

Due to the skill required and intensive nature of ballistics identification, law enforcement agencies around the world have expressed great interest in the application of ballistics imaging identification systems to both introduce reliability (or repeatability) to the process, and also to greatly reduce the time for a positive identification. Several ballistics identification systems are available either in a commercial form or in a beta-test state. A Canadian company, Walsh Automation, has already developed a commercial system called "Bulletproof", which can acquire and store images of projectiles and cartridge cases, and automatically search the image database for particular striations on projectiles but not impressed markings or striations on cartridge cases. This inherent limitation of the system with respect to cartridge cases of the system has prohibited its use. The Edith Cowan University of Australia, in conjunction with the Western Australia Police, has developed a prototype database called FIREBALL **[4]**. It has the capability of interactively obtaining position metrics for the impression of firing-pin mark, ejector mark, and extractor mark and also of storing and retrieving images of cartridge cases heads. The limitation of the system is that the position and shape of the impression images must be located and traced manually by users.

The papers on the automatic identification of cartridge cases are hardly to be found. Le-Ping Xin **[5]** proposed a cartridge cases based identification system for firearm authentication. His work was focused on the cartridge cases of center-fire mechanism. And he also provided a decision strategy from which the high recognition rate would be achieved interactively. Chenyuan Kou et al. **[6]** described a neural network based

model for the identification of the chambering marks on cartridge cases. But no experiment results were given in their paper.

In this paper, the method proposed, is a system for identifying the firing pin marks of cartridge cases images automatically using a hierarchical neural network model. The main focus is on the consideration of rim-firing pin mark identification. The system will also make a significant contribution towards the efficient and precise identification of cartridge cases in the further processing, such as the locating and coding of ejector marks, extractor marks and chambering marks of cartridge cases. In Section 2, the SOFM neural network and the methods of image processing in our study is described briefly. The capturing and preprocessing of cartridge cases images are presented in Section 3. The model based on a hierarchical neural networks for identification of cartridge cases images is proposed in Section 4. Section 5 gives a numeric experiment. Finally, the conclusion is presented in Section 6.

SOFM and Image Processing

2.1 SOFM Neural Network

We pick the Self-Organizing Feature Map (SOFM) neural networks as the basic classifying units in our identification system. This system has been applied to the study of complex problems such as speech recognition, combinatorial optimization, control, pattern recognition and modeling of the structure of the visual cortex **[7], [8], [9]** and **[10]**. The SOFM we used is a kind of un-supervised neural network models, it in effect represents the result of a vector quantization algorithm that places a number of reference or codebook vectors into a high-dimension input data space to approximate defined between the reference vectors, the relative values of the latter are made to depend on ate to its data set in an ordered fashion. When local-order relations are each other as if there neighboring values would lies along an "elastic surface". By means of the self-organizing algorithm, this "surface" becomes defined as a kind of nonlinear regression of the reference vectors through the data points **[11]**.

We employ the standard Kohonen's SOFM algorithm summarized in Table 1, the topology of SOFM is shown in **Fig.1**.

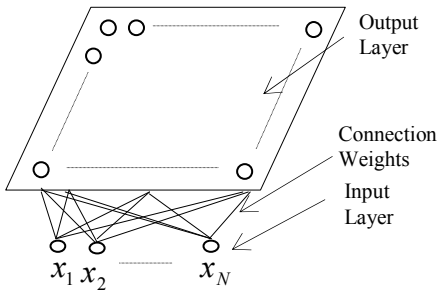


Fig.1. The topology of SOFM

2.2 Image Processing, Feature Extraction

Contrast Enhancement. One of the general functions in image preprocessing is the contrast enhancement transformation **[12]**, and function is expressed in Equation (1). Low-contrast images can result from poor lighting conditions, lack of dynamic rang in the imaging sensor, or even wrong setting of a lens aperture during image acquisition. The idea behind contrast enhancement is to increase the dynamic range of the gray levels in the image being processed. The image shown in **Fig.2b** is transformed by contrast enhancement.

Polar Transaction. During the stage of image preprocessing, polar transformation is also a useful tool. In our study, the polar transformation can bring us some advantages: In the test phase (see Section 4), we only move the detecting windows over the testing images in direction of horizontal and vertical rather than rotating the testing images or the detecting windows. This will decrease the numerical error and increase the

efficiency. Under the Polar Systems, we can get more informations about the testing images. Some images that have similar shapes may be different in shapes and be distinguished in Polar Systems.

Table 1. The Unsupervised SOFM Algorithm

Step1. Initialize the weights for the given size map. Initialize the learning rate parameter, neighborhood size and set the number of unsupervised learning iterations.
Step2. Present the input feature vector $x=[x_1, x_2, \cdots, x_n, \cdots, x_N]$ in the training data set, where x_n is the n th element in the feature vector.
Step3. Determine the winner node c such that $\|x-w_c\| = \min_i \{\|x-w_i\|\}$
Step4. Update the weights, w_i 's, within the neighborhood of node c , $N_c(t)$, using the standard updating rule: $w_i(t+1) = w_i(t) + \alpha(t)[x_n - w_i(t)]$, where $i \in N_c(t)$.
Step5. Update learning rate, $\alpha(t)$, and neighborhood size, $N_c(t)$. $\alpha(t+1) = \alpha(0)\{1-t/K\}$; $N_i(t+1) = N_i(0)\{1-t/K\}$, where K is a constant and is usually set to be equal to the total number of iterations in the self-organizing phase.
Step6. Repeat 2-5 for the specified number of unsupervised learning iterations.

$$f(x) = \begin{cases} \frac{y_1}{x_1}x, & x < x_1 \\ \frac{y_2 - y_1}{x_2 - x_1}(x - x_1) + y_1, & x_1 \le x \le x_2 \\ \frac{255 - y_2}{255 - x_2}(x - x_2) + y_2, & x > x_2 \end{cases} . \quad (1)$$

Feature Extracting. In the recogniton system, feature extracting plays an important role. In the real application, the time consuming of feature extracting technique is also a crucial factor to be considered. So we pick up the morphological gradient **[12]** of the images processed by the two steps mentioned aboved as the images features. We deal with digital image functions of the form $f(x,y)$ and $b(x,y)$, where $f(x,y)$ is the input image and $b(x,y)$ is a structuring element , itself a subimage function.

Gray-scale dilation of f by b , denoted $f \oplus b$, is defined as

$$(f \oplus b)(s,t) = \max\{f(s-x, t-y) + b(x,y) \mid (s-x), (t-y) \in D_f; (x,y) \in D_b\} \quad (2)$$

where D_f and D_b are the domains of f and b , respectively.

Gray-scale erosion of f by b , denoted $f \ominus b$, is defined as

$$(f \ominus b)(s,t) = \min\{f(s+x, t+y) - b(x,y) \mid (s+x), (t+y) \in D_f; (x,y) \in D_b\} \quad (3)$$

where D_f and D_b are the domains of f and b , respectively.

The morphological gradient of an image, denoted g , is defined as

$$g = (f \oplus b) - (f \ominus b). \qquad (4)$$

a b c

Fig.2. Low-contrast image, a. Result of contrast enhancement, b. Result of threshold, c.

3 Cartridge Cases Images

There are two general types for the firing mechanism: the firing pin is either rim-firing mechanism or center-firing mechanism, as shown in **Fig.3**. The firing pin mark of cartridge case is formed when the bullet is fired. It is one of the most important characteristics for identifying the individual firearm. A variety of firing pins marks have been used in the manufacture of firearms for the rim-firing cartridge cases. In our study, the cartridge cases belonged to six guns can be classified into six types by shape of firing pin marks (shown in **Fig.4**).

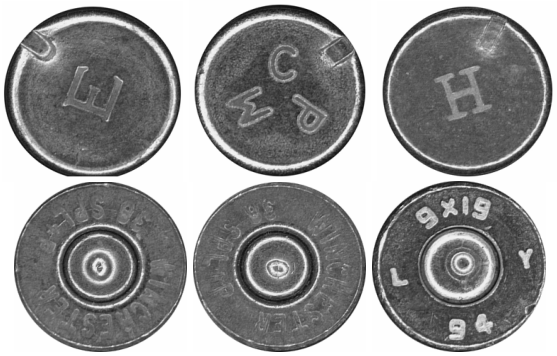


Fig.3. Rim-firing, first row; Center-firing, second row.

All the images of cartridge cases are obtained through the optical microscope in the real application. So some information such as the depth of the impression will be dismissed. Other factors such as the lighting conditions, the material of cartridge cases, and the stamp letters of manufacturer can bring strong noise into the cartridge cases images or damage the shapes of the cartridge cases images. These would all bring many difficulties to feature extracting and identifying. The lighting conditions for the image capturing of cartridge case is crucially importance. In order to produce high contrast of striation (firing-pin mark) on the cartridge cases, the illuminator must be installed at an angle of greater than 45 degree from normal to the plane of the head of the cartridge [1].

The 150 rim-fire cartridge cases, which are belonged to six guns, provided by the Western Australia Police are captured through the optical microscope, one image for each, formed 150 BMP files in gray scale size by 244×240 pixels, and classified into six types by shape of firing pin marks. They are: 1. U-shaped pin mark, 2. Axe-head pin mark, 3. Rectangular (Short) pin mark, 4. Rectangular (Long) pin mark, 5. Square pin mark, 6. Slant pin mark. Examples of the six types are shown in **Fig.4** (The numbers below these figures labeled the class number associated with each cartridge cases). We choose 50 images including the images of all the six guns randomly to form the set C_0 and form the testing set T for the rest images. Then, the images of set C_0 are processed through the image processing and feature extraction stage (shown in **Fig. 5**) discussed in Section 2.2.

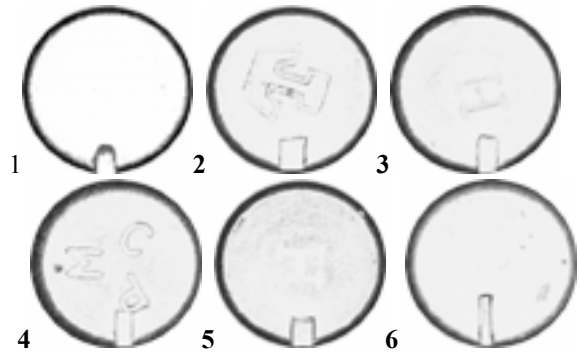


Fig.4. Six type of cartridge cases images

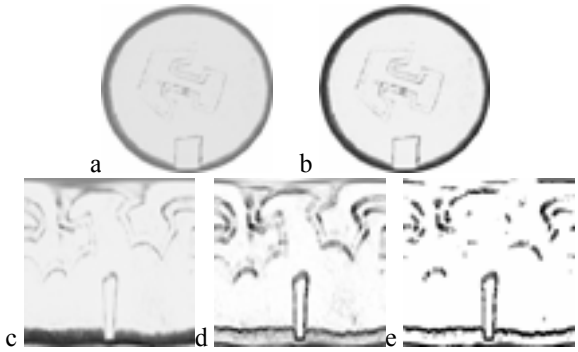


Fig.5. The original image a, the contrast stretching b, the polar transformation c, the morphological gradient d, the threshold e.

Now that the above transformations for the images of every type is finished, we need a "window" operation:

First, windows, size by $n_i \times m_i$ pixels, are used to copy the sub-images---the firing pin marks of the cartridge cases images processed before, where i stands for the label of the class to which the firing pin marks belong. The sizes of six type windows associated with six type firing pin marks are as follows in **Table2**. Second, the images (the firing pin marks) within these six type windows are copied into windows with size normalized by 48×196 pixels to meet the need of having unified input units of SOFM. The process is shown in **Fig.6**. In addition, we process part of images obtained as mentioned above in the manners: a. Shifted up to two pixels by the direction left, right, up, and down. b. Scaled by factor 0.95 and 0.90, this is in order to make our model have some robustness to subtle changes in the testing cartridge cases images. All the images we obtained through these processing above, with the number of 350, are combined into a training set C for the model based on SOFM, which will be discussed in the following section.

Table2. The Size (in pixels) of Six Type Windows

Type 1	20×96	Type 2	20×96
Type 3	20×120	Type 4	24×116
Type 5	20×120	Type 6	24×168

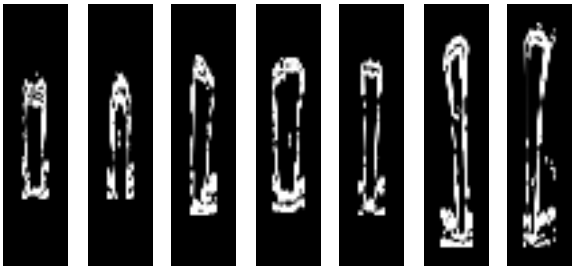


Fig.6. Six type of firing pin marks within windows with size normalization. The first row shows the six firing pin marks within six type windows. The second row shows the firing pin marks within windows with size normalization.

4 Hierarchical Identification Model

In this section, a hierarchical firearm identification model based on cartridge cases images is proposed. The structure of the model, the training, testing of SOFM, and decision-making strategy is given in details in following parts, respectively.

Identification Model. The system proposed is comprised of three stages as shown in **Fig.7**, the preprocessing stage mentioned in Section 2 and Section 3, the classification stage based on neural networks involving two levels SOFM neural networks and the decision-making stage. In our study, the two levels SOFM neural networks are:

The first level, which has one SOFM neural network (as shown in **Fig.1**) labeled by $SOFM_0$ acting as a coarse classifier among the training (or testing) patterns presented to it. The training or learning processing is the same as that mentioned in Section 2, which belongs to the type of unsupervised learning.

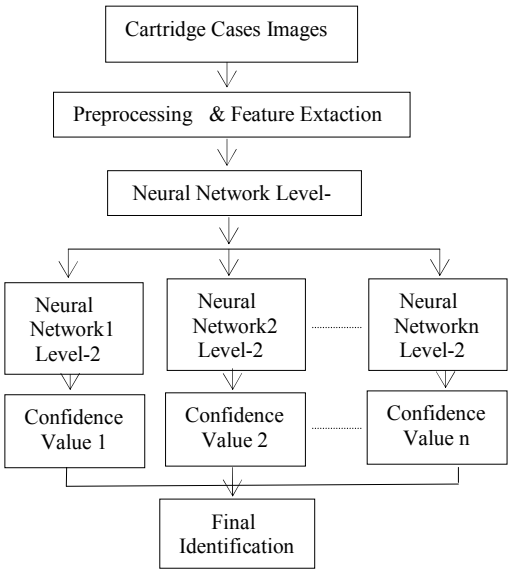


Fig.7. The proposed identification system

The second level neural networks are composed of several child SOFM networks denoted by $SOFM_i$ $i = 1, 2, \cdots, n$, where n is the number of child SOFM networks, making fine identification among the patterns classified by $SOFM_0$ (or the output of $SOFM_0$).

Training. In our study method, The training or learning processing for $SOFM_0$ is as same as that mentioned in **Table1**, which belongs to the type of unsupervised learning (we use the images of *C* to train the $SOFM_0$. The number of neurons in input layer is 48×196, corresponding to the size of windows normalized mentioned before). In the training phase, a neuron can be removed from the network, when the neuron of output layer is inactive for a period of time. A neuron may be considered inactive if it is not chosen frequently as the winner over a finite time interval. After being trained, the neurons, which are active with

high output value in the output layer of $SOFM_0$, stand for the classes to which the training images (or the testing specimens) belong. In our study, the training set C has been parted into several subsets by the result of classification of $SOFM_0$. Combination of these subsets in proper manners achieve training sets for the SOFMs of second level. The second level SOFM neural networks are generated when the positions of two classes in the output layer are very close or overlapping. The training sets are formed by combining the two of class patterns those are close or overlapping. The training processing is as same as $SOFM_0$.

Testing. The testing procedure for firearm identification system is as follows:

Step1. Select a testing cartridge case image from the testing set T , and present this testing pattern to the first stage of identification system--the preprocessing stage.

Step 2. Select a type of window from all types in turn, then move this window over the testing pattern processed in Step1 at every location by every pixel horizontally and vertically, pick up the sub-images.

Step3. Present all the sub-images to the $SOFM_0$ in turn, and then to $SOFM_i$ by the result of $SOFM_0$, and calculate the confidence values with Formula (5) for each sub-image. Return Step2 until all type windows are used up.

Step4. Present these confidence values to the third stage, the decision-making stage, and calculate the finnal result for the testing cartridge case image by Formula (6) and (7).

Decision-making Strategy. Due to the reasons of noise, lighting conditions, and the trademarks on the head of cartridge cases images, the following situation could generally be encountered in the testing phase:

a. More than one sub-image under this type window is classified to include a firing pin mark; for a testing cartridge case image, when a type of detecting window is used over the image.

b. For a particular testing cartridge case image, when all types of windows are used over the pattern, more than one sub-image under the different windows is classified to include a type of firing pin mark.

We use a final decision-making mechanism in decision-making stage to solve these problems mentioned above and improve the performance and accuracy, defining a Confidence Function $D(i, j)$ for the testing pattern i to the jth class which measures the ratio between the testing pattern distance to the weight vectors and the average distance of training patterns to the weight vectors, as follows:

$$D(i, j) = D(j) / D(i, j), \qquad (5)$$

where $dist(j)$ is the average distant when all the training patterns, which belong to the jth class, are tested with the jth type window, $dist(i, j)$ is the distant resulted when the ith testing pattern is tested using the jth type window. Defining a decision-making rule as follows: $i \in$ Class K , if

$$D(i,k) = \min_j \{D(i, j) > \Delta_j\}, j = 1, 2, \cdots n, \qquad (6)$$

where $\Delta_j \quad j = 1, 2, \cdots n$, is an appropriate threshold selected for the class j by experiments. In generally, for the unbalanced distribution of training patterns we get in the pattern space, results the unbalance in the neural network for each class. Hence, Δ_j for every class is not unique.

Defining a rejection rule as follows, testing pattern i is rejected by all classes, if

$$D(i, j) < \Delta_j, j = 1, 2, \cdots n, \qquad (7)$$

where $\Delta_j \quad j = 1, 2, \cdots n$, is same as in Formula (6).

5 Experimental Results

In our study, we use the following experimental parameters (shown in **Table 3**) for $SOFM_0$, Level 2 SOFMs and get experimental results over Training set C .

The neurons of the output layer of $SOFM_0$ are divided into six areas separately, through which the specimens of each class are represented, when the training phase is finished. The training set C is divided into three subsets for the training three sub-networks of the second level by the fact: the distribution area of each class is not balanced, some classes are near, and others are apart, in following manner:

Subset c_1 including images labeled with class 1 and 2 is selected as the training set of $SOFM_1$,

Subset c_2 including images labeled with class 3 and 5 is selected as the training set of $SOFM_2$,

Subset c_3 including images labeled with class 4 and 6 is selected as the training set of $SOFM_3$.

Table3. Experimental Parameters for $SOFM_0$, Level 2 SOFMs and Results over Training set C

	Input Layer	Output Layer	$\eta(0)$	$\Lambda_i(0)$
$SOFM_0$	48×196	9×9	0.60	7
$SOFM_1$	48×196	3×3	0.05	2
$SOFM_2$	48×196	5×5	0.05	3
$SOFM_3$	48×196	5×5	0.05	2
	training pattern	right rate	error rate	rejection rate
	350	100%	0%	0%

We have the experiment results over testing set T as follows:

Table 4. Experiments Results

Testing pattern	Right rate
100	97.0%
Rejection rate	Error rate
3.0%	0%

Analysis of experiment results

From the results of **Table 4**, we can see: Identification model in our study can make the combination of location and identification of firing pin mark of cartridge case images into one stage. It shows that the model proposed has high performance and robustness for the testing patterns in aspects as follows: Having high accuracy in location and identification of firing pin marks. Some testing results under Cartesian co-ordinates are shown in **Fig.8**. Having robustnesses to the noise patterns, to the damaged and deformed patterns shown in **Fig.8(8-13)**. Having some robustnesses to the scaled patterns.

We also see that there still are rejections for some patterns, and we found that the rejection is caused mainly by the following reasons: the high noise on the cartridge images; the letters of trademark on the cartridge images; the similitude of one pattern with others in some location.

Further Work. In order to improve our model to achieve higher performance, we will do some further researching in following aspects:

To improve the quality of image capturing and preprocessing. To extract some fine features with more complex techniques to represent the patterns (training or testing). To integate multiple classifier combination using different features sets.

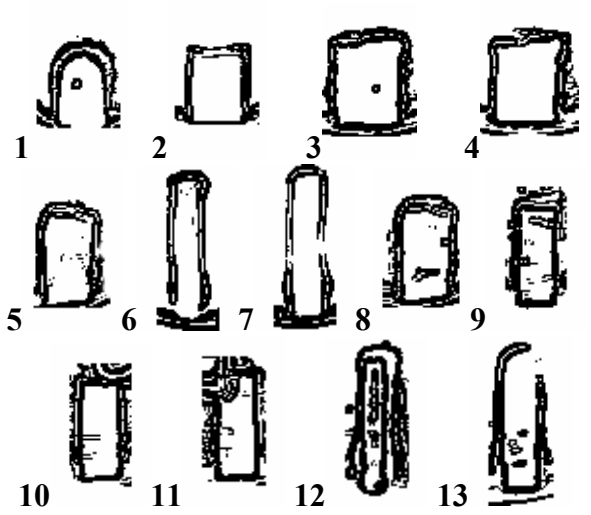


Fig.8. Some right identification results of testing set ***T***

6 Conclusion

In this paper, we have mainly focused on the consideration of rim-firing pin mark identification. Using a hierarchical neural network model, this study is investigating a system for identifying the firing pin marks of cartridge cases images automatically. The identification model in our study can make the combination of location and identification of firing pin mark of cartridge case images into one stage. It shows that the model proposed has high performance and robustness for real testing patterns. The efficiency of this system will also make a significant contribution towards the efficient and precise identification of ballistics specimens in the further processing, such as the more efficient and precise identification of cartridge cases by combination with more characteristics on cartridge cases images.

Reference

[1] C.L. Smith, and J.M. Cross, (1995) "Optical Imaging Techniques for Ballistics Specimens to Identify Firearms". Proceedings of the 29th Annual 1995 International Carnahan Conference on Security Technology, pp. 275-289, Oct. 1995, England.

[2] R. Saferstein (ED), Forensic Science Handbook: Volume 2. Englewood Cliffs: Prentice Hall, 1988.

[3] G.Burrard, (1951) "Identification of Firearms and Forensic Ballistics". London: Herbert Jenkins,1951.

[4] C.L. Smith, J.M. Cross, and G.J. Variyan, (1995) "FIREBALL: An Interactive Database for the Forensic Ballistic Identification of Firearms". Research Report, Australian Institute of Security and Applied Technology, Edith Cowan University, Western Australia, 1995.

[5] Le-Ping Xin, (2000) "A Cartridge Identification System for Firearm Authentication", Signal Processing Proceedings, 2000. WCCC_ICSP 2000. 5th International Conference on Volume: 2, P1405-1408.

[6] Chenyuan Kou, Cheng-Tan Tung and H. C. FU, (1994) "FISOFM: Firearms Identification based on SOFM Model of Neural Network", Security Technology, 1994. Proceedings. Institute of Electrical and Electronics Engineers 28th Annual 1994 International Carnahan Conference on , 12-14 Oct. Pages: 120-125.

[7] T. Kohonen, (1990) "Self-organising Maps", Springer, Berlin, 1995, The self-organising Maps, Proc. IEEE 78 (9) (1990) 1464-1480.

[8] S.B. Cho, "Pattern Recognition with Neural Networks Combined by Genetic Algorithm", Fuzzy Set and System 103(1999) 339-347.

[9] T.M. Ha, H. Bunke, "Off-line Handwritten Numeral Recognition by Perturbation Method", IEEE Trans. Pattern Anal. Mach.Intell. 19(5) (1997)535 -539.

[10] P.N. Suganthan, "Structure Adaptive Multilayer Overlapped SOMs with Partial Supervision for Handprinted Digit Classification ", Proceedings of International Joint Conference on Neural Networks, WCCI'98, Alaska, May 1998.

[11] T. Kohonen, Tutorial Notes, (1993) International Symposium on Artificial Neural Networks, pp. 9-15, Dec.20-22, 1993.

[12] Rafael C. Gonzalez, Richard E. Woods, Digital Image Processing, Second Edition, Beijing: Publishing House of Electronics Industry, 2002, 7, 519-566.

A General Model for Oblivious Transfer

Hossein Ghodosi

School of Mathematics, Physics and Information technology
James Cook University, Townsville, QLD 4811, Australia
E. mail: hossein.ghodosi@jcu.edu.au

ABSTRACT

Naor and Pinkas (1999a) pointed out that 'this is of interest, given the possibility of implementing Oblivious Transfer using physical means, e.g., via a noisy channel or quantum cryptography. However, some of these reductions are not particularly efficient...'.

Using tamper-proof smart cards, we propose some radically different approaches to the Oblivious Transfer problem. Despite the simplicity of our model, we achieve unconditional security and privacy for both the sender and receiver. In contrast to all known oblivious transfer protocols, which require extensive computations, the proposed schemes are very computationally efficient.

INTRODUCTION

An Oblivious Transfer (OT) protocol, as it has been introduced by Rabin (1981), allows two parties, Alice as a sender and Bob as a receiver, to interact in such a way that Bob has a 50% chance of receiving the information sent by Alice. The additional constraint of the protocol is that Alice has no knowledge as to whether Bob has received the message or not. This setting has been generalized, by Even et al. (1982), to a scenario in which Alice has two strings M_0 and M_1, and Bob wishes to obtain one of these strings such that Alice cannot figure out which one Bob has obtained. In addition, the protocol must not allow Bob to receive any information about the other string. This setting, also known as a 1-out-of-2 oblivious transfer (OT_1^2), has been the subject of investigation by several researchers, e.g. Crepeau (1987, 1994), Crepeau et al. (1988a, 1995), and Brassard and Crepeau (1997).

Several scenarios of oblivious transfer have been studied in the literature. Bellare and Micali (1989) have suggested a 2-out-of-3 oblivious transfer that can be generalized to a 't-1 out of t' oblivious transfer, for any integer of t. Naor and Pinkas (1999a) studied 1-out-of-N Oblivious Transfer (OT_1^N) in which the sender has N values and the receiver would like to obtain knowledge about one of them. Naor and Pinkas (1999b) have also discussed the scenario in which the sender has N values and the receiver would like to learn K of them; this is known as the 'K-out-of-N' Oblivious Transfer (OT_K^N). Mu et al. (2002) also studied (OT_K^N) and presented a scheme, but Ghodosi and Zaare-Nahandi (2006) have shown that their scheme is not secure. Distributed oblivious transfer, in which Alice's task is distributed among a set of servers, has been discussed by Naor and Pinkas (2000). In this setting, the receiver must contact a predetermined number of servers in order to learn the one (out of two) chosen message. Ghodosi (2007) has shown that the distributed oblivious transfer scheme proposed by Naor and Pinkas is not secure. Note that a setting of oblivious transfer must satisfy the following two requirements:

- **The Sender's Security:** At the end of the protocol, the receiver must learn nothing more than what he was supposed to learn.
- **The Receiver's Privacy:** At the end of the protocol, the sender cannot figure out what the receiver has learned.

Different oblivious transfer protocols have been designed in order to serve different scenarios.
That is, the practicality of each protocol is limited to a particular scenario. In this paper, we present a general model that can be employed in all scenarios of oblivious transfer in a similar manner. In spite of the simplicity of our system, we achieve unconditional security and privacy for both the sender and receiver

assuming that a tamper-proof device exists. Moreover, our oblivious transfer model leads to the implementation of one-time use systems, which prevents multiple/parallel execution of the protocol.

The organization of this paper is as follows. In the next section we will give the motivation for our work. Then we introduce our model, and present a hardware-based implementation of our model. This implementation utilizes tamper-proof devices such as smart cards. The following section is devoted to the software-based implementation of our model. Finally, we introduce an oblivious transfer scheme that uses no cryptographic tools.

Motivation

The originality of the oblivious transfer is due to Rabin (1981). Informally, Rabin's proposed protocol for the oblivious transfer is as follows. Alice and Bob both know the number M and Alice knows the factors of M. Alice and Bob wish to interact in such a way that with probability ½ Bob learns the factors of M, and with probability ½ Bob is not able to factor M. In addition, Alice should not be able to figure out whether Bob has learned the factors of M or not. Rabin's protocol contains three steps:

i. Bob sends $Y = X^2 \bmod M$ to Alice, where $0<X<M$ is a random integer.
ii. Alice responds with a random square root Z of Y mod M (if no square root exists, then Alice does nothing);
iii. Bob checks that $Z^2 = Y \bmod M$, and if not outputs 'c' for cheating.

Since Y has four square roots mod M, with probability of ½ the value of Z will be X or -X, which gives no information to Bob in order to factor M. However, with probability ½ the value of Z is different from X, -X and in this case Bob learns a factor of M using the greatest common divisor of M and X-Z. Fischer et al. (1996) have shown that the above protocol is not rigorous and they enhanced this work.

Even et al. (1982) generalized Rabin's idea to the OT_1^2 in which Alice has two strings and Bob wishes to learn one of them. This scenario has been studied extensively and generalized to a wide variety of models including OT_1^N, OT_{N-1}^N, and OT_K^N. In an OT_1^N the sender has N strings and the receiver wishes to obtain one of these strings. Naor and Pinkas (1999a) suggested a protocol based on a protocol for OT_1^2. They have also proposed protocols for OT_K^N in which the sender has N strings and the receiver wishes to obtain K ($K < N$) of these strings. Hence, OT_{N-1}^N can be considered as a special case of the general OT_K^N.

The major concern in the implementation of an oblivious transfer protocol is to preserve the secrecy and the privacy of both parties. Rivest (2000) pointed out the following fact from Damgard et al. (1999): 'It is well known (and easy to see) that in a two-player scenario with only noiseless communication, OT [Oblivious Transfer] and BC [Bit Commitment] with information-theoretic security is not possible, even if only passive cheating is assuming, and players are allowed infinite computing power.' Rivest (2000) has shown that unconditional secrecy can be achieved with the help of a *trusted initializer*. In Rivest's protocol, at the initialization step, the trusted party sends some information to both the sender and the receiver. In the message transfer stage, the receiver sends some request to the sender and the sender replies with some information. That is, the protocol has three steps and works with the help of a trusted initializer.

Another important issue in the implementation of an oblivious transfer protocol is the efficiency of the system. Indeed, as has been pointed out by Naor and Pinkas (2001) 'The result of Impagliazzo and Rudich (1989) implies that it is unlikely that oblivious transfer could be based on more efficient one-way functions or other private-key cryptographic primitives.' As a result, all known oblivious transfer protocols require public-key operations, and therefore their performance is computationally intensive. Several works have been done in order to improve the efficiency of many oblivious transfer protocols, e.g. Brassard and Crepeau (1997), Naor and Pinkas (1999b, 2001). Dodis and Micali (1999) have discussed the computational cost of several oblivious transfer protocols.

Our motivation is to devise oblivious transfer schemes that achieve unconditional security with no help from a trusted party; neither individual trust nor distributed trust.

Note that in Rivest's scheme if the trusted party cooperates with the receiver then the receiver can obtain all Alice's secrets. Our scheme prevents such collaboration, since it does not require any interaction between the sender and the receiver at all.

THE MODEL

We observe that OT_1^2, OT_1^N, and OT_{N-1}^N are special cases of the OT_K^N, where K<N (for any K and N). In OT_K^N, Alice has N strings, $M_1 \ldots M_N$, and Bob wishes to obtain K of these strings. A trivial solution could be if Bob sends the indices of the messages he wishes to learn to Alice and then Alice sends the requested messages to Bob. But, in this way, Alice knows what information is obtained by Bob. The other trivial solution could be if Alice sends all strings to Bob and then Bob chooses the strings he wishes to learn. Clearly, this does not prevent Bob from learning the other strings. A slightly better scheme could be if Alice gives all N strings to a trusted party and Bob asks for a subset, K, of such strings. In this way, Alice cannot figure out what information is given to Bob, and Bob cannot learn anything more than the K request strings. This scheme is a weaker version of the one presented by Rivest (2000). In fact, the advantage of the Rivest scheme is that the trusted party learns nothing about Alice's secret information. The common shortcoming of these two schemes, however, is the existence of a trusted party, which is normally unacceptable in cryptographic protocols.

A logical solution to this problem could be if Alice transfers all information to Bob in such a way that learning the desirable portion of the transmitted information implies sacrificing the information that is not supposed to be learned by Bob. As a basic solution, we suggest the use of any tamper-proof device (e.g., smart-cards). By a tamper-proof device, we mean any devices that can be used only in a particular manner; otherwise the device will be corrupted and its content will not be accessible any more. In some situations, devices can be devised in such a way that they can destroy themselves.

Our basic model for achieving oblivious transfer is very simple and works as follows. Alice loads each device with some information (depending on the required OT model) and gives these devices and an access code (password) to Bob. Bob can operate with these devices using the access code given to him. The device, however, is designed in such a way that it destroys some part of its information in each state of operation. That is, Bob can learn the desirable information, if he accepts sacrificing the information that was not supposed to be learned. The tamper-proof device, which we will use, has the following two functions:

1) **GetKey** - This function allows the user to ask for the key (the input parameter to GetMessage function). If this function is called, the device either returns the key (if it contains the key) or answers ‘no-key’ (if no key is stored in the device). After this function is called, no matter which case occurs, the device will be corrupted.

2) **GetMessage** - This function requests a key as the input and if the key is authorized then it returns the message stored within; otherwise it returns ‘error’. In any case, after this function is called, the device will be corrupted.

HARDWARE-BASED IMPLEMENTATION

We assume that S is the key (the input parameter to the function GetMessage) in order to obtain the message stored in the device. Note that S is known to Alice only, but Bob knows which string is stored in which device, i.e., he knows the indices of the messages associated with each device. We also assume that S is an element of Z_p, where p is a prime.

A General Hardware-Based Model

In the general scenario of oblivious transfer, Alice has N strings, $M_1 \ldots M_N$, and Bob wishes to learn K (K < N) of these strings. In order to implement an oblivious transfer for this scenario, Alice employs the Shamir (1979) threshold scheme, where the threshold parameter is N-K. That is, she splits the secret key into N

pieces such that any set of at least N-K of these pieces can reconstruct the original secret. More precisely, she performs the following:

a) Alice secretly chooses (independently at random) N-K-1 elements of Z_p, denoted $a_1 \ldots a_{N-K-1}$ and forms the polynomial $f(x) = S + a_1 x + \ldots + a_{N-K-1} x^{N-K-1}$
b) For $i = 1 \ldots N$, Alice computes s_i, where $s_i = f(i) \bmod p$
c) Alice loads device d_i with key s_i, as the key value, and message M_i, as the message value.
d) Alice gives all devices and an access code to Bob.

After receiving N devices there is no contact between Alice and Bob, i.e. Bob can obtain K messages (out of N) in his choice. Let, w.l.o.g. $M_1 \ldots M_K$ be the K messages that Bob wishes to learn. Bob consults with N-K devices $d_{K+1} \ldots d_N$ and performs the GetKey function on all these devices. Hence, he obtains N-K shares associated with the polynomial f(x). Knowing N-K points of a polynomial of degree at most N-K-1 enables Bob to reconstruct the polynomial (e.g., using Lagrange polynomial interpolation) and learn the secret access code S. As a result of this process, all devices except those containing the message that Bob wishes to learn are corrupted. Now, having the access code S, Bob can obtain the messages $M_1 \ldots M_K$, by performing the GetMessage function on devices $d_1 \ldots d_K$.

As can be seen, the above scheme works for all models of oblivious transfer in a similar manner. It is worth mentioning that the implementation of OT_1^2 is even much easier, since the polynomial f(x) will be of degree zero, and thus there is no need for any computation for share calculation. That is, in this case, Alice devises two tamper-proof devices and loads one of them with S, M_0 and the other one with S, M_1 as their key and message, respectively. Bob, who wishes to learn the message stored in one of these devices, can consult with the other device and obtain the access key, S, which enables him to get the message.

Back to the Origin

Now we recall the original scenario of oblivious transfer, due to Rabin (1981). Let Alice and Bob both know the number M and Alice know the factorization of M. Alice and Bob wish to interact in such a way that with probability ½ Bob learns the factors of M, and with probability ½ Bob is not able to factor M. In addition, Alice should have no knowledge as to whether Bob has learned the factorization of M or not.

For this scenario, Alice devises two tamper-proof devices such that one of them contains a factor of M as its message, while the other device contains S as its key but contains no message. Alice gives these two devices to Bob. Note that in this scheme devices are indistinguishable from each other, that is, Bob has no idea which device contains the key and which device contains the message. Bob can learn the factors of M if he knows the key. On the other hand, Bob can obtain the key if he performs the function GetKey on the device that contains the key. Otherwise, he loses the factors of M --although he can get the key later but he cannot learn the factors of M. Since two devices are the same, Bob has only a 50% chance to contact the correct device first (to obtain the key without losing the message). That is, with probability ½ Bob learns the factors of M, and with probability ½ he is not able to factor M.

It is worth mentioning that Crepeau and Kilian (1988a, 1988b) have extended Rabin's scheme for other fractions of probabilities (e.g., ¼ etc.). In particular, they have shown that a binary symmetric channel can be used to implement OT_1^2 with unconditional security, but their scheme is not very efficient. One can see these extensions are easily approachable using our proposed scheme. For example, let Alice devise three tamper-proof devices, d_1, d_2, d_3 such that only one device contains a factor of M as its message and the other two devices have no message. Alice performs the Shamir (2, 2)-threshold scheme and generates two shares s_1, s_2 regarding the access key S. Then Alice loads the key information of those devices that have no message with s_1 and s_2 respectively. Hence, two devices contain only a key and one device contains only factors of M. Giving these three devices to Bob, with probability 1/3 Bob can learn the factors of m and with probability of 2/3 he is not able to factor M. One can see, comparing to the previous works in this regard, our scheme is not only very simple and efficient but also very rigorous.

Remark: For the sake of clarity of the system we assumed that each device contains a pair of functions (GetKey and GetMessage). It is reasonable that the whole scheme be devised within a single device. That is, the sender provides to the receiver a single device that contains N pairs of functions, ($GetKey_1$,

$GetMessage_1$) … ($GetKey_N$, $GetMessage_N$) such that each pair of functions works independently, i.e. performing each function will corrupt only the relevant key and/or message.

SECURITY ANALYSIS

***Theorem**:* Assuming that tamper-proof devices exist, the proposed oblivious transfer protocols preserve the unconditional security of the sender.

Proof: In the proposed oblivious transfer protocols, messages are stored in tamper-proof devices. The receiver has no way to learn the message except by obtaining the key. The key, however, is distributed among the devices in such a way that the receiver can obtain the key if and only if he sacrifices all messages that he is not supposed to learn. This is because Shamir's secret sharing scheme is perfect, i.e. knowing less than a predetermined number of shares gives absolutely no idea about the secret. Therefore, the receiver cannot learn anything more than the information that is supposed to be learned.

Theorem: Assuming that tamper-proof devices exist, the proposed oblivious transfer protocols preserve the unconditional privacy of the receiver.

Proof: In the proposed oblivious transfer protocols, there is no interaction between the sender and the receiver, that is, no information will leak via the protocol itself. Moreover, after providing all devices to the receiver, there is no way for the sender to figure out which devices the receiver has used, and therefore, the sender has no knowledge about the information that has been obtained by the receiver (note that at the end of the protocol all devices are corrupted).

SOFTWARE-BASED IMPLEMENTATION

Designing software systems equivalent to hardware systems and vice versa, is a common practice in the Computer Science and Computer Engineering fields.
It is not difficult to implement a software system that is equivalent to the proposed hardware-based model. For example, consider the following piece of code, which is written in C++ programming language.

```
// Program file -- GetKey1.cpp
#include<stdlib.h>
#include<iostream.h>
int main(int argc, char *argv[ ])
{
	// return the stored key value
	// ...
	system ("rm GetMessage1");
	system ("rm GetKey1");
	return 0;
}

// Program file -- GetMessage1.cpp
#include<stdlib.h>
#include<iostream.h>
int main(int argc, char *argv[ ])
{
	// If the given key (i.e., argv[1]) is correct
	// then return the relevant data.
	// ...
	system ("rm GetKey1");
	system ("rm GetMessage1");
	return 0;
}
```

Assume that the above function codes are compiled and the executable codes are stored as GetKey1 and GetMessage1 programs respectively. Running the GetKey1 program will return the stored key and will then remove both the GetMessage1 and GetKey1 programs (because of the execution of the 'system call commands'). On the other hand, if the GetMessage1 program is called, it first checks the validity of the given key (entered in the 'command-line') and if the given key is correct then it returns the relevant data. After this statement, it does not matter whether the data is returned (i.e. the given key was correct) or not, the system will discard the GetKey1 and GetMessage1 programs. Therefore, the above pair of functions satisfies the requirements defined in the hardware-based model.

A General Software-Based Model

Recall the general scenario in which Alice has N strings, $M_1 \ldots M_N$, and Bob wishes to learn K ($K<N$) of these strings. As with the hardware-based model, Alice employs the Shamir (N-K, N)-threshold secret sharing scheme in order to split the secret access code into N pieces such that any set of at least N-K of these pieces is sufficient to reconstruct the original secret. That is, Alice generates N pairs of programs ($GetKey_i$, $GetMessage_i$, $i = 1 \ldots N$) and loads the program $GetKey_i$ with share s_i (drawn from the secret access code) and loads the program $GetMessage_i$ with message M_i.

As in the hardware-based scheme, if Bob wishes to learn a particular set of K messages, then he must sacrifice the rest of N-K messages (by running the relevant $GetKey_i$ programs in order to reconstruct the secret code, and therefore losing the relevant messages). Obviously, this scheme works for all scenarios of the oblivious transfer in a similar manner. In particular for the OT_1^2 scheme, Alice does not need to perform the Shamir scheme for share distribution.

Back to the Origin

The original scenario of the oblivious transfer can also be implemented in a similar way. That is, Alice can generate two pairs of programs (GetKey1, GetMessage1) and (GetKey2, GetMessage2), such that in one of these pairs the GetKey function contains no key, but the relevant GetMessage function contains factors of M as its message. In the other pair, the GetKey function contains the access key while the relevant GetMessage function contains no message. In this implementation, there is a 50% chance that Bob will learn the factors of M, and there is also a 50% chance that Bob is not able to discover factor M. Moreover, in a similar manner, the scheme can be extended for other fractions of probabilities (e.g. 1/3, 1/4 etc.).

SECURITY ANALYSIS

The receiver, Bob, cannot learn any information about the access code without obtaining N-K shares of the secret --this is due to the fact that Shamir's (N-K, N)-threshold scheme is perfect.
On the other hand, the system is designed in a way that learning such N-K hares implies the loss of relevant N-K messages. That is, Bob cannot learn the access key, and therefore the desirable K messages, without sacrificing the remaining N-K messages. Hence, Bob cannot learn more than what he was supposed to learn.

After the protocol is completed, all functions are discarded (similar to the hardware-based model in which all devices are corrupted) and thus there is no way for the sender to figure out what information has been discovered by the receiver. Note that in both hardware and software based protocols; we assume that the receiver performs the protocol privately (i.e. the sender does not know the order of devices/functions which the receiver examines). In order to avoid any redundancy, the proofs of the following two theorems, which are similar to those in the hardware-based scheme, are omitted:

Theorem: Assuming that the protocol runs privately, the proposed oblivious transfer protocol preserves the unconditional security of the sender.

Theorem: Assuming that the protocol runs privately, the proposed oblivious transfer protocol preserves the unconditional privacy of the receiver.

Another Issue in Oblivious Transfer

In existing oblivious transfer schemes, the receiver decides what information he wishes to learn and a protocol is devised such that if it has been followed properly, the receiver obtains the desired information. That is, the protocol guarantees that the receiver will obtain a *predetermined fraction* of the information, not a *predetermined section* of the information. For instance, in the OT_1^2 protocols the receiver can choose whether he wishes to obtain M_0 or M_1, and then, based on his input 'c' from the set {0, 1}, he will obtain M_c. That is, if the receiver performs the protocol twice (with different inputs), he can learn both messages, M_0 and M_1. This may not be a problem when Alice participates in the protocol as the sender (e.g. in the original OT scheme due to Rabin) or when a trusted party plays the role of the sender (as in the Rivest scheme). However, in many other schemes, where the information is stored in some servers, the main question is what mechanism prevents multiple/parallel performance of the protocol? To the best of the knowledge of the authors, this problem has never been considered in the literature. The only scheme that discusses a relevant problem is the
Naor-Pinkas (2000) distributed oblivious transfer. In this scheme, in order to ensure that the chooser does not obtain more than K shares (that will help the chooser to learn more than he was supposed to learn), their scheme utilizes a sub-protocol that prevents the chooser from contacting more than K servers. Our observation is that utilizing such sub-protocol enables us to implement a general oblivious transfer protocol without using any cryptographic tool. In fact, we can devise the following oblivious transfer protocol that can be served in all scenarios.

The Protocol:

i. Each server S_i is loaded with a message M_i (i = 1 … N).
ii. The chooser/receiver contacts relevant servers and obtains the information stored in each server.

This simple protocol provides unconditional security and privacy for both the sender and receiver.

CONCLUSIONS

We studied oblivious transfer schemes and proposed general models than can be employed, in a similar manner, for all oblivious transfer scenarios. Note that the proposed scheme not only provides a solution to OT_K^N, but also can be used in the cases where the receiver obtains the desired information with some predetermined probability (e.g. as in the original oblivious transfer protocol due to Rabin (1981), or the work of Crepeau and Kilian (1988a, 1988b). That is, our protocols can also be employed in the original scenario of oblivious transfer.

The subtle difference between our proposed hardware-based model and existing oblivious transfer protocols is that in our schemes, it is impossible to perform the protocol more than once, and therefore our schemes are not subject to the problem of multiple performances of the protocols. Moreover, assuming that a mechanism exists that can prevent multiple/parallel performances of the protocol; we have suggested a software-based model for constructing an oblivious transfer scheme without using any cryptographic tools.

References

Bellare M. and Micali S. (1989) Non-Interactive Oblivious Transfer and Applications, in *Advances in Cryptology - Proceedings of CRYPTO '89* (Brassard G. ed.), vol. 435 of Lecture Notes in Computer Science, pp.547-559, Springer-Verlag.

Brassard G. and Crepeau C (1997) Oblivious Transfer and Privacy Amplification, in *Advances in Cryptology - Proceedings of EUROCRYPT '97* (Fumy W., ed.), vol.1233 of Lecture Notes in Computer Science, pp.334-347, Springer-Verlag.

Crepeau C. (1987) Equivalence Between two Flavours of Oblivious Transfer, in *Advances in Cryptology – Proceedings of CRYPTO '87* (Pomerance C., ed.), vol. 293 of Lecture Notes in Computer Science, pp.350-354, Springer-Verlag.

Crepeau C. (1994) Quantum Oblivious Transfer, in *Journal of Modern Optics*, vol. 41, pp.2455-2466.

Crepeau C. and Kilian J. (1988a) Weakening Security Assumptions and Oblivious Transfer, in *Advances in Cryptology - Proceedings of CRYPTO '88* (Goldwasser S., ed.), vol. 403 of Lecture Notes in Computer Science, pp.2-7, Springer-Verlag.

Crepeau C. and Kilian J. (1988b) Achieving Oblivious Transfer Using Weakened Security Assumptions, in *the 29th IEEE Symposium on the Foundations of Computer Science* (FOCS), pp.42-52.

Crepeau C., Graaf J. van de, and Tapp A. (1995) Committed Oblivious Transfer and Private Multi Party Computation, in *Advances in Cryptology - Proceedings of CRYPTO '95* (Coppersmith D., ed.), vol. 963 of Lecture Notes in Computer Science, pp.110-123, Springer-Verlag.

Damgard I., Kilian J., and Salvail L. (1999) On the (Im)possibility of Basing Oblivious Transfer and Bit Commitment on Weakened Security Assumption, in *Advances in Cryptology - Proceedings of EUROCRYPT '99* (Stern J., ed.), vol. 1592 of Lecture Notes in Computer Science, pp.56-73, Springer-Verlag.

Dodis Y. and Micali S. (1999) Lower Bounds for Oblivious Transfer Reductions, in *Advances in Cryptology - Proceedings of EUROCRYPT '99* (Stern J., ed.), vol. 1592 of Lecture Notes in Computer Science, pp. 42-54, Springer-Verlag.

Even S., Goldreich O., and Lempel A. (1982) A Randomized Protocol for Signing Contracts, in *Advances in Cryptology - Proceedings of CRYPTO '82* (Rivest R., Sherman A., and Chaum D., eds.), pp.205-210, Plenum Press.

Fischer M., Micali S., and Rackoff C. (1996) A Secret Protocol for the Oblivious Transfer, in *Journal of Cryptology*, vol. 9, no. 3, pp.191-195.

Ghodosi H. and Zaare-Nahandi R. (2006) Comments on '*m* out of *n* oblivious transfer', in *Information Processing Letters*, vol. 97, no. 4, pp.153-155.

Ghodosi H. (2007) On insecurity of Naor-Pinkas' distributed oblivious transfer, in *Information Processing Letters*, vol. 104, no. 5, pp.179-182.

Impagliazzo R. and Rudich S. (1989) Limits on the Provable Consequences of One-Way Permutations, in Proceedings of the 20th ACM Symposium on the Theory of Computing, (STOC), pp.44-61.

Mu Y., Zhang J. and Vardharajan V. (2002) *m* out of *n* oblivious transfer, in *ACISP2002 – Proceedings of Australasian Conference in Information Security and Privacy* (Batten L. and Seberry J. eds), vol. 2384 of Lecture Notes in Computer Science, pp.395-405, Springer-Verlag.

Naor M. and Pinkas B. (1999a) Oblivious Transfer and Polynomial Evaluation, in *31st Symposium on the Theory of Computing* (STOC), pp.245-254.

Naor M. and Pinkas B. (1999b) Oblivious Transfer with Adaptive Queries, in *Advances in Cryptology - Proceedings of CRYPTO '99* (Wiener M., ed.), vol. 1666 of Lecture Notes in Computer Science, pp.573-590, Springer-Verlag.

Naor M. and Pinkas B. (2000) Distributed Oblivious Transfer, in *Advances in Cryptology - Proceedings of ASIACRYPT 2000*, vol. 1976 of Lecture Notes in Computer Science, pp.205-219, Springer-Verlag.

Naor M. and Pinkas B. (2001) Efficient Oblivious Transfer Protocols, in Proceedings of (SIAM) Symposium on Discrete Algorithms.

Rabin M (1981) How to exchange secrets by oblivious transfer, Technical Report TR-81, Aiken Computation Laboratory, USA.

Rivest R. (2000) Unconditionally Secure Commitment and Oblivious Transfer Schemes Using Private Channels and a Trusted Initializer, manuscript. Available: http://theory.lcs.mit.edu/\~~rivest/publications.html.

Shamir A. (1979) How to Share a Secret, in *Communications of the ACM*, vol. 22, pp.612-613.

A Non-Interactive Multiparty Computation Protocol

Hossein Ghodosi [1] Rahim Zaare-Nahandi [2]

[1]School of Mathematics, Physics and Information Technology
James Cook University, Townsville, QLD 4811 Australia
E-mail: hossein.ghodosi@jcu.edu.au

[2]Department of Mathematics and Computer Science
University of Tehran, Tehran, Iran
E-mail: rahimzn@khayam.ut.ac.ir

ABSTRACT

Multiparty computation considers the design of protocols such that a set of n users $u_1 \ldots u_n$, each with their own secret input x_i, can compute a function $Y=F(x_1, \ldots, x_n)$. In this paper, we study a requirement for devising non-interactive multiparty computation protocols and present the first non-interactive multiparty computation protocol, which has the following advantages:

a) The amount of computation required by each participant is equal to the amount of computation in an *ideal* model (i.e. the proposed scheme is optimal).
b) In the presence of passive adversaries, no set of less than n users can learn more than what is allowed in the underlying secret sharing scheme (other than the function value).
c) In the presence of active adversaries (i.e. Byzantine faults are allowed), no set of less than $n/2$ users can either learn more than what is allowed in the underlying secret sharing scheme, nor can they disrupt the computation.
d) The partial security provided in our scheme is *unconditional*. That is, the proposed scheme does not rely on any non-proven cryptographic assumption.

INTRODUCTION

There are many situations in which several parties collaborate to achieve a common goal (e.g. signing a contract, electronic voting, etc.). Multiparty computation, in particular, considers the design of protocols such that a set of n users $u_1 \ldots u_n$, each with their own secret input x_i, can compute a function $Y=F(x_1, \ldots, x_n)$. In secure multiparty computation protocols, an essential requirement is that the protocol must compute the correct value for Y, while keeping input values of users as private as possible. This problem has been investigated extensively and several models have been introduced in the literature. From the security point of view, these models can be classified into two categories:

Computationally secure multiparty computations: In this model, the system is secure based on the assumption that the adversary is polynomially bounded. More precisely, breaching the security of the system implies to solve a problem that is believed to be intractable, and thus an adversary with limited computing power cannot do so (e.g. assuming that one-way functions exist, or the discrete logarithm problem is hard, and so on). In this setting, the communication network does not need to be secure, since with the help of cryptographic protocols, privacy and authentication problems can be managed easily.

Unconditionally secure multiparty computations: In this model, the system is intrinsically secure. That is, no matter how much time and/or computing power is available to the adversary, they cannot break the system. This notion of security is clearly much stronger than computational security. A basic requirement,

however, is that secure channels must exist amongst the users (to provide privacy and authenticity for the messages being transferred amongst the users).

Background

The notion of multiparty computation was first introduced by Yao (1982), in the context of two-party computation, and then generalized by Goldreich et al. (1987) for any number of users. Their model works based on the assumption that *one-way functions* with trapdoor exist (i.e. their model is computationally secure). They have shown that in the presence of passive adversaries (i.e. all parties follow the protocol, and no fault occurs) any function can be computed by n users, in such a way that no subset of less than n users can learn any useful information more than the function value. They have also shown that if Byzantine faults are allowed, any function can be computed by n collaborating users, as long as the majority of participants are honest.

Ben-Or et al. (1988) and Chaum et al. (1988), independently studied unconditionally secure multiparty computation of a function $F(x_1, \ldots, x_n)$ by a set of n collaborating users. They have shown that:

i. If no faults occur, no set of size $t<n/2$ of participants learn any additional information, other than the function value.
ii. If Byzantine faults are allowed, no set of size $t<n/3$ can learn any additional information, or disrupt the computation.

Rabin and Ben-Or (1989) suggested to improve the bounds for case of Byzantine fault by tolerating a negligible error probability. Their conjecture is that some cryptographic assumptions are needed, even with a secure channel, to handle t faulty users for $n/3 \leq t < n/2$. Kushilevitz (1989) has shown that for every $1 \leq g(n) \leq 2^{n+1}$ there are functions that can be privately computed with $g(n)$ rounds of computation, but not with $g(n)-1$ round of computation, i.e. the communication cost of private protocols are exponentially higher than the communication cost of non-private protocols.

In subsequent years, many other researchers investigated multiparty computation schemes with different constraints. For example, multiparty computation in an asynchronous network has been studied by Ben-Or et al. (1993) and Canetti (1995). Hirt and Maurer (2000) studied multiparty computation protocols with general structure (in all previous works the access structure is threshold).

Issues in Devising Multiparty Computation Protocols

There are two main issues (namely *security* and *efficiency*) in devising any multiparty computation protocol. The pioneer work on multiparty computation is due to Goldreich et al. (1987). Their scheme achieves security in computational model, but is very time consuming and impractical. Classical results in unconditionally secure multiparty computation are due to Ben-Or et al. (1988), and Chaum et al. (1988). Although the protocol in Ben-Or et al. (1988) is more efficient than Chaum et al. (1988), it is still not very practical for many applications. In order to illustrate the problem, let us give a brief review of their scheme in the presence of a passive adversary. The scheme is divided into three steps: (i) initialization, (ii) computation, and (iii) obtaining the result.

Initialization: Each user u_j $(j = 1 \ldots n)$, using Shamir's (1979) secret sharing scheme, distributes its secret input, x_j, amongst all users. More precisely, u_j chooses (independently at random) t elements of Z_p, denoted $a_1 \ldots a_t$ and forms the polynomial $f_j(x) = x_j + a_1 x + \ldots + a_t x^t$ (here, p is a prime, x_i is an element from Z_p, and all computations are done on Z_p). Then u_j privately gives share $s_{i,j} = f_j(i)$ to user u_i $(i = 1 \ldots n)$.
If all users properly follow the sharing protocol, the values of x_i are shared amongst them in such a way that subsets of t or fewer users cannot learn any useful information about the other users' secret input. This is because the underlying Shamir secret sharing scheme is perfect

Computation: Let $s_{i,k}$ and $s_{i,m}$ be u_i's shares (associated with polynomials $f_k(x)$ and $f_m(x)$) from the secrets x_k and x_m, respectively. Computation of any linear function is straightforward. For example, in order to compute $x_k + x_m$ each cooperating participant, u_i, computes $s_i^{k+m} = s_{i,k} + s_{i,m}$, which is the share of u_i from polynomial $h(x) = f_k(x) + f_m(x)$. Since $h(x)$ is a polynomial of degree at most t, a set of at least $t+1$ users can

reconstruct the polynomial and thus retrieve the constant term of $h(x)$, which is x_k+x_m. Similarly, $c\ x_k$, where c is an element of Z_p is a known scaler, can be computed by $t+1$ participants (i.e. each participant u_i calculates $c\ s_{i,k}$ as its share from $c\ x_k$. That is, there is a non-interactive protocol for computing any linear function $F(x_1, \ldots, x_n)$.

Computing non-linear functions, however, needs some consideration. Assume we want to compute $x_k\ x_m$. Although u_i can compute $s_i^{k\,m} = s_{i,k}\ s_{i,\,m}$, where $s_i^{k\,m}$ is the share of u_i from $x_k\ x_m$, there are two problems. The first problem is that $s_i^{k\,m}$ is the share of u_i from a polynomial $h(x) = f_k(x)\ f_m(x)$, where $h(x)$ is a polynomial of degree, at most, $2t$. The second problem is that $h(x)$ is not a random polynomial (e.g. it is not irreducible). To overcome these two problems it has been suggested to perform two sub-protocols:

1) The degree reduction protocol.
2) The randomization protocol.

The purpose of these two sub-protocols is to modify the shares $s_i^{k\,m}$ $(i=1 \ldots n)$, such that they be associated with random polynomial of degree at most t. This implies at least $2t+1$ performance of share distribution of Shamir scheme. Moreover, if $n < 3t$, after each multiplication, the above two sub-protocols must be executed, otherwise further multiplications will raise the degree, and if the degree passes n, there will not be enough points for the interpolation. One can see that even if all users are honest, a secure multiparty computation protocol is highly interactive and time consuming. Obviously, in the presence of Byzantine faults, the situation is much worse. For example, all users must be convinced that all shares they have received are actually derived from polynomials of degree at most t. In order to achieve this goal, the underlying secret sharing scheme must be armed with extra features that enable the users to verify their shares. In an unconditionally secure environment, this usually implies to corporate some zero-knowledge protocol in the system, which is a highly interactive and time consuming task. Indeed, Hirt et al. (2000) pointed out that the most efficient unconditionally secure protocols among n players, tolerating cheating by up to $t<n/3$ of them, requires communicating $O(n^6)$ field elements for each multiplication of two elements, even if only one player cheats. In their work, they have presented a method that requires communicating $O(n^3)$ field elements per each multiplication.

Obtaining the Result: At the end of the computation stage, a set of participants who properly followed the protocol, collectively has shares of a polynomial $g(x)$, where $g(0)=Y$, and the degree of $g(x)$ is less than the number of participants. Hence, the collaborating participants can interpolate the polynomial and retrieve the value $Y=F(x_1, \ldots, x_n)$.

Motivation

Let consider an *ideal-model* for multiparty computation. In ideal-model, each participant gives their secret input to a trusted third party, who computes the function $Y=F(x_1, \ldots, x_n)$, privately, and returns the results to corresponding participants. This setting implies *no interaction amongst the users* and requires the *lowest possible amount of computation*. However, this model is unacceptable in cryptographic environment, since all parties may not agree on a trusted party. In order to overcome this problem, an *ideal-multiparty-computation* protocol can be thought of in the following way.

Initialization: Each user u_i distribute its input secret x_i amongst all users, according to an agreed secret sharing scheme.

Computation: Each user u_i privately calculates $Y_i= F(s_{i,1}, \ldots, s_{i,n})$, where $s_{i,\,j}$ is the share of participant u_i from input x_j $(j = 1, \ldots, n)$.

Obtaining the Result: All users pool their partial results Y_i, and compute the function value, $Y=F(x_1, \ldots, x_n)$. This step is similar to the secret reconstruction phase of the underlying secret sharing scheme.

In this model, there will be no interaction during the computation. Furthermore, each participant performs the same amount of computation that a trusted party does in an *ideal-model*. That is, the above design provides the most efficient possible solution to multiparty computation in the absence of a trusted party. Implementation of such *ideal-multiparty-computation* protocols will be straightforward, if the underlying secret sharing scheme is compatible with addition and multiplication. Note that Shamir's secret sharing

scheme is compatible with addition, and thus any linear function can be computed with no interaction. But it is not compatible with multiplication, i.e. multiplication of two shares associated with t degree polynomials is the share of multiples of their corresponding secret that is associated with a $2t$ degree polynomial. Current results in unconditionally secure multiparty computation indicates that implementation of secure ideal-multiparty-computation is impossible. In this paper we will show by tolerating some level of security loss it is possible to achieve an ideal model for computing any function.

Our Contribution

1) We introduce a new secret sharing schemes which is compatible with addition, multiplication and scalar product, i.e. if $s_1, \ldots, s_n$ and $k_1, \ldots, k_n$ are shares of users $u_1, \ldots, u_n$ from secret values s and k (respectively) and c is any scalar, then:
   i. $(s_1 + k_1), \ldots, (s_n + k_n)$ are shares of $u_1, \ldots, u_n$ from $s + k$.
   ii. $(s_1 . k_1), \ldots, (s_n . k_n)$ are shares of $u_1, \ldots, u_n$ from $s . k$.
   iii. $(c . s_1), \ldots, (c . s_n)$ are shares of $u_1, \ldots, u_n$ from $c . s$.
2) Utilizing the proposed secret sharing scheme, we will show that every function of n inputs can be collaboratively computed by n users in such a way that:
   i. If no fault occurs, no set of size $t=n-1$ or less participants learn any additional information, other than allowed in the underlying secret sharing scheme.
   ii. If Byzantine faults are allowed, no set of size $t < n/2$ users can learn any additional information, other than allowed in the underlying secret sharing scheme, nor can they disrupt the computation.
3) The partial security provided in our scheme is *unconditional*. That is, no matter how much time and/or computing power is available to the adversaries, they cannot learn more than that is allowed in the underlying secret sharing scheme.

The structure of this paper is as follows. In the next section we introduce our new secret sharing scheme, and study the compatibility properties of this secret sharing scheme. Then we present an ideal-multiparty-computation in the presence of passive adversaries. Then we will show how to construct a VSS scheme based on the proposed secret sharing scheme. Finally, we present our robust ideal-multiparty-computation protocol.

A New Secret Sharing Scheme

In this section we introduce a new (n, n)-threshold secret sharing scheme. Let $0 \le s < P$ be a secret that we want to be shared amongst n users, $u_1, \ldots, u_n$. Also let $P=p_1.p_2. \ldots .p_n$, where $2^k < p_i < 2^{k+1}$ (for some positive integer k) are pairwise co-prime, i.e. $gcd(p_i, p_j)=1$, for all $i \neq j$.

Share Distribution: The dealer, who knows the secret s, privately gives to user u_i the share $s_i = s \bmod p_i$

Secret Reconstruction: Using Chinese Remainder Theorem (CRT), the secret can be reconstructed as,

$$s = a_1 q_1 s_1 + a_2 q_2 s_2 + \ldots + a_n q_n s_n \mod P,$$

where $a_1 q_1 + a_2 q_2 + \ldots + a_n q_n = 1 \mod P$, with $q_i = P/p_i$. Note that p_is and q_is, and hence a_is are public.

Efficiency Considerations

This secret sharing scheme is more efficient than the Shamir's polynomial approach.
In Shamir's scheme, a random $n-1$ degree polynomial is associated with an (n, n)-threshold scheme. Therefore, in the secret reconstruction phase, performing Lagrange interpolation formula, Shamir's scheme requires $O(n \log^2 n)$ operations. In our proposed modular approach, secret reconstruction requires only $O(n)$ operations.

Compatibility Properties

Theorem: if $s_1, \ldots, s_n$ and $k_1, \ldots, k_n$ are shares of users $u_1, \ldots, u_n$ from secret values s and k (respectively) and c is any scalar, then:

i. $(s_1 + k_1), \ldots, (s_n + k_n)$ are shares of $u_1, \ldots, u_n$ from $s + k$.
ii. $(s_1 . k_1), \ldots, (s_n . k_n)$ are shares of $u_1, \ldots, u_n$ from $s . k$.
iii. $(c . s_1), \ldots, (c . s_n)$ are shares of $u_1, \ldots, u_n$ from $c . s$.

Proof: These are in fact consequences of the module structure of integers modulo some natural number. We prove the first two. The proof of the third item is similar.
For each $i = 1, \ldots, n$, by hypothesis, $s - s_i = b_i p_i$ and $k - k_i = c_i p_i$ for some integers b_i and c_i.
Thus, $s + k - (s_i + k_i) = (b_i + c_i) p_i$, that is, $s_i + k_i = s + k \mod p_i$.
For the second part, observe that,

$$s k - s_i k_i = s (k - k_i) + k_i (s - s_i) = (s c_i + k_i b_i) p_i,$$

that is, $s_i k_i \equiv s k \mod p_i$.

Corollary: If $s_1, \ldots, s_n$ are shares of users $u_1, \ldots, u_n$ from secret s, and c is a positive scalar, then $s_1^c, \ldots, s_n^c$ are shares of users $u_1, \ldots, u_n$ from s^c.

AN IDEAL MULTIPARTY COMPUTATION PROTOCOL

Communication Model

We employ a non-cryptographic setting for communication (although we do not need interaction amongst the users, the initialization phase requires secure channels for share distribution). That is, we consider private channels between each pair of users in a synchronous model (i.e. a message send by a party in a clock 'tick' will be received by other party in the next 'tick').

The Protocol

There are n users $u_1, \ldots, u_n$, each with a public module p_i, and their own secret input x_i, such that $0 \leq x_i < P$, where $P = p_1 p_2 \ldots p_n$. The users wish to compute a function $Y = F(x_1, \ldots, x_n)$.
We assume that the adversary is passive (i.e. all parties follow the protocol, and no faults occur), and all computation are done on Z_P.

Initialization: Each user u_i ($i = 1, \ldots, n$) privately sends $s_{j,i} = x_i \mod p_j$ to u_j ($j = 1, \ldots, n$).
Computation: Each user u_i computes function $Y_i = F(s_{i,1}, \ldots, s_{i,n})$, where $s_{i,j}$ is u_i's share from the secret x_j. Note that each user performs computation on their public module, e.g. u_i's computation is on Z_{pi}. According to the compatibility properties of our underlying secret sharing scheme, the computed value Y_i, by user u_i, is in fact the share of user u_i from the value Y, i.e. $Y_i = Y \mod p_i$.

Obtaining the Result: All users pool their partial results Y_i, and compute the function value Y, using the CRT.

Efficiency

Obviously, the proposed multiparty computation protocol is the most efficient possible multiparty computation protocol. Indeed, the amount of computation required by each participant is equivalent to the amount of computation in the *ideal* model.

Verifiable Secret Sharing (VSS) Scheme

A primary goal of a secret sharing scheme is that the shareholders must be able to reconstruct the correct secret. The silent assumption of original secret sharing schemes was that the dealer is honest and distributes the shares of the secret properly. It is also assumed that in the secret reconstruction phase, shareholders will contribute their genuine shares and the original secret will be reconstructed. Tompa and Woll (1988) have shown that Shamir's scheme is subject to cheating by malicious users. They have shown a (or a group of)

cheater(s) can contribute false data and learn the correct value of the secret, while honest participants learn an incorrect value as the secret.

In verifiable secret sharing (VSS) schemes, the assumption is that the dealer is not trustworthy and the shareholders may be malicious. This implies that, in share distribution phase, shareholders must be able to verify the consistency of their shares (e.g. in Shamir scheme, shares are associated with a t degree polynomial). Also, in secret reconstruction phase, shareholders must be able to verify whether or not an information piece contributed by a participant is genuine. VSS schemes have been studied extensively, and several *computationally* and *unconditionally* secure VSS schemes have been presented in the literature --see e.g. Feldman (1987), Pedersen (1991), and Stadler (1996). To the best of the knowledge of authors, all VSS schemes in the presence of faulty dealer are based on the Shamir secret sharing scheme. In this section, we utilize the validity checking method suggested by Assmuth and Bloom (1983), in order to construct a VSS scheme in the presence of faulty shareholders/dealer.

Share Distribution Phase: There are n users in the system. Also there exists an active adversary that can corrupt up to t users, where $t<n/2$. The dealer, who is trustworthy or not, wants to distribute a secret s amongst these users in such a way that any set of $t+1$ or more users can uniquely determine the secret, but any set of t or less users fail to do so. The dealer performs the followings:

a) Chooses n integers $q_1, \ldots, q_n$, and $c = C^n_{t+1}$ integers $r_1, \ldots, r_c$, such that they are pairwise co-prime, and $q_1<q_2< \ldots <q_n$ have almost the same size, i.e. $2^k<q_i<2^{k+1}$ for some positive integer k.
b) Constructs n modules $p_1, \ldots, p_n$, subject to the following:
   i. $p_i = q_i \cdot r_{i1} \cdot r_{i2} \cdot \ldots r_{ik}$, such that q_i appears in p_i, and each r_j $(j=1, \ldots, c)$ appears in $t+1$ modules p_i,
   ii. $lcm(p_1, \ldots, p_{t+1}) > lcm(p_n, p_{n-1}, \ldots, p_{n-t+1})$, where 'lcm' stands for the Least Common Multiple.
   iii. Privately gives the share $s_i = s \bmod p_i$ to user u_i $(i=1, \ldots, n)$, where $0 \le s<P$, and $P=lcm(p_1, \ldots, p_{t+1})$.
c) In order to check whether the shares of participants are consistent (i.e. in the secret reconstruction phase, a set of honest participant can reconstruct a unique value in the range $[0,P[$), participants take part in the following protocol:
   i. u_i privately sends to each user u_j the value $v_{j,i} = s_i \bmod gcd(p_i, p_j)$.
   ii. u_j complains if $v_{j,i} \neq s_j \bmod gcd(p_i, p_j)$.
   iii. If more than t users complain then the dealer is disqualified otherwise the secret is shared.

Secret Reconstruction Phase: All shareholders pool their shares. A set of $t+1$ mutually compatible shares are chosen, and the secret is reconstructed.

Robust Ideal Multiparty Computation

In this scenario, the adversary is active and thus all information in the system must be verified.
In our system, however, we only verify the initial values and the partial results. This is because if a set of honest shareholders accept their shares (as they mutually compatible with at least t other honest participants' shares) then their partial results must also compatible. That is, participants do not need to have any interaction during the computation stage.

Initialization: All participants agree on the public modules $p_1, \ldots, p_n$ in the form that they have been created in the previous section. We assume that $0 \le x_i<P$, where $P= lcm(p_1, \ldots, p_{t+1})$.
User u_i $(i=1, \ldots, n)$ performs:

a) Privately sends to each user u_j the value $s_{j,i} = x_i \bmod (p_j)$.
b) After all participants received their shares from the secret x_i, each user u_j privately sends to user u_k the value $v_{k,j} = s_{j,i} \bmod gcd(p_j, p_k)$.
c) u_k complains if $v_{k,j} \neq s_{k,i} \bmod gcd(p_k, p_j)$.
d) If more than t users complain then u_i is disqualified. If u_i is disqualified all shares $s_{j,i}$ are discarded, u_i is eliminated from the system, and a public integer c_i is accepted by all users as the secret value x_i.

Since t corrupted users cannot conspire in order to disqualify an honest user, at the end of the initialization phase, there exists at least $t+1$ users in the system.

Computation: This phase is similar to the case that the adversary was passive, but computation of corrupted users may be faulty.

Obtaining the Result: This stage is similar to the secret reconstruction phase in our VSS scheme, that is:

1. All users pool their partial results Y_i (corrupted users may contribute with faulty results or may not co-operate).
2. A set of $t+1$ mutually compatible partial result is chosen, and the function value Y, is computed.

CONCLUSIONS

The main results of this paper are:

a) We studied *ideal-model* for multiparty computation, and pointed out the homomorphism requirement of the underlying secret sharing scheme.
b) We presented a secret sharing scheme that possesses the required homomorphism properties. Although our secret sharing scheme does not provide perfect secrecy, we are not aware whether a perfect secret sharing with required homomorphism properties can be constructed or not (current results in multiparty computation indicates that is unlikely to have such a scheme).
c) Utilizing the proposed secret sharing scheme, we presented the first non-interactive multiparty computation, which also is the most possible efficient multiparty computation protocol.

References

Asmuth C. and Bloom J. (1983) A Modular Approach to Key Safeguarding, in *IEEE Transactions on Information Theory*, vol. IT-29, pp.208-210.

Ben-Or M., Canetti R., and Goldreich O. (1993) Asynchronous Secure Computation, in *Proceedings of the 25th ACM Symposium on the Theory of Computing*, (STOC93), pp.52-61.

Ben-Or M., Goldwasser S., and Wigderson A. (1988) Completeness Theorem for Non-Cryptographic Fault-Tolerant Distributed Computation, in *Proceedings of the 20th ACM Annual Symposium on the Theory of Computing* (STOC88), pp.1-10.

Chaum D., Crepeau C., and Damgard I. (1988) Multiparty Unconditionally Secure Protocols, in *Proceedings of the 20th ACM Annual Symposium on the Theory of Computing* (STOC88), pp.11-19.

Canetti R. (1995) *Studies in Secure Multiparty Computation and Applications*, in *PhD Thesis to Weizmann Institute of Science*.

Feldman P. (1987) A Practical Scheme for Non-interactive Verifiable Secret Sharing, in *28th IEEE Symposium on Foundations of Computer Science*, pp.427-437.

Goldreich O., Micali S., and Wigderson A. (1987) How to Play any Mental Game, in *Proceedings of the 19th ACM Annual Symposium on the Theory of Computing*, (STOC87), pp.218-229.

Hirt M. and Maurer U. (2000) Player Simulation and General Adversary Structures in Perfect Multiparty Computation, in *Journal of Cryptology*, vol. 13, no.1, pp.31-60.

Hirt M., Maurer U., and Przydatek B. (2000) Efficient Secure Multi-party Computation, in *Advances in Cryptology - Proceedings of ASIACRYPT 2000* (Okamoto T., ed.), vol. 1976 of Lecture Notes in Computer Science}, pp.143-161, Springer-Verlag.

Kushilevitz E. (1989) Privacy and Communication Complexity, in *30th IEEE Symposium on the Foundations of Computer Science* (FOCS), pp.416-421.

Pedersen T. (1991) Non-Interactive and Information-Theoretic Secure Verifiable Secret Sharing, in *Advances in Cryptology - Proceedings of CRYPTO '91* (Feigenbaum J., ed.), vol. 576 of Lecture Notes in Computer Science, pp.129-140, Springer-Verlag.

Rabin T. and Ben-Or M. (1989) Verifiable Secret Sharing and Multiparty Protocols with Honest Majority, in *Proceedings of the 21st ACM Annual Symposium on the Theory of Computing*, (STOC89), pp.73-85.

Shamir A. (1979) How to Share a Secret, in *Communications of the ACM*, vol. 22, pp.612-613.

Stadler M. (1996) Publicly Verifiable Secret Sharing, in *Advances in Cryptology - Proceedings of EUROCRYPT '96* (Maurer U., ed.), vol. 1070 of Lecture Notes in Computer Science, pp.190-199, Springer-Verlag.

Tompa M. and Woll H. (1988) How To Share a Secret with Cheaters, in *Journal of Cryptology*, vol. 1, no.2, pp.133-138.

Yao A. (1982) Theory and Applications of Trapdoor Functions, in the 23rd IEEE Symposium on the Foundations of Computer Science, pp.80-91.

China's Information Security Policy Inadequate To Protect Its Netizen's Personal Rights

Fenyu Zeng and Pan Hua

School of Management & Humanity
Shanghai University of Electric Power
2103, Pingliang Rd. Shanghai, 200090, China
E-mail: Fengyuzeng@hotmail.com
Stevepan2005@hotmail.com

Abstract

It is widely accepted that the vast majority of security breaches are a result of human errors rather than technological flaws. Awareness of the risks and available safeguards are first line defenses for the security of information systems and networks. In both the public and private sectors, information security challenges must be met with a combination of factors, namely: technology, people, processes and policies or legislations. Regulations were designed to protect internet and network security, but it failed to accomplish this goal. Based on Public Choice Theory, this paper explains why it is the government's responsibility to create a trusted and secured information environment, to spread awareness among its citizens, and to protect privacy while promoting E-Governance and E-commerce for the overall benefit of the society in this Digital Age. After comparing with other countries like America, European Union, and Australia, which have set forth and undertaken a variety of public policies or legislations in raising awareness among its citizens to reduce the security risks and protecting their privacy, this essay analyzes the whole of China's existing information security policies and regulations. Our conclusion is: China's existing information security policy is inadequate to protect its Netizen's personal rights.

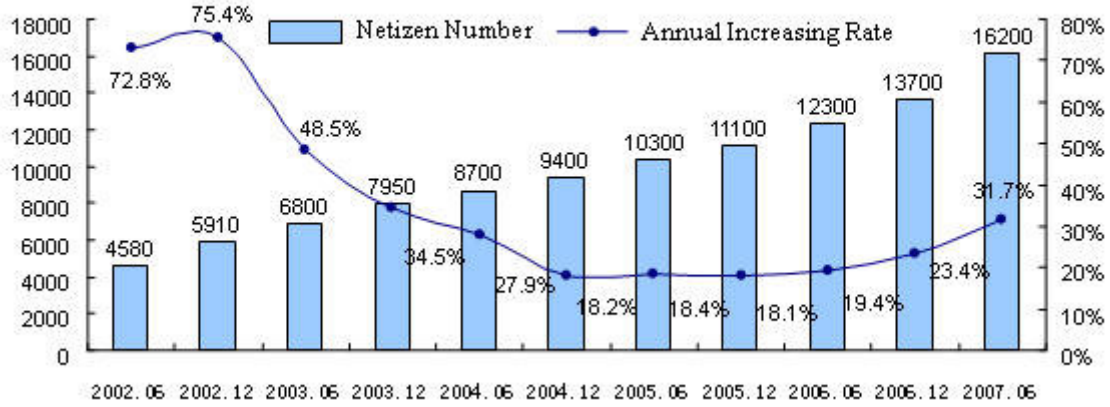
Key words: Insecurity of Internet; Netizen's unawareness; Information Security Policy; Inadequate

Introduction

♦ **The profile of Internet users in China.**

In recent years, Information and Communication Technologies (ICT) have become increasingly important for citizens of China, who are becoming more and more dependent on the use of internet and information networks services in their daily lives. According to the latest survey report on internet development in China released by the China Internet Network Information Center ("CNNIC"), as on June 30, 2007, the number of Internet users in China reached 162 million, a number which ranks the second in the world after America (211 million). Further, compared to the result of the last survey dated December 31, 2006, the number of internet users has soared by 25 million in the past six months, a rise of 31.7% (presented in Figure 1). This means by end 2008, China would surpass US to become the country that would house world's maximum Netizens. At present among these netizens, there are 51.2% under the age of 25, and 70.6% of the netizen under the age of 30. China's netizen on an average spend 18.6 hours each week online, which is higher when compared with other countries. More than 73.8% people use the internet at their home, and 31.2% at work, 37.2% at internet cafe. These data explains that young home users or single users cover the major portion of Netizens in China.

Figure 1:China Netizen Number and Annual growth rate



♦ **Current Internet Security is Failing us**

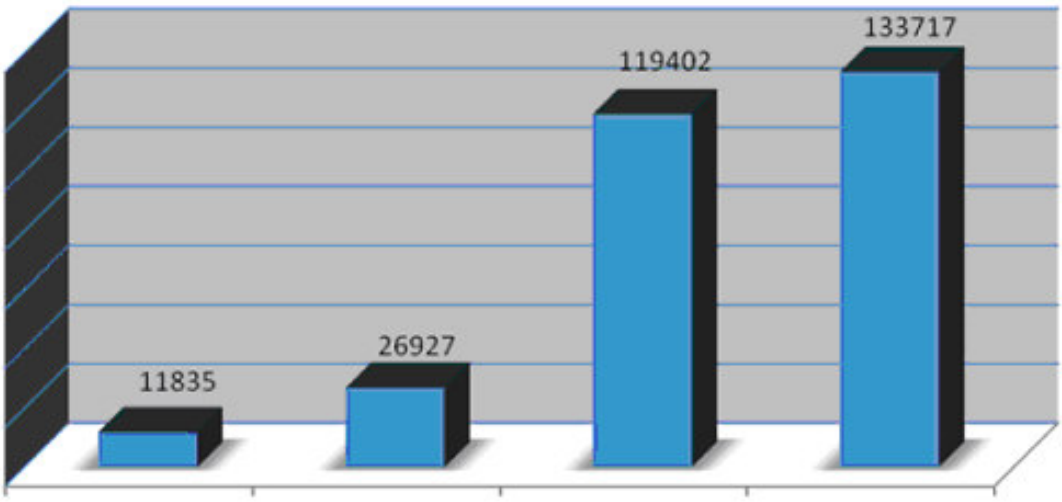
While there is a soaring growth of internet users in the country, the network environment is becoming more hostile and vulnerable to attack due to fast changing technologies, unawareness of security risk among users and lack of policies/legislations on information security. The attacks are becoming targeted, i.e. aimed at a particular company, organization, or individual users. The attacks are also much more sophisticated, using a mixture of techniques, for instance triggering users to install software which turns out to be malicious. The attackers are always looking for the weakest links.It could be weak passwords, careless users, unpatched servers, or vulnerable applications …. The attacks cannot always be detected, as because these criminals want money not the notoriety. So many users do not even know that their crucial informations are stolen. According to Beijing Rising International Software Co. Lit "China Virus and Internet Security Report in the first half year 2007", Some more new kind of virus attacks has reached to 133717 numbers in the first six month of 2007 (presented in Figure 2). These attacks, compared with the same period in 2006, is an increase by 11.9%. More than 35 million computers had suffered attacked in 2006. Information Security has become a serious problem to internet users and their trust on the network system weathering away. According to CNNIC report, only 29.2% of net users feel safe when they are online. The 2006 E-Crime Watch Survey from CSO Magazine reveals Insider Threats are on rise. The most effective e-Crime fighting technologies include State full firewalls (87 percent), electronic access or control systems (86 percent), password complexity (80 percent), network-based anti-virus (74 percent) and encryption (74 percent). It estimated that the junk mail bring at least $ 1.0 billion loss in China each year. As per CISA report, loss due to spams is approx $58 billion

♦ **Information security is by no means just technology**

Of all the scary things that can sabotage a network, human beings are by far the deadliest. In fact, a survey has showed that "*up to 60% of security problems are caused by PEOPLE"* and it is widely accepted that the vast majority of security breaches are a result of human errors rather than technological flaws. In both the public and private sectors, information security challenges must be met with a combination of factors, namely: Technology, People, Processes and Policy or Legislations. No amount of technology can reduce the human element in protecting information.

Figure 2: The first half year from 2004-2007 new virus

From : *Beijing Rising International Software Co. Lit*



♦ **Netizen's unawareness of security risk:**

The OECD Guidelines for the Security of Information Systems and Networks state that *"Awareness of the risks and available safeguards is the first line of defense for the security of information systems and networks"*. But, the fact that netizens are unaware of their information security risk is quite common around the world.

- According to "*China Personal Online Security Research Report 2007*", the individual net security threats are mainly from virus and malware, including spyware, adware, hijacker and Trojan. Users are more than often infected with these virus, spyware, etc.
- Most users mainly use outdated versions of anti-virus and fire wall. They seldom use an updated and original effective antivirus system. Seldom have they paid attention to regular up-dating of their anti-virus software. In 2005-2006, amongst the whole e-Security events, over 73% security failure were due to the leak of protection or unpatching/un-updated software.
- Users, seldom change their password. When asked, 34.50% net users admitted they have no password protection. 13.61% users used password but the password were stolen. Only 39.59% users update their anti-virus software each day, and 25.84% update each week. Moreover very simple /weak selection of passwords or storing them in another file or noting them somewhere which can be accessed by others or sharing passwords with other colleagues/friends have caused several security breaches.
- Many users (mostly fresh/young users, fresh earners) before being fully aware of basic security of online banking, online shopping, e-Commerce or Credit card fraud barge on to using them.
- Net friendships and chat room problems (net friend is real or not, esp. for young people) lead to sharing of personal information to unknowns

♦ **Information Security has become a serious threat: loss of trust on government**

- With the ever increasing incidence of net crimes there is no doubt of loss of trust on government as these netizens carry an unprotected feeling. e-Governance and its many benefits would remain a far away realization in such a case. Among China's 162 million Internet users only 24.69% are using online banking. 61% users still doubt the security of online banking.
- Internet market has met with dilemma now: While on one hand the internet market is expanding stupendously, on the other hand internet crimes quickly overspreading. Information security has become a serious threat. The internet market seems to have failed in balancing the growth and security. According to public choice, the government should avoid any regulatory interventions unless there is evidence of a significant market failure and that the actions of government could remedy efficiently.

Market failures as a rationale for government intervention

- **Public choice.**

Public choice is the economic study of non-market decision making, especially the application of economic data and its analysis in public policy making. Public choice theory recognizes that government must perform certain functions that the marketplace is unable to handle; i.e. it must remedy "market failures". The concept of market failure was originally presented by economists as a normative explanation of why the need for government expenditures might arise. Gradually, the concept has taken on to a form of a full-scale diagnostic tool frequently employed by policy analysts to determine the exact scope and nature of government intervention. Public choice theory is based on the market failure.

- **Market failure.**

A **market failure** can be defined as "the inability of a market system to provide certain goods and/or services either wholly or at the most desirable or at 'optimal' level." It has been identified as an underpinning rationale for public funding to support social economy development, in particular in terms of the provision of National Aid. Market failure is an imperfection in the market mechanism that prevents optimal outcomes. Externalities and Public goods are the two main forms of market failures: **Externalities.** Externalities occur when one person's actions affect another person's well-being and the relevant costs and benefits are not reflected in market prices. An externality arises as an unintended by-product of behavior of individual or firms. Consumption externalities occur when the consumption of an individual imposes costs or benefits on other individuals that are not accurately transmitted through a market. Externalities are both positive and negative externalities. A negative externality arises when one person's actions harm another. Most often the individual or the firm does not have an economic incentive to minimize the negative external costs. Just because competitive markets are inefficient when externalities are present, governments often take policy decisions in an attempt to correct, or internalize the externalities.

Public goods occur when the market cannot provide public goods, because their costs exceed their value to any single buyer, and a single buyer would not be in a position to keep non-buyers from using it. Public goods are neither excludable nor rival. That is when people can not be prevented from using a public good and one person's enjoyment of a public good does not reduce another person's enjoyment of it. Public Goods could not be provided by the free market because of their two main characteristics:

Non-excludability where it is not possible to provide a good or service to one person without it being made available for others to enjoy; **Non-rivalry** where the consumption of a good or service by one person will not prevent others from enjoying it. Thus, the government has to provide public goods and public service. The government provides public goods when the private market can not produce an efficient quantity on its own. Internment security service was regarded as kind of public goods. The fact has shown the internet market itself could not provide efficient service, it needed the government to intervene.

- **Government intervention.**

Market failure establishes the basis for government intervention. Government intervention occurs when markets are not working optimally. Both externalities and public goods are recognized as market failures and a justification for government intervention. Government would intervene either by regulating the activities that produce externalities or by supplying goods that the market fails to supply.

Evidence shows that there are also failures in Internet market. Since the internet information security disasters appeared and they are expending consistently, more and more people wonder whether market forces can solve the security problems. Throughout the 90s, we as a society delegated public safety and national security to market forces. But they're not designed to do it well. Thus the government has to use control policies or regulations or even laws designed to enhance the information security.

"Regulation" refers to two types of legislation - Acts and Regulations:

Acts of Parliament (or Statutes) are the laws made by Parliament. They describe what types of things can and can't be undertaken in a country, and what types of actions will be seen as breaking that law,what requirements need to be met for approved activities and what are the penalties for such offences .

Regulations (or Statutory Regulations) generally deal with more technical details associated with Acts (such as forms, fees and administrative procedures), and are usually easier to change than Acts.

Due to the issue of information security becoming more and more crucial everyday, governments of different countries have adopted different methods to protect their information security.

How other countries do? (Policies, Acts and e-Security awareness programs)

Developed nations like US or countries of EU or Australia have brought in many new acts, policies and have carried out several regular amendment of existing acts to counter the ever increasing and ever changing threats to information security. The aim is to make sure that cyber crimes are minimized and cyber criminals are punished like any other criminals. In addition to technological research, development and implementation of preventive security measures, the policy making bodies are working at sealing the loop wholes those exist in the system through law making/amendment and establishing greater co-ordination between various departments through an integrated approach to information security and also establishing co-ordination with other nations to fight the cyber criminals across boarder.

- **Major Information Security and Privacy Laws of US:**

CFAA (Computer Fraud and Abuse Act of 1984) and **DMCA (**The Digital Millennium Copyright Act) imposes both civil and criminal liability for…. . **ECPA (**Electronic Communications Protection Act) prohibits the unauthorized and unjustified interception, disclosure, or use of communications, including electronic communications. **The Computer Security Act** aims at to protect Federal computer systems. **EEA (**The Economic Espionage Act) addresses the threat by authorizing the imposition of sentences for those who commit or conspire to commit thefts of trade secrets prohibited by the act with clear punishments/penalization. Some other Acts, like **FISMA**(Federal Information Security Management Act), **The USA patriot Act 2001, The Homeland Security Act 2002, Fair and Accurate Credit Transaction Act of 2003,** mainly new acts for information security programs which were acts/ammendments made responding to the September 11,2001 attack.

American federal government adopted many other Acts to control the security of private information, which include : **Fair Information Practices 1973, Right to Financial Privacy Act of 1978 (RFPA), Health Insurance Portability and Accountability Act 1996 (HIPAA), The Gramm-Leach-Bliley Act of 1999 (GLBA), Family Education Rights and Privacy Act 1974, Cable Communications Policy Act 0f 1984 (CCPA), Video Privacy Protection Act 1988 (VPPA), Electronic Communication Privacy Act of 1986 (ECPA), Children's Online Privacy Protection Act of 1998 (COPPA).** Barring these there are several other federal acts and many state laws of each states those protect the private information of US citizens**.** More than this, many US states governments have set Cyber Security Awareness Programs**.**:**State of Alabama** set policy which requires that each agency have an awareness and training plan that is approved by each agency head. **Commonwealth of Pennsylvania** conducts an annual "Security Awareness Day" at the state capital to raise awareness among legislators and citizens. **State of Michigan** held a town hall meeting on cyber security and protecting children online in 2007. More cost-effective methods, such as online videos or even conference calls, provide mechanisms for spreading the IT security awareness message. **State of Wisconsin** has conducted awareness-raising presentations to educators, principals, and others on teachers' in-service days. The state is currently examining possibilities for rolling awareness activities into school children's computer classes. **State of Florida** has developed the C-Safe in- person security awareness and training program that can be tailored to individual business' and organization's needs. **National Cyber Security Awareness Month (Every October)** commemorated by the U.S. Congress, supported by major global corporations and coordinated by the National Cyber Security Alliance (NCSA), the month aims to heighten public awareness of the critical role each citizen plays in protecting information assets and implement online safety programs at work, home and school. Moreover, **Kids Safe Online Interactive Play Webcast (Oct. 17), Internet Safety Night (Oct. 23)** and **Computer Security Day (Nov. 30)** are three important national days for the country observed to spread awareness on identity theft and other internet-related security issues.

- **Major Information Security Systems of European Union:**

ENISA (The European Network and Information Security Agency) is a European Union Agency created to advance the functioning of the Internal Market by advising and assisting Member States and the EU bodies to ensure a high and effective level of security. ENISA serves as a centre of expertise that facilitates information exchange and cooperation. ENISA and many Member States have already realized the importance of their role in creating a culture of greater self-awareness of information security. ENISA presently

- Offer an insight into the types of problems currently being faced by countries with regards to information security.
- . Illustrate examples of campaigns and other awareness-raising initiatives that have been run or are planned to run in Member State countries.

- Provide examples of high level non technical messages conveyed in a typical campaign.
- Contribute to the development of an information security culture in Member States by encouraging users to act responsibly and thus operate more securely.

With the theme of a Safer Internet, the Netherlands has celebrated the **Safer Internet Day on February 8th 2005**. One of the central premises was for children to teach adults. The Safer Internet Day took place across Europe and world-wide. It was celebrated by 65 organizations in 30 countries across the world from Australia to Iceland, and Russia to Singapore. Safer Internet Day 2005 features an Internet Magic story-telling contest from 16 countries across Europe.

♦ **What other countries do - Australia**

The Australian Government established the **E-Security National Agenda (ESNA)** in 2001 to create a secure and trusted electronic operating environment for both the public and private sectors. The development of three new priorities to address concerns and to assist in achieving the original objective of ESNA, to:

- reduce the e-security risk to Australian Government information and communications systems
- reduce the e-security risk to Australia's national critical infrastructure, and
- Enhance the protection of home users and SMEs from electronic attacks and fraud.

TheCybercrime Act 2001 (Cth) (**Act**) adds a new part 10.7 to the Criminal Code Act 1995 (Cth). In 2006, the Attorney-General, the Minister for Communications, Information Technology and the Arts, the Minister for Defense and the Special Minister for State announced a review of the ESNA to ensure that Australia's policy and operational framework continues to be responsive to the changing e-Security environment. A new whole-of-government interdepartmental committee, the e-Security Policy and Coordination (ESPaC) Committee, is being established to coordinate e-Security policy throughout the Australian Government. The Department of Communications, Information Technology and the Arts (DCITA) aims to raise the awareness of home users and SMEs by providing them with the knowledge and skills to improve their computer defense and their online behaviors.

Hong Kong Clean PC Day (Nov. 28) Jointly organized by the Office of the Government Chief Information Officer (OGCIO), China Hong Kong Computer Emergency Response Team Coordination Centre (HKCERT) and the Hong Kong Police Force (HKPF), It aims to achieve the broadest penetration of a "Three Smart Tips" campaign through various promotional activities and materials to bring internet safety and security awareness.

New Zealand Global Security Week (the week leading up to September 11th each year) is an opportunity to join forces with other security professionals worldwide and help raise awareness of security issues and techniques.

From above analysis, we found that the government of the countries starting from United States to EU, from Australia to Hong Kong, have paid a much higher attention to information security laws while raising their citizen's awareness on internet security.

What are China's existing policies on personal information security

In China, the governmental bodies at various levels and at various times have repeatedly pronounced that the Internet and networking services will be legally deployed. So far, China has made many complex regulations and unclear policies on the Internet security. However, we found that among these existing internet regulations and policies there are not many those can adequately deal with the protection of individual information security and privacy rights. The inadequacy of information security is at various levels which manifest not only in number of policies, but also in contents of these policies.

♦ **No adequate rules or regulations to protect personal information security**

From its first focus on the connection to the Internet, China's government, since long has made efforts to control the Internet security and made enormous strides to shape a regulatory framework for the Internet. China made many rules and regulations, like *Measures for Managing Internet Information Services* (2000); *Provisional Rules for the Administration of the Operation of News Publication Services by Web Sites* (2000); *Rules for the Administration of Internet Bulletin Board System Services* (2000), to aim at the security of the

internet. However, an analysis and result of China's 107 major legislations on information security issued or amended between 2000 to 2007, can be seen in Figure -1

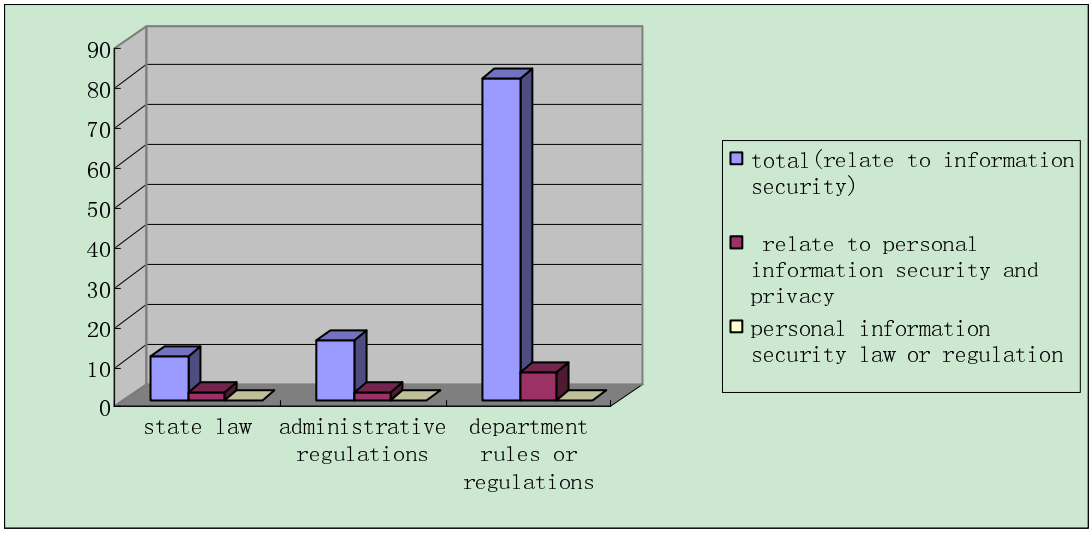


Figure1:

This clearly shows that, laws or regulations interrelated to citizens information security and privacy is a very small part only. What is scarier is that specific laws or regulations primarily aimed at citizen's information security and privacy are completely absent. Policy making and bringing in new legislations or specific amendment of exiting ones on individual information and privacy protection has taken hardly any step forward. This is the one of the many reasons why penalizing a cyber criminal is seldom easy. If a citizen is attacked by hacking or Trojan virus or any other hacker/spam/spayware etc, he/she can't stick up to his/her own rights via existing laws and put up a complaint to any cyber law enforcing authority against those websites spreading such malicious acts. Similarly, if the personal and private information collected by any agency for providing certain services are given away to tele marketers or other organizations who are not authorized to receive such private information, there is no specific law or department through which the individual can take up his/her case. On many occasions, the victims of cyber crimes having found no legal way to protect their privacy and information security, accept it as a fact of life and such crimes seldom gets registered anywhere. So, such incidents do not come into light. The development of E-commerce is also crying for laws or regulation on citizen's private information security. Similar situation or problems are observed in other fields such as telecommunication and finance

♦ **Regulations or laws issued many years ago unable to challenge today's cyber threats.**

Most of the laws or regulations interrelated to citizen information security were issued many years ago. Its interrelated structure can't meet the need of new age problem with many different and emerging threat perceptions. Almost all of them need amendment keeping in view the fast technological changes happening in China and around the world. Laws such as General Rule of Civil Law or Criminal Law in China which involve articles related to citizen information security and individual privacy are very difficult in their application to a cyber crime and fix responsibility of crime. In a age where where the cyber crimes have advanced far ahead with the changing technologies and the Laws have remained static.

With the development and usage of internet in our daily lives, the demand involving flow of individual private information has changed greatly. Where as China's constitution presently prescribes definite right to citizen's icon , right of reputation and so on, but it has limited and specific personal privacy information protection laws. There are hardly any articles about the cyber crime in existing *Criminal Law,* except only two terms related with computer crimes. This cannot cover the widespread and large verities/types of net crime. New and newer threats are emerging with technological changes. Many existing laws which relates to individual information or privacy protection, but were issued many years ago must be amended to address the new type of crimes and methodologies used to commit such crimes in today's Hi-tech information technology age. To stop/limit the occurrence of cyber crime, it's an absolute necessity now to have specific legislations with clearly specified deterrence (punishment and/or penalization). Then only the laws can

protect citizen's individual information effectively. In this study (presented in Figure 2), it can be observed that only one of the laws is revised between 2000 to 2007 and no other rules of law or regulation which were revised between this period are related or interrelated to personal information security..

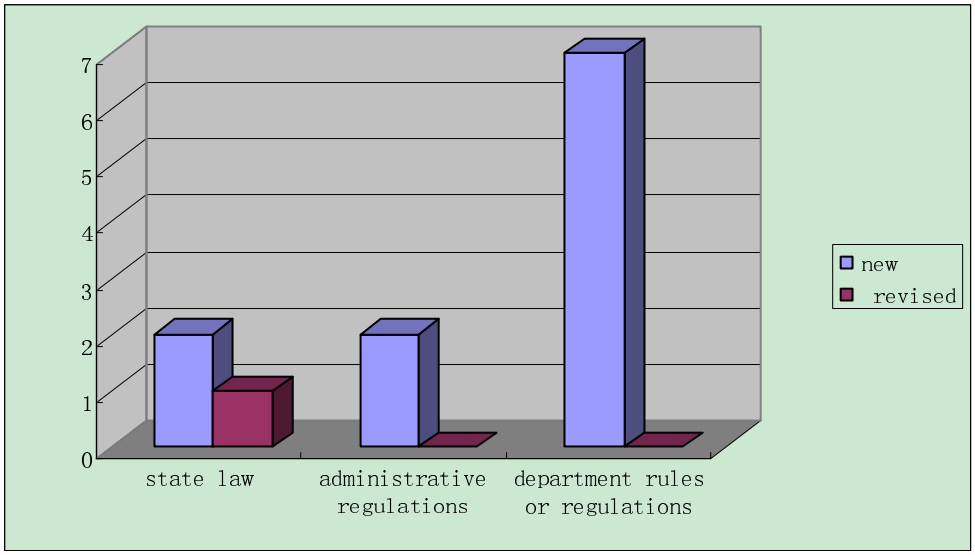


Figure2:

It has been observed that in quite a few cases the words or expressions in existing legislations on individual information privacy are quite vague and unclear i.e. they are either having more than one meaning or impracticable to apply and prove or redundant to this age. For example, one of the statutes in "*Regulation of the People's Republic of China for Safety Protection of Computer Information Systems*" says that "no units or individual can utilize computer information system to be engaged in activities that harm the benefit of country and collectivity and personal legitimate right or endanger security of computer information system". An Internet user has no way of knowing what topics might be considered as injurious. Such generalized legislations can sometimes punish an innocent for a very minor incident which he/she has done it without any intention of committing a crime. At the same time it can sentence another person much liberally of a very serious offence. Any such vague and generalized expression must be amended with specific statements of what's illegal. A law is most effective when it's clearly understood by the people. It's the government's responsibility to takes steps to make its people aware of the laws, the punishment they carry and remove any ambiguity and vagueness for uniform application of these laws. Sometimes, crimes are committed since people/organizations do not know what's legal or illegal, specifically applicable to cyber crime & information and privacy protection

♦ **Lack of integrated approach in preparation of policies and legislations:**

Until now, there has been little systematic and integrated approach to formation and enactment of private information protection laws in China. Many policies and regulations are made by different administrative departments. Besides the Ministry of Public Security, the other bodies, which hold authority to enact regulations or rules, include the State Information Office (in charge of online news), Ministry of Culture (in charge of Internet Cafe), Ministry of Information Industry (In charge of connection and general administration of internet), State Administration for Industry and Commerce (in charge of the internet companies registration and online advertising), the State Administration for the Protection of Secrets (in charge of State Secrecy and Encryption), the State Administration for Press and Publications (in charge of online publications), Ministry of State Security (in charge of national security) and the State Administration of Radio, Film and Television (in charge of online video programs). That's where policies or regulations about information security from different departments often conflict with each other. Moreover these rules grant various government authorities full power to monitor organizations and individuals on the Internet/network systems often inflicting on the privacy rights of individuals. So, it will not be out of place to state, that while every organization regulates or monitors the network information, there is seldom any protection available to individual's information security.

More than this, the local government, local administrative department also promulgates laws or regulations. Conflicts also occur between the central government and local government rules, between the central department and the local department, between the different administrative departments. There are few common or uniform agreements across the country. As for lacking of a centralised monitoring and evaluating institution for information policy making, it has spurred violent and equal criticism from the scholars and the policy analysts as well.

- **Lack of national campaign or training plans to raise citizen e-Security Awareness**

Unlike Unite States, or EU, or Australia, there are no wide and specific campaigns or training plans by the central, state or local governments to raise citizen e-Security Awareness. Many of the awareness and training programmes in China are just limited to a handful of civil servants and employees in big companies. Most of the IT companies or anti-virus software companies do their best to raise the user's awareness about the threat of the internet for the need of their products selling, which are more of a marketing gimmick than raising public awareness on the security aspect. There are no systematic education or special training/lessons in universities or in schools for young users those who contribute to more than half of China's internet user's population. No special security slogan from mass media or communication and advertisement organizations, all of which have been proved effective in raising awareness in other developed countries. Realizing the criticality of cyber security awareness, developed countries like US have gone ahead in organizing specific and regular programs for raising safety and awareness among the children.

Conclusion

Considering various facts/figures collected and presented in this paper and comparing China's present information security and awareness policies with developed nations, it can be stated that, China's existing information security policy is inadequate to protect its Netizen's personal rights. China has a much longer way to go in creating adequate policies, legislations and e-Security awareness programs, which will provide its netizens a secured, trusted internet/network system where their information privacy rights are not violated, where China's Netizen Community are made well aware of the security threats, legal and illegal actions pertaining to use of internet/network systems, so that they behave like any other responsible and well aware citizens of developed nations. The fact that in coming years China will have the highest number of internet users in the world, it's high time for the central, state and local governments of China to review all existing policies pertaining to information security and privacy rights by setting up a centralized department with experts from public & private sectors, law & policy making bodies to combat the impending information security threat and protect privacy of its Netizens.

References

- Australian Computer Crime and Security Survey, 2006, http://www.auscert.org.au/render.html?it=2001
- Brian Dollery and Joe Wallis (2001) The Theory of Market Failure and Policy Making in Contemporary Local Government, http://www.une.edu.au/febl/EconStud/wps.htm
- Bian xiang-xu (2004). The Development of Information Policies in America. Journal of Jilin Province Economic Management Cadre College, 18(3), pp.65-67
- Cybercrime Act 2001, http://scaleplus.law.gov.au/html/comact/11/6458/pdf/161of2001.pdf
- DanielJ.Solove,George, A brief History of Information Privacy Law, http://papers.ssrn.com/sol3/papers.cfm?abstract_id=914271
- ENISA(European Network and Information Security Agency)(2005):Raising Awareness in Information Security–Insight and Guidance for Member States
- ENISA (European Network and Information Security Agency(2007)):Information security awareness initiatives report: current practices and measurement of awareness
- Jiang chao-hui and Xu qing-shi (2007). The Status Quo, Problem and Countermeasure of Information Management in China, http://www.ececec.cn/html/zw/20070426/217.html
- Report on Survey of China Information Net Security and Computers Virus. Network and Computer Security，2006 , http://www.11cn.net/article.asp?id=229&page=27

- Report On China Computer Virus Epidemic Situation & Net Security in The First Half Year of 2007. http://tech.163.com/special/000915RB/2007report.html
- Steven Robinson, US Information security Law(Part1,2,3,4), http://www.securityfocus.com/infocus/1669
- The Twentieth Stat. Report of Net Development Status in China (2007). http://9.douban.com/site/entry/22669509/
- Wang yi-qun. The Management the Factors of Person, Machine and Environment in Net Information Security. China Science and Technology Information, 2006,7, pp.195-197
- Xin Hua Net (2007). Information Security Incident Have Been Increasing for Later Three Years in Chin a. http://news.xinhuanet.com/newscenter/2007-09/25/content_6791381.htm
- Zhang xi-jie (2007), Analysis on Strengthen Network Information Security, http://www.topoint.com.cn/Html/wenku/xxaq/8281501110720_2.html
- Zhang cai-xia (2007). Personal Information Security Expecting for Protection of Laws in Net Age. http://www.honor365.com/Html/special/inforsec/inforsec_5/2007-1/3/165652482.html

Sequencing Clusters of Spatial Join Operations Using Weighted Match

Jitian Xiao

School of Computer and Information Science, Edith Cowan University,
2 Bradford Street, Mount Lawley, WA 6050, Australia
Email: j.xiao@ecu.edu.au

Abstract

Spatial join queries in spatial databases usually access a large number of spatial data. As spatial objects can be very large in size, they are usually stored in secondary storage, such as disks. To process a spatial join operation, the referred objects need to be fetched into the main memory for processing. The I/O cost can be very high for a single spatial join operation. The I/O cost can be reduced by clustering joinable spatial objects and then scheduling the join operations such that the number of times the same objects to be fetched into memory can be minimized. One of the key issues behind this approach is how to produce a good sequence of clusters to guide the join operation scheduling. In our previous work, we proposed a cluster scheduling method to reduce the I/O cost in spatial join processing in (Xiao et al, 2000). This paper presents a new efficient algorithm that generates better cluster sequence than the previous algorithm does in the sense that more fetching time used for fetching those overlapping objects of clusters can be saved. Experiments have been conducted to demonstrate the saving of I/O cost in spatial join processing by using the cluster sequences generated by the new method to guide the scheduling of clustered spatial join operations.

Keywords: Spatial databases, Spatial join processing, Scheduling, Match.

INTRODUCTION

Spatial objects can be very large in size (Abel *et al*, 1995). For instance, a large polygon object in a road map may have up to tens of thousands of edges that may need several megabytes of storage (Ooi, B. C., 1990). In large spatial databases, objects are stored in secondary storage, such as disks. To process a spatial operation, the referred objects need to be fetched into the main memory for processing.

Spatial join queries in spatial databases usually access a large number of spatial data. They are the common spatial query type that requires a high processing cost due to the large volume of spatial data and the computation-intensive spatial operations. To reduce the CPU and I/O costs for spatial join processing, most spatial join processing methods are performed in two steps, i.e., *filter-and-refine* approach (Abel, 1989). The first step chooses pairs of data that are likely to satisfy the join predicate. The second step examines the predicate satisfaction for all those pairs of data passing through the filtering step. During the filtering step, a conservative approximation of each spatial object is used to eliminate objects that cannot contribute to the join result, and a *weaker* condition for the spatial predicate is applied on the approximations. This step produces a list of *candidates* that is a superset of the joinable candidates. These candidates are usually represented as pairs of object identifiers. All candidates are then checked in the refinement step by applying the spatial operation on the full descriptions of the spatial objects to eliminate the "false drops". The join cost can be reduced this way because the weaker condition is usually computationally less expensive to evaluate and the approximations are small in size than the full geometry of spatial objects. Generally, the refinement cost consists of two parts: one is the cost for fetching objects from the database, and the other is the cost for checking spatial relationship of pairs of objects using computational geometry algorithm. The former cost dominates the refinement cost when the spatial objects are small (e.g., tens to hundreds of

vertices), and the later cost dominates for large spatial objects. In this paper, we focus on reduction of former cost in the refinement step.

A spatial join operation may involve many (large) objects which can not be all fetched into the main memory at the same time to complete the join. In such a case, the join is divided into many sub-join, each joins a part of joinable objects. To reduce the I/O cost in the refinement step, researchers proposed two-phase join strategy (Abel *et al*, 1995, Xiao *et al*, 2000), i.e., *clustering* candidate objects into clusters and then *joining* these clusters pairwisely. The clustering phase tries to cluster spatial objects such that they join as many other objects as possible within their cluster and join as few objects as possible across clusters (Xiao *et al* 2000, Zhou *et al* 1998). In the joining phase, these clusters are scheduled in a sequence such that a maximum number of overlapping objects between consecutive clusters can be reused in the memory when processing next cluster (i.e., the overlapping objects do not need to be fetched into memory again because they are already there). In this way, a significant reduction on disk accesses has been achieved and demonstrated through simulations.

We proposed a method to generate the scheduling sequence of spatial (object) clusters in (Xiao *et al*, 2000). The method maps cluster scheduling tasks to a weighted graph where nodes represent spatial clusters and the edges represent the object overlapping status between clusters. The weight of each edge is the total size of the overlapping objects between the two clusters the edge linked to. A cluster scheduling sequence is then generated over the graph so that the total weight of overlapping objects reaches the maximal. The issue of scheduling clusters was discussed in three steps (Xiao *et al*, 2000). Firstly, a concept, i.e., *maximum overlapping (MO) order*, was defined in a graph model. Secondly, it was proved that the problem of finding an MO order in an arbitrary graph is *NP*-complete. And finally, a heuristic was developed to produce an approximation to the MO (AMO) order for an arbitrary graph. The AMO order was used for scheduling clusters in the spatial join processing. The cluster sequencing method proposed in (Xiao *et al*, 2000) was late improved using a match based algorithm (Xiao, 2003).

This paper is to improve the above algorithms by using a new version of weighted matching approach (Cook and Rohe, 1999). We will prove that the new algorithm generates better cluster scheduling sequence in the sense that the total overlapping weight of generated AMO order is generally higher than the one generated by the method proposed in (Xiao *et al*, 2000), and the complexity of the new algorithm is lower than that of (Xiao, 2003) while keeping the same quality of AMO order generated by it.

The rest of the paper is organized as follows: Section 2 reviews our previous work by formalizing the problem using a graph model. Some concepts and properties about MO order are also reviewed in the section. In Section 3, the new algorithm is described. An analysis to the algorithm follows and the performance of the algorithm is evaluated through examples. Experimental results are presented in Section 4. And finally, Section 5 concludes the paper.

OUR PREVIOUS WORK

Problem Definition

Suppose, for a given spatial join operation, the candidate set has been produced in the filtering step, and the candidate objects have been clustered in the clustering phase.

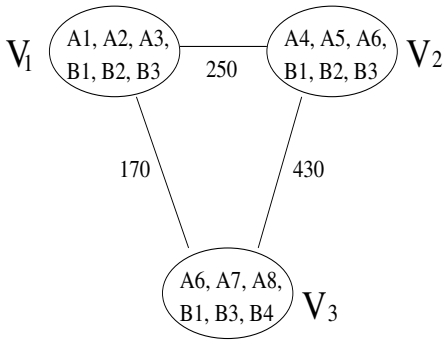
Let $\mathbb{V} = \{v_1, v_2, ..., v_k\}$ be the set of spatial objects referenced in the *candidate* set, and $V_1, V_2, \ldots, V_n$ the clusters of $\mathbb{V}$. For each i ($1 \le i \le n$), $V_i = \{ v_{i_1}, v_{i_2}, ..., v_{i_m} \}$ ($m \ge 1$), $v_{i_j} \in \mathbb{V}$ ($1 \le j \le m$). That is, $\bigcup_{i=1}^{n} V_i = \mathbb{V}$ and $V_i \neq \phi$ for each i ($1 \le i \le n$). For convenience, we define *size*(V_i) as the sum of the sizes of objects in V_i, i.e., $size(V_i) = \sum_{v \in V_i} s(v)$ where $s(v)$ is the size of object v.

We introduce a weighted graph $G = (V, E, w)$, upon $\mathbb{V}$, called *cluster overlapping* (*CO*) graph, to represent the overlapping relationships between data clusters. The node set $V = \{V_1, V_2, ..., V_n\}$ is a set of

clusters, and the edge set E is defined as: for each node pair V_i and V_j ($i \neq j$), there is an edge $E_{ij} = (V_i, V_j)$ if $w(V_i, V_j) = size(V_i \cap V_j) \neq 0$. Here $w(V_i, V_j)$, also denoted as $w(E_{ij})$, is the weight of edge E_{ij}. As an example, let the spatial object set involved in a given spatial join operation be $\mathbb{V}$ = {A1, A2, A3, A4, A5, A6, A7, A8, B1, B2, B3, B4, B4}, the candidate set be F = {(A1, B1), (A2, B1), (A3, B2), (A3, B3), (A4, B3), (A5, B1), (A6, B2), (A6, B4), (A7, B1), (A8, B3), (A8, B4)}, and its three clusters be V_1 ={(A1, B1), (A2, B1), (A3, B2), (A3, B3)}, V_2 ={ (A4, B3), (A5, B1), (A6, B2)} and V_3={ (A6, B4), (A7, B1), (A8, B3), (A8, B4)}. Based on the object sizes given in Figure 1 (a), Figure 1 (b) shows the CO graph corresponding to the above clusters.

oid	obj size
A1	200
A2	80
A3	40
A4	30
A5	32
A6	260
A7	18
A8	60
B1	60
B2	80
B3	110
B4	76

(a) Object size



(b) cluster overlapping graph

Figure 1. An example of CO graph

At refinement step, if the object clusters are joined in the sequence of $V_1, V_2, \ldots, V_n$ (i.e., no scheduling), then the total I/O cost is:

$$C_{I/O} = \sum_{i=1}^{n} size(V_i) - \sum_{i=1}^{n-1} size(V_i \cap V_j) \tag{1}$$

When processing cluster V_{i+1}, objects in $V_i \cap V_{i+1}$ are already in memory just after processing V_i. There is no need to load these objects again. Generally, for a schedule π which determines the processing sequence of $V_1, V_2, \ldots, V_n$ as $V_{\pi_1}, V_{\pi_2}, \ldots, V_{\pi_n}$, where $V_{\pi_i} \in V$ and $V_{\pi_i} \neq V_{\pi_j}$ for $i \neq j$, $1 \leq i, j \leq n$, the I/O cost for schedule π is

$$C^{\pi}_{I/O} = \sum_{i=1}^{n} size(V_{\pi_i}) - \sum_{i=1}^{n-1} size(V_{\pi_i} \cap V_{\pi_{i+1}}) \tag{2}$$

When the clusters are given, $\sum_{i=1}^{n} size(V_{\pi_i})$ is a constant. Let y be:

$$y = \sum_{i=1}^{n-1} size(V_{\pi_i} \cap V_{\pi_{i+1}}) \tag{3}$$

The goal of cluster sequencing is to find a schedule π such that y is maximized, which is the case that $C^{\pi}_{I/O}$ is minimized.

Maximum overlapping order

The concept of *maximum overlapping (MO) order* was introduced in (Xiao *et al*, 2000) to recognize better schedules. Given a CO graph $G = (V, E, w)$ with $V = \{V_1, V_2, \ldots, V_n\}$, an MO order among sets $V_1, V_2,$

..., V_n is a sequence ($V_{i_1}, V_{i_2}, ..., V_{i_n}$) such that $\sum_{l=1}^{n-1} size(V_{i_l} \cap V_{i_{l+1}})$ reaches the maximum among all permutations of *V*. In other words, an MO order in a CO graph *G* is a permutation of nodes in *G* such that the total size of overlapping objects between adjacent nodes reaches the maximum. For example, (V_1, V_2, V_3) is an MO order in the CO graph in Figure 1 (a), and the total size of overlapping objects between adjacent nodes in the order is 680.

A simplest algorithm to find an MO order is to check all permutations of *V* to see which one makes the $\max\{\sum_{l}^{n-1} size(V_{i_l} \cap V_{i_{l+1}})\}$. The complexity of such a method clearly has factorial order and is certainly not practical.

Although an MO order exists for each CO graph *G*, it is impossible to find an MO order in polynomial time. However, the task of finding an MO order can be reduced to the case where *G* is a connected graph. In fact, the following facts can be proved (Xiao *et al*, 2000).

- *The problem of finding an MO order in a CO graph is NP-complete*.
- *Let G be a CO graph. If each component G_i of G gets an MO order $V_{i_1}, V_{i_2}, ..., V_{i_{m_i}}$ (i = 1, 2, ..., p), then the order $V_{1_1}, V_{1_2}, ..., V_{1_{m1}}, ..., V_{p_1}, V_{p_2}, ..., V_{p_{m_p}}$ is an MO order of G.*

Finding an Approximation of MO Order in G

A maximum spanning tree (MST) based algorithm was developed in (Xiao *et al*, 2000) to produce an approximation to MO order (*AMO* order) of relative 'high' overlapping weights in the sense that the weight of the AMO order produced by the algorithm is always greater than or equal to half the weight of an optimal MO order. The algorithm consists of three steps: The first step produces a maximum spanning tree T of the CO graph G; the second step conducts a *depth-first search* (DFS) on T and, in the third step, an AMO order is built, which is the traversal order of the DFS on T. The complexity of the algorithm is $O(m^2 \log_2 m)$, where $m = max(|V|, |E|)$.

However, the quality of the produced AMO order is sensitive to the selection of the starting cluster in the algorithm. Different starting points may result in different AMO orders that can have significant impact on the spatial join scheduling performance. In addition, the sum of the overlapping weights for a large proportion of produced AMO orders is much less than that of an optimal MO order. As an extension to this work, an efficient algorithm was proposed in (Xiao, 2003). This algorithm is based on Edmond's matching algorithm (Lawler, 1976), and can produce better AMO order. However the complexity of the algorithm was not as good as that of (Xiao *et al*, 2000), i.e., algorithm complexity is $O(n^3)$ where *n* the number of clusters.

THE MAXIMAL MATCH BASED CLUSTER SEQUENCING ALGORITHM

We now improve the algorithm proposed in (Xiao *et al*, 2000) and (Xiao, 2003). The new algorithm not only generates AMO orders better than the ones generated by (Xiao *et al*, 2000), it also has a lower complexity than that of the algorithm in (Xiao, 2003) while keeping the same quality of AMO order generated. For simplicity, we limit our discussion to connected CO graphs. According to (Xiao *et al*, 2000), the algorithm and the related discussion can be easily extended to the case of unconnected CO graphs.

The following terminologies (Lawler, 1976) are frequently used in the rest of this paper.

A *simple path* of a graph is a path in which all nodes are distinct. A *path graph* $G = (V, E)$ with *n* nodes is a graph in which nodes in *V* can be listed as a sequence $v_1, v_2, v_3, ... v_{n-1}, v_n$ such that $(v_1, v_2), (v_2, v_3), ... (v_{n-1}, v_n)$ are the only edges of *E*. A *match* of a graph is a set of edges, in which any two of them are not incident to the same node. A match is *maximal* if any edge in the graph that is not in the match has at least one of its endpoints matched, and the sum of the edge weights of the match is maximal among all matches of the graph. A *weighted matching* (WM) problem is, for a given (edge weighted) graph *G*, to find a match of *G* such that the sum of the edge weights of the match is maximal. The WM problem was solved by Edmonds

(1965) and the complexity of his algorithm is $O(n^2m)$, where n and m denote the number of nodes and edges in the graph respectively. Since then Edmonds' algorithm has been studied by a number of researchers. The fastest implementation of the Edmonds's algorithm is due to Cook and Role (1999) with a time complexity of $O(nm \log n)$. For any graph, these algorithms output a *maximal match* of the graph.

Our new algorithm is maximal match based, and we employ Cook and Role (1999) implementation as a component of the new algorithm.

The Algorithm

The basic idea of our new algorithm is (1) to divide the CO graph into sets of disjoint path graphs such that the sum of the edge weights of the longest paths in the path graphs reaches the maximum, and (2) to link these paths using maximal match among the endpoints of the paths.

The algorithm is a recursive one containing three main steps: In the first step, a maximal match M of G is produced using Cook's *maximum matching* algorithm (Cook and Role, 1999). Each edge, together with its connected nodes, in M is taken as an initial graph path. That is, G is conceptually divided into sets of path graphs, each consists of a pair of matched nodes and the edge connecting them (each unmatched node of G forms a special path graph, i.e., one without an edge). In the second step, the graph G is coarsened by collapsing the matching nodes (or, endpoints of the longest paths in path graphs). At this step, each pair of matching nodes (or, endpoints of the path in individual path graph) are combined to form a single node of the next level coarser graph $G'= (V', E', w')$. Nodes in V' are all in the form of either $v = \{v_i, v_j\}$, where $v_i, v_j \in V$ are matched in M (i.e., $(v_i, v_j) \in M$), or $v = \{v_i\}$, where v_i is a unmatched node of M.

Intuitively, each node $v=\{v_i, v_j\} \in V$ represents a path graph in G where v_i and v_j are the endpoints of its longest path. A node v of form $\{v_i, v_j\}$ in V' is referred to as a *t-node*, and a node v of form $\{v_i\}$ is referred to an *s-node*. A *multinode* can be either a t-nod or an s-node.

E' and w' are then defined such that the edge between any pair of multinodes v' and v'' corresponds to an edge in E whose two endpoints are in v' and v'', respectively, and whose weight is maximal among all edges connecting nodes in between the multinodes v' and v'' (i.e., the endpoints of the path graph represented by $\{v_i, v_j\}$), if such an edge exists.

After graph G' is built, Cook's maximal matching algorithm is applied to G' again to produce a maximal match M'. By this point, the next level of coarser graph $G'' = (V'', E'', w'')$ can be built following the same procedure (as described above). The above matching and collapsing process continues until no further matching can be found.

At this point, each t-node $\{v_x, v_y\}$ in the last coarser graph can be stretched to a path graph, with the two endpoints as v_x and v_y, respectively. The AMO order is produced by printing nodes in all path graphs (i.e., the nodes in each path graph are listed in an order in its longest path), one after another.

In summary, we conceptually take a pair of matching nodes and the edge between them as a path graph at the end of first round of matching and collapsing process (i.e., each with a path of length 1 or 0), then from the second round of matching and collapsing process on, the longest paths in these path graphs are concatenated pairwisely using edges of maximal weight between the endpoints of the paths. With the matching and collapsing process going on, paths are linked using the maximal matching on levels of coarser graphs until a set of disconnected path graphs is reached. At this stage, a sequence of nodes of the longest path for each path graph was output. Any order of these sequences can be taken as an AMO, because the produced path graphs are disjoint with each other, and each node of the original CO graph belongs to one and only one path graph.

The algorithm can be formally described using pseudo-code:

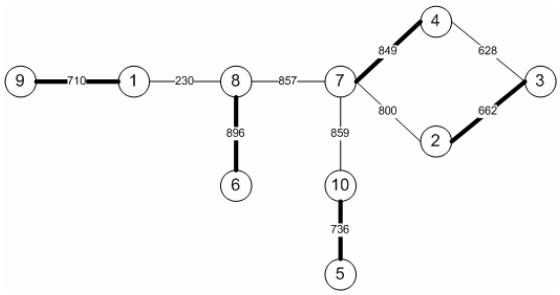
Algorithm *MaxMatchBasedAMO*(G)

Input: $G = (V, E, w)$; // A CO graph with $V = \{V_1, V_2, ..., V_n\}$.

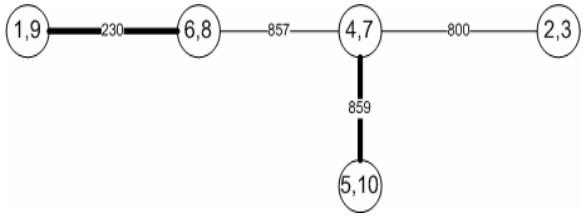
Output: $V_{i_1}, V_{i_2}, ..., V_{i_n}$; // AMO order of *G, a permutation of nodes in V.*

{
[1] Find a maximal match *M* of *G* using *Cook*'s algorithm; // see (Cook and Role, 1999)
[2] ***if*** no matching was found
[3] {**For** each isolated multimode
[4] output its nodes in the order in its longest path; //output the AMO order of one path graph
[5] **Return**};
[6] Coarsen *G* by collapsing matching nodes of *M* to produce a coarser graph *G'*;
[7] *MaxMatchBasedAMO* (*G'*);
[8] ***return***;}

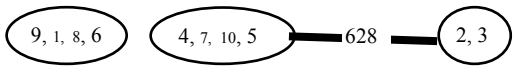
Figure 2 shows how the AMO of a given CO graph is produced step by step.



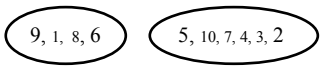
(a) The CO graph G with its first round of matching



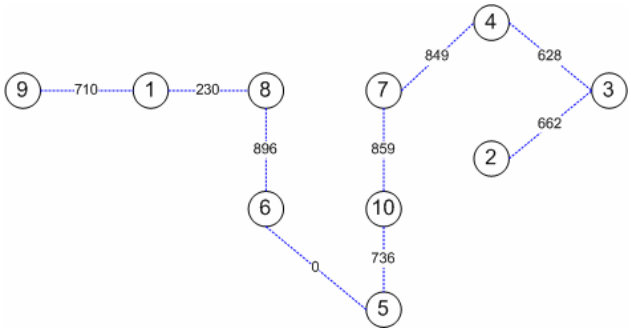
(b) The coarser graph G' and its matching



(c) Next level coarser graph and its matching



(d) No further matches can be found



(e) An AMO order produced (with a total overlapping weight of 5570)

Figure2: Execution of the MBM algorithm

Algorithm analysis

Now we analyse the complexity of the algorithm *MaxMatchBasedAMO*. Suppose that the CO graph have *n'* nodes and *m'* edges. For ease of analysis, let $n=\max\{n', m'\}$. Then line 1 needs $O(n'm' \log n') \le O(n^2 \log n)$ running time (Cook and Role 1999). Lines 3~5 would be executed once only (i.e., when no more matching can be found) and needs at most $O(n^2)$ time as it scans at most once for each node to find one of the endpoints in each path graph, and then use constant time to find each one linked node after that. Lines 6 has

the complexity of $O(n^2)$ because, for each matched node, it needs no more than once scanning to combine to its matched one to form a multinode of the next level coarser graph. So the total complexity of lines 1~6 is limited by $O(n^2 \log n)$. Line 7 completes the recursive execution of the algorithm.

Let $g(n)$ be the complexity function of the algorithm. We analyse both the best case and the worst case. First, consider the best case of the execution (i.e., all nodes in G were matched in step [1]). Denote $f(n)$ as the complexity of this case. For an ideal matching, each node is matched, thus the next level of coarser graph has $n/2$ nodes. In this case, we can get a recurrence formula for the complexity function as

$$f(n) = \begin{cases} 0 & \text{if } n \le 1 \\ \alpha \cdot n^2 \log n & \text{if } n > 1 \text{ and no further match exists} \\ \alpha \cdot n^2 \log n + f(n/2) & \text{if } n > 1 \text{ and further match exists} \end{cases}$$

where α is a constant. As the collapsing reduces half the number of available (multi)nodes, either $n \le 1$ or ($n > 1$ *and no further match exists*) will become true after running the algorithm recursively for some rounds. Therefore, for a large n, according to the recurrent property of $f(n)$, we have

$$\begin{aligned} f(n) &= \alpha \cdot n^2 \log n + f(n/2) = \alpha \cdot n^2 \log n + \alpha \cdot (n/2)^2 \log (n/2) + f(n/4) \\ &= \ldots = \alpha \cdot (n^2 \log n + (n/2)^2 \log (n/2) + (n/4)^2 \log (n/4) + \ldots + 1) + f(1) \end{aligned}$$

This equation is valid for any n that is a power of 2, say $n = 2k$. Thus, we have

$$\begin{aligned} f(n) &= \alpha n^2 \cdot (\log n + 1/2^2 (\log n - \log 2) + 1/2^{2 \cdot 2} (\log n - \log(2^2)) + 1/2^{3 \cdot 2} (\log n - \log(2^3)) + \ldots + 1/2^{(2k-1)} (\log n - \log(2^{k-1}))) + f(1) \\ &= \alpha n^2 \log n \cdot (1 + 1/2^2 + 1/2^{2 \cdot 2} + 1/2^{3 \cdot 2} + \ldots + 1/2^{2(k-1)}) + f(1) - \alpha n^2 \log 2 \cdot (1/2^2 + 2/2^{2 \cdot 2} + 3/2^{3 \cdot 2} + \ldots + (k-1)/2^{2(k-1)}) \end{aligned}$$

Recalling that $f(1) = 0$, we get $f(n) = 4/3 \cdot \alpha n^2 \log n - \alpha\beta n^2 \log 2$, where α *and* β are constants, or

$$f(n) = O(n^2 \log n) \tag{4}$$

If n is not a power of 2, there must exist k such that $2k < n \le 2(k+1)$. Therefore, we have $4/3 \cdot \alpha n^2 \log n - \alpha\beta n^2 \log 2 \le f(n) \le 4/3 \cdot 4\alpha n^2 \log n$, which still leads to formula (4).

Secondly, consider the case where some unmatched (multi)nodes produced in step [1] in the algorithm. Denote $F(n)$ as the complexity of the algorithm in this case. Without loss of generality, we assume that the average number of matching pairs is $n/4$ ($\approx (1 + 2 + \ldots + n/2)/(n/2)$). After collapsing, the next level of coarser graph will have $3n/4$ multinodes. In this case, we can get a recurrent function as

$$F(n) = \begin{cases} 0 & \text{if } n \le 1 \\ \alpha \cdot n^2 \log n & \text{if } n > 1 \text{ and no further match exists} \\ \alpha \cdot n^2 \log n + F(3n/4) & \text{if } n > 1 \text{ and further match exists} \end{cases}$$

where $F(3n/4)$ is the complexity of the matching and collapsing process on the next level coarser graph. In the worst case, every recursive execution of the algorithm would produce a (next level) coarser graph whose number of multinodes is about four thirds of that of the current graph. After k times of recursive execution, either the number of the (multi)nodes of the graph becomes 1, or no further match can be found from the graph. So, for simplicity, we can assume that $n \cdot (\frac{3}{4})^k = 1$, or $n = \left\lfloor (\frac{4}{3})^k \right\rfloor$ for some k. According to the recurrent relation of $F(n)$, we can derive

$$F(n) = \alpha \cdot n^2 \log n + F(\frac{3}{4}n) = \alpha \cdot n^2 \log n + \alpha \cdot (\frac{3}{4}n)^2 \log (\frac{3}{4}n) + F(\frac{3^2}{4^2}n) = \ldots$$

$$\leq \alpha \cdot n^2 \log n \left(1 + \frac{3^2}{4^2} + \frac{3^4}{4^4} + \frac{3^6}{4^6} + \ldots\right)$$

$$= \frac{16}{7} \cdot \alpha \cdot n^2 \log n$$

Or

$$F(n) \leq O(n^2 \log n) \tag{5}$$

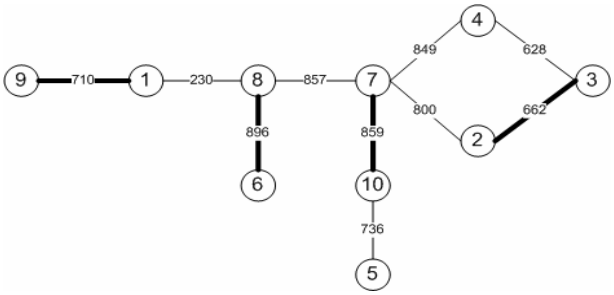
Following the discussion above, it is not difficult to prove that formula (5) holds for every large n (the proof is omitted here).

As the complexity of the algorithm is always greater than or equal to $f(n)$, and less than or equal to $F(n)$, i.e., $O(n^2 \log n) = f(n) \leq g(n) \leq F(n) \leq O(n^2 \log n))$, we get

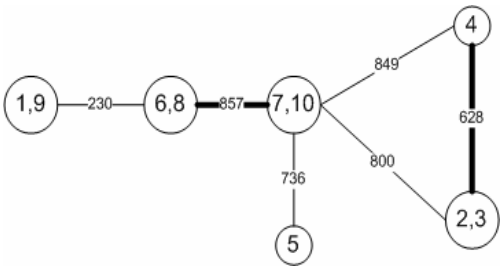
$$g(n) = O(n^2 \log n).$$

3.3 The Greedy Matching Based Algorithm

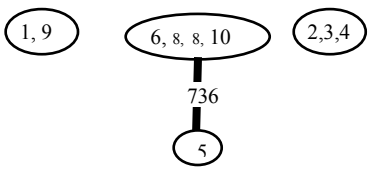
In online spatial application, users are usually expected a much faster response for their spatial searching. Due to this, we also implemented a 'light' version of the above algorithm, in which the Cooks' maximal matching algorithm in the above implementation is replaced with the *greedy matching algorithm* (Doratha & Hougardy, 2003). The greedy matching algorithm is an approximation algorithm for solving the weighted matching problem and has a performance ratio of ½ to optimal matching (Doratha & Hougardy, 2003). It can be implemented with a running time $O(|E| \log |V|)$ if the edges of G are sorted in a pre-processing step by decreasing weight.



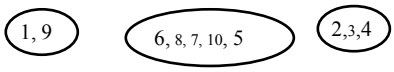
(a) The CO graph G with its first round of matching



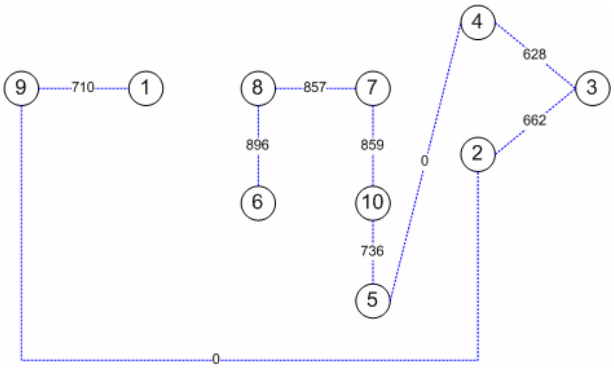
(b) The coarser graph G' and its matching



(c) Next level coarser graph and its matching



(d) No further matches can be found



(e) An AMO order produced (with total overlapping weight of 4632)

Figure 3. Execution of the Greedy Maximal Match based algorithm

If comparing the complexities between the Cook's maximal matching algorithm and the greedy matching algorithm (i.e., $O(|V|\cdot|E|\ log\ |V|)$ vs. $O(|E|\ log\ |V|)$), it is no doubt that the greedy matching algorithm runs faster, thus the greedy matching based AMO algorithm will be much faster than that of maximal matching based AMO algorithm, although the quality of the AMO orders generated by them may differ. As a comparison, Figure 3 shows the execution of the greedy matching based AMO algorithm, with a total overlapping weight of 4632 (comparing with the total overlapping weight 5570 for the AMO order generated by *MaxMatchBasedAMO*(), as in Figure 2).

4. EXPERIMENT EVALUATION

The experiments were conducted to demonstrate the reduction of the I/O costs in spatial join processing by using the AMO orders to guide the scheduling of processing of clustered join operations. We compare the quality of the AMO orders generated by various methods in term of the overlapping weight of the AMO order. The new cluster sequencing method (i.e., Cooks' *maximal match based method,* or *CMM*) is simulated against other methods, namely *Maximum spanning tree (MST) based method* (Xiao *et al*, 2000), Edmonds' maximal matching (EMM) based methods (Xiao, 2003), and greedy matching method (GMM). As EMM and CMM always produce almost the same overlapping weights for all clusters, the EMM results were not included here for comparison.

In the experiments, most spatial datasets are generated while a small portion of datasets is from real spatial applications. The object sizes change from tens to hundreds of vertices. At each simulation point, the simulation runs 10 times. Since every object needs to be fetched into the memory for the refinement operation, for simplicity, we measure the I/O cost in terms of the total size of the overlapping objects that are fetched repeatedly into the memory for processing (i.e., y value in formula (3)).

Table 1 shows the experiment results with ten clusters/nodes in the CO graphs. There were ten experiments conducted with a different number of edges connecting the clusters. For example, for ten edges, the total overlapping weights produced by MST, GMM, and CMM methods are 4721, 5348 and 5570, respectively. Thus, GMM outperforms MST by 13.28% and CMM method outperforms MST and GMM by 17.98% and 4.15%, respectively. The average results showed that CMM method can potentially produce 5.59% and 16.91% more total overlapping weight comparing to GMM and MST respectively.

Table 3: Results of experiment with 10 clusters

Cluster Number	Edge	MST	GMM	CMM	CMM over GMM	CMM over MST	GMM over MST
10	10	4721	5348	5570	4.15%	17.98%	13.28%
10	15	3869	4027	4302	6.83%	11.19%	4.08%
10	20	5596	6405	6648	3.79%	18.80%	14.46%
10	22	5416	6122	6536	6.76%	20.68%	13.04%
10	25	5624	6782	7144	5.34%	27.03%	20.59%
10	30	5774	6015	6532	8.60%	13.13%	4.17%
10	33	5575	6120	6300	2.94%	13.00%	9.78%
10	35	6542	7442	7882	5.91%	20.48%	13.76%
10	40	6686	7035	7548	7.29%	12.89%	5.22%
10	45	6944	7583	7910	4.31%	13.91%	9.20%
Average:		5674.7	6287.9	6637.2	5.59%	16.91%	10.76%

Table 2 shows the summary result of the experiments. For each cluster number, we conducted ten experiments and the average results are shown in the table. For example, for ten clusters, the average of total overlapping weight produced by MST, GMM, and CMM method are 5674.7, 6287.9 and 6637.2, respectively, and the average percentage of performance comparison for each method is also shown in the table (detailed experiments are omitted here).

Table 4: Summary of experiment results

Cluster number	MST	GMM	CMM	CM over GMM	CMM over MST	GMM over MST
10	5674.7	6287.9	6637.2	5.59%	16.91%	10.76%
20	9509.8	10401.4	11038.9	6.12%	16.18%	9.49%
30	14068.9	15103.9	15708.7	4.05%	12.11%	7.77%
40	17592.4	18980.3	19819	4.34%	13.28%	8.60%
50	21539.5	23530	24431.6	3.75%	14.09%	10.00%
60	31011.8	33060.1	34424.5	4.11%	11.05%	6.67%
70	38011	41211.5	42448.2	3.03%	12.28%	8.98%
80	38738.5	41552.1	42870.6	3.20%	10.69%	7.28%
90	41586.7	44306.5	45569.5	2.78%	9.78%	6.82%
100	43001.9	47588.9	48517.4	1.94%	12.95%	10.80%
Average				3.89%	12.93%	8.72%

Figure 4 shows the average total amount of I/O cost that can be saved by each method in a line chart. As expected, the proposed method performs all the time better than the other two methods. On average, there are 12.93% saving comparing with the MST and 3.89% saving comparing with the GMM. For example, when the cluster number is 50, the average saving of I/O cost by MST and GMM are 21539.5 and 23530, respectively, while it is 24431.6 by using the CMM method.

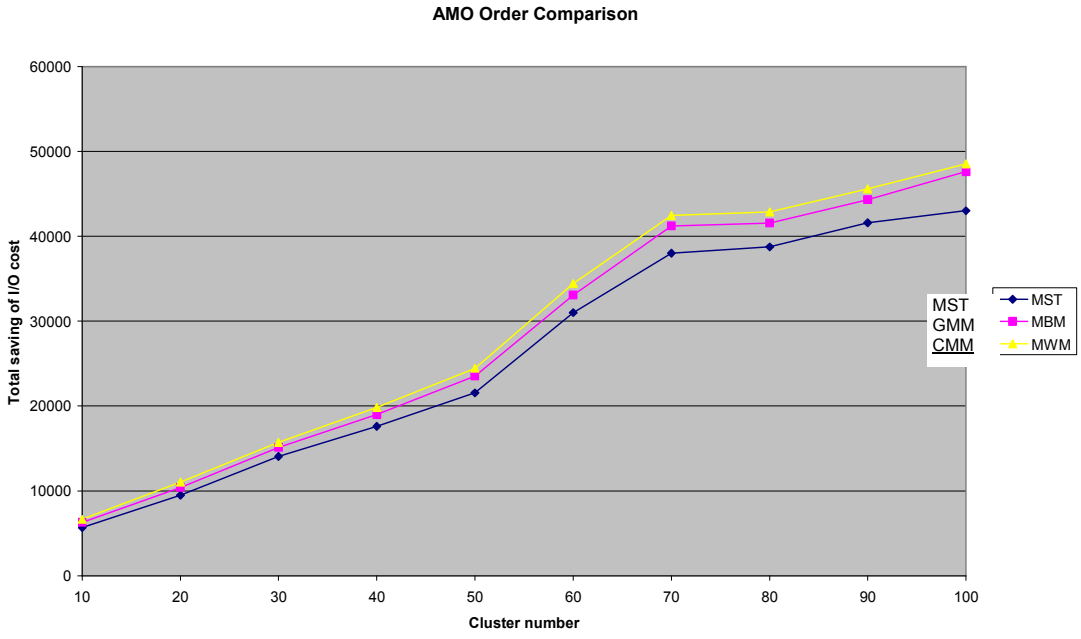


Figure 4. Comparison of total saving of I/O cost

5. CONCLUSION

In spatial join processing, spatial objects are usually clustered and then are processed cluster by cluster. Since two clusters may have overlapping, the overlapping objects may be repeatedly loaded into memory. We proposed a method in (Xiao *et al*, 2000) and (Xiao, 2003) that schedule the processing of the clusters in such a sequence that two consecutive clusters in the sequence have higher number of overlapping objects. Thus, there is no need to load those overlapping objects when processing the next cluster because they are already in the memory. The I/O cost can, therefore, be reduced.

The key issue behind this method is how to produce a better cluster sequence to guide the scheduling. This paper proposed a new algorithm that generates a better cluster sequence in the sense that the average overlapping weight of the cluster sequence produced by the new algorithm is much greater than that of the MST based algorithm (Xiao *et al*, 2000). Experimental results have shown that, if the spatial join operations are processed cluster by cluster according to the cluster sequence produced by the new algorithm, about 13% of the fetching time used for fetching those overlapping objects can be saved when compared to the case where the MST method was used. In addition, while the new method generates cluster sequences of same quality to the ones generated by the Cook's maximal match (CMM) based algorithm (Xiao, 2003), the complexity of the new algorithm is much lower than CMM based algorithm.

ACKNOWLEDGEMENT

The author thanks Mr Usman Farroog and Mr Husen Husen for their contribution in implementation of the algorithms and conducting experiments.

REFERENCES

Abel, D. (1989). SIRO-DBMS: A Database Tool Kit for Geographical Information Systems. International. J. of Geographical Information Systems, Vol. 3, No. 2, pp.103-116.

Abel, D., Ooi, B. C., Tan, K-L., Power, R., Yu, J. X. (1995). Spatial Join Strategies in Distributed Spatial DBMS. Proceedings of Fourth International Symposium on Large Spatial Databases, Portland, Maine, pp.348-367.

Abel, D., Gaede, V., Power, R. and Zhou, X (1997), Resequencing and Clustering to Improve the Performance of Spatial Join. Technical Report, CSIRO Mathematical and Information Sciences, Australia.

Cook, W. J., & Rohe, A. (1999). Computing Minimum-Weight Perfect Matchings. INFORMS journal on computing, 11(2), 138.

Doratha, E. D., & Hougardy, S. (2003). A Simple Approximation Algorithm for the Weighted Matching Problem. Information Processing Letters, 85(4), 211-213.

Edmonds, J. (1965). Path, Tree, and Flower. Journal of Math, 17, 449-467.

Lawler, E. L (1976), Combinatorial Optimization: Networks and Matroids. Holt, Rinehart and Winston, New York.

Ooi, B. C. (1990), Efficient Query Processing in Geographic Information Systems. Lecture Notes in Computer Science (471).

Xiao, J., Zhang, Y. , Jia, X. and Zhou, X. (2000), A Schedule of Join Operations to Reduce I/O Cost in Spatial Database Systems, Data & Knowledge Engineering, Elsevier Science B.V, Vol. 35, pp299-317.

Xiao, J., Zhang Y. and Jia, X (2001). Clustering Non-uniform-sized Spatial Objects to Reduce I/O cost for Spatial-join Processing. The Computer Journal, Vol. 44 No.5, British Computer Society, pp384-397.

Xiao, J. (2003), An Efficient Algorithm For Scheduling Spatial Join Operations In Spatial Database Systems, Proceedings of 7th IASTED International Conference on Software Engineering and Applications, Marina del Rey, CA, USA, Nov. 3-5, pp48-53.

Zhou, X., Abel, D. and Truffet, D. (1998), Data Partitioning for Parallel Spatial Join Processing. GeoInformatica 2:2, Kluwer Academic Publisher, pp175-204

Mobile Agent for Electrical Power Infrastructure Protection

Dr. Michael W. David

GISSET, Fukuoka, Japan

INTRODUCTION

The massive power outage that occurred in Canada and the US on 14 August 2003 provided an indication of what can happen when part of the critical infrastructure fails [30]. The Electric Power Research Institute (EPRI) began its own Infrastructure Security Initiative (ISI) in mid-2002 that was scheduled for development and initial delivery by mid-2004 [2]. EPRI's ISI envisions a massive effort to create a "mega infrastructure" of real-time information and power exchange. Essentially, this would create a private communication network outside the Internet. However, this is likely to be a daunting challenge since California alone has 55,000 power grid nodes that would need to be monitored [31]. To meet this need, we propose the concept of a Critical Network Infrastructure Analysis Center (CNIAC) to provide better coordination and dissemination of information, improve incident prevention and detection, database analysis and real time network monitoring and surveillance systems. This will use a combination of human analysts supported by mobile agents to continually audit, monitor, assess and protect the networks. From one perspective, this can be viewed as defensive information warfare (IW) designed to protect the information infrastructure [32]. However, castle walls and static fortifications like the Maginot Line have shown that fixed defenses are not always the best. Therefore, the concept proposes a mobile agent or dynamic defense.

An agent is a computer program that acts autonomously on behalf of a person or organization. Currently, most agents are programmed in an interpreted language (for example, Tcl and Java) for portability. Each agent has its own thread of execution so tasks can be performed on its own initiative. A software agent can be defined as a software entity which functions continuously and autonomously in a particular environment and which is able to carry out activities in a flexible and intelligent manner that is responsive to changes in the environment.

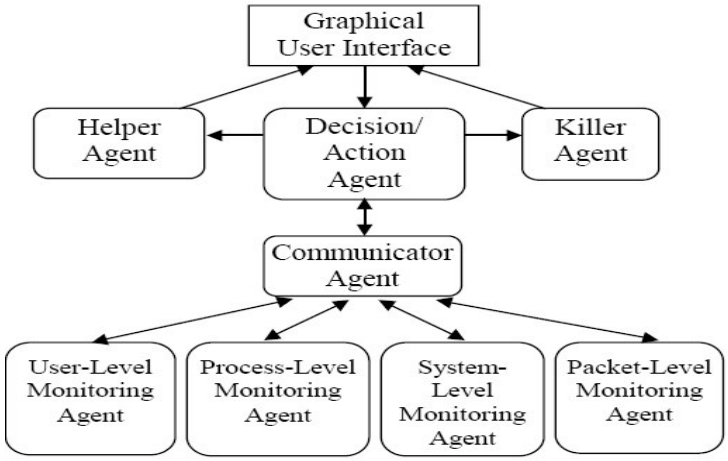


Figure 1: Generic Mobile Agent System [9]

Ideally, an agent that functions continuously would be able to learn from its experience, to communicate and cooperate with other agents, and to move from place to place in doing so. An agent system is a platform that can create, interpret, execute, transfer, and terminate agents. Like an agent, an agent system is associated with an authority that identifies the person or organization for which the agent system acts [25].

Critical Network Infrastructure Protection

In general, these are networks that carry information relevant to national security, safety or financial value. They are generically called Critical Network Infrastructure (CNI) [7]. The President's Commission on Critical Information Protection (PCCIP) in October 1997 identified the basic mission and objectives of an Information Sharing and Analysis Center (ISAC) [26]. Dr. Denning saw the ISACs initially focusing on gathering strategic information about threats, vulnerabilities, practices and resources to enable effective analysis to better understand the cyber dimension of the infrastructure [12]. The ISAC for the electrical power infrastructure is the North American Electric Reliability Council (NERC).

Although each sector, like electric power, has the phenomenal ability to model their own individual infrastructures, there is less of an ability to predict and model the interdependence among infrastructures [21]. The National Infrastructure Simulation and Analysis Center (NISAC) is working with the Idaho National Engineering and Environmental Laboratory (INEEL) to model potential weaknesses in supervisory control and data acquisition (SCADA) systems. SCADA systems are becoming more efficient and cost-effective, but arguably less secure because the architecture and interfaces between scattered SCADA systems have become more open through advances in public networks. NISAC is also testing agent-based simulation, where an agent is an encapsulated piece of software that acts as a decision-making piece of the physical infrastructure. For example, in an electric power plant, the agents could be generator, transmission or SCADA objects working separately but impacting on the whole [21].

In September 2003, the Department of Homeland Security (DHS), National Cyber Security Division (NCSD) created the U.S. Computer Emergency Response Team (US-CERT) to lead all CNI related incident prevention, warnings and response efforts across the country. The existing Federal Computer Incident Response Center (FedCERT) and the Carnegie Mellon University Computer Emergency Response Team Coordination Center (CERT/CC) will supply US-CERT's core capabilities. US-CERTs mission is to be the coordination point for prevention, protection and response to cyber attacks across the Internet and support the existing National Infrastructure Protection Center (NIPC) [16].

Electric Grid Architecture and Function

Overview

The transmission system is the central trunk of the electricity grid; thousands of distribution systems branch off from this central trunk and fork and diverge into tens of thousands of feeder lines reaching into homes, buildings, and industries. The power flow to the distribution systems is largely determined by the power flow through the transmission systems, and in fact when most people talk of the power "grid," they're often referring to the transmission system.

The transmission system truly is a grid; transmission lines run not only from power plants to load centers, but also run from transmission line to transmission line, providing a redundant system that helps to assure the smooth flow of power. If a transmission line is taken out of service in one part of the power grid, the power can usually be rerouted through other power lines to continue delivering the power to the customer [14].

Controlling the Power Grid

On the continental US there are about 150 Control Area Operators (CAO) using computerized control centers to dispatch generators as needed. These generators are divided into three main categories: baseload, peaking and intermediate power plants. Nuclear plants for instance are nearly always operated as baseload plants because their systems are the most stable at full power.

The CAOs run the grid within their control areas. However, the responsibility for the electric grids traditionally lies with utility companies. Due to competition, some states have passed the control of the gird to Independent System Operators (ISO).

CRITICAL NETWORK INFRASTRUCTURE ANALYSIS CENTER (CNIAC)

In an earlier work, we proposed a Cyber Intelligence Analysis Center (CAC) to provide the so-called essential elements of enemy information (EEEI): who, what, when, where and how someone will attack an objective or system [10]. We tentatively call this the Critical Network Infrastructure Analysis Center (CNIAC), and suggest one for each major critical infrastructure ISAC like air traffic control, emergency services, etc [6]. The example we will focus on is the ISAC for the electric power grid, NERC and its interface with the Control Area Operators (CAO), regional Independent Systems Operators (ISO) and Regional Transmission Organizations (RTO).

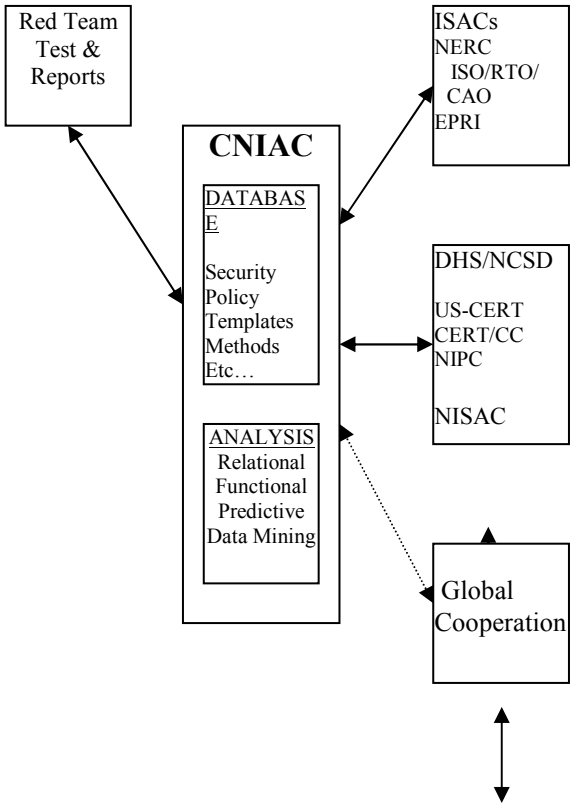


Figure 2: Generic outline of the proposed Critical Network Infrastructure Analysis Center (CNIAC)

The proposal is for the analysts at the CNIAC to work as teams that concentrate on specific groups of critical infrastructure networks. It should have a staff on-call or available 24/7 if possible to anticipate, deny or pre-empt a failure or attack. This is because human experts are still the key tool for identifying, tracking and disabling new attacks. This often requires experts from many organizations to work together and share their knowledge and hypothesis. That is, an important part of the discovery, analysis, and defense against new attacks is based on the cooperation between experts across different organizations [4,15].

We are not experts on power generation plants, but we do propose a conceptual structure in Figure 4.1 as a basis for planning. The CNIAC teams would become familiar with the normal pattern of activity of the networks they support, and provide a human element of expertise in noticing deviations or anomalies that could indicate an attack was being planned, or was underway [3]. Analysts would also perform more detailed research and reporting based on data mining. Data mining would be conducted offline, and create knowledge or intelligence. The Data mining processes search for hidden patterns based on previously undetected intrusions to help develop new detection templates. In addition, data mining focuses on new hidden patterns in old data to create previously unknown knowledge or intelligence [5].

The CNIAC's analysts should become the experts in their sectors, and provide updated indicators, warnings and advice to the information security / assurance community. The actual analyst roles/positions could be an extension of, or migration by the EPRI's ISI Red Teaming Project experts. These Red Teams conduct mock assaults on selected computer systems of volunteer utilities, probing for weaknesses in a manner similar to the Federal Aviation Administration's Red Teams [28]. They could provide a core of expertise for, or supplement the CNIAC as depicted in Figure 3 on an ad hoc basis.

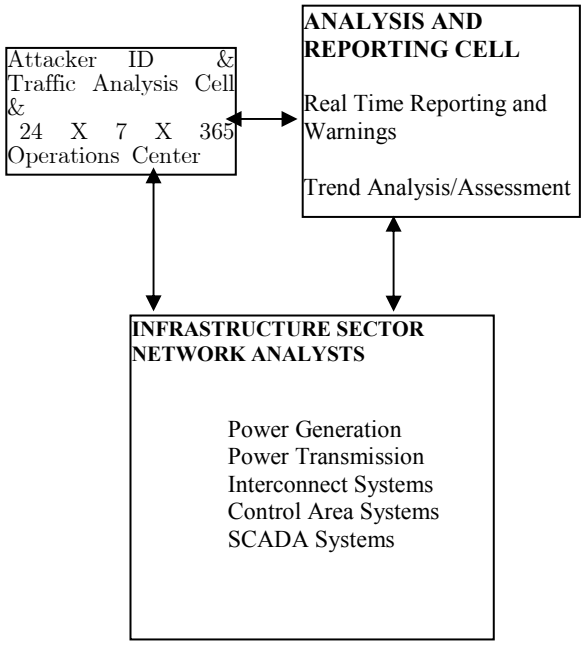


Figure 3: Functional Depiction of a Possible Electric Power CNIAC.

PROPOSALS FOR METHODOLOGIES AND SYSTEMS

Preventive Measures

Intrusion or attack prevention is an area of growing interest. For example the National Institute of Standards (NIST) and the National Security Agency (NSA) are continually issuing Information Security guidelines. One such example is the NIST-NSA "Guide to Secure Configuration and Administration of Microsoft SQL Server 2000". Reportedly, organizations that had implemented these guidelines, and audited compliance with those policies for every database on their network prevented exposure to the SQL Slammer worm. At least one firm, Preventsys, has capitalized on this and created a Network Audit and Policy Assurance System (NAPAS). This system provides automated security auditing for large corporate and government networks and the Internet. It audits for policy violations across the network. A unique feature is that it is implemented without the deployment of any agents or monitors in the network [27]. This system or something similar could serve to audit and update all the networks within the power grid.

Detection Systems

Traditional anomaly detection approaches build models of normal data and detect deviations from the norm in observed or monitored data. Anomaly detection for intrusion detection has been active since it was originally proposed by Denning [13]. However, anomaly detection schemes suffer from a high rate of false alarms. This is because new, but legitimate, behaviors are evaluated as anomalies, thereby raising an alarm [24]. Fortunately, a work on comparative anomaly detection schemes indicates that local outlier factor (LOF) based detection systems are very successful in identifying novel attacks, and minimizing the number of false alarms [23].

A proposal for improved interdomain routing also looks promising to help work around the inability of the Border Gateway Protocol (BGP) to provide a way to identify the source of bad or malicious data. The new protocol is designed to detect and mitigate accidentally or maliciously introduced faulty routing information [17]

In addition, new work on Denial of Service (DoS) attacks indicates that while no single method can fit all possible scenarios related to DoS, there is hope. Research indicates that network-monitoring techniques can be used to detect service violations by measuring the SLA parameters and comparing them against the contracted values between the user and the network provider. It also suggests that monitoring techniques have the potential to detect DoS attacks in early stages before they cause serious damage [18]

However, it is probably infeasible to have one national, centralized real-time IDS system. Therefore, we propose that a combination of these techniques could be implemented at the power plant and transmission facility network levels, and be part of the front line of defense against attack from malicious or faulty data.

The data from these systems could be part of the package picked up by mobile agents sent out by the CNIAC. The returned data could assist the CNIAC, NERC and US-CERT in tracking down the source of the attack or faulty data. In traditional IDS terminology, the CNIAC would be the command and control, the CAO/ISOs could be the aggregation nodes and the plants, interconnect and SCADA systems would be the primary collection nodes.

Dynamic Decision Making

Although it is difficult to conduct offensive operations against potential threats, it is possible to implement dynamic or mobile systems to be more vigilant and responsive to potential attacks.

M.H. Kuo has proposed an interesting design for a Dynamic Information Security Decision Model (DISDM). Kuo has proposed an agent-based DISDM that automatically integrates all distributed information security systems to achieve efficient defensive information warfare. The DISDM includes five kinds of agents (inspection, evaluation, messaging, command and defensive). It can dynamically execute sound security strategies and recover from attacks over a 24-hour period. The concept is reportedly being evaluated using IBM Aglets [22].

Monitoring and Analysis

Some of the difficulties related to catching intruders arise because (1) existing systems logs are too large; (2) they do not contain all the information that could be useful in tracking down an intruder; and (3) attacks often come from multiple sources and spawn processes at multiple destinations. Currently, it is difficult to coordinate logs across administrative domains [4].

We propose a dynamic electric power mobile agent system (EPMAS) be designed to support secure communications interface, update security policy, collect and audit IDS related data and provide status reports across domains in the electric power grid.

The EPRI's ISI Secure Communications Project has done a Security Methodology assessment of the costs of implementing security measures for utility communications systems, "A Scoping Study on Security Processes" (057144 March 2003). This will most likely provide the basis for next generation security protocols to protect SCADA and other critical utility communications systems [29]. Communications mobile agents from the CNIAC could monitor and update this protocol. As noted above, something along the lines of the Preventsys NAPAS product might be useful. However, this product does not use mobile agents, and may be best suited for use at the power plant and transmission facility level [27]. The CNIAC mobile agents could be monitoring the NAPAS and make sure that all the many elements within a plant are being contacted and updated on a regular basis by interfacing with the NAPAS. Essentially, these security agents would be watching the watcher.

The CNIAC analysts would work to support the NERC in its efforts to prevent, detect and respond to malicious attacks or failures. The source of the mobile agents would originate from a server in CNIAC. Mid-level managers at both objects of management operations and sources of their own mobile agents would reside in the ISO/RTO centers. The managed resources would be at the power plant and transmission facilities level as outlined in Figure 4. There would be multiple types of agents that would be related to such areas as: security policy, data collection, monitoring, detection, response and repair.

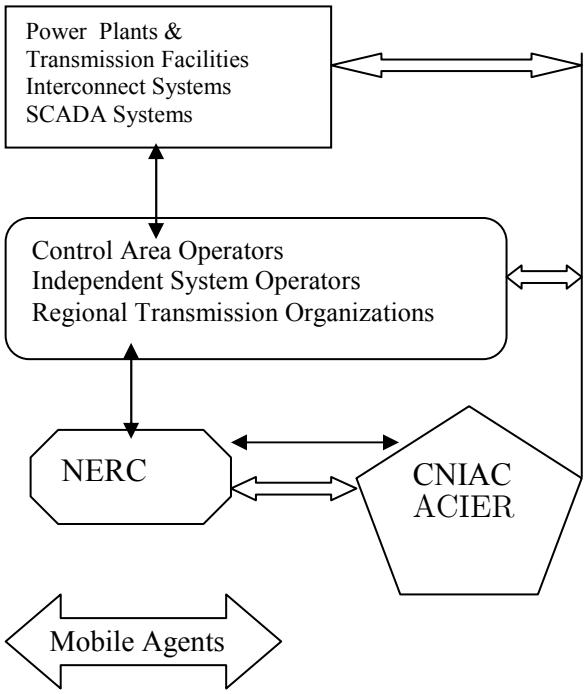


Figure 4: EPMAS Outline for the NERC-CNIAC.

Dasgupta proposed an interesting Security Agents for Network Traffic Analysis (SANTA) distributed agent architecture that seems applicable for use by a CNIAC. This architecture is immunity-based and involves collaboration by a combination of multiple autonomous agents: monitoring, communicator, decision/action, helper and killer agents. A detailed report on the SANTA implementation indicates the system can emulate some mechanisms of the human immune system [8,9].

We suggest this may be ideal for a large mix of networks like the electric grid. The SANTA or SANTA-type agents could support the EPRI's Automated Critical Incident Event Reporter (ACIER) application, which is currently under development. This program is designed to provide a semi-automated means of recognizing, tracking and reporting critical incidents. Within this context, it should help to improve the national capability to distinguish anomalies from attacks [1].

Of course, the security of the agents themselves is an important factor, and considerable work has also been done in the area of agent confidentiality, integrity and availability through the use of cryptographic methods [19, 20, 28]. The agents launched by the CNIAC, ISO and RTO servers must be authenticated upon arrival, and checked for authorization and access. The integrity of the mobile agent code will be monitored at the CNIAC, ISO/RTO and plant and facilities level so that any changes to an agent are detected. In any event, they will need to be capable of transmission within the EPRI's developing Secure Communications Project mentioned above.

EPMAS ACTIVITIES

Agent Functions and Missions

The agents would travel to a CAO and also be focused on plant types and systems. They would conduct their missions in their domains; perform their assigned tasks, regenerate or return [15]. Some of these may be database specific like security policy. Others will be program specific, like SCADA, ICS and power generation. The agents can be considered to have a short life span since they must continually represent new instructions, data or decisions. Some will be general like the plant level, while others will be more specific like the SCADA instructions.

One of the most important functions the EPMAS can play in supporting the CNIAC is event correlation. Since the EPMAS spans geographic regions, functional roles and different networks, they will be able to

provide CNIAC analysts with the ability to discover patterns or identify anomalies across domains and hierarchies somewhat similar to data mining in information-based warfare [32]. Having agents visit data repositories and mine results is an ideal role, well suited to the ability of mobile agents to transfer their computations to the data they collect. Besides reducing network load, this is conducive to having specialized agents focused on specific classes of intrusions [20]. This can be applied to their local generation plant or SCADA system within the CAO.

For example, the EPMAS infrastructure can capture and access security logs, help correlate data from the security logs and support analysts trying to formulate attack hypothesis [4]. The CNIAC analysts would be able to work on correlation of these logs and other information. One of the primary goals of this correlation should be to identify penetration or internal attacks. Data is meaningless unless it is processed, analyzed and evaluated for an assessment or report. The EPMAS can provide the data and the CNIAC will provide the final assessments and reports/recommended actions to the NERC. The greatest potential for the EPMAS provide the data and the CNIAC will provide the final assessments and reports/recommended actions to the NERC. The greatest potential for the EPMAS, however, may lie with their ability to provide local response versus detection. This is because response can be initiated from nearly anywhere in the system [20].

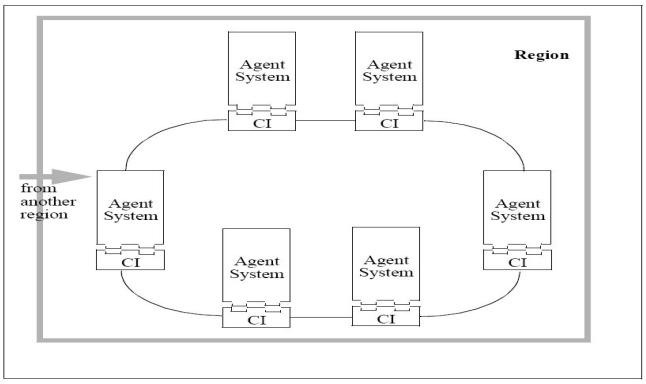


Figure 5: Generic Regional Mobile Agent System Architecture [25]

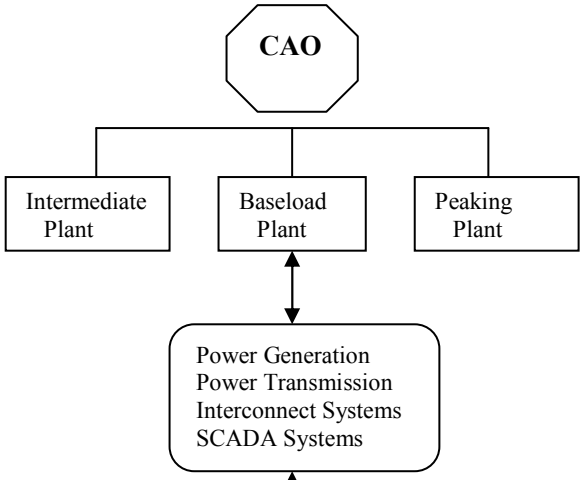
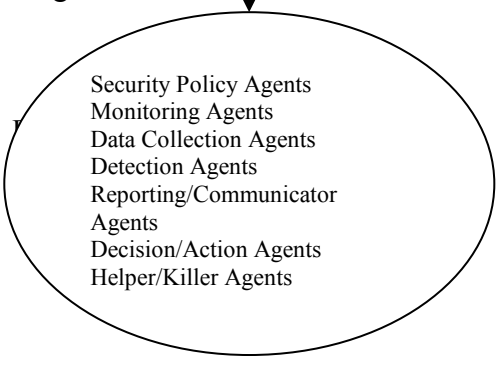


Figure 4.6: EPMAS Activities

Figure 6:



Generic Sector / Area

As noted above, the key functions of the EPMAS would be: security policy, data collection, monitoring, detection, response and repair. This is depicted in the general flow in Figure 3. This would be across regions as depicted in the generic regional agent system in Figure 4. The EPMAS agents could be distributed based on several factors: geographic location or region, type of power generation and nature of the information/data network they work in. A more detailed outline is provided for a hypothetical Control Area Operator (CAO) in Figure 5, which takes into account the types of power generation: baseload, peaking and intermediate power plants.

CONCLUSION

The Critical National Infrastructure (CNI) includes the electric power grid, and the information networks that help to operate and manage it [6]. Since infrastructure systems and sub-systems are increasingly interrelated, they cannot easily be separated or isolated for damage containment. If electric power is disrupted, so are all the systems that use it [11]. We have proposed a structure called the CNIAC and EPMAS that would integrate and coordinate the use of the various elements of a security system: security policy, intrusion detection, incident monitoring/reporting and analysis/assessment. This includes the use of human analysts along with automated distributed mobile agents to continually audit, monitor, assess and protect the system.

There is much more detailed work to be done to attempt to see how this concept can support the EPRI's Infrastructure Security Initiative (ISI) and its various programs associated with vulnerability assessment, red team attacks, secure communications and automated incident reporting [1,2,28,29]. However, we believe the concept is sound, and can make a positive contribution to the security of information networks within the CNI.

References:

[1] ACIER, "Automated Critical Incident Reporter (ACIER) Application Implementation", URL: http://www.epri.com/D2004/

[2] Amin, M., "Infrastructure Security Initiative", Electric Power Research Institute (EPRI), URL: http://www.epri.com, July 2002

[3]Anderson, R.J. Security Engineering – *A Guide to Building Dependable Distributed Systems*, Wiley 2001

[4] Aslam, J., Cremonimi,, M., Kotz, D., Rus, D., "Using Mobile Agents for Analyzing Intrusion in Computer Networks", URL: http://www.kerk.cs.dartmouth.edu

[5] Bass, T. "Intrusion Detection Systems and Multisensor Data Fusion." *Communications of the ACM* Vol. 43, No. 4, April 2000, 99-105.

[6] CIAO. URL: http://www.ciao.gov

[7] CNI/03, "Introduction to Critical Network Infrastructures", *ITU Workshop on Creating Trust in Critical Network Infrastructures*, Korea, 20 May 2002

[8] Dasgupta, D. "Immunity-Based Intrusion Detection Systems: A General Framework", Proceedings *of the 22nd NISSC*, URL: http://issrl.cs.memphis.edu/nissc-99.pdf, 18 – 21 Oct 1999

[9] Dasgupta, D., Hall, B. "Mobile Security Agents for Network Traffic Analysis", *IEEE Computer Society Press, Proceedings of DISCEX-II*, June 2001

[10] David, M., Sakurai, K. "Combating Cyber Terrorism: Countering Cyber Terrorists Advantages of Surprise and Anonymity", *Proceedings of AINA 2003:* pp716-722, Mar 2003

[11] Dearth, D., "Critical Infrastructures and the Human Target in Information Operations", *Cyberwar 3.0*, Editors Campen, A., Dearth, D. AFCEA Press, Fairfax, Va. 2000, Page 203-210.

[12] Denning, D. *Information Warfare and Security*. Addison Wesley Longman Inc., Reading, Ma. 1999

[13] Denning, D. "An Intrusion Detection Model", *IEEE Transactions on Software Engineering*, SE-13:222-232, 1987

[14] DOE / Office of Energy Efficiency Renewable Energy: www.eere.energy.gov/der/grid_ach_function.html
Energy: www.eere.energy.gov/der/control_pwrgrid.html

[15] Foukia, N., Hulaas, J., Harms, J., "Intrusion Detection with Mobile Agents:, URL: http://www.isoc.org/isoc/conferences/INET/01/

[16] Frank, D., "DHS creates emergency response team.", URL: http://www.fcw.com/2003/0915/web-dhs-09-15-03
[17] Goodell, G., Aielllo, W., Griffin, T., Ioannidis, J., McDaniel, P., Rubin, T. "Working Around BGP: An Incremental Approach to Improving Security and Accuracy in Interdomain Routing", URL: http://www.isoc.org/isoc/conferences/ndss/03/proceedings/
[18] Habib, A., Hefeed, M.H., Bhargava, B.K., "Detecting Service Violations and DoS Attacks", URL: http://www.isoc.org/isoc/conferences/ndss/03/proceedings/
[19] Humphries, J.W., Pooch, U.W., "Secure Mobile Agents for Network Vulnerability Scanning", URL: http://www.itoc.usma/edu/workshop/2000
[20] Jansen, W. "Intrusion Detection with Mobile Agents", URL:http://csrc.nist.gov/mobileagents/projects/html
[21] Jones, J. "Models of Mayhem: The government wants to simulate the ripple effects of critical infrastructure attacks." URL: http://www.fcw.com/supplements/homeland/2002/sup3/hom-models1-09-30-02.asp
[22] Kuo, M.H. "An Agent Based Dynamic Information Security Model in Information Warfare", URL: http://www.dodccrp.org/activities/symposia/7thICCRTS
[23] Lazarevic, A., Ozgur, A., Ertoz, L., Srivastava, J., Kumar, V. "A Comparative Study of Anomaly Detection Systems in Network Intrusion Detection", URL: http://www.cs.umn.edu/research/minds/MINDS/papers
[24] Northcutt, S., Novak, J. *Network Intrusion Detection – An Analyst's Handbook*, New Riders Publishing, Sep 2000
[25] OMG, "Mobile Agent Facility Specification", 2000. http://www.omg.org/cgi-bin/apps/doc?formal/00-01-02.pdf

[26] PCCIP. "Critical Foundations: Protecting America's Infrastructures, The Report of the President's Commission on Critical Infrastructure Protection" URL: http://www.pccip.gov, Oct 1997
[27] Preventsy: URL: http://www.preventsys.com_16 Sep 2003
[28] Reiser, H., Vogt, G. "Threat Analysis and Security Architecture of Mobile Agent Based Management Systems", *IEEE/IFIP Proceedings of NOMS 2000*, Apr 2000
[29] Sobajic, D., "Infrastructure Security Initiative Early Phase Completed", URL: http://www.epri.com_16May03
[30] Sweet, W., "Blackout 2003 Special" *IEEE Spectrum*, http://www.spectrum.ieee.org/webonly/special/aug03/black03.html
[31] Verton, D., "Power Industry Unveils $100b Upgrade Plan", URL: http://computerworld.com/securitytopics/security/recovery/story/0,10801,84329.00.html, 25 Aug 2003
[32] Waltz, E. (1998). *Information Warfare: Principles and Operations*. Norwood, MA: Artech House, 1998

The Evaluation of Security Systems: Testing Biometric and Intelligent Imaging Systems

Keynote Address: The Sixth International Workshop for Applied PKC (IWAAP2007)

Clifton L Smith

Centre for Security
Edith Cowan University
Joondalup, Western Australia
clifton.smith@ecu.edu.au

INTRODUCTION

Effective security strategies are achieved through the risk management process applied to the security management plan. The components of the security management plan each contribute equally in the protection of assets for an organisation. Thus *missed elements* of components in the risk management plan have the potential to impact on the quality of security established for high–end security such as national infrastructure and the commercial facilities. It is important that any degree of alteration in relevant variables in the components of the security management plan will require the entire plan to be reassessed in order to ensure security strategies that adequately cover the security risks identified in the analysis. This activity should occur even if it disrupts the schedule for completion of the risk assessment components. As well, the security strategies developed to operationalise the security management plan should be subjected to continued scrutiny through analysis and study.

Not only is the reassessment of relevant components of the risk management process essential to avoiding the occurrence of security deficiencies in security strategies, but it is also imperative for the reassessment of the security technology testing process. The criteria for the testing and evaluation of security equipment are determined from the security strategies development from the security management plan. The risk management process that determines the security management plan can be derived from the Australian Standard/New Zealand Standard AS/NZS 4360:2004 – *Risk Management* which specifies the procedures for effective assessment. Security equipment testing and evaluation is only as effective as the risk management process itself, as it part of the process of risk assessment.

The application of Australian Standard/New Zealand Standard AS/NZS 4360:2004 – *Risk Management* is critical to the development of effective security strategies. These strategies will provide parameters for the assessment of security equipment appropriate to adequately protect assets. The role of security technology in the protection of assets of an organisation is determined by the security strategies, which are derived from an organisation's security management plan.

The evaluation of the reliability and validity of security technology to be applied within security strategies from the security management plan will be determined by comprehensive testing methodologies, executed within a proposed testing management model (Jones & Smith, 2005). The AS/NZS 4360:2004 risk management process together with relevant international standards, proposes a formal process for testing and evaluating security technology for specific security strategies. Without a formal framework, such as that provided by AS/NZS 4360:2004, the testing and evaluation of security equipment for the protection of assets could have specified inappropriate security equipment that inadequately protects against risks, poorly integrate with organisational operations, is rejected by its users, and potentially impacts on overall organisational performance.

This paper will discuss the proposed formal frame work for the proposed testing regime and specify the testing function as a component in the risk management process. The functions of testing will be presented in order to determine criteria for testing protocols, and principles of reliability and validity will be applied to

the test function. Examples of the test function will be applied to security technology such as biometric systems and intelligent imaging systems.

CONCEPT OF EVALUATION

The purpose of evaluation of security technology is to determine whether the equipment has the capability to perform its protection of assets function. The evaluation of the function of security technology requires the design and analysis of a test protocol that supports the protection objectives that the designed system must satisfy (Garcia, 2001, p5).

The evaluation of the security technology is a component of the countermeasures for the protection of assets according to the security risks that have been assessed for the assets. These countermeasures can be considered as a combination of security equipment, security policies, and security procedures. Hence the evaluation of security technology in the security environment of the assets contributes a major function in the treatment of these security risks.

There is no clear consensus in the testing community about which group within an organization should be responsible for performing the testing function. Some organizations make performance testing a responsibility of the development group who initially developed the technology. The rationale for this approach to the evaluation of the technology is that these people are most familiar with the functionality of the technology and hence will be most able to detect and alleviate errors. Alternately, there is a concern that the application developers are too closely involved with the technology to be able to identify flaws in the system. Hence the testing function should reside within the scope of a separate testing group that is independent of the development staff. Again, there is an opinion that testing should be the right and responsibility of the end-user group who will actually use the technology. It is considered that the end-user is the best judge of whether the technology meets the necessary requirements for the protection of assets in the security environment of the organisation. As the testing function is performed to support the goals and objectives of the organisation, then the responsibility for testing resides with the organisation that will be most familiar with the unique circumstances and resources in its security environment.

While the objective of security equipment testing is to satisfy the necessity for the protection of assets, the aim of the evaluation of the technology is to produce outcomes that are both reliable and valid. The concepts of reliability and validity are distinctly different but are somewhat linked. While high reliability does not warrant validity, results cannot achieve validity without reliability.

Reliability

Reliability is the degree of confidence in which a testing regime can repeatedly yield the same results. The objective of reliability in security equipment testing, is to measure whether a set of results is reliable enough to be accepted as accurate, consistent and dependable; hence the ability to be able to reproduce the same result time again, ensuring that they are not statistical outlier data (Jones & Smith, 2005).

Validity

Validity is the degree of confidence in which findings from a study assess what they purport to assess (Haslam & McGarty, 1998). That is, validity is concerned with the characterization of the issues to be assessed, and hence display a measure of confidence in the quality of the results or outcomes achieved. Testing for validity is crucial for the evaluation of security technology to be approved as part of the countermeasures for the assessed risk in the protection of assets. Before results can be considered valid, these must first be shown to be reliable. Hence, results can be reliable and valid, reliable and invalid, unreliable and invalid; but never unreliable and valid. Thus the aim of security equipment testing and evaluation is to produce results that are both reliable and valid. If reliability and validity cannot be demonstrated, then the testing and evaluation process will be defective and should be discarded.

Reliability and validity are crucial to the credibility and effectiveness of test results. This view has been supported by Wegner & Spyridakis (1989), who address the issues of testing methodologies which have

been developed by equipment usability testers that have not addressed the reliability and validity of data produced by their testing methodologies.

This investigation has raised the concern over the credibility and effectiveness of the outcomes produced in the testing process. It is suggested that *the concepts of reliability and validity are important to useability testing*, and proposed that the *concern for reliability and validity will enhance the credibility and effectiveness of useability testers* (Wegner & Spyridakis, 1989). These comments extend to security equipment testing and evaluation.

Reliability and Validity in Testing Protocols

The two independent issues of reliability and validity of the performance of the security technologies, and the reliability and validity of the testing methodologies should be considered. Each of these two issues has a crucial role in the quality of the performance of the security equipment, through the outcomes of the testing regimes.

The reliability of the security technology is a measure of the equipment's ability to repeatedly produce the same results under the same circumstances and in the same environment time and again. Whereas the validity of the testing process is the ability of the equipment from performance tests to produce the outcomes that have been specified by the manufacturer.

Again, the reliability of the testing methodologies is the degree of confidence in which a testing regime can repeatedly yield the same results. That is, how reliable is the testing protocol from which the security technology will be assessed? Thus the robustness of the testing procedure will be under scrutiny for reliability to be demonstrated. Then the validity of the testing methodologies will be assessed by the degree of confidence in the findings from the testing procedures, and confidence in what they purport to assess from the evaluation process.

Testing and Evaluation Model

This model for the testing and evaluation of security technology was developed by Jones and Smith (2005) to regularize the testing process according to reliability and validity criteria. The model requires the processes to be undertaken at two different testing levels. Level 1: Laboratory Environment, which determines if the security equipment is technically and physically capable of delivering the clients needs, and Level 2: Operational or Simulated Operational Environment, which assesses the suitability of the equipment to perform within its intended environment and in its intended role.

Level 1 testing is conducted within a controlled laboratory environment where environmental conditions, operational activities, equipment operation, platform of operation, length of operational time, subject diversity, and other factors that are variable in operational environments are largely controlled. Laboratory environments are ideal for preliminary testing and evaluation, as the controlled conditions allow specific functions and characteristics of security equipment to be isolated and repeatedly tested, ensuring reliability and validity.

The primary objective of Level 1 testing under the Jones & Smith model (Jones & Smith, 2005) is for the measurement of the reliability and validity of security equipment. Level 1 testing takes place under controlled laboratory conditions where system elements, characteristics, and functions can be isolated and tested repeatedly to ensure reliability and validity of the security equipment.

Reliability and validity assessments are determined for reliability through repeatedly of test results and validity through the production of results matching the manufacturer claims. The laboratory environment testing will not inform the client if the security equipment is suitable for application within its proposed environment, however the testing will inform them of the degree the security equipment performs under controlled conditions.

Level 2 testing is dependent on the outcome of Level 1 testing (Jones & Smith, 2005). The client can decide whether to terminate testing at the controlled laboratory stage, or to proceed with testing in the operational environment for the security technology (Figure 1).

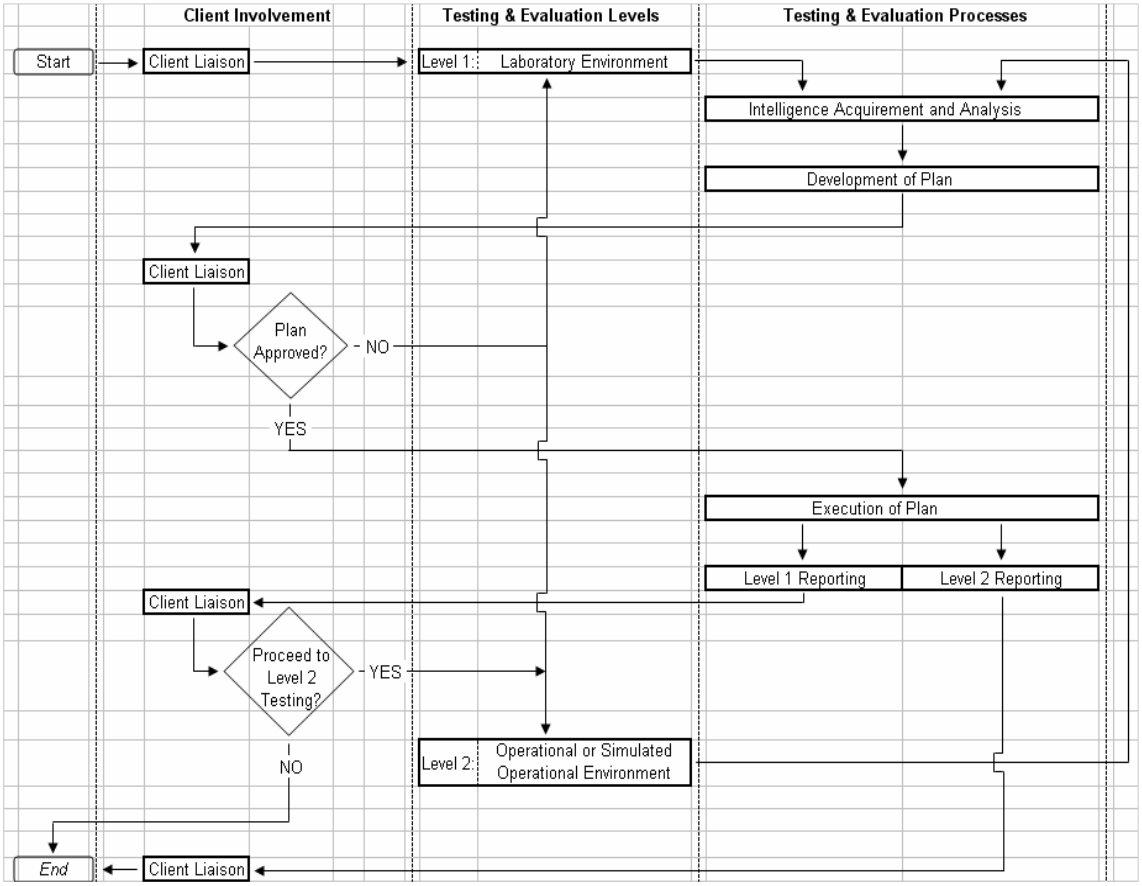


Figure 1: Model of Security Technology Testing and Evaluation (Jones & Smith, 2005)

The entire testing process is repeated for Level 2 testing with fresh intelligent information gathered for actual environment, and the testing plan designed to function according to manufacturers' specifications in the environment. All levels of testing require consultancy with the client, and full reporting to the client.

Jones and Smith (2005) then propose that the Australian Standard/New Zealand Standard AS/NZS 4360:2004 – *Risk Management* could be modified to include the security technology testing process as a component of the risk management process. Applying the testing model for security technology to the AS/NZS 4360:2004 risk management model produces a composite approach for the treatment of risk in a security environment Figure 2). Thus this proposed model becomes an essential element of effectively treating security risks integral to the risk management process.

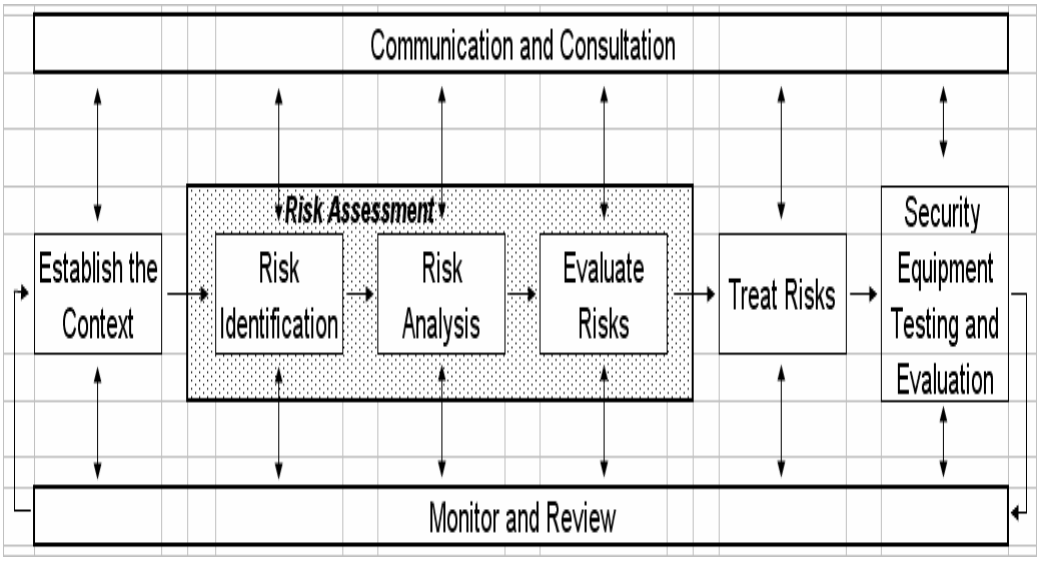


Figure 2. The modified Jones & Smith (2005) version of the AS/NZS 4360:2004 Risk Management Process – Overview model.

The modified security risk management model shown in Figure 2 has included the testing and evaluation component for the management of risk. This security equipment testing and evaluation component emphasizes the need for consideration of as part of the risk management process. That is, the management of risk in the protection of assets is incomplete if the security technology applied in the security strategy has not been shown to exhibit both reliability and validity of operation.

Performance Testing

The testing of security technology in the actual environment can be considered to be performance tested, when the purpose of the testing is to replicate the manufacturers' specifications. That is, the performance claims of the manufacturer should be able to be replicated in the environment in which it is required to operate. Within tolerances of performance and tolerances of environmental conditions, the expectation of replication of performance is appropriate. Lindamae Peck and Lacombe (2007) at the US Army cold weather test site in New Hampshire has been testing sophisticated open ground and perimeter detection systems for twenty years. Similarly, the Home Office Scientific Development Branch in UK [REFERENCE] has been testing these systems in their rough weather site in Wales.

Attack Testing

However attack testing, sometimes called defeat testing, seeks to exploit weaknesses in the design and operation of security systems in order to penetrate the detection barriers of an intrusion detection system strategy. The operations of the security technologies and the management of the security technologies can be exploited in order to defeat the protection barriers of the assets. Intrusion attacks on security technology often targets limitations in design and performance of the equipment with effective outcomes. Intrusion by stealth is the most effective means of penetration of a facility, as the weaknesses in the security management strategies are not disclosed. Attack testing is performed by national agencies, and police and defence groups who with need to develop intrusion skills in the national interest.

BIOMETRICS

People have always relied on their natural abilities to recognize other people by biometric characteristics such as faces, voices and writing patterns. In order to recognize or authenticate their identity, an individual's unique physical or behavioral characteristics can be measured. Physiological biometrics, such as fingerprints or hand geometry are physical characteristics, and behavioral biometrics such as signature or voice quality are appropriate identifiers of people (Figure 3).

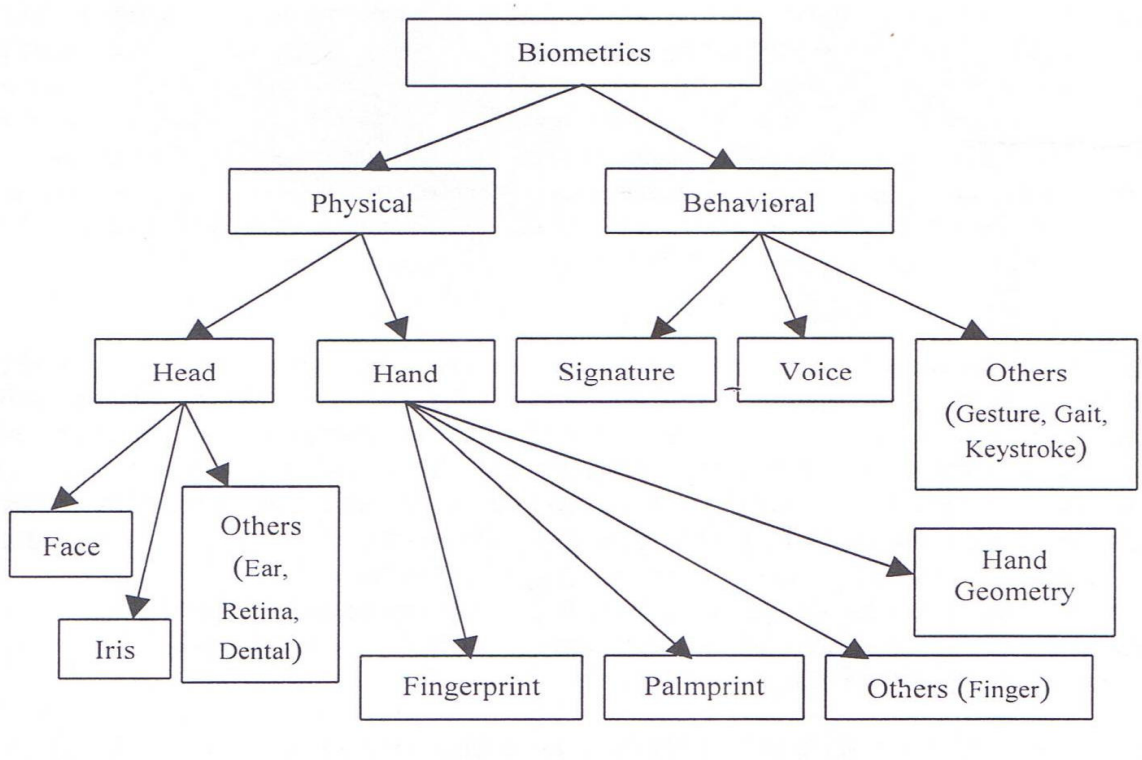


Figure 3: Classification of Biometrics

In the access control, biometric technology has become a familiar application to restrict access of persons to facilities, information and authorised areas, as well as a tool of forensic identification. However, in order for the biometric system to be both reliable and valid, the selected biometric character to identify/verify an identity should possess requisite properties:

- Universal: every person should have that characteristic
- Unique: no two people should have exactly the same properties of that characteristic
- Permanent: invariant with time
- Collectable: can be measured quantitatively
- Reliable: must be safe and operate at a satisfactory performance level
- Acceptable: non-invasive and socially tolerable
- Non-circumventable: how easily the system is fooled into granting access to impostors

The application of biometrics systems requires both the enrolment process and identification/verification process (Figure 4). The biometrics system functions either to identify or verify biometric signatures. In security, the systems may verify a person through matching of one-to-one biometric signatures, or the identification of a person through the matching of one-to-many signatures.

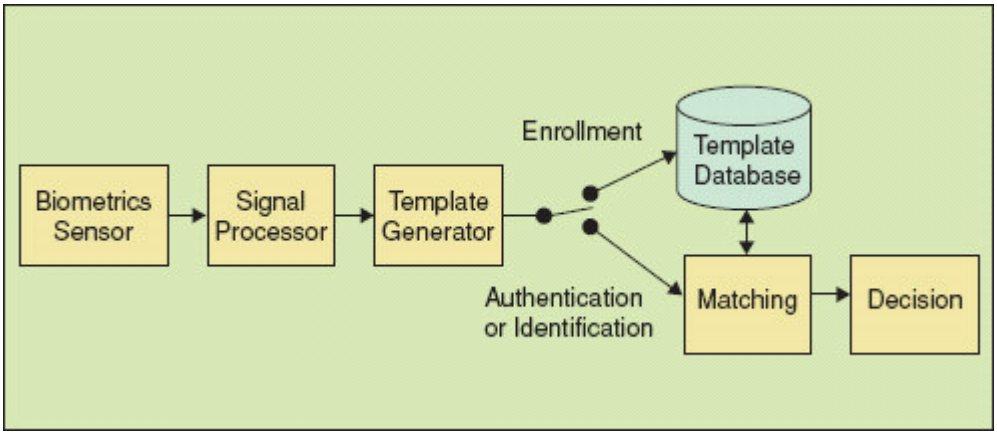


Figure 4: Biometric system operation diagram (Qinghan Xiao, 2007)

Testing Biometric Systems

The schematic of the template for the process for biometric access control is shown in Figure 4 where the components of the process are presented as stages in the identification of the person seeking access. These same processes are again shown in Figure 5 where the template of the generic biometric system is presented. However in Figure 5, the attack points for breaching the biometric system are shown where entry can be gained to defeat the system.

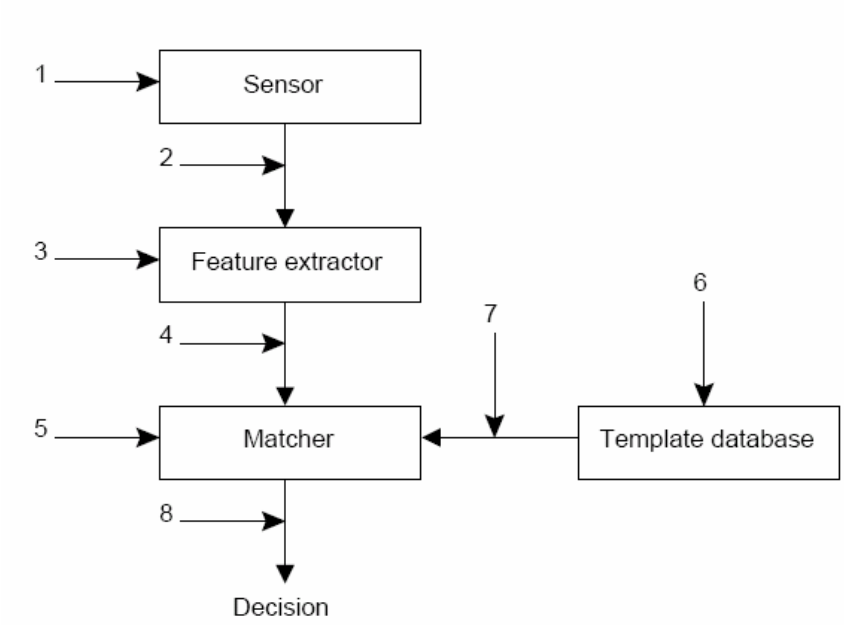


Figure 5: Generic biometric system attack points. (Uludag, 2006)

These attack points in the generic biometric system can be penetrated according to the following procedures:
Attack point 1: An unauthorised person may present a fake fingerprint at the sensor. Replica fingerprints made from wafer-thin silicone can be attached to a finger for presentation to the sensor. Often the match decision will allow the adversary access to the authorised area.

Attack point 2: Digitally pre-recorded fingerprint signals can be electronically presented to the input sensing system. A recorded signal can be replayed to the automated system, bypassing the sensor.

Attack point 3: The feature extraction process can be overridden with fake data. The system could be attacked with a Trojan horse so that it produces feature sets pre-selected by the adversary. These pre-selected sets can be developed with co-operation from authorised users of the system.

Attack point 4: After the features have been extracted from the input signal, the features are subjected to replacement with a different tampered feature set. This process is based on the assumption that the attacker knows the techniques for encoding the biometrics feature set into the representation.

Attack point 5: The fingerprint template matcher can be manipulated so that it always produces artificially high or low match scores. This can be achieved by corrupting the matcher configuration algorithm.

Attack point 6: Attack on the database will result in stored templates being manipulated or corrupted. Unauthorised individual may try to modify the biometric signature representations which can result in authorisation of fraudulent individuals or denial of service to authorised users.

Attack point 7: It is possible that the channel between the stored templates in the database and the matcher can be attacked by modifying the contents of the templates before it is are received by the matcher. This process is based on the assumption that the attacker knows the techniques for encoding the biometrics feature set into the representation.

Attack point 8: The authentication system can also be subjected to corruption as the final matching result can be intercepted, altered and illicitly overridden.

A range of attack techniques can be applied to defeating biometric access control systems with the intention of unauthorized intrusion. The Figure 6 shows a schematic mapping of the techniques which may be applied to defeating biometric systems engaged in the protection of assets.

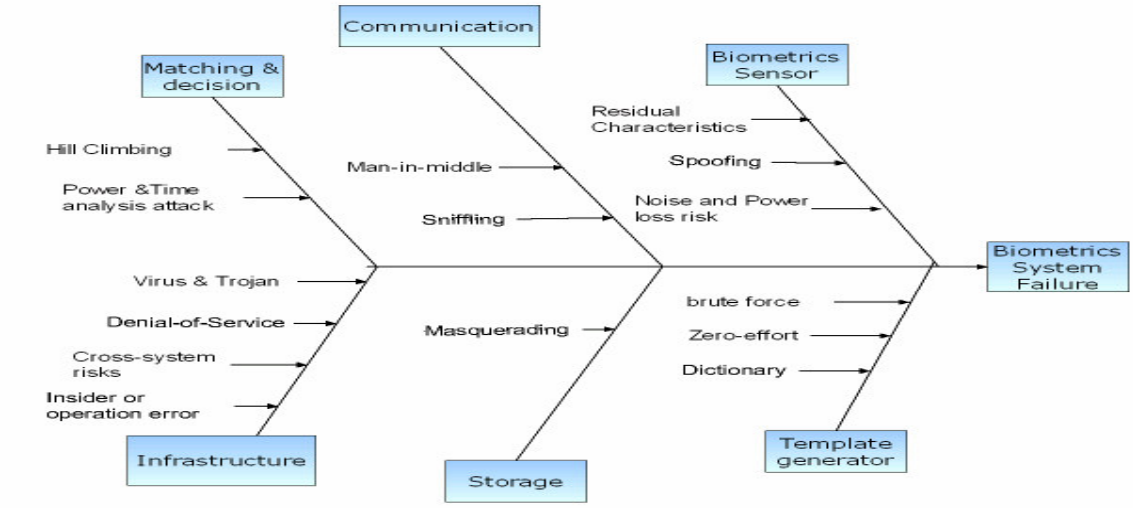


Figure 6: A fishbone diagram illustrating the attack techniques on locations of biometrics systems.

The attack techniques presented in the Figure 6 are applied to the major components of sensors, communications, matching, infrastructure, storage and signature or template generator. These components can be subjected to physical, logical and electronic attack in order to defeat the biometric system.

Authorised person identification is a major application of the biometric technique with further emerging applications in the security field. However it should not be assumed that biometric security technology is the key to asset protection but a component of security protection strategy. Even though the biometric technology drives the future direction of strong authentication, its use will require time to mature due to the need to balance all aspects of system integration, privacy and security, reliability and countermeasures to attack as well as the accompanying problems of testing and evaluation, operating standards, ethical issues and cost benefit analysis of installing the applications.

INTELLIGENT IMAGING

The security industry is currently experiencing a massive transformation as many of the traditional analogue technologies that have been applied to the protection of assets, are replaced by the digital innovation. The conversion of analogue signaling from sensors to digital outputs allows comprehensive signal processing with the quantum step in level of intelligence that can be incorporated in the system. This transformation has set the scene for future major improvements in quality of processing of signals from sensors, providing levels of intelligence in the capacity of system to discriminate intruder from an unwanted alarm condition (Smith, 2006).

All modes and types of security systems have and will benefit from fundamental improvements in the applications of physics and engineering on the detection of objects, the transmission of signals, and processing of the signals for an anomalous output. These security systems include internal and external intrusion detection systems, access control systems, integrated systems, smart buildings and information security systems. The improvement in quality and performance of security systems is increasing exponentially (Smith, 2004) with exceptional performance currently delivered by a selection of systems (Cano, 2004).

Intelligent CCTV

The development of enhanced CCTV systems through automation and intelligence has converted these systems into powerful and responsive management tools. There is a trend for CCTV to be further integrated with other surveillance and access control technologies to provide an effective multi-functional security system. The integration of CCTV images and computer processing is developing a powerful tool for the detection of intruders.

A form of intelligent CCTV is the video motion detection (VMD) application that uses a camera system as a means of intruder detection. The earliest intruder image processing algorithms employed frame differencing, where the previous frame is compared with the current frame, is a technique that still forms the cornerstone of many of the current sophisticated surveillance systems. The technique requires stored sample frames from the camera input which are compared to subsequent frames of the camera to identify changes in picture detail. Some problems that occur with this approach in VMD are:

- No concept of image understanding.
- No exploitation of domain knowledge.
- No knowledge of the target and its movement.
- Susceptible to camera movement and weather effects.

However, a VMD system has the capability of presenting several detection zones simultaneously on screen. A change in one or more of these zones can activate an alarm and cause the system to commence recording. Although VMD is more suitable to indoor applications as changes in light, shadows or small moving objects in the internal application, as the exterior applications have a tendency to cause false alarms. However, exterior applications have been developed to minimise the negative features of VMD by filtering out unwanted alarm conditions.

The Advanced Exterior Sensor (AES) system (Ashby and Pritchard, 2004) integrates the three sensor technologies of thermal infrared imaging, visible light imaging, and microwave radar in a system that scans a full rotation in about a second. Surveillance of wide areas is possible from the three sensors, with images from the infrared and visible detectors and the radar range data being upgraded each rotation (Figure 7). The range information has a resolution of about a metre, and the panoramic imagery is examined for change. This system has the capacity to be applied at airport runways, oil refineries and gas facilities, and other infrastructure locations where large open areas are present.

Figure 7: Advanced Exterior Sensor (AES) system integrates the three sensor technologies of thermal infrared imaging, visible light imaging, and microwave radar.

Multiple video camera surveillance systems can be applied to monitor an environment where there is a limit to the perceptions of human observers; that is, many events are occurring simultaneously. The application of automated systems require the development of image processing and computer vision to detect, locate, and track targets as they move through the field of view. Multiple cameras can be applied in several ways to the surveillance of the locality by:

- Spatially adjacent cameras extend the coverage of the surveillance.
- Overlapping views provide redundancy of the images, so as to minimise the ambiguities of occlusion, and maximise the accuracy of position determination.

Tracking objects that appear simultaneously as images in two or more cameras can be used to minimise the effects of occlusion (unlikely to occur in both views at the same time). A mapping function relates the location of a pixel in one view with the same pixel in the view from the other camera. This pixel mapping function is called homography and can be determined by locating a minimum of four equivalent points in two views. However, the points must lie in the same plane, but this condition is often satisfied for surveillance systems in constructed environments, where tracked objects will share a common ground plane (Smith, 2006).

The trend to develop specialised video surveillance tools has seen a range of approaches being applied to the detection of people. A single or multi-camera system has the capacity to detect, identify, classify, and track objects moving through a surveillance field. Particularly, the system has the capacity to detect people traveling the wrong way in crowded environments, such as airport security exits. The intelligence in the system has the ability to distinguish between potentially threatening activities and environmental events such as trees blowing in the wind, waves on the shore, water reflections, and tidal movements (Smith, 2006).

Testing Imaging Systems

Intelligent CCTV and imaging systems are complex technologies that require dedicated testing protocols and procedures in order to determine the effectiveness of the system in real time. The characteristics of these systems are such that some degree of ineffectiveness can be exploited in order to defeat the system. Experience has shown that environmental conditions are a major contributor to the failing of intelligent CCTV in real imaging situations where the monitoring of people and activity are the principal objective of the imaging system. Some issues of concern with intelligent CCTV are:

- Wind blowing vegetation.
- Low contrast in poor lighting conditions.

- Water flowing across surfaces and fountains.
- Movement of small objects in field of view.
- Fast moving objects in field of view.

These activities are typical of those that cause problems for intelligent CCTV systems, and unnecessarily produce alarm conditions. In sterile situations, intelligent CCTV generally functions according to specifications, but in real situations a multitude of conditions can produce alarm conditions.

CONCLUSION

The evaluation of the reliability and validity of sophisticated security systems is crucial in the application of a risk model to the protection of assets of an organisation. The application of Australian Standard / New Zealand Standard AS/NZS 4360 – *Risk Management* is critical to the development of protection of assets strategies for a facility, and establishment of levels of security required sustaining an attack on an asset. These strategies will provide the parameters for the quality and quantity of security technology appropriate to protect the assets according to the security management plan. The role of security equipment in the protection of assets of an organisation is determined by security strategies applied in the security management plan.

The reliability and validity of the security technologies will be determined by a comprehensive testing programme where the effectiveness and appropriateness of security systems are evaluated within the bounds of national and international standards. The concept of applying national and international standards to the testing and evaluation of security equipment must be considered in the context of a proposed model developed which outlines a formal process for testing and evaluating security equipment for specific security strategies formulated by the client as a result of the 'risk treatment' element of the AS/NZS 4360 risk management process. AS/NZS 4360 is appropriate for commercial and national infrastructure facilities, and as such has been considered as a suitable model for establishing the security strategy criteria within which security technologies are tested and evaluated against for reliability and validity. Testing of security technology for performance and for defeat is a crucial element in the assessment of the effectiveness of sophisticated security systems.

REFERENCES

Ashby R. and Pritchard, D.A. (2004) Seeing Beyond the Perimeter: The Advanced Exterior Sensor (AES). Proceedings of the IEEE 38th Annual 2004 International Carnahan Conference on Security Technology, pp182- 188.

Cano, L.A. (2004) Smart Sensor Integration into Security Networks. Proceedings of the IEEE 38th Annual 2004 International Carnahan Conference on Security Technology, pp82- 84.

Garcia, M.L. (2001) The Design and Evaluation of Physical Protection Systems. Boston: Butterworth-Heinemann.

Haslam, S.A. and McGarty C. (1998) Doing Psychology. Wiltshire: Sage Publications Ltd.

Jones, D.E.L. and Smith, C.L. (2005) The Development of a Model for Testing and Evaluation of Security Equipment within Australian Standard / New Zealand Standard AS/NZS 4360:2004 – *Risk Management*. Recent Advances in Counterterrorism Technology and Infrastructure Protection. Proceedings of the 2005 Science, Engineering and Technology Summit Canberra, 2005

Qinghan Xiao (2007) Biometrics–Technology, Application, Challenge, and Computational Intelligence Solutions. IEEE 2007(5). 5-25.

Peck, L. and Lacombe, J. (2007) Seismic-based Personnel Detection. Proceedings of the IEEE 41st Annual 2004 International Carnahan Conference on Security Technology, pp169-175.

Smith, C.L. (2004) The Development of a Security Systems Research and Test Laboratory at a University. Proceedings of the IEEE 38th Annual 2004 International Carnahan Conference on Security Technology, pp111- 115.

Smith, C.L. (2006) Handbook of Security: Trends in the Development of Security Technology. New York: Palgrave Macmillan.

Uludag, U. (2006) Graduate Psychology: Secure Biometric Systems. *Michigan State University.*

Wegner, M.J. and Spyridakis, J.H. (1989) The Relevance of Reliability and Validity to Usability Testing. *IEEE Transactions on Professional Communication,* 32(4), pp265-72.

Triangulation Based Static Wide Angle Laser Scanning For Obstacle Detection

K. Sahba[1], K. E. Alameh[2] and C. L. Smith[3]

[1]Western Australia Centre of Excellence for MicroPhotonic Systems
School of Engineering and Mathematics
Edith Cowan University, Australia
E-mail: k.sahba@ecu.edu.au

[2]Western Australia Centre of Excellence for MicroPhotonic Systems
Edith Cowan University, Australia
E-mail: k.alameh@ecu.edu.au

[3]School of Engineering and Mathematics
Edith Cowan University, Australia
E-mail: clifton.smith@ecu.edu.au

ABSTRACT

This paper demonstrates discrete laser spot projection over a wide angle using a novel cylindrical quasi-cavity waveguide, with no moving parts. Furthermore, the distance to each spot is calculated using active laser triangulation. The triangulation arrangement and the trajectory of principal rays are modelled using a system of linear equations based on optical geometry. Linear algebra is used to derive the unique baseline and outgoing angle of every projected beam. The system is calibrated by finding optimal values for uncertain instrumental parameters using constrained non-linear optimization. Distances calculated indoors result in accuracies of over 93%.

INTRODUCTION

High power, high pulse rate, fibre and diode eye-safe lasers has allowed terrestrial laser scanning (TLS) to become a highly advanced optical detection device within the defence-in-depth (DiD) model, while also being utilised in civil and military remote sensing. Incorporating beam deflection units, TLS makes multiple range measurements, producing a three-dimensional (3D) reconstruction of the scanned area. In terms of perimeter security, this implies monitoring the perimeter contour and checking for differences in 3D shape between the latest scan and that of a reference, or unperturbed perimeter.

Currently, TLS laser beam deflection is achieved by electro-mechanically driving single or multi-faceted rotating polygon mirrors, sometimes in combination with galvanometric/resonant mirrors. Examples are given in Figure 1. The system depicted in Figure 1(a) contains a motor driven rotating polygonal mirror and a sweeping mirror driven by an actor. The outgoing beam is first deflected vertically by the sweeping mirror and then impinges on a polygonal mirror fact which then deflects the beam horizontally, thus producing a raster scan of the surroundings as described in U.S. Patent No. 6480270 (2002). Figure 1(b) shows a single sided, or 'monogon' mirror deflecting a laser beam as it is rotated by a motor, in conjunction with an angle encoder (SICK, 2003). In Figure 1(c), a fibre optic cable is aligned with the input aperture of a beam expander which guides the beam to the oscillating mirror. The resonant motor oscillates the mirror in the order of 100Hz. A sinusoidal scan pattern occurs because the resonant scanner assembly and window are moving simultaneously as one entity, in azimuth and elevation. The stepper motor within the upper

portion of the scan head rotates the head at a rate of 1-2Hz. Figure 1(c) is drawn from U.S. Patent No. 6985212 B2 (2006) and Ray, Evans & Jamieson (2005).

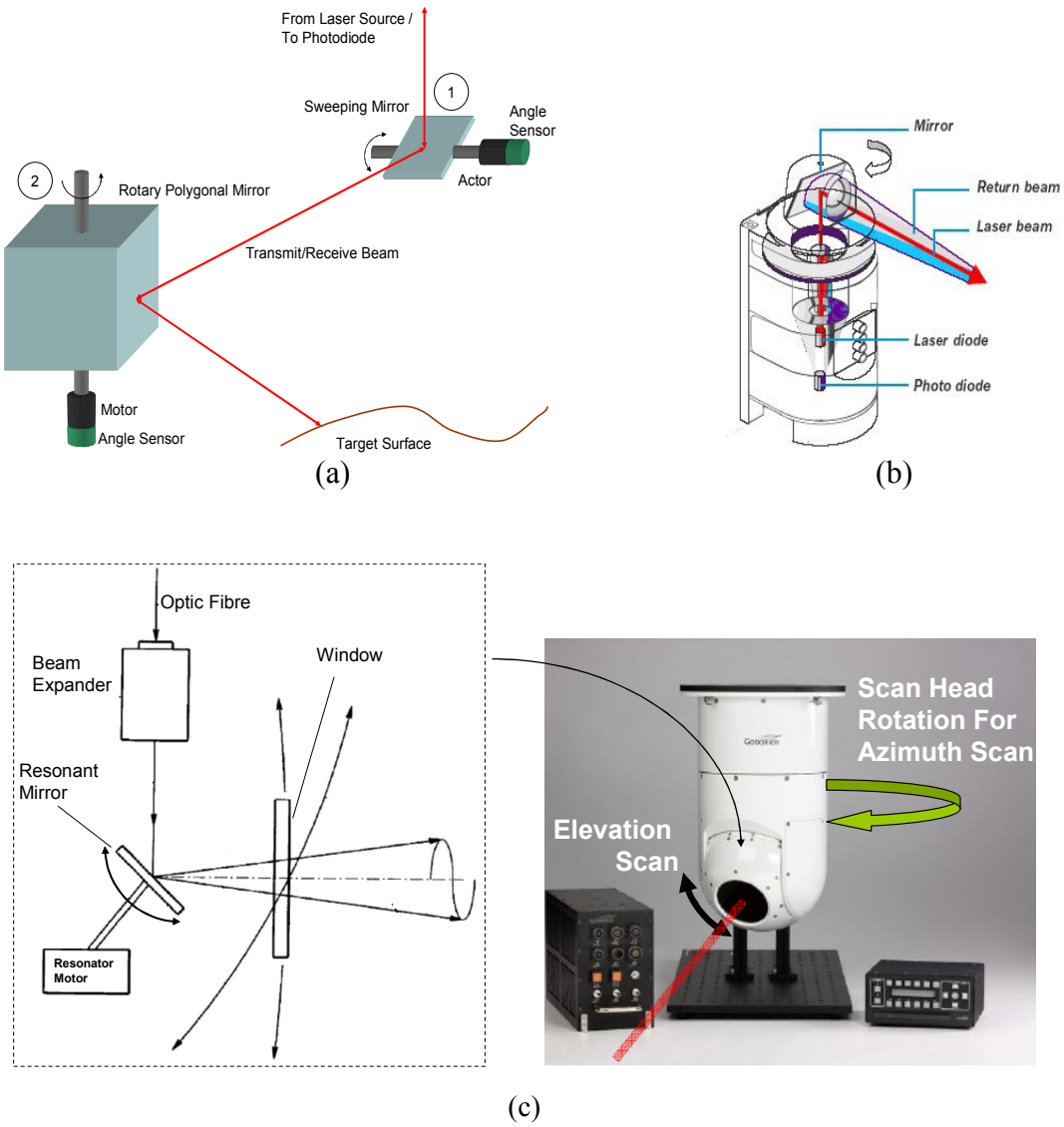


Figure 1. Examples of current beam deflection methods.

The intrinsic motion of the mirror and servo motor causes scanning problems which are the hardest to quantify and correct. Acceleration time taken to reach the constant scanning speed from a stationary position can result in range measurements being stored at the wrong corresponding points in the scanned image. Additionally, all mirrors and measurement components must be synchronized exactly. The performance of polygonal scanners, especially with respect to maintaining an accurate deflection beam path, is prone to manufacturing tolerances. Dynamic track and jitter errors are caused by tolerances for polygon machining errors, mounting errors, mounting-hub errors, random wobble caused by ball bearings, motor cogging and torque variations (Stutz, 2005).

Cogging, torque and facet flatness variations cause errors in the actual scan line. Other problems listed with rotational scanners are (Stutz, 2005):

1. Synchronization with other time-dependent elements in the system is rather difficult.
2. Motor stability and durability at higher rotation speeds also present problems.

3. There is an upper limit to the rotation speed due to the tensile strength of the mirror material. The mirror must not disintegrate at the maximum rotation speed.

Vibrations and shock of the whole scanner housing also cause errors in range measurements as the rotating mirror/s become out of phase. In general, a multi-beam stationary optic approach is much less sensitive to vibration (Taylor, et. al. 1998). Also, mirror device scanners are slow, bulky and expensive (Elkhalili, 2004) and being inherently mechanical they wear out as a result of acceleration, cause deflection errors and require regular calibration (Schnadt & Katzenbeißer, 2004). A comprehensive description of rotary mirror scanning errors can be found in Marshall (2004).

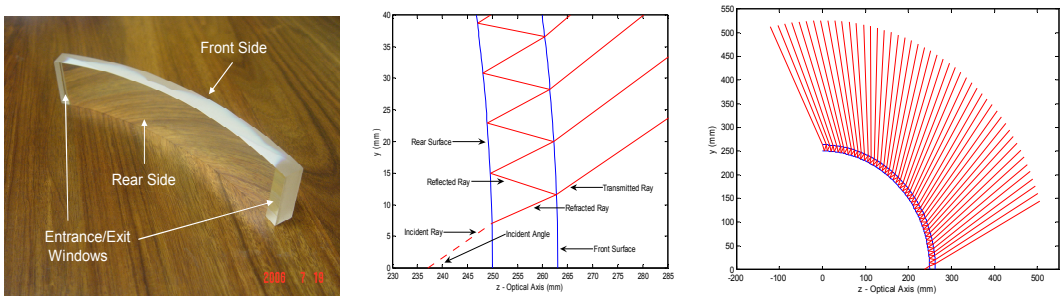
In this paper, an approach to generating multiple beams over a wide angle with no moving parts and deriving the range to the corresponding laser spots falling on the surrounding perimeter is demonstrated, within a proof-of-concept phase. A novel optical piece in the form a quasi-cylindrical cavity with a 45º curvature has been fabricated and used to perform structured laser light projection in the form of a spot array. This component acts as a waveguide as an incident beam undergoes multiple reflections within its two dielectric cylindrical surfaces which share the same centre. By depositing a partially transmissive nano-layered thin film, a percentage of light is transmitted at every intersection with the outer interface.

Ray propagation modeling using linear algebra is demonstrated and applied to predict the unique baseline location and outgoing angle of every outgoing beam in the arrangement. The arrangement is modelled using a system of linear equations and the experiment's principal rays and components are plotted using equations for straight lines and circles. The concept is demonstrated experimentally by adding the quasi-cylindrical optical cavity to a conventional active laser triangulation layout, imaging each spot with a CCD imager and a TV lens. Each spot's sub-pixel peak intensity position is estimated using an appropriate Gaussian peak estimator algorithm. Coupled with the modelled beam angle and baseline parameters, the forward distance to each spot is estimated.

MULTI-LASER BEAM GENERATION USING A QUASI-CYLINDRICAL OPTICAL CAVITY

The custom fabricated concentric concave-convex cavity of 45º curvature is shown in Figure 2(a). It comprises an inner and outer dielectric interfaces of radii R_1 and R_2, respectively, separated by a BK-7 glass medium of thickness $d = R_2-R_1$, and non-coated entrance and exit windows. Light transmission is achieved by depositing nano-layered thin film coatings on both interfaces. The rear side is deposited with a highly reflective coating (R≥99%) and the front side with a partial transmission coating (T≤13%), effective over the 600 – 900nm waveband.

At every reflection with the outer interface, a fraction of the light is transmitted through the cavity thus generating a laser spot. The reflected power undergoes further reflections within the cavity to generate subsequent laser spots as illustrated in Figure 2(b). A plot illustrating the ray trace for a 90°-cavity is shown in Figure 2(c) for R_1=0.25m and R_2=0.263m. The inter-beam spacing and angular resolution are adjustable, according the incident angle of the injected laser beam through the entrance window.



(a) (b) (c)

Figure 2. (a) quasi-cylindrical cavity, (b) entrance close-up plot and (b) ray trajectory over 90º.

To trace the path of the reflected and transmitted rays, optical geometry is used with no approximation. Given the radii of curvature of the two interfaces, the slope and ray intercept with the inner interface, the two surfaces and rays are modeled using a system of basic linear equations. Rays are plotted as straight lines, given by

$$y = mz + b, \qquad \text{Eq. 1}$$

where m is the gradient and b is the y-intercept. The two interfaces are plotted as arcs where

$$y = \sqrt{R_n^2 - z^2}, \qquad \text{Eq. 2}$$

where R_n is the interface radius.

Intersections between rays and the surfaces are represented by algebraic solutions to equations (1) and (2). Internal interfaces are modeled as cylindrical mirrors; hence the incidence and reflection angles of a ray are equal.

Gaussian beam behaviour in terms of the spot size evolution has been reported in (Sahba, 2007).

ACTIVE TRIANGULATION GEOMETRY FUNDAMENTALS

Figure. 3 illustrates the geometry principle of active laser triangulation. An imaging device of focal length f is positioned in line with the Z-axis and has the X-axis running through its lens centre. To the left of the lens, at a baseline distance β, a laser source launches a light beam at a variable angle w. The image point lying on the x plane, together with β, w and f determine the X and Z coordinates of the illuminated target point P. The horizontal and vertical distances, X and Z, to the projected laser spot from the lens centre are given by Hartrumpf & Munser, (1997):

$$X = x \cdot \frac{f \tan(w) - \beta}{x - f \tan(w)}, \qquad \text{Eq. 3(a)}$$

$$Z = f \cdot \frac{x - \beta}{x - f \tan(w)}. \qquad \text{Eq. 3(b)}$$

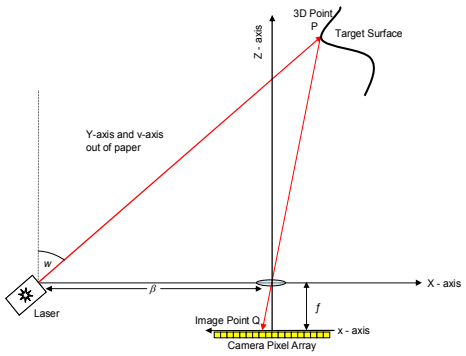


Figure 3. Basic principle of active laser triangulation.

LASER TRIANGULATION SETUP INCLUDING THE OPTICAL CAVITY

Figure 4 illustrates principal rays of the novel triangulation system incorporating the quasi-cylindrical optical cavity. Adopting the linear algebraic ray tracing method from (Sahba, 2007), cavity interfaces and rays are plotted using equations for circles and straight lines, respectively. Figure 4 indicates that each outgoing laser beam has unique β and w values in relation to the rotated lens line, L.

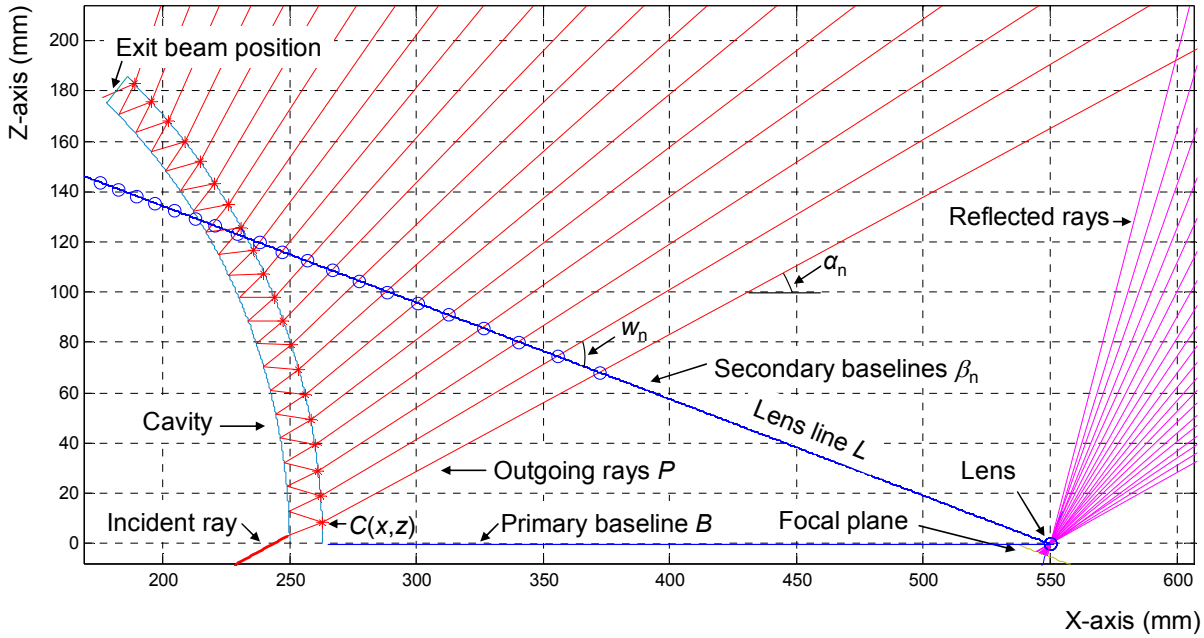


Figure 4. Principal ray plot of the arrangement, showing outgoing and reflected rays, baselines, the quasi-cavity, objective and imager.

In calculating the angle between the nth emerging ray and the X-axis, α_n, it was found that relying on the law of reflection and Snell's law produced theoretically correct values but did not predict the real angles. This could be attributed to several factors, namely, a non-homogeneous cavity substrate, an imperfect cylindrical shape or non-uniform dielectric thin film coatings. To derive accurate a_n values, the laser incident angle and cavity index defined in the model were altered so the exit beam position, shown in Figure 4, coincided with the real point of exit at the end of the cavity, within 1mm accuracy. For any outgoing ray, the predicted coordinates of beam intersection with the outer cavity surface, $C(x,z)$ were recorded.

The coordinates $S(x,z)$ of the corresponding laser spot on the laboratory wall were recorded manually by hand. In this case, the angle of a ray with respect to the X-axis is given by:

$$\alpha = \tan^{-1}\left(\frac{C(z)-S(z)}{C(x)-S(x)}\right). \quad \text{Eq. 4}$$

Each ray's outgoing angle with respect to the lens line of slope L_m is:

$$w = \tan^{-1}\left(\frac{\alpha - L_m}{1+(L_m \cdot \alpha)}\right). \quad \text{Eq. 5}$$

The baseline distance, β is calculated as:

$$\beta = \sqrt{(\beta(x)-L(x))+(\beta(z)-L(z))}, \quad \text{Eq. 6}$$

where

$$\beta(x) = \frac{L_b - P_b}{P_m + L_m}, \quad \text{Eq. 7}$$

and

$$\beta(z) = \frac{(P_m \cdot L_b)-(L_m \cdot P_b)}{P_m \cdot L_m}. \quad \text{Eq. 8}$$

P_m, L_m and P_b, L_b are the slope and y-intercept of a projected ray and the lens line, respectively.

Figure 5 shows the experimental setup that was used to demonstrate active triangulation using the quasi-cylindrical optical cavity of 45° curvature. Twenty laser spots were generated when an incident laser beam was injected through the entrance window. The cavity's orientation and position were adjusted using a tilt and precision translation stage, respectively.

The incident beam was produced by a 632.3nm, 1mW, HeNe laser, which was mounted onto a rotating stage with a 0.5° step. To image the projected laser spots, a ½" interline transfer CCD imager was employed, comprising of 768(H) × 494(V) pixels of size 8.4 × 9.8μm. A C-mount TV lens of focal length f = 12.5mm, focused at infinity collected the reflected laser light. The lens iris was adjusted appropriately to avoid saturation of the imaged spot. Images from the camera were digitized in 12-bit form using a Spiricon Plug and Play PCI frame grabber.

Both the quasi-cavity and camera were staged on a common rail, the distance between them defining the primary baseline, B = 0.3m, as shown in Figure 5.

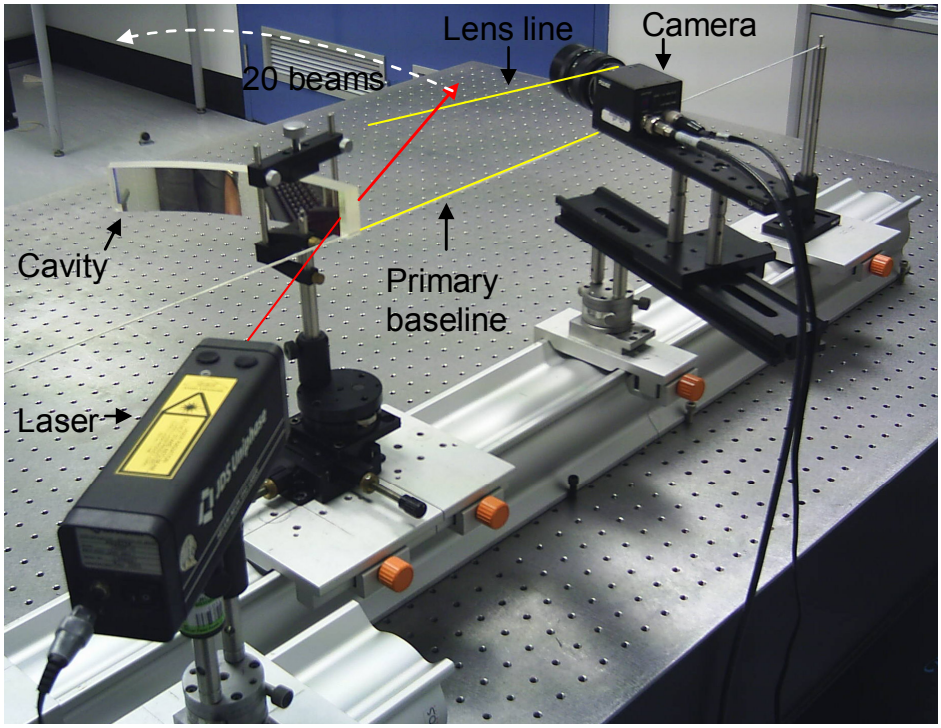
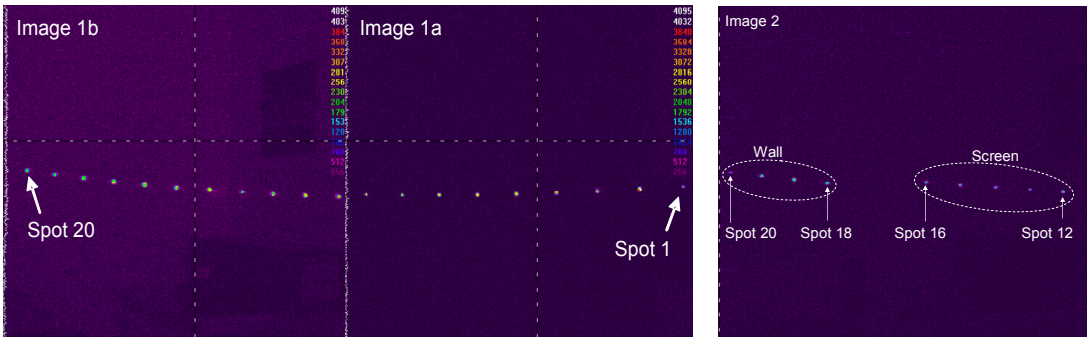


Figure 5. Photo of the experimental setup for active triangulation using the quasi-cavity.

Two sets of range measurements were taken to validate the system. Firstly, all the laser spots were projected onto the laboratory walls and imaged. Since the field of view (FOV) of the camera lens is not wide enough to capture all spots instantaneously, the camera was rotated about the lens center in order to acquire two images containing 9 and 11 laser spots, which are shown in Figure 6(a). The camera angle, L_{θ}, was set at 46º and 69º for images 1a and 1b respectively, with respect to the X-axis.

For the second set, a screen was placed 2.5m away from the camera lens centre. Thus spots 17 to 20 were projected onto the screen and the remainders were projected onto the wall. One image was taken of spots 12 to 20 to demonstrate that a closer object's range, with respect to the wall, could be accurately determined. L_{θ}, was set at 64.5º. The image is shown in Figure 6(b).



(a) (b)

Figure 6. The laser spot arrays produced by the quasi-cylindrical optical cavity.

For each image, an intensity profile was taken across every spot and the digital pixel array processed to find the peak intensity pixel position using the Gaussian sub-pixel peak position estimator algorithm defined by Fisher & Naidu (1996) as:

$$\delta = \frac{1}{2}\frac{\ln(f(x-1)) - \ln(f(x+1))}{\ln(f(x-1)) - 2\cdot\ln(f(x)) + \ln(f(x+1))}, \quad \text{Eq. 8}$$

where x is the pixel position of the observed peak sensor reading with an intensity of $f(x)$. The peak position of a laser beam spot imaged at pixel position x is offset by the estimated pixel fraction δ.

To calibrate the system, constrained nonlinear optimization was used to find the optimal values of parameters subject to uncertainties. The optimization was based on the minimization of the least square error, namely:

$$\sum_{i=1}^{20}\left(Z_i - \hat{Z}_i\right)^2, \quad \text{Eq. 9}$$

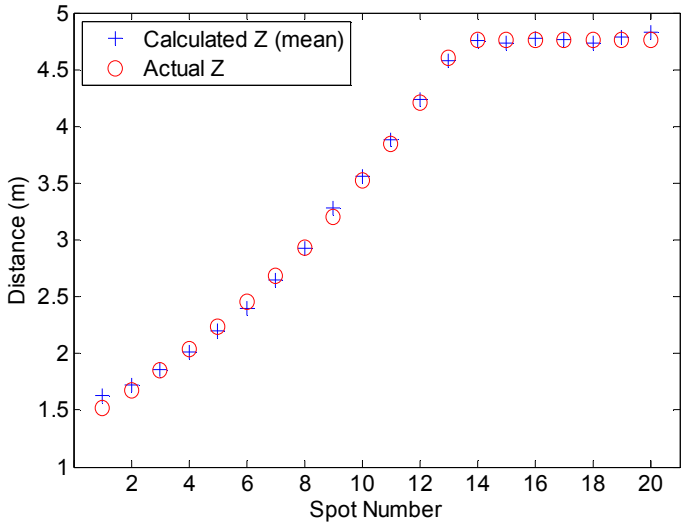
where Z_i is the actual range and $\hat{Z}_i$ is the calculated range for the i^{th} laser spot. Note that for the second set of ranges, $i = 12$, since only one image was acquired of spots 12 to 20. The two most significant uncertainties in the experimental setup were:

1. Imager pixel size. Although already manufacturer-specified, the pixel pitch was not known, thus producing error in calculating the captured ray's physical position on the image sensor. Hence, a pixel size scaling factor, η, was used.
2. L_θ is not totally accurate since the camera's rotating stage was fixed onto the rail using a single bolt, hence making it subject to rotational movement due to slight knocks or vibrations.

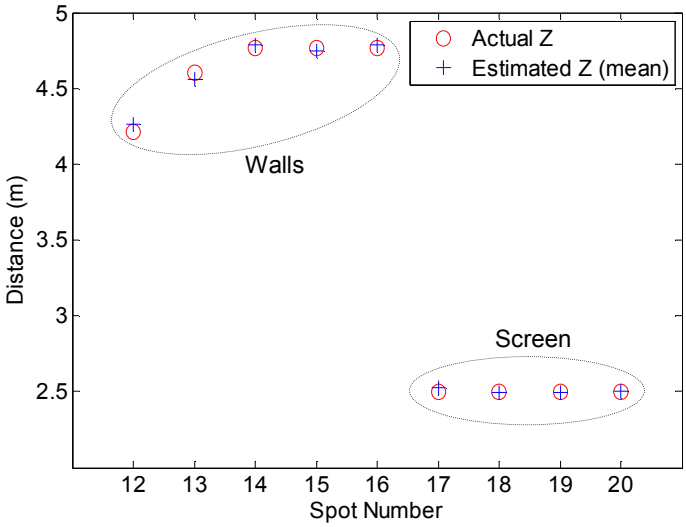
Other uncertainties included the exact position of the lens centre along the optical axis within the complex TV lens system and the centre-to-centre alignment of the image sensor and lens. An alignment error attributed to the primary baseline, B, running directly through the lens centre can also exist. Note that if the camera is displaced vertically, the lens will not be in the same plane as B, and this leads to inaccuracy in range measurements.

RESULTS AND DISCUSSION

Four frames were taken at each L_θ angle and the mean forward range, Z, to every spot obtained as shown in Figure 7. The first set of estimated range measurements, for spots 1 to 20 projected on the laboratory walls, is shown in Figure 7(a). The second set, from spots 12 to 20, is shown in Figure 7(b).



(a)



(b)

Figure 7. Range measurement results. (a) shows the estimated ranges for 20 laser spots falling on the laboratory walls. (b) shows the estimated ranges for spots 12 to 16 projected onto the walls and spots 17-20 onto a perturbing screen.

Note that Figure 7 shows very good agreement between the measured and calculated ranges, validating the method of deriving β and w using linear algebra as described in previously. Note also that Standard deviation of the mean Z remained below 7.68cm and 0.525cm for measurement sets 1 and 2 respectively, demonstrating system stability. The minimum ranging accuracies achieved were 93.63% and 98.81% for spot 1 of measurement set 1 and spot 12 of measurement set 2, respectively.

Table 1 shows estimated and optimized η and L_θ. Constrained nonlinear optimization of these parameters was carried out using Matlab®'s `fmincon` search function in the Optimization Toolbox.

Table 1. Optimized triangulation parameters.

	Image 1a		Image 1b		Image 2	
	η	L_θ	η	L_θ	η	L_θ
Min.	0.5	42	0.5	42	0.5	62
Max.	1.5	52	1.5	52	1.5	72
Estimate	1	46	1	72.5	1	68.5
Optimal	0.7744	50.107	0.9944	72.974	0.9825	69.507

CONCLUSION

This paper has demonstrated a novel method for wide angle laser pattern projection in the form of a spot array through multiple internal reflections and refractions. The effect of an off-axis optical system with respect to ray trajectory has been demonstrated. By conjoining as many quasi-cavities as needed, the scanned angle can be extended to 360°. Furthermore, scanning in elevation can be achieved by stacking the quasi-cavities vertically.

Accurate triangulation-based ranging to multiple laser spots generated over a wide angle by the custom quasi-cylindrical optical cavity has been demonstrated. A system of linear equations has been used to simulate the principal rays and components of the triangulation system and obtain the unique baseline distance and outgoing angle of every beam. It has also been shown that the imaging device can be rotated about its lens centre in order to triangulate to a spot array wider than its instantaneous FOV.

Ranging accuracy is heavily dependent on precise system instrumentation of the system arrangement. Calibration has been achieved by non-linearly optimizing values for uncertain instrumental parameters within realistic constraints.

Potential implications of this novel scanning architecture include a longer mean time between failure, and virtually no wear and tear as there are no moving parts. This is particularly beneficial in applications where robustness and mean time between failures is critical, such as perimeter security, collision avoidance and robot navigation.

REFERENCES

Elkhalili, O., Schrey, O. M., Mengel, P., Petermann, M., Brockherde, W. and Hosticka, B. J. (2004). A 4 X 64 pixel CMOS image sensor for 3-D measurement applications. Institute of Electrical and Electronic Engineers, Journal of Solid-State Circuits, 39, 1208-1212.

Fisher, R. B., Naidu, D, K., "A Comparison of Algorithms for Subpixel Peak Detection," in Sanz (ed.) *Advances in Image Processing, Multimedia and Machine Vision* (Springer-Verlag, 1996).

Hartrumpf, M., Munser, R., "Optical three-dimensional measurements by radially symmetric structured light projection," Appl. Opt. 36 (13) (1997), http://www.opticsinfobase.org/abstract.cfm?URI=ao-36-13-2923.

L. F. Marshall, Handbook of Optical and Laser Scanning (Marcel Dekker Inc., 2004), Chap. 4.

Ray, M. D., Evans, O., & Jamieson, J. R. (2005). Three dimensional laser radar for perimeter security. Electro-Optical Remote Sensing, Proc. of SPIE, 5988, (598806).

Ray, M. D & Jamieson, J. R. (2006). *Laser Perimeter Awareness System*. U.S. Patent 6985212 B2.

Sahba, K., Alameh, K. E., Smith, C.L., and Paap, A. (2007). Cylindrical quasi-cavity waveguide for static wide angle pattern projection. Optics Express. vol. 15, pp. 3023-3030.

Schnadt, K. and Katzenbeißer, R. (2004). Unique Airborne Fiber Scanner Technique for Application-Oriented LIDAR Products. Proceedings of the International Society of Photogrammetry and Remote Sensing Working Group VIII/2, 36.

SICK AG, Division Auto Ident. (2003). LMS 200/LMS 211/LMS 220/ LMS 221/ LMS 291 Laser Measurement Systems Technical Description. URL: www.sick.com.

Studnicka, N. & Ullrich, A. (2002). *Method for monitoring objects or an object area*. U.S. Patent 6480270 B1.

Stutz, G. (2005). Guiding Light, in *SPIE's oemagazine*, *5*, 25-27.
URL: http://oemagazine.com/fromTheMagazine/apr05/tutorial.html

Taylor, C., Barlett, D., Chason, E. and Floro, J. (1998). A Laser-Based Thin-Film Growth Monitor. *The Industrial Physicist.*4, pp 26-30.

Improving Security for Vision Impaired ATM Access via Dynamic Patterns

D. Veal

School of Computer and Information Science
Edith Cowan University (ECU)
Australia
Email: d.veal@ecu.edu.au

KEYWORDS

Blind, Visual Disability, Automatic Teller Machines (ATMs), Americans with Disabilities Act (ADA), Web Access for the disabled, Adaptive Technology, DPS, GUI design.

ABSTRACT

There is an access problem for many bind and deaf-blind users of ATMs, online banking and the Web. This paper looks at a proposed potential solution for those users who, though classified as legally blind can still recognise on screen patterns representing textual characters, words or symbols. Such a system, the Dynamic Pattern System (DPS), has been developed to enable the matching of such patterns to a visually disabled person's remaining vision system. This paper also considers some of the security problems posed by the blind and the deaf-blind undertaking routine banking operations and use of the DPS as a potential aid.

INTRODUCTION

Even for the fully sighted Automatic Teller Machines (ATMs) are a potential security hazard. Introna cites cases of a loop of tape fitted into the card slot of an ATM whereby a rogue user queuing behind the victim notes that they had a similar problem recently and just entered their key code and it worked. When the victim re-entered password it did not work and the card was still held in the machine by the loop of tape. However the victim's key code was noted by the rogue user who later recovered the card and reused it after the victim had left the scene (Introna and Whittaker, 2005). However, there are additional potential security problems with some visually disabled person's access to ATMs or Web transactions that are not experienced by many fully sighted or non-disabled individuals. The visually disabled may need to depend upon the assistance of non-blind friends, relatives or neighbours. This means that there is a potential for fraud to occur in such transactions when those passwords and other details that should remain secret are shared with a helper. Such sharing confidential information may not be necessary because although most people legally registered as blind have some residual vision (RBS, 1996). Although, this residual vision may not be sufficient to read enlarged print (DoJ, 1994) they may be able to see large on screen objects on a computer monitor. Hence there exists the possibility of such users being able to undertake there own banking operations by adapting the visual output to utilise this ability to see large on screen objects by translating convention text into such objects. The Dynamic Patterns System (DPS) is a system that uses this principle.

Some people are not only legally blind but may also be deaf. This can lead to extra challenges in accessing information. The problems of providing access to information for the deaf-blind have been the subject of literature as long ago as the nineteenth (Bell, 1883) and early twentieth centuries (Keller, 1903). The availability of cheap modern powerful personal computers has led to much research in this area. The importance of GUI design for visually disabled users has been investigated by Fraser (Fraser and Gutwin, 2000). Furthermore, in

many developed economies equality legalisation such as the Americans with Disabilities Act (ADA) in the U.S. (DoJ, 1994) may require that employers and providers of services make available facilities and special equipment, if needed, for the disabled. The need to utilise existing skills and talents, has led to more research into this area. Yet Stevens notes that: *'The importance of the World Wide Web for information dissemination is indisputable. However, the dominance of visual design on the Web leaves visually disabled people at a disadvantage'* (Yesilada et al., 2007). There is also a difficulty of blind and deaf-blind users not only using web browsers for online access (Goble et al., 2000) but also using high street ATMs. Instead of using standard text for ATMs there exists a possibility of using patterns enabling users that can perceive coloured areas of a screen could also use such patterns as a substitute for textual information.

THE DYNAMIC PATTERN SYSTEM (DPS)

The DPS has been designed to enable the presentation of text as a sequence of on-screen patterns to a visually handicapped user (Veal and Maj, 1998). These patterns can be matched to the individual's remaining visual ability thereby permitting the user to select the size and colour of the elements that constitute each pattern. Each of these patterns can represent a character. As visual span (Legge et al., 1997) may well be narrow each pattern is presented singularly on the screen for a given time. A sequence of such patterns presented to the user can represent a word. The time between words can be greater than that between patterns as is the case in Morse code (Ching-Hsing and Ching-Hsing, 1997). The presentation time between words represented by a pattern for the white space whose time can also be varied. Such a variation allows words and sentences to exhibit a form of rhythm and so have some commonality with music. As a pattern could represent a letter, a word or an event, therefore DPS could also be said to have some commonality with marine signalling flag systems: '*These flags are international signals used by ships at sea to spell out short messages, or more commonly, used individually or in combination they have special meanings'* (Maineharbors, 2005). The DPS has an associated GUI and hence allows a sighted facilitator to adjust the timing, and size of patterns to align with the visually disabled user's preferences (Figure 1) and also to hide the controls for effective testing (Figure 2).

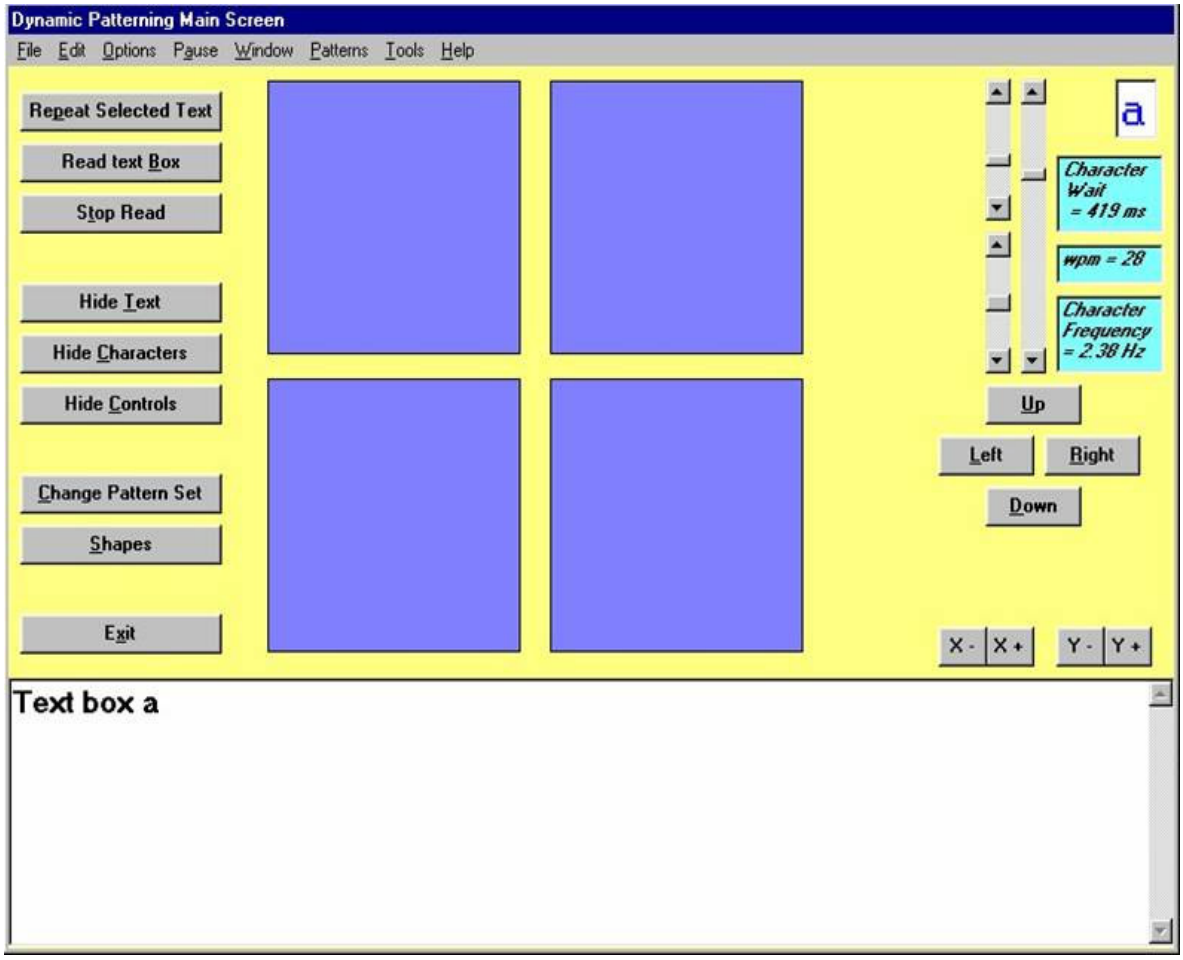


Figure 1 A typical DPS pattern control screen

The DPS provides timing and pattern element size separation readings to assist effective data collection and a user's preferred reading speed (Veal and Maj, 1998). This information can be saved along with the user's preferred set of patterns and be reloaded automatically. This set of patterns is known as a patternset. An advantage of such a system is that it uses an existing computer and monitor and does not require any additional technology to be provided (Veal and Maj, 2001). The current version is DPS restricts the delivery rate of patterns to be presented to disallow those that could possibly induce photosensitive epilepsy (Tokyo, 1997).

However, the memorisation of a whole patternset, which includes alphanumeric characters and some punctuation (Figure 3), is an arduous task and would only be worth undertaking should the person not be able to read enlarged text. Automatic pattern generation matching an individuals' patternset to their remaining vision system could help to reduce the time spend on matching patternsets to individual users. Singh has noted the redundancies in the English language mean that not all characters need to be recognised for the meaning to be conveyed correctly (Singh, 1966).

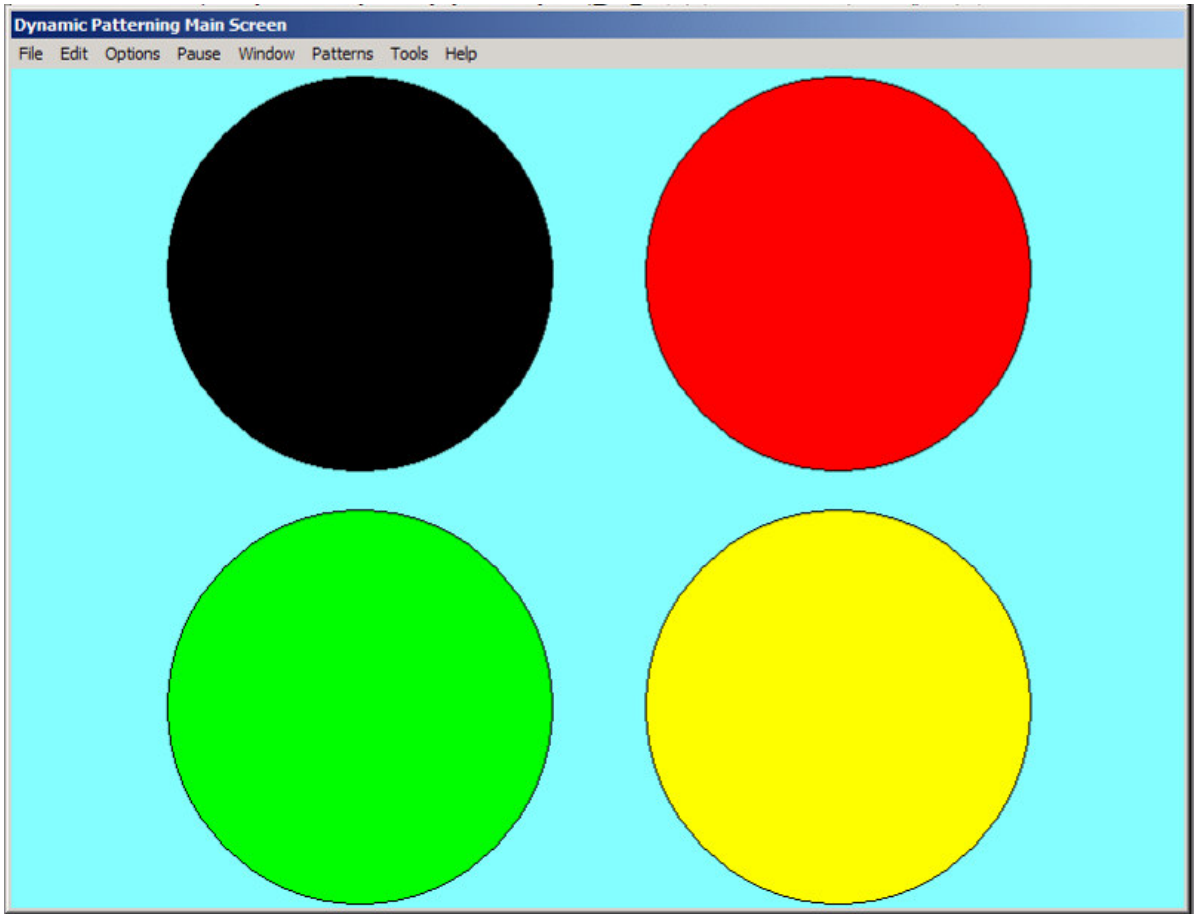


Figure 2 Typical DPS user screen

The patterns of DPS could be said to be a very simple form of pictures and Seifert has noted the ability of users to memorise pictures faster than words (Seifert, 1997). Series of patterns can also be used to represent just a single letter. This sequence could be said to represent movement or a gesture. Segan has found the use of gesture based interfaces useful for the visually handicapped (Segen and Kumar, 1998). There has also been developments into gesture interaction with information systems via wearable devices (O'Neil, 2006).

The DPS, within its present form, limits the number of patterns that a user is required to learn. For example, upper and lower case letters have the same pattern representation and hence cannot be distinguished from each other. Note that the letter being presented is available to be seen for training purposes by a fully sighted instructor. This enables the instructor to adjust the system in accordance with the users' wishes (Veal and Maj, 2005).

The appropriate use of colours is important for low vision users when using ATMs (Manzke, 1998). However, Fraser notes that *'Changes in colour are very helpful for some low vision users, but of no benefit to others. High contrast colour schemes will be useful for certain types of visual disabilities, but will be even more difficult for users with other types of visual disabilities'* (Fraser and Gutwin, 2000).

A fully developed DPS system would require help screens available both in standard text and the individual user's pattern set. The DPS system does give the option for the help files to be re-read by the system to enable users' access to the help facility. The delivery speed of the characters and repeating text options may be adjusted on screen. With user assistance the patterns in the form of the element colours for each pattern and background colours are chosen using the screen as shown in Figure 3. These may then be saved and loaded with the

users patternset choices which will also include presentation timing pattern element spacing and inter-word wait times. Volunteers who were blind or deaf-blind were able to recognise simple sentences using DPS (Veal and Maj, 2001). Some other possible areas of DPS application include use for engineering drawings by vision handicapped students (Veal et al., 2004). The increasing use of larger standard monitor screens could assist some users to more readily perceive sequences of patterns.

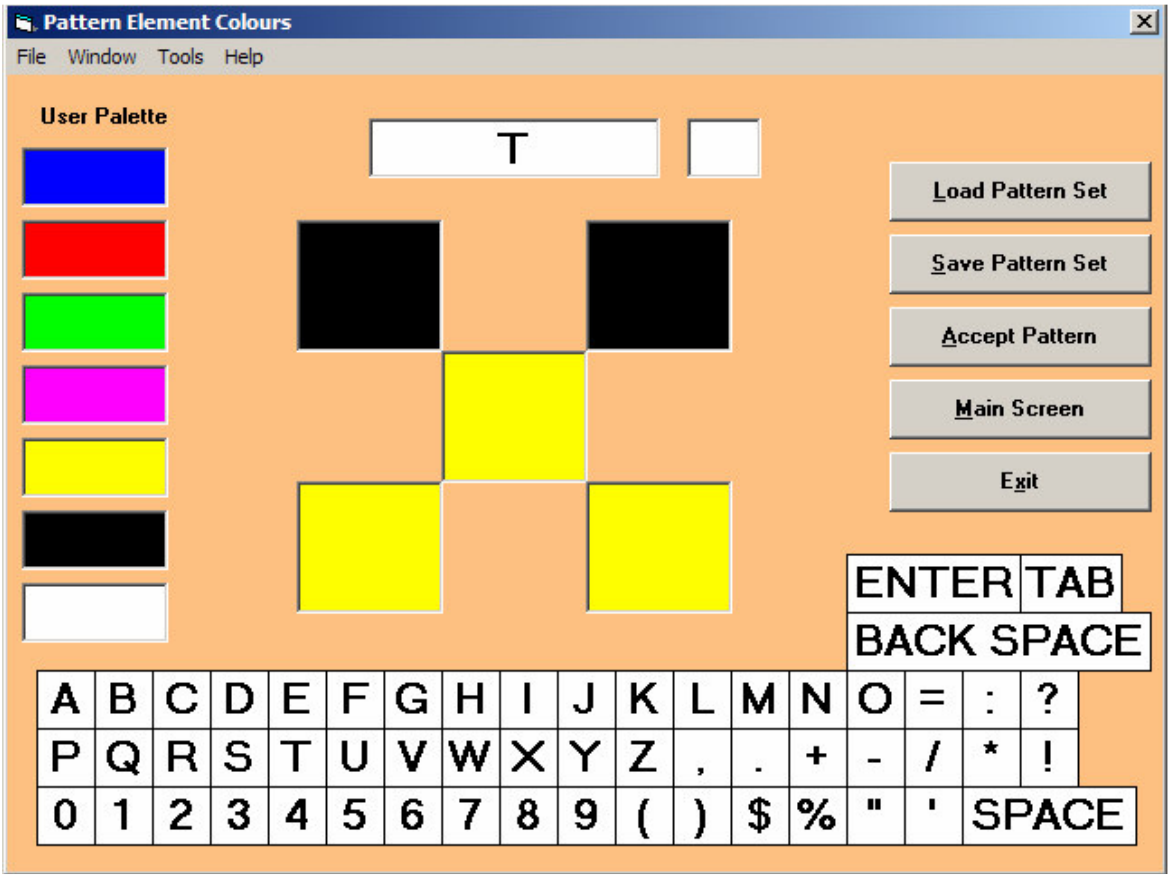


Figure 3. The DPS pattern selection screen

A vision impaired user could be presented with an on-screen DPS pattern as shown in Figure 2. Here the controls, text box and reading have been hidden and only the pattern and menu bar are showing to allow for an uncluttered screen when user testing takes place. Figure 3 shows the DPS pattern selection screen where the patterns are set to represent selected textual character and can be stored in the user's own patternset file which can be automatically loaded when the DPS program is run.

AN ATM WITH A DPS OPTION

The US ADA act specifically made mention of access to ATMs (DoJ, 1994). There have been great strides in fitting some ATMs with audio interfaces for the blind requiring the use of an earpiece or headphones (Kobayashi et al., 2000). However, Introna has noted that there are still many ATMs which the blind find difficult to operate (Introna and Whittaker, 2005). There are many of the deaf-blind bank customers who are unable to use ATM sound systems. Gill notes that: *'It is not obvious that those working on speech technology appreciate that 35% of the visually disabled population have a hearing deficit. Therefore, there is an important role for users organisations to ensure that research workers are aware of the present and future unmet needs of visually disabled people'* (Gill, 1995). The DPS could be of potential use and may not require a major redesign of ATMs.

The envisioned use of a DPS system is as a potential additional component to an ATM. The DPS program could operate as an option via the customers' identity which would be gained via their bank card and password. As such this would not necessarily require an upgrade of existing ATM provision. A more limited use of a DPS patternset not requiring users to learn a whole set of alphanumeric representation patterns is to need only characters representing 0-10 and other patterns that could represent specific ATM functions.

The individual's patternset could also be used for email and web access therefore they would have the necessary practice to memorise their patternset. Training packages for the use of ATMs based on individuals patternsets could be delivered via the user's PCs and hence allow drill and practice sessions in a given range of scenarios before attempting to use a live DPS enabled ATM. Such an ATM would appear the same as a convention ATM until used by a visually disabled customer when the screen would show changing coloured patterns instead of convention text. Although a rogue user may be more able to see the larger patterns than conventional text they would have a lot of difficulty in interpreting such varying pattern sequences. Improvements in ATM displays now commonly utilise multi-coloured graphics which means that the DPS requirements should be possible to display without needing to radically alter the display.

The number of legally registered blind within the US alone was put at 7 to 10 million in 1994 out of a population of 262 million by the American Federation for the Blind (AFB, 1995). The Australian Commonwealth Government note that: *'... there are about 300,000 Australians who are blind or have some form of vision impairment , approximately ninety percent of people classified as legally blind have some usable vision'* (Commonwealth, 2001). By 2021 Vision Australia note this figure could increase to over 420,000 (Vision, 2004). With so many people classified as blind yet having some residual vision, then there is a possibility of a large potential need for systems such as the DPS.

CONCLUSIONS

Research specifically using user's legally classified as blind or deaf-blind but who have some residual vision has shown that they can recognise simple sentences delivered in the form of large on-screen patterns representing textual information. This could form the basis of an ATM readout and input option. Web access and online banking can also be difficult and can involve the blind and deaf-blind to extra security risks such as fraud by their assistants as well as restrict them from participation in many online activities. DPS has a potential to allow the users themselves to access ATMs and online banking and avoiding the need to rely upon third party assistance. However, more research needs to be undertaken in more fully developing and testing the DPS program. Access to ATMs and the Web is also important in many countries anti discrimination legislation and could enable more visually disabled people to participate, or continue to participate in the workforce. There is a large number of people who could benefit from such a system should it be found to work successfully.

REFERENCES

AFB (1995), Employment statistics for people who are blind or visually impaired. American Foundation for the Blind, Washington DC USA. Available from: <www.afb.org/Section.asp?SectionID=15&DocumentID=1529&Mode=Print> [Accessed [September 23, 2007].

Bell, A. G. (1883). Upon a method of teaching language to a very young congenitally deaf child. American Annals of the Deaf and Dumb, xxvii, pp.124-139.

Ching-Hsing, S. and Ching-Hsing, L. (1997). A morse-code recognition system with LMS and matching algorithms for persons with disabilities. International Journal of Medical Informatics, Elisvier, Ireland, **44,** pp. 93-102.

Commonwealth, G. (2001) What is a disability? In Commonwealth Disability Strategy, Commonwealth Government, Canberra ACT Australia. Available from <www.facs.gov.au/disability/cds/fs/fs_03.htm> [Accessed September 13, 2007].

DoJ (1994), ADA Standards for Accessible Design, U.S. Department of Justice, Washington, DC, USA. Available from: <www.usdoj.gov/crt/ada/adastd94.pdf>. [Accessed September 18 2007].

Fraser, J. and Gutwin, C. (2000). A framework of assistive pointers for low vision users. In ASSETS'00 ACM Press, Arlington, VA, USA, pp. 9-16.

Gill, J. (1995) Technology for visually disabled persons: The Widening Gap. In Symposium on Communication, Disability, Compensation and Development - Scientific and Technical Reports, Linkopeing, Sweden, pp. 12-18.

Goble, C., Harper, S. and Stevens, R. (2000). The Travails of Visually Impaired Web Travellers. In Hyptertext 2000 ACM Press, San Antonio, TX, USA.

Introna, L. and Whittaker, L. (2005), Power, cash and convenience: the political space of the ATM. Lancaster University Management School, Lancaster UK. Available from: <www.lums.lancs.ac.uk/publications/> [Accessed September 11, 2007].

Keller, H. (1903, 2005), The story of my life. American Foundation for the Blind.

Available from: <www.afb.org/mylife/book.asp?ch=HK-ded> Accessed [July 23, 2007].

Kobayashi, I., Iwazaki, A. and Sasaki, K. (2000) An inclusive design of remittance services for the blind user's operation of automatic teller machines. In CUU'00ACM Press, Arlington, VA,, USA. pp. 153-154.

Legge, G. E., Ahn, S. J., Klitz, T. S. and Lueber, A. (1997). The visual span in normal and low vision. Vision Research, **37,** pp. 1999-2010.

Maineharbors (2005), Marine Signal Flags. Marine Harbors, Maine MA. USA, Available from: <www.maineharbors.com/> Accessed [July 24, 2007].

Manzke, J. M. (1998) Adaptation of a cash dispenser to the needs of blind and visually impaired people. In Proceedings of the Third International ACM Conference on Assistive Technologies ACM Press, Marina del Rey, CA, USA, pp. 116 - 123.

O'Neil, E. (2006) Can we do without GUIs? Gesture and speech interaction with a patient information system. Ubiquitous Computing**,** pp. 269-283.

RBS (1996) When even glasses don't help: A study of the needs of people who are blind or vision impaired Royal Blind Society (RBS) of NSW, Sydney Australia, pp. 43-52.

Segen, J. and Kumar, S. (1998) Gesture VR: Vision-based 3d hand interface for spatial interaction. In Proceedings of the sixth ACM International Conference on Multimedia ACM Press, Bristol, UK, pp. 455-464.

Seifert, L. S. (1997) Activating representations in permanent memory: Different benefits for pictures and words. Journal of Experimental Psychology, **23,** pp.1106-1117.

Singh, J. (1966) Great Ideas in Information Theory, Language, and Cybernetics., Dover Publications Inc, New York, USA.

Tokyo (1997) Japanese probe sickening cartoon. In The West Australian December 20 Perth WA Australia, pp. 24.

Veal, D. and Maj, S. P. (1998) Dynamic patterns as an alternative to conventional text for the partially sighted. ACM Special Interest Group in Computers and the Physically Handicapped (SIGCAPH)**,** pp. 11-15.

Veal, D. and Maj, S. P. (2001) A computer interface for the "blind" using dynamic patterns. In International Conference on Computing and Information Technologies (ICCIT'2001) Montclair State University, NJ, USA.

Veal, D. and Maj, S. P. (2005) A new visual communication system for deaf-blind tertiary education students. In 4th Global Colloquium on Engineering Education.

American Society for Engineering Education (ASEE) and the Australasian Association for Engineering Education (AaeE), Sydney, NSW, Australia.
Veal, D., Maj, S. P. and Kohli, G. (2004) A narrow bandwidth GUI for diagram recognition by the blind. In American Society for Engineering Education (ASEE), 2004 Annual Conference ASEE, Salt Lake City, UT, USA.
Vision, A. (2004), Eye Conditions. Vision Australia: Blindness and Low Vision Services, Canberra ACT Australia. Available from: <www.visionaustralia.org.au/info.aspx?page=1136&template=PrintReg> Accessed [September 30, 2007].
Yesilada, Y., Stevens, R., Harper, S. and Goble, C. (2007) Evaluating DANTE: Semantic transcoding for visually disabled users. ACM Transactions on Computer-Human Interaction, **14**.

DAVID VEAL

David Veal has a BA honours degree in Physics from the University of York and a general ordinary degree from the Open University UK. David has a Post Graduate Certificate in Education from the University of Keele UK where his specialist subject areas were Physics, Mathematics and Computing. He has a Graduate Diploma in Computing Science from Curtin University in Perth and a PhD in Computer Science from ECU where his research areas were competency-based assessment and models of computers and computer networks. He now lives in Western Australia where he has taught computing, mathematics and physics at high school level. David lectures in computing science at ECU in Perth, Western Australia. His areas of research include: Graphical user interfaces for the partially sighted, competency-based assessment techniques in computing science, and the modeling of computers and networks to aid student understanding.

David first thought of the idea of using coloured areas for communication at the age of seventeen whilst experimenting with analogue light beam communication. This system used an amplifier and modulated flashlight bulb as a transmitter and a telescope with modified form of photoelectric transistor and an audio amplifier as a receiver. Chromatic aberration in the telescope's lens system meant that the received light signal displayed coloured rings around received light source which varied according to the light output of the transmitter. David's further investigations led to a system of audio amplifiers and filters based upon capacitors and inductors and sets of coloured flashlight bulbs grouped together as a 'screen'. Lack of finances precluded further development of this early speech to coloured pattern system in 1965.

COPYRIGHT

State Model Diagrams and Home Security System Control

D. Veal & S. P. Maj

School of Computer and Information Science
Edith Cowan University
Australia
Email: d.veal@ecu.edu.au

KEYWORDS

Home Security Network Education, State Model Diagrams, Abstract Models, Power Line Control, X-10, A10, UPB, INTSEON. Domotics, Smart Homes, Home Automation, Objective Assessment.

ABSTRACT

There are many problems in delivering a constantly changing computer networking syllabus to students. These problems are exacerbated by the increasing number of complex protocols that may have to be learnt in a short period of time. Abstract models, known as State Model Diagrams (SMDs), may help to address some of these problems. The authors wished to test whether SMDs could also be used for teaching networking protocols that were very different from those previously taught to students studying networking units. Examples of such different protocols are those used by home security devices. Although the main aim focus of this research was on improving teaching and leaning outcomes of necessity an understanding of the context of home security devices was also considered important and some a selection of these have been included in this paper.

In an attempt to ascertain the effectiveness of SMDs for teaching home security device protocols a group of students who had previously studied computer networking via SMDs were introduced to a standard home security protocol and their opinions were subsequently sought as to whether the SMDs aided their learning and understanding of this protocol and the operation of the security device network demonstrated to them. The results of this research are presented in this paper.

INTRODUCTION

There are many different but equally valid ways of teaching computer networking e.g. algorithmic, engineering etc There is a large amount of material to cover in such units and the models can assist in this task (ACM/IEEE, 2001). The importance of abstraction has also been noted by the ACM who state that: *"... the use of abstraction in managing complexity, structuring systems, hiding details, and capturing recurring patterns; the ability to represent an entity or system by abstractions having different levels of detail and specificity"* (ACM, 1991). High level abstract models known as the State Model Diagrams (SMDs) have been designed to aid student learning (Maj and Veal, In press, Maj and Kohli, 2004). SMDs are based upon the layered model approach known as the OSI Seven Layer Model and TCP/IP four layer model. Tests have been undertaken on a range of students from Universities, Technical And Further Education (TAFE)s and high schools some who had previously or were currently undertaking Cisco Network Academy Program (CNAP) (Murphy et al., 2004) studies. The CNAP is a vendor based education programme where the curriculum is provided via a recognised company in the field (Veal et al., 2005). Currently Cisco is a major international company developing and supplying devices to be used within computer networks and on the Internet (Cisco, 2007). The use of SMDs were found to aid students' learning (Maj and Kohli, 2004). An example of an SMD is shown in Figure 1.

This SMD may seem complex but it is much simpler than the many screens of readout produced as the result of commands entered into typical Cisco and other vendors' networking or internetworking devices.

SMDs can also be used in troubleshooting where network addresses and numeral results expected from operational networks can be readily compared with results when networks are malfunctioning (Maj et al., 2006). Objective self assessment and examination questions can also be based upon such outputs where SMDs can be used to enable different scenarios to be checked and misunderstanding detected effectively. In this respect SMDs may be said to have similarities with multiple choice tests (Veal et al., 2001a) but avoiding easy guessing (Farthing et al., 1998) as there are very many possible answers to enter into the boxes in the SMD.

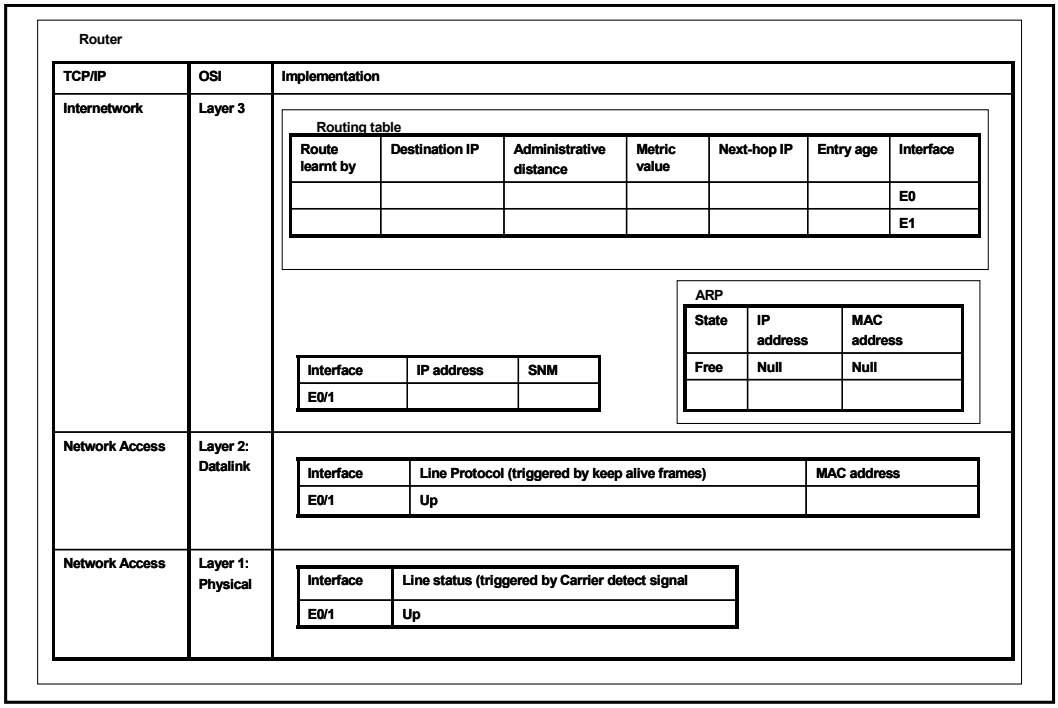


Figure1. A typical SMD for a router

The authors were interested in testing the applicability of SMDs to networks other than standard computer based networks. Home security devices can be networked and this was envisioned as a useful test bed for SMDs

HOME SECURITY SYSTEMS

There are now many Do It Yourself (DIY) home security kits. These are often included as part of a smart home, intelligent home or home automation concept. This is also known as *'Domotics'* Hannik notes that: *'DOMus informatics (information technology in the home) (domus is Latin for home). Although remote lighting and appliance control have been used for years, domotics is another term for the digital home, including the networks and devices that add comfort and convenience as well as security'* (Hannink, 2006).

Power Line Control systems (PLCs) use mains power lines as their transmission media to send and receive signals to control equipment plugged into PLC control devices that are themselves plugged into the mains power supply. Safety considerations are of the upmost importance as it should not be forgotten that the signal transmission medium uses the electrical mains system which is potentially lethal.

Home security systems can range from the relatively inexpensive wired equipment to kits using PLCs to Wireless security systems such as Zigbee which has a routing protocol (Heile, 2006), Z-Wave or combinations of Wireless PLC and Infra Red (IR) control. There are now keypads from the simple to the complex and many windows based systems that can control multiple protocol PLC systems (Howard, 2007).

The pre-programmed automatic opening and closing of curtains may seem to be useful only for the disabled yet this can also be an important security feature because curtains opened or closed at non-standard times can help to alert potential intruders of empty premises. Such devices can readily work with standard home security offerings or general home automation controllers. These can also be used to dim the room used as a home theatre (WinPlus, 2007). Yasmin Hashmi of the online security residential technology trade magazine 'Hidden Wires' notes that *'The number of home automation system manufacturers is increasing, and if they have anything to do with it, home automation will no longer be the preserve of the wealthy technophile. The technology is heading for the mainstream ...'* (Hashmi, 2007).

X-10

One of the oldest successful and readily available home PLC systems is X-10 (Kingery, 1999). The X-10 signals consist of binary values that are composed of 1 ms burst of 120 kHz signals at the mains zero crossing point. This is the point where the mains voltage in a home power line goes to zero. With 50Hz mains, as in Australia and the U.K. this zero crossing point occurs twice very every one fiftieth of a second or 20ms. If a burst had been sent during the first half cycle indicating a 1 then no burst should be sent during the second half cycle as in Figure 2. Conversely if on burst was sent during the first half cycle, indicating a binary 0 then a burst should be sent during the second half cycle. Here a 1 followed by a 0 equals a binary 1 value and a 0 followed by a 1 equals a binary 0 value. This is shown in the second row of values in Figure 2 where each of the values listed are derived from the values on the above row.

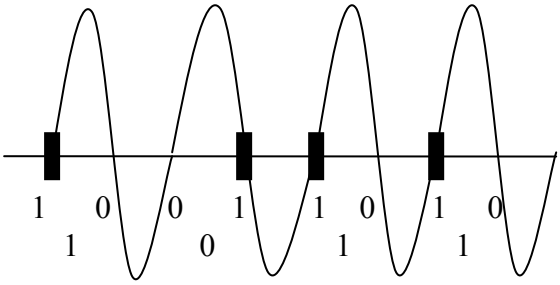


Figure 2. A bit pattern derived from the 120 kHz pulses

Figure 2 shows how 1011 is derived from 120 KHz pulses at the mains power line zero points. The start of frame 1110 code does not have the second half cycle inverted repetition and so this appears as shown in Figure 3.

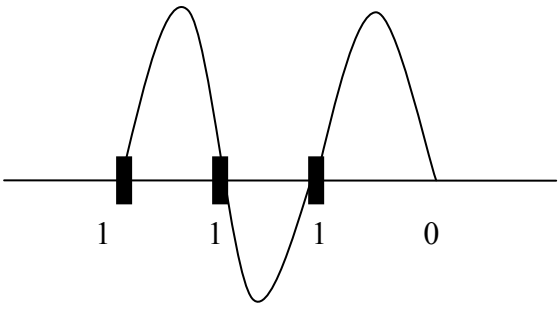


Figure 3. The binary code representation of 1110

After the start of frame code is received the next 4 bits define the house code with a possible 16 houses. This help to prevent codes from one residence 'leaking' into another residence and controlling their X-10 devices (Belcourt, 2003). However, these codes do not have to be in separate residences and can be used to control separate areas within a residence such as the garden, garage and floors. Kingery notes that: '*The basic X-10 message consists of 13 bits. This consists of a 4 bit start code, a 4 bit house code, a 4 bit unit function code and a 1 bit function bit. The function bit indicates whether the previous 4 bits should be interpreted as a unit code or a function code'* (Kingery, 1999)..

Kingery also notes that : *'To turn on an X-10 device will require two 13 bit messages, one to transmit the house and unit code and one to transmit the command'* (Kingery, 1999). The power cycles between addresses and commands should be 6 zeros and every X-10 command should be transmitted twice to increase reliability (Kingery, 1999). This results in 24 power cycles to send a single 4 bit command. This is an almost an 8 bits per second data rate to send a command assuming a 50Hz mains power frequency. Yet for many security functions such as alarms and light switching and dimming this rate is sufficient. X-10 gives the possibility of controlling up to 256 devices provided that neighbours on the same powerline are not also using X-10.

X-10 ENHANCEMENTS

There have been many improvements to the X-10 system but as its patents have lapsed it has become a de-facto standard. Systems such as A10 are compatible with it but give a higher 120 KHz burst voltage and can detect weaker signals. Importantly A10 is compatible with X-10 (Envioustechnonolgy, 2005).

A limitation of the initial version of X-10 is was that it had a command set of only 16 instructions however, X-10 also allows for an extended command set. Smarthome notes that *'Although not used in most commonly available X-10 devices, by specifying the extended code or extended data code in the function code field, an additional 20 bits of data can be appended. This capability dramatically opens up X-10 functionality…. In fact this technology has been used successfully in many OEM applications to transmit temperature data, security system functions, energy load shedding functions, window covering functions and more'* (Smarthome, 2006).

FURTHER PROBLEMS AND LIMITATIONS OF X-10

Using a transmission media that was not designed for such a purpose inevitable presents problems. These can sometimes result in some unreliability due to interference on noisy power lines. PLC systems can also experience problems with multiphase systems of mains power wiring where devices wishing to communicate are on different phases. X-10 repeats the signal at the 60 and 120 degree points in each half cycle to send the signal at the zero crossing points on the other two phases of the three phase system. Couplers may also be employed to send the signal across house power lines on different phases (Envioustechnonolgy, 2005).

Problems may also exist due to the low impedance effects of transformers and other devices shorting the higher frequency 120 kHz signals. In some situations these can be overcome by using filters to help cut out the noise but allow the mains frequency and signal to pass. Additionally repeaters may be used that amplify the signal to avoid attenuation effects. Couplers may also be used to boost communication between different power line phases (Belcourt, 2003). Furthermore bridges may also be used to enable inter-protocol PLC system communication. Should an X-10 frame collide or not reach its destination there is no resending of the frame with X-10 apart from the fact that as part of the protocol each frame is

sent twice. There is no error checking in X-10 fames unlike Ethernet and other computer network protocols. Other PLC systems such as those based upon Intseon and Universal Power Bus (UPB) have error checking (Smarthome, 2006).

INSTEON AND UPS

Universal Power Bus (UPB) allows both UPB signals and X-10 signals to coexist on the same power line (Be-Home, 2006). The Insteon system is an improvement on the X-10 system and Insteon and the X-10 standard are also compatible. Insteon employs error checking and using acknowledgements can 'multicast' commands to groups of devices as well as 'simulcast' where each device sends on the original frame to help ensure its arrival at its destination. To avoid an OSI level 2 broadcast storm this simulcasting is only repeated through three devices (Smarthome, 2005).

With respect to the UPB Be-Home notes that '*UPB uses low frequency, spread spectrum technology which produces a very strong signal. In addition, UPB uses true 2-way communications, allowing devices to reveal their status with each other, providing feedback that commands have been successfully executed'* (Be-Home, 2006). Some Original Equipment Manufacturers (OEMs) also have dual systems that send details via both PLC and wireless. (Smarthome, 2006). This can also be the case in many modern based X-10 systems (WinPlus, 2007). Despite such problems X-10 is still used extensively and has a large range of products and suppliers internationally across both the 60Hz 120 V used in US systems and the 240 V 50Hz used in Australasian and European electrical power systems. There are now limitations on the spectrum range of PLC systems and many European countries use the CENELEC standard (OFCOM, 2005), with similar standards enforced elsewhere. This restricts the bands allowed to be used in powerlines and control signal transmission power levels to specific uses analogous to the restrictions placed upon radio frequency transmissions.

SMDS APPLIED TO X-10

The ready availability of X-10 systems and its greater dissimilarity with standard computer networking protocols made it a suitable choice to test the applicability of SMDs across different networking models. In the X-10 model postulated by the authors the physical layer and datalink layer are those familiar from the OSI seven layer model, namely OSI Layer 1 which is the physical layer and is concerned with bit formation, timing, voltages, media etc. Layer 2 is concerned with frames access to the media and frame addressing.

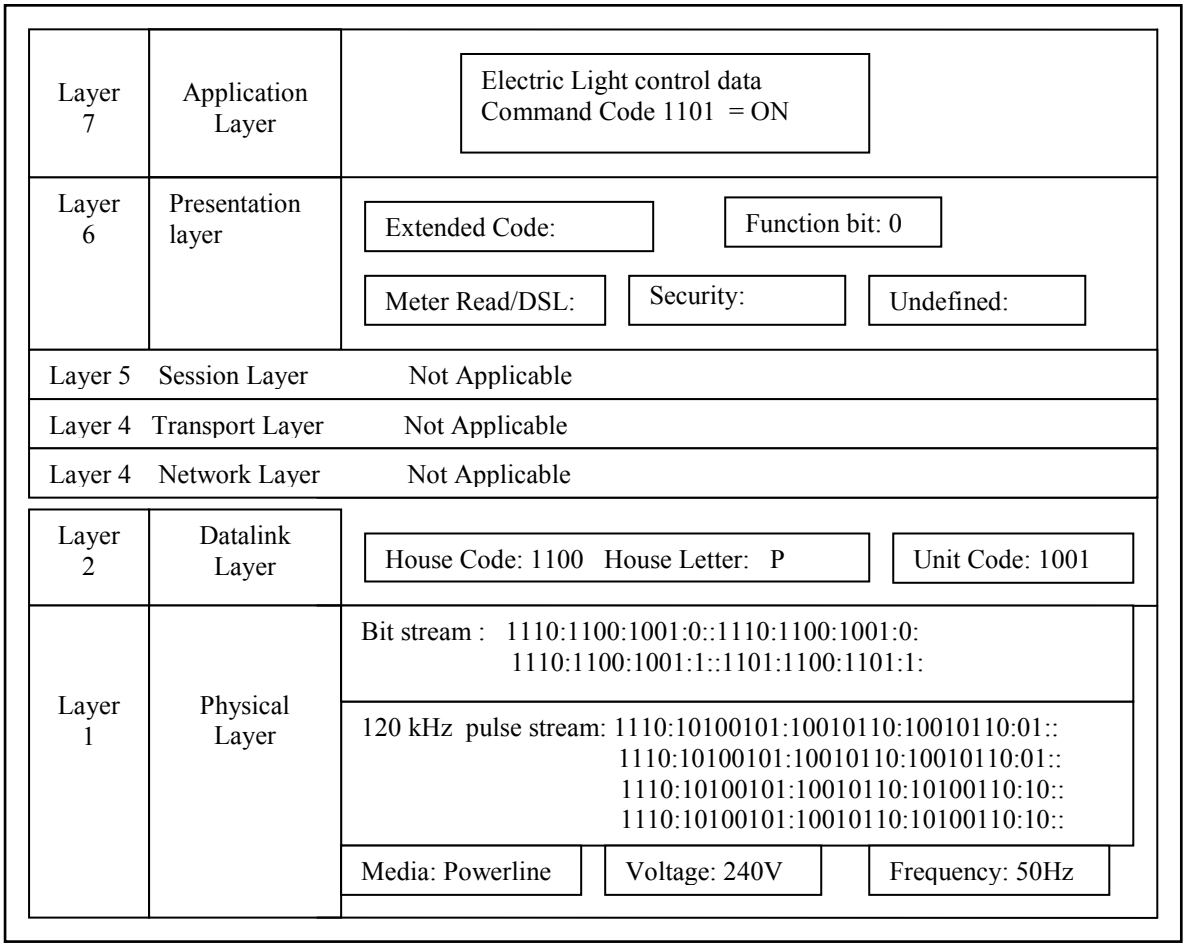


Figure 4. An SMD for X-10

The authors derived a possible SMD for X-10 devices where the network, transport and session layers are not applicable. Layer 6 the presentation layer was concerned with whether the function code bit was 1 or zero in order to determine the use of extended code whereby more sets data could follow to extend the command set. Figure 4 shows an X-10 SMD for switching on an electric light. The bit stream is derived from the 120kHz pulse stream. The single colon ':' is used to separate parts of the steam, the double colon '::' is used to separate repeated parts of the stream and to separate the command and addressing sections of the streams. This method of separating out streams is non standard but may assist in readability and understanding.

THE EXPERIMENT

As many systems are compatible with X-10 and because this protocol is least similar to computer network protocols it was decided to use X-10 to test the hypothesis that SMDs could be helpful in teaching home security networked device protocols.

Masters students who had previously been taught computer networking and internetworking using SMDs were given a lecture and demonstration of a home security networked devices based upon X-10 with an extended command set. A short demonstration of the X-10 system in operation was also given to complement the talk. The importance of hands-on components for learning has been noted with respect to previous units offered by the authors (Veal et al., 2001b). The explanation was based upon SMDs.

EXPERIMENTAL RESULTS

The questions shown below were asked via handouts and were handed in anonymously with no student name or identification to assist unbiased feedback:

Question 1: State Model Diagrams assisted in quickly understanding basic operation of X-10.
Question 2: State Model Diagrams assisted in learning about TCP/IP protocols
Question 3: State Model Diagrams assisted in understanding X-10 because I have used state models before:
Question 4: If I needed to know more about X-10 I would like to receive instruction and workshops based upon SMDs.

The results were as follows from 19 respondents:

	Strongly Agree %	Agree %	Neutral %	Disagree %	Strongly Disagree %
Question 1	21.0	63.2	15.8	0	0
Question 2	36.8	57.9	5.3	0	0
Question 3	26.3	52.6	15.8	5.3	0
Question 4	15.8	63.2	21.1	0	0

Figure 5. The results from the student questionnaire

Figure 5 indicates that the majority either agreed or strongly agreed with the statements in the questions.

CONCLUSIONS

It was found from the student questionnaire that there is strong indication that the SMDs are also a useful as a learning tool in non TCP/IP based networks. More research needs to be undertaken using other home security device networking protocols such as Insteon, UPB and others to further investigate the use of SMDs in this field. The greater complexity of some home security protocols means that SMDs can be used that require all or many of the layers of the OSI seven layer model. Such complex protocols may mean that the use of SMDs are of greater value due to their ability to handle complexity through abstraction, information hiding and leveling. Further research is needed to include other groups of students from other institutions such as TAFEs where such protocols are taught as part of security installation and management courses and to also include university level security course students. Research also needs to be undertaken using groups that have not previously been exposed to SMDs to ascertain their options about the efficacy of using SMDs for teaching home security protocols.

REFERENCES

ACM (1991). ACM/IEEE Joint Curriculum Task Force. Available from: <http://www.computer.org/educate/cc1991/eab1.htm> [Accessed 20 November 1998].

ACM/IEEE (2001). Computing curricula 2001 ACM/IEEE Joint Task Force Computer Science Final Report: ACM Press.

Be-Home (2006), UPB Information, Available from: <www.be-home.com.au/Parts/parts_content/upb%20products/UPB_overview.htm> [Retrieved August 10, 2007].

Belcourt, L. (2003). Home automation explained. Available from: <www.haliving.com.au/catalog/ha.php?osCsid=352e85c24f00e4bbc4905d6401ad3f7f> [Accessed 11 August 2007].

Cisco (2007), Cisco Networking Academy Program. San Jose, CA. Available from: < http://www.cisco.com/web/learning/netacad/index.html> [Accessed August 10, 2007].
Envioustechnonolgy (2005), The A10 product range Available from: <http://www.envioustechnology.com.au/support/x10.php> Accessed August 10, 2007,
Farthing, D. W., Jones, D. M. and McPhee, D. (1998). In iTiCSE ACM Press, Dublin, Ireland, pp. 81-85.
Hannink, E. (2006), Domotics. Available from: <www.squidoo.com/domotics> [Accessed 2 August 2007].
Hashmi, Y. (2007), Industry Opinion: Home Automation - Trends and Developments. HiddenWires. Available from: <http://hiddenwires.co.uk/resourcesarticles2007/articles20070806-01.html> [Accessed 15 August 2007].
Heile, B. (2006). Wireless sensors and control networks: Enabling new opportunities with ZigBee. Available from: <www.zigbee.org/en/resources/presentations.asp> [Accessed 31 July 2007].
Howard, D. W. (2007), Power Home Quick Start Guide Version 1.03.4.11 Revision 5, Available from <www.power-home.com> [Accessed August 30 2007].
Kingery, P. (1999), Digital X-10. Available from: <www.hometoys.com/htinews/feb99/articles/kingery/kingery13.htm#Digital%20X-10> [Accessed 17 July 2007].
Maj, S. P. and Kohli, G. (2004). Journal of Issues in Informing Science and Information Technology, **1,** 385-392.
Maj, S. P., Tran, B. and Veal, D. (2007). State model diagrams - A systems tool for teaching network technologies and network Management. International Joint Conferences on Computer, Information, and Systems Sciences, and Engineering: Engineering Education and Instructional Technology, Assessment and E-learning, pp. 355-360. Bridgeport, CT. USA.
Maj, S. P. and Veal, D. (In press) State model diagrams as a pedagogical tool - An international evaluation. IEEE Transactions on Education.
Murphy, G., Kohli, G., Veal, D. and Maj, S. P. (2004). An examination of vendor-based curricula in higher and further education in Western Australia In American Society for Engineering Education (ASEE), 2004 Annual Conference ASEE, Salt Lake City, UT, USA.
OFCOM (2005), PLC Communication. Federal Office of Communication, Switzerland. Available from <www.bakom.ch/themen/technologie/01157/index.html?lang=en> [Accessed August 20, 2007].
Smarthome (2005), Insteon: The details. Smarthome Technology. Irvive CA USA.
Smarthome (2006), Designing your own distribution system. Available from: <http://www.smarthome.com/remote_entire_home.html> Accessed August 8 2007.
Veal, D., Engel, A. and Maj, S. P. (2001a). Multiple choice questions on a computer installation and maintenance unit. In UICEE 5th Baltic Region Seminar on Engineering Education(Ed, Pudlowski, Z. J.) The UNESCO International Centre for Engineering Education (UICEE),, Gdynia, Poland.
Veal, D., Kohli, G. and Maj, S. P. (2005). Cisco based curricula in university units. In 4th Global Colloquium on Engineering Education American Society for Engineering Education (ASEE) and the Australasian Association for Engineering Education (AaeE), Sydney. NSW. Australia.
Veal, D., Maj, S. P. and Duley, R. (2001b). Assessing hands-on skills on CS1 computer and network technology units. In ACM SIGCSE 32nd Technical Symposium on Computer Science Education (Ed, Russell, I.) ACM. SIGCSE., Charlotte, NC, USA. pp. 381-385.
WinPlus (2007), Home Theatre Application., Available from: <http://www.winplus.com.au/images/hometheatre.pdf Retrieved> [August 5, 2007].

DAVID VEAL

David Veal has a BA honours degree in Physics from the University of York and a general ordinary degree from the Open University UK. David has a Post Graduate Certificate in Education from the University of Keele UK where his specialist subject areas were Physics, Mathematics and Computing. He has a Graduate Diploma in Computing Science from Curtin University in Perth and a PhD in Computer Science from ECU where his research areas were competency-based assessment and models of computers and computer networks. David has worked for the UK Marconi Company as a trainee telecommunications technician and was also a radio amateur. He taught physics and mathematics at South Devon College where was the faculty IT advisor. After migrating to Australia in 1992 and he has taught mathematics, physics and computer science at schools and colleges in the Perth area. David is now a lecturer, tutor, and unit coordinator on the Cisco Certified Network Associate (CCNA) and Cisco Certified Network Professional (CCNP) based units in Computer Science at ECU. David has had an interest in safety since nearly electrocuting himself in 1963 at the age of fifteen whilst experimenting with analogue powerline communication systems.

PAUL MAJ

Associate Professor Dr S. P. Maj is a recognized authority in the field of industrial and scientific information systems integration and management. He is the author of a text book, '*The Use of Computers in Laboratory Automation*', which was commissioned by the Royal Society of Chemistry (UK). His first book, *'Language Independent Design Methodology - an introduction'*, was commissioned by the National Computing Centre (NCC). Dr Maj has organized, chaired and been invited to speak at many international conferences at the highest level. He has served on many national and international committees and was on the editorial board of two international journals concerned with the advancement of science and technology. As Deputy Chairman and Treasurer of the *Institute of Instrumentation and Control Australia (IICA*) educational sub-committee he was responsible for successfully designing, in less than two years a new, practical degree in *Instrumentation and Control* to meet the needs of the process industries. This is the first degree of its kind in Australia with the first intake in 1996. It should be recognized that this was a major industry driven initiative. Paul has undertaken research using Cisco equipment, has lectured on CNAP based units and has developed postgraduate level units based upon Cisco equipment.

COPYRIGHT

Study on Whole Lifecycle Protection of Digital Contents

MEI Xue, XINMIN Geng

School of Computer and Information Science
Shanghai University of Electric Power, China
E-mail: Xue Mei, qq.snow@163.com; GENG Xinming, jimgeng@sina.com

ABSTRACT
With the development of computer and network technology, there are more and more applications of digital contents which are easy to be copy, modify and spread. Because of these features, digital contents are also always under the risk to be filched, juggled, accessed without authorization and illegally distributed. Although remarkable achievements have been fulfilled in this field, we haven't found any sufficiently effective solution for current research deficiencies: the dominant protection model in many applications can't ensure security for the whole lifecycle; little attention has been paid to the differences in various application areas of digital contents, while each area has its unique requirements; the existent protection model cannot provide all life-long tracing. In this paper, we present the Content Lifecycle Protection Concept (CLPC), which mainly employs 'Box' to encapsulate content so that management and protection is feasible in each phase of the lifecycle. Then, we give out a detailed description of its core concept-BCO. At the end of this paper, we present three key services based on PKC to explain how the BCO works.

KEYWORDS
content protection，Whole Lifecycle，Security Rules，All Lifelong Tracing，Distribution

PREFACE

With the development of computer and network technology, there are more and more applications of digital contents which are easy to copy, modify and spread. Because of these features, the digital contents are also always under the risk to be filched, juggled, accessed without authorization and illegally distributed. The digital contents can hardly be protected effectively in the network environment due to its large-scale disordered distribution, which has hindered the development of the applications based on the digital contents. Therefore, the research on the protection of digital contents has drawn great interests from both the industrial and academic communities. It has already become a hot spot in scientific study. This paper is mainly intended to address Information Content Security issues. Although remarkable achievements have been fulfilled in this field, we haven't found any good solution to current research deficiencies: the dominant protection model in many applications can't ensure security for the whole lifecycle; little attention is paid to the differences in various application areas of digital contents, while each area has its unique requirements; the existent protection model cannot provide all life-long tracing and embedded-protection. So the focus of this paper is about how to implement high-quality, whole-life-cycle protection for digital contents.

This paper is organized as the following: in the first section, we analyze the current security techniques to build a good base for the proposition of Content Lifecycle Protection. In the second section, aimed at improving the deficiency of current research we put forward Content Lifecycle Protection Concept (CLPC) ,which mainly employs 'Box' to encapsulate content so that management and protection is applicable in each phase of the lifecycle. In the third section, we presents the core concept of CLPC—BCO and discussed its logical structure, physical structure and physical coupling structure. In the fourth section, on the basis of CLPC technology and theory, the experimental system CLSS (Content Lifecycle Security System) is

implemented and good results have been obtained through experiments. Especially focus on the BCO key services relevant to PKC.

ANALYSIS OF SECURITY TECHINIQUES

There are various technologies in the information security field, such as Encryption, Firewall, Anti-virus, and Information Audit and so on. The protection concepts on which these techniques are based can be classified into two categories: one is like providing something like checkpoint protection, which embodies the common characteristics of technologies like Firewall, Invade Detect, and the other is providing hull protection, which embodies the common characteristics of technologies like Encryption, Watermark. Based on protection concepts, digital content protecting models can be divided into two types:

(1)Event-oriented Protection Model (EPM)
EPM is a model that sets up a checkpoint to ensure the security of contents that pass by, and it's triggered by the event. When the same event happens to the various contents, EPM takes the same protection measures according to the security policy predefined for the contents. This model ensures that contents meet the predefined security requirements, and it's featured with point-to-point security, which ensures the security from the pre-check point to the post-check point. Techniques like Firewall, Invasion Detect and Access Control belong to this protection model. Exampled with Firewall, when various contents are transmitted across two networks, Firewall will inspect every data packet and judge whether the packet is satisfying the security policy despite which content the packet belongs to.

(2)Content-oriented Protection Model (CPM)
CPM is a content-oriented model which is focused on security of content itself and controlling the access to the content. It is attaching a single hull to ensure the security of a single content. Techniques like Encryption, DRM, and Watermark belong to CPM. With Encryption as an example, symmetric encryption encrypts content, and the content is recoded so as to ensure its transmission security. Even though it is filched during transmission, the original content can not be decoded without decryption key. For another example, Watermark is to embed a mark into content, and the mark exists from the content's creation to its destruction. With this mark we can judge content's legality to solve law disputation.

Although these two models above provide content protection to a certain extent, they are insufficient to protect the whole lifecycle of digital contents. Firstly, both of them only ensure content security at a certain phase without considering threats to other phases. For example, the subsequent operations are ignored after conforming whether the content was properly-used in the client side. Secondly, they both rarely concern different protection policies in various application circumstances. For example, we allow digital contents flowing among the client sides, while at a government department the content flowing is highly restricted. Therefore there is a great demand for a life-long protection solution for the content, which can be customized according to the application circumstances.

CONTENT LIFECYCLE PROTECTION CONCEPT (CLPC)

The whole lifecycle of a digital content is a continuous and dynamic process, which can be decomposed into four phases: creation, distribution, usage and destruction.

Although CPM has certain limitations, it sets up a good basis for whole lifecycle protection. In the content's lifecycle, there are four statuses of 'original information resource', which are 'content-created', 'content-distributed', 'content-used' and 'content-destructed'. Different statuses face their own threats: filched after its creation, juggled during distribution; illegally used during usage and so on, so the whole lifecycle protection have to defend all these threats. Our strategy for Content Lifecycle Protection is: when a digital content is created, a protective

‘Box’ is created for the content at the same time, which serves as a bodyguard for the content to transform, inspect and administrate the activities in the sequential phases and keeps the content under protection until it is destructed. Life-long tracking is supported in this protection.

CLPC RELATED CONCEPTS

Definition and Function of Box Object (BO)

To satisfy the content protection requirements mentioned above, BO should have the corresponding functions and characteristics, which are as follows:

Be able to transform the content to ensure its confidentiality, so as to make sure stealer can’t obtain the original content.

Be able to produce signature for content creator, which ensures the content receiver received is published by the one declared as its manufacturer.

Be able to verify content’s integrity to ensure content un-juggled during its transmission.

Be able to record content’s use rules, so as to control customer’s usage, assure legal customer’s rights and prevent the illegal customer’s operation beyond his authorization.

Be able to validate customer’s identity.

Be able to control customer’s operation according to use rules predefined for the content.

Be able to record every status of the content and build audit information.

BO is an object with above features. We can define BO as:
BO= (ID,A,M)
ID: unique ID of BO, A: attribute aggregate of BO, M: function aggregate of BO
BO can be described as:

```
Object BO
{
ID id;
Attribute attr[];
Method method[];
}
attr[]=
{SR//use rules of content
BOCreateInfo// information about BO’s creation
}
Method[]=
{ConverseC()//transform content;
ValidateSign()//verify signature;
ValidateIntegrity()//verify integrity;
ValidateUser()//verify user’s identity;
ValidateRights()//verify user’s rights;
ControlUsage()//control user’s usage based on use rules;
CollectTrackmsg()//collect track information;
BuildAuditMsg()//build audit information;
RecordRightsOwnerMsg()//record creator's copyright;
Unit()//bind with content;
ModifyAttri()//modify BO itself attributes;
}
```

Content Object (CO) Description

CO is an entity with a unique ID and not a method. It can be described as:
CO = (ID, A)
ID: the unique ID of CO, A: attribute aggregate of CO.

CLPC Diagram

CLPC can be expressed in a two-dimensional figure1:

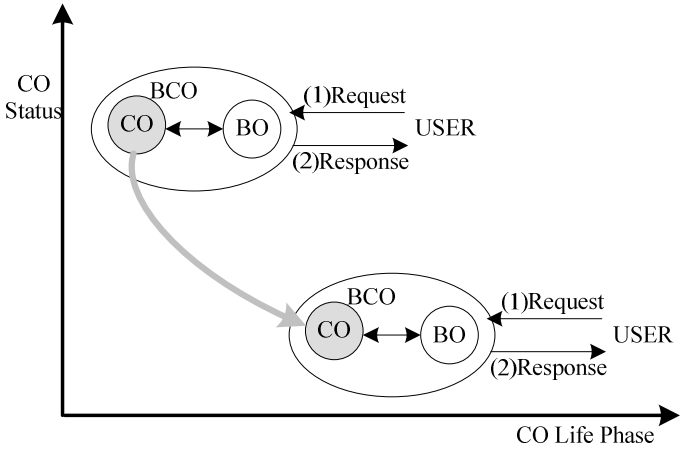


Figure 1:Content Lifecycle Protection Concept (CLPC) Diagram

X-coordinate denotes various CO's life phase. Y-coordinate denotes various CO statuses. CLPC means that CO in different status caused by phase change is always under BO's protection, which is a life-long content-oriented protection. During the process, BO control user's usage of CO when user has requests to access to CO. Moreover, BO enables life-long tracing of distribution and usage which helps to locate the root of invalid operation so that reduces the disorder of distribution. From the figure we know that BO co-exists with CO, so they can be regard as an integrated object logically, and we call it BCO in this paper. BCO is the core of CLPC. It is formed as the CO created, and it embodies the BO's whole lifecycle protection on CO.

CLPC CORE---BCO

BCO Physical Structure

It's clear that the security of CO ensured by BO, which is bound with CO to form BCO, a logical integrated formation. In this section, the points we concern are: what physical components does BCO contain? Whether its physical form is integrated as its logical form? If not, how the relations among these components are? All these need us to analyze in detail.

Based on the CLPC, we can conclude that the BO protection function contains two parts: (1) taking protection measures on content, (2) controlling user's operation according to use rules. They are respectively shown in the left and right part in figure2. Also, we can find that the protected UI (execution environment for use rules) and protection measures on various CO have common characteristic---they both relatively stable and don't need move with content's movement. They are often in the form of executable software. While the use rules of CO maybe variable in its lifecycle, it can be various with CO and CO's phase. Therefore, from the viewpoint of changeable or not, we can classified the BO protection functions into two types, as shown in figure2, (1) Fixed Part, including execute environment for use rules and protection measures on CO;(2) Variable Part, including use rules of CO and CO itself and some necessary key information for protection implementation, such as validation code for integrity, signature etc.

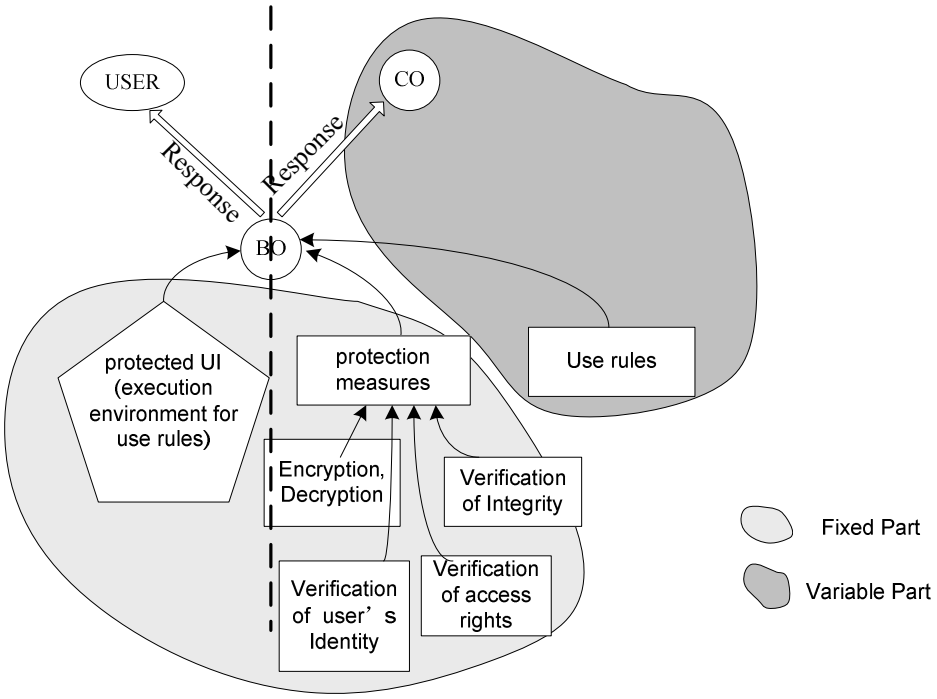
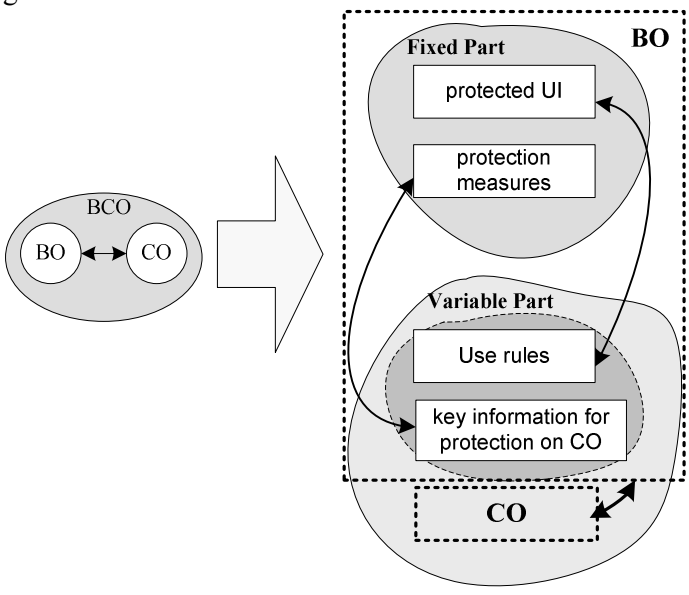


Figure2:BO protection function

Therefore, the mapping from BCO's logical structure to its physical structure is shown as figure3:



BCO Logical Structure BCO Physical Structure

Figure3:mapping from BCO's logical structure to physical structure

BCO Physical Components

In BCO logical structure, BCO=(BO, CO)

Based on the analysis above, BCO physical structure can be concluded as follows:
BCO=(Fixed Part, Variable Part)
Fixed Part= (Executing Environment based on use rules, Protection Measures for CO)
Variable Part=(use rules of CO, key information for protection on CO, CO)

BCO physical structure can be regard as a combination of three components: VM, SI and CO, VM performs protection function based on corresponding SI, and SI is related with certain CO. they can be expressed as follows:
VM →Fixed part
SI → (use rules of CO, key information necessary for protecting implementation)
CO →CO

Mapping from BCO physical structure to three physical components is shown in the figure4:

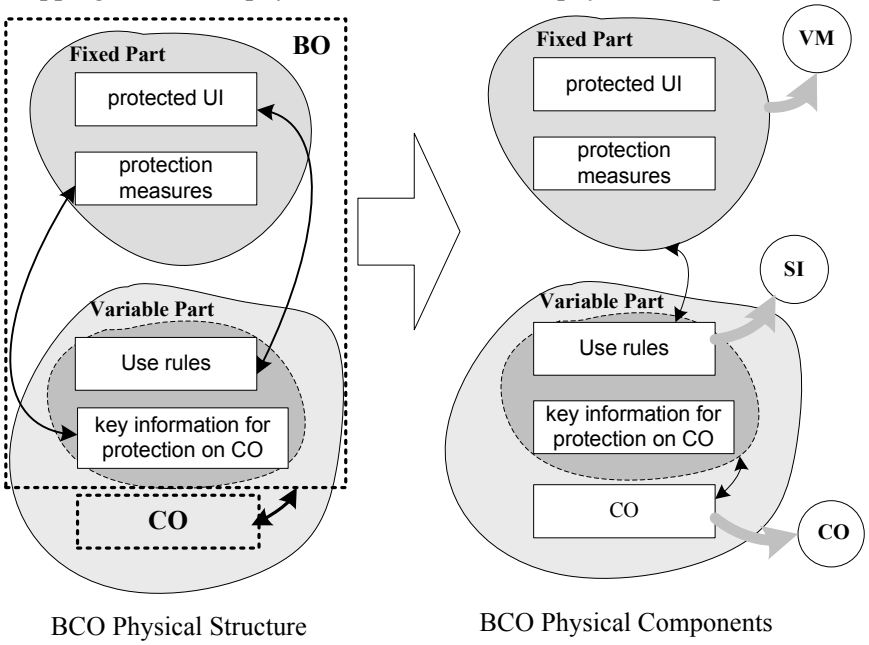


Figure4:BCO Physical Components mapping

Therefore, BCO physical structure with three components is:
$BCO_{component}$=(VM, SI, CO)
Mapping from BO logical structure to its physical structure is:
BO → (VM, SI)

The detail description of BCO's three components is as follows:
CO: content protected form transformed by BO, often it is encrypted. In addition to the content itself, it contains the information of creator, creation time etc in it. CO structure can be described in XML is:

```
<CO>
<Content>
</Content>
<Info_about_Creation>
<author>
</author>
<abstract>
</abstract>
</Info_about_Creation>
</CO>
```

SI: composition of two parts: use rules for CO (or we can call it security rules, denoted as SR), and necessary key information of CO to implement protection (denoted as SelfProtectInfo). SR is the standard for controlling user access to the CO. It can be expressed as access rights list. SelfProtectInfo contains integrity verification code and creator's signature etc. There are

several tools to describe SR, such as Extensible Access Control Markup Language (XACML) [X05], OASIS Rights Language Technical Committee [O], MPEG Rights Expression Language[M-1][M-2], Digital Property Rights Language (DPRL), Open eBook Forum (OeBF) [O03], Extensible Rights Markup Language (XrML) [C01], IEEE LTSC DREL Project (Digital Rights Expression Language), Open Digital Rights Language (ODRL) [O02] .The core concept reflected in these tools is that principal owns the right on resource if and only if he meets certain condition, which can be described as a quaternion:
SR=<principal, resource, condition, right>

The XML structure of SI is:

```
<SI>
<SR>
<principal>
</principal>
<resource>
</resource>
<condition>
</condition>
<right>
</right>
<SR>
<SelfProtectInfo>
<integrity_para>
</integrity_para>
<signature_para>
</signature_para>
</SelfProtectInfo>
<SI>
```

VM: providing a safety executing environment for operations on CO, and meanwhile takeing corresponding protection measures on CO, such as encapsulating content, controlling user's usage, verifying user's identification and verifying content's integrity. VM can be software or hardware, and majority is software. They have the following capabilities (1)ability to build the necessary key information SelfProtectInfo for CO.(2)ability to validate CO's security based on SelfProtecfInfo, such as judging juggled or not during transmission.(3)ability to control user's operation on CO according to SR.

BCO Physical Coupling Structure

In previous section we analyzed BCO physical structure which is composed of VM, SI and CO. To construct BCO, there should have certain coupling relations among three components.

Generally, there are two coupling relations between two elements:
(1) Independent, two elements are independent of each other, denoted as $f_{independent}$
(2) Binding, two elements are bound with each other, denoted as $f_{binding}$

Based on the above coupling relations, there are four BCO physical coupling structures among three elements:
(1) Highest-coupled fixed control mode: VM, SI and CO binding together. None is moveable.

$BCO_{structure}=VM\ f_{binding}\ SI\ f_{binding}\ CO\ldots\ldots(1)$
(2) Higher-coupled fixed control mode: VM binding with SI, and CO independent of the two. CO is moveable.

$BCO_{structure}=(VM\ f_{binding}\ SI)\ f_{independent}\ CO\ldots\ldots(2)$
(3) Middle-coupled flexible mode: SI binding with CO, and their combination independent of VM. The combination is movable.

$BCO_{structure}$= VM $f_{independent}$ (SI $f_{binding}$ CO) ……(3)
(4)Low-coupled flexible control mode: VM, SI and CO are independent of each other, and SI and CO are moveable.

$BCO_{structure}$ = VM $f_{independent}$ SI $f_{independent}$ CO……(4)

We can conclude from above that BCO Physical Coupling structure is:

$BCO_{structure}$={(1),(2),(3),(4)}
Mapping among BCO logical structure, physical components and physical coupling structure is shown in the figure5:

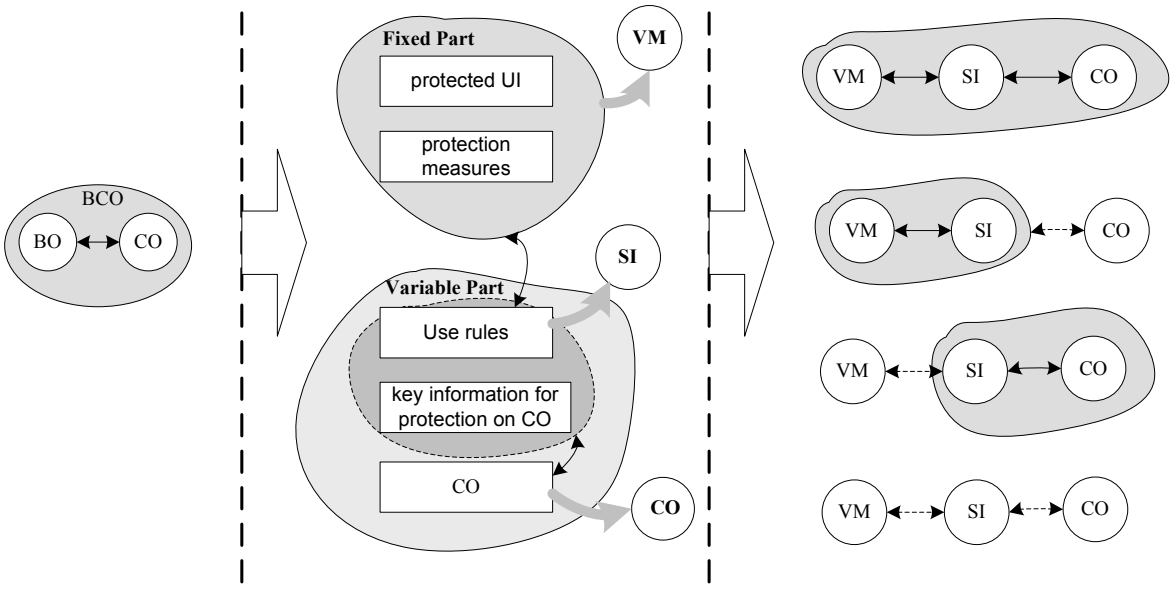


Figure 5:Mapping among BCO logical structure, physical components and physical coupling structure

Among them, the Digibox of InterTrust covers (3)(4) coupling forms, while Cryptolope of IBM cover (3) coupling form. They both belong to the implementation of BCO. BCO physical coupling structures provide different content protection levels. Various content applications can choose proper BCO physical coupling structure based on their factual requirements.

CLPC SECURITY ANALYSIS

From the viewpoint of security, there are two kinds of attacks: identified attack (known attack) and unidentified attack (unknown attack). Figure6 displays the logical diagram of attack and protection. In CLPC, CO is encapsulated into BCO, a protected form, which can protect the content from some known attack (such as filched, access beyond rights and copy etc.) and some unknown attack. As shown in the figure, there still have some unknown attacks, which will break BO's protection on CO in BCO and get access to CO. In case of this unknown attack, There's method to record relevant track information (BCO whole lifecycle track characteristic), and we can analyze these records and identify the unknown attack and defend it as early as possible. In the figure, the grey arrow means the protection provided by whole lifecycle tracking. If the attacker destroys the track information, the possibility of identifying this unknown attack will decrease. The down arrow shows the attack of destroying track information.

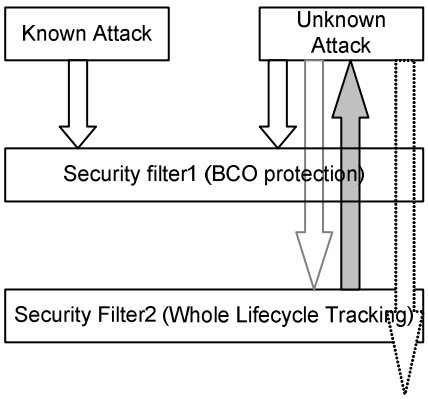


Figure6 Logical diagram of attack and protection

EXPERIMENT

Based on the CLPC above, we built a Content Lifecycle Security System (CLSS). The realization techniques include XML Web Service, WSSecurity, ACTIVEX, RBAC, Encryption etc. To satisfy the requirements from the application circumstance, we deployed the fourth BCO physical coupling structure. Three key services, which are relevant to PKC, are listed in the following:

BCO Creation Service Description

Denotation:

CAS: client application software
GS: Grant Server
CS: Credential Server
TS: Track Server
IDCreator: Creator ID
Igoods: Information work
D(igoods):Digest of igoods
IDigoods: ID of igoods
SRigoods:SR of igoods
Copyrightigoods:Copyright of igoods
Kigoods: key of encapsulating igoods
EKpp: Encrypt with the public key delivered by PP
CAS ⇒ GS: {M}: CAS send msg to GS
Kcreator_pr : private key of creator
Kcreator_pu : public key of creator

```
Service Build_BCO
{
        CAS ⇒CS: check IDcreator's credential
         CAS: Adding  Copyrightigoods, EKigoods（igoods）and OperInfoCreator  into CO
         CAS:  Adding  IDigoods, SRigoods, Kigood, D(igoods) and  OperInfoCreator  into BO
         CAS ⇒GS: { EKpp (BO)}
         CAS ⇒ IDPublisher :
        CAS ⇒TS {OperInfocreator}
}
```

BCO Usage Service Description

```
Service Consume_CO
{
        CAS ⇒CS: check IDconsumer's credential
        if the identification is valid then
        {
            CAS ⇒GS: Ekconsumer_pr(Request(IDigoods, right))
            GS: check if consumer meets the condition of SRigood
            If meet then
            {
              GS ⇒CAS: {EKconsumer_pp( IDigoods,SRigoods, Kigoods,Digest(igoods) )}
              CAS: DKconsumer_pr(EKconsumer_pp(SRigoods, Kigoods,Digest(igoods)))
              CAS: check if Digest(Dkigoods(EKigoods(igoods))?=Digest(igoods)
                    if equal then
                    {
                       igoods was not attacked
                 control consumer's operation
                 update the OperInfoCreator in the CO
                  reEncapsulate the Copyrightigoods, EKigoods（igoods）and OperInfoCreator
                          into CO after consumer's operation
                    }
                    else
                    {
                       igoods was attacked
                         refuse
                     }
              CAS ⇒TS {OperInfocreator}
            }
        }
}
```

BCO Track Service Description

```
Service Get_AuditInfo
{
        CAS ⇒TS: {Request(IDigoods, IDcreator)}
    TS ⇒ CAS: {AuditInfo}
}
```

CONCLUSION

Content Lifecycle Protection Concept (CLPC) proposed in this paper supports the life-long protection of digital contents, which will enhance relevant content applications. This paper is of significance to provide vital suggestions for future studies on digital content protection.

ACKNOWLEDGEMENTS

Projected supported by Shanghai Leading Academic Discipline Project, China(Grant No. P1303).

REFERENCE

[X05] Extensible Access Control Markup Language (XACML), http://docs.oasis-open.org /xacml/2.0/XACML-2.0-OS-NORMATIVE.zip 2005
[O02] Open Digital Rights Language (ODRL), http://www.w3.org/TR/odrl/ 2002
[M-1] MPEG-21 ,Part 5: Rights Expression Language (REL). http://mpeg. telecomitalialab. com/standards/mpeg-21/mpeg-21. btm
[M-2] MPEG-21, Part 6: Rights Data Dictionary (RDD). http://mpeg. telecomitalialab. com/standards/mpeg-21/mpeg-21. Htm
[O] OASIS Rights Language Technical Committee. http: // www.oasis-open.org/ committees/rights/
[O03] Open eBook Forum (OeBF), http://www.openebook.org/specifications/ rrwgcoordinated.htm
[C01] ContentGuard Extensible Rights Markup Language (XrML), http://www.xrml.org/, 2001

Data Mining and Genetic Algorithm Application In Bioinformatics With Microarray

C. Y. Jiao [1] and D. G. Li [2]

[1]School of Computer and Information Science
Edith Cowan University, Australia,
E-mail: cjiao@student.ecu.edu.au

[2]School of Computer and Information Science
Edith Cowan University, Australia,
E-mail: d.li@ecu.edu.au

ABSTRACT

Genetic Algorithm in the bioinformatic can be categorised into many differet groups by the implementations they were instructed. Each approach has its own pros and cons.

When making a choice between GA approaches to data mining, it is important that the scientist knows the advantages and disadvantages of each approach. In this project a suitable GA strategy will be identified and refined, based on microarray data mining.

The research applies the improved GA approach to a life microarray database, provided by a biological research institution to analyse and visualise the results so as to assist further development of biological strategies for identifying disease and evaluating drug applications.

INTRODUCTION

DNA study is one of most rapidly developed areas arising from the merging of computing area biological sciences that created bioinformatics. This is a field that draws on a range of specific sciences to solve complex biological problems, especially at the molecular level, such as the human genome. Many computer science methodologies are applied to DNA data analysis, as well as traditional mathematics. Hardware, including specialised silicon chips such as microarrays, has been developed to support hybridising of genes samples producing images that can be measured into sets of data for further analysis. Very large numbers of genes may be present and this poses data analysis problems. Soft computing technologies are then used to analyse the data translated by microarrays to yield information that is meaningful to the biologist. This study is aim to further develop a microarray analysis method base on computer science.

MOLECULAR BIOLOGY AND DNA

Biological, Mathematical and Computing Disciplines

It is the biologists' task to explain how a study of human being's health can be transformed as molecular micro study. For the purpose of this study, the diagram 1.1 is constructed to explain the relationship between DNA and molecules.

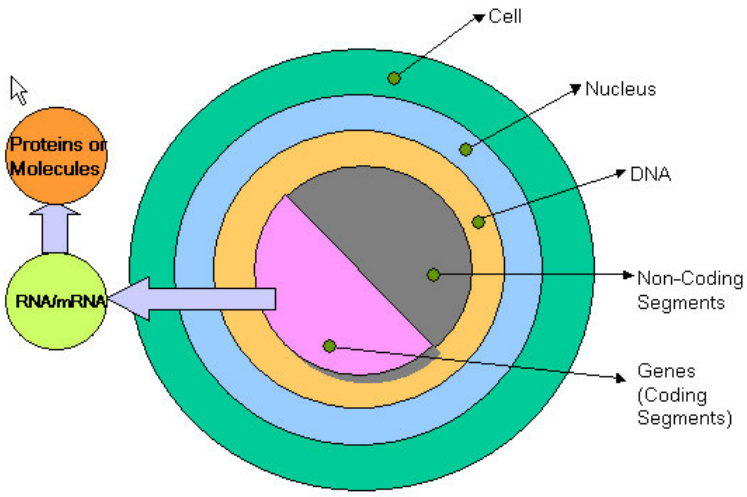


Figure 1 and DNA
Molecules

The most interested part of a DNA molecule is its coding segments. We often call these "genes". Because genes determine which proteins (or molecules) are produced in a human organism, over in turn which other molecules are synthesized, with the help of proteinaceous enzymes, they are sources of information of diseases.

MICROARRAY

One of the most common microarrays is the Affymetrix GeneChip® microarray. Dyed biological samples are placed in the grips of the silicon where that hybridise to the chip. After being processed under defined conditions (temperature etc.), the microarray slide is washed and scanned with a laser beam into an image. The image is then translated into a set of data based on measurement of the image by probes. A mathematic normalisation process will remove unwanted systematic variability from the data (Stekel, 2003). The data can then be either distributed into a public data bank (often in the internet) to be shared for the analysis, or be analysed directly for the presence and abundance of fluorescent dye labelled nucleic acids (Stekel, 2003).

MICROARRAY DATA MINING

Microarray-generated-data go through the processes of imaging, translation and normalisation before it is ready to be analysed it representation of biological or clinical means. The purpose of microarray data mining is to allow exploration of biological mechanisms. The analyses of microarray data are often described into following stages.

1) Analysis of differentially expressed genes

This analyse how each gene in a biological sample expresses itself and this differs from other genes. These will assists a biologist to develop a biological theory in comparison, for example, between samples from healthy and ill objects.

2) Analysis of relationships between genes

This analyses how each gene in a biological sample expresses itself and this differs from other genes. These will assists a biologist to develop a biological theory arising frame comparison, for example, between samples from healthy and ill subjects. (Wang and Fu, 2005)). This is a challenging and for microarray data analysis.

MICROARRAY DATA MINING

Microarray data mining, as part of the bioinformatics, is set to “extend the possibilities of applying computational analysis and data mining to aid research in biology and medicine” (Piatetsky-Shapiro and Tamayo, 2003).

Data Mining relies greatly on AI and when applied to microarray data, should help detect specific patterns within very large genetic data sets. For example, one type of data mining unsupervised learning, was used in forecasting protein interactions (COHEN, 2004) . Data mining strategies can be applied powerfully to microarray data mining. By using seven key phrase: (Han and Kamber, 2006)

1. Data cleaning (to remove noise and inconsistent data)
2. Data integration (to combine related data sources)
3. Data selection (to search and collect data that relevant to the analysis task)
4. Data transformation (to convert data to be mined)
5. Data mining (to extract data patterns by using intelligent methods)
6. Pattern evaluation (to identify the patterns representing knowledge)
7. Knowledge presentation (to visualization mined knowledge to the user)

AI and GA

It is not surprising that the powerful artificial intelligent (AI) based data mining methodology for can be hence microarray data mining (MDM). Among a range of AI methods including neural networks, fuzzy logic and genetic algorithm GA, GA has been identified as a promising method for searching and grouping gene data arising from the microarray.

Stage	Purposes	Data Mining Concept	Difficulties	Methods	Challenge
;ene Selection Phrase 1-3)	Search the genes most strongly related to a particular class Find differential molecular behaviour relevant to a given biological problem	Samples are treated as instants and the genes are treated as attributes	Not enough knowledge about the normal biological variation	False Positive - use classified normal genes to search for the gens that different from the majority Gene-Ranking - use p-values method to outperform the t-test Prototype-based feature selection	Most methods evaluate each gene in isolation without evaluation of gene to gene correlation
Classification (Phrase 4)	Classify diseases Predict treatment outcomes	Analysis of gene express patterns For the low-level analysis the intensities of the probe data sets Preliminary analysis of data sets based on probe outputs	Many more attributes than records; Random chance could lead to false discovery	ArrayAnalyzer® Nearest shrunken centroid Loss-based with cross-validation selection Machine learning	Integration of biological and technical knowledge
Clustering (Phrase 5-7)	Find new biological classes Refining existing classes Identify groups of co-expressed genes Recognize coherent expression patterns; finding gene networks and gene interaction	Abstract genes / attributes that belongs to the same class	Interpretation depends on the biological knowledge Difficult to fully automate the process	Coherent pattern index graph Graphical Gaussian model Integrative genomics	To produce more meaningful information to the biologists so that they can make the use of the analysis results

Table 1 Microarray Data Analysis Types and Their Data Mining Strategies Related to 7 Phrases

GA is a search method for solving the optimization problems posed by the specific research topic (Holland, 1975). It is similar to Darwinian evolution in that the mathematic algorithm uses strategies that are they used to akin to "increasingly beneficial adaptations" (Jones, 2005). These are essential to the survival of species with "fittest" genes being passed on to the next generation's chromosomes. In GA "chromosomes" represents each individual from a population within the problem being studied, and the term "fitness" represents the measurements made to generated data set. Figure 1.2 illustrates the core idea behind the algorithm (Jones, 2005).

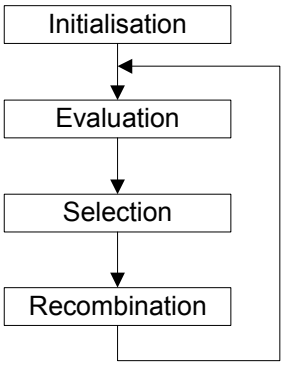


Figure 2 Genetic algorithm high-level flowcharts (After Jones, 2005)

Table 2 explains each step's purpose and explained in microarray data analysis term.

	Purpose	Microarray Data Terms
Initialisation	Create/Collect initial population (chromosome sets)	Collect gene data sets
Evaluation	Calculate the fitness of each chromosome with defined formula	Calculate each set of genes' fitness
Selection	Select fitter chromosomes to reproduce future smaller population	Select a group of genes that interests the study
Recombination	Combine pairs of chromosomes (cross-over or mutation) to produce new chromosomes	Combine two groups of genes to make a new group
Go back to step 2	Until the exit conduction is reached	Until a small set of meaning full groups are formed

Table 2 – GA procedures and Microarray Data Analysis Terms

Study 1 - ArrayMiner (Falkenauer and Marchand, 2001)

This research was a joint development between Brussels University and Optimal Design Ins. It was developed based on the realization that the commonly used k-means clustering method was highly unreliable. In the research GA was implemented for both clustering and classification. It believed that the k-means clustering yields high-quality solutions with low probability after GA applied.

Study 2 - Case-Based-Reasoning (Perner, 2002)

The study developed systematic architecture to mine the microarray expression data and build a genetic network model to characterize diseases. The model used a import facility to collect the microarray data provided by data banks / laboratories, a GA tool to analyse the expression, a data mining tool to establish relationships to an existing pubic database, and to visualize the presentation of the genetic network. In the microarray data expression analysis, the study used a AI Neural Network method, based on name adaptive resonance theory (ART), to effectively create a neural pattern recognition machine.

Study 3 - GA/KNN (Li et al., 2001)

This was similar to study 1, in that it sought to improve the quality of the classification. The GA/KNN method was developed to "select a subset of predictive genes" from laboratory microarray data provided. It is believed that the method can remove more variants and noise among the microarray data so that the classes can be targeted more accurately. The study combined both GA and KNN methodologies to enhance the outcomes.

Study 4 – GEPAT (Weniger et al., 2007)

The study was based on the theory of dimension reduction in microarray data analysis. The routine developed integrated the analysis of the results and their biological interpretation so that the biologists can access a more comprehensive data directory. Essentially, it combined available methods into a processing line for the microarray data.

Study 5 – Hybrid Intelligent Systems (Bosl, 2007)

In this novel approach a hierarchical computational model was built for a hybrid intelligent system that modeled molecular pathways and mined knowledge from the microarray data. It highlighted the power of AI methodologies in the area of microarray data analysis.

Table 3 listed above studies / developments with the main attributes interest to this study.

Study No	Name	Year	Paper Title	Question to be Answer	Method	Advantage	Disadvantage
1	ArrayMiner	2001	Using K-Means? Consider Array Miner	What are the genes that can differentiate diseases if a patient's DNA microarray data is provided for the analysis?	- Genetic Algorithms	- Tree view window - Increases the number of clusters	No details of clustering method were provided
2	Case-Based-Reasoning	2002	Genomic Data Explosion – The Challenge for Bioinformatics?	How is a disease represented in the genomic data?	- Neural Network - Case-based Learning	- Efficient for understanding the dynamics of pathogenic processes	It is a theory for other to develop into a strategy
3	GA/KNN	2002	Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the GA/KNN method	Can the method produce better results classification in sample, gene identification? Can genes been jointly discriminated?	- GA Search - KNN Classification	- Selecting subsets of predictive genes to enable the analysis of the gene's relationships - Multivariate approach - Repeatability of selection in independent runs	Gene selection may be less robust than classification
4	GEPAT	2007	Genome Expression Pathway Analysis Tool – Analysis and visualization of microarray gene expression data under genomic, proteomic and metabolic context	Can an integrated microarray data analysis package be used as a toolkit that handles the whole progress of microarray data analysis and interpretation?	- Hierarchical Statistical methods - k-means - PCA clustering - Linear model - t-test	- Selecting subsets of predictive genes to enable the analysis of the genes' relationships	Not compatible to the no-linear work flow Relies on established data analysis packages
5	Hybrid Intelligent	2007	Systems biology by the rules: hybrid intelligent systems for pathway modelling and discovery	Can soft computing technologies be integrated to solve bioinformatics problems?	- Fuzzy logic - Neural nets - Genetic algorithms - Statistical analysis	- Allows biologists to build complex models without mathematical involvement - Dynamics	It is a blue print that has yet to be achieved

Table 3 – the Comparison Between Five Studies

FURTHER DEVELOPMENT

The challenges in bioinformatics data mining, different from the mainstream commercial data mining, are initiated by the nature of the biological data. The data gained from a device like microarray do not have a large number of entities but many attributes. That is while the number of the samples (entities, harder to collect), is fewer, the number of the genes (attributes) is as large as thousands. That is why many developed data mining tools are particular suitable for microarray data mining as they are face the possibility of presenting “false positives” (Perner, 2002). This study is aim to develop a specific data mining strategy for the needs of microarray data mining.

REFERENCES

BOSL, W. J. (2007) Systems biology by the rules: hybrid intelligent systems for pathway modeling and discovery. *BMC Systems Biology,* 1:13.

COHEN, J. (2004) Bioinformatics—An Introduction for Computer Scientists. *ACM Computing Surveys (CSUR),* 36**,** 37.

FALKENAUER, E. & MARCHAND, A. (2001) Using k-Means? Consider ArrayMiner. *2001 International Conference on Mathematics and Engineering Techniques in Medicine and Biological Sciences (METMBS'2001).* Las Vegas, Nevada, USA.

HAN, J. & KAMBER, M. (2006) *Data Mining,* Bostan, Morgan Kaufmann Publishers.

HOLLAND, J. (1975) *Adaption in Natural and Artificial Systems,* Ann Arbor, The University of Michigan Press.

JONES, M. T. (2005) *AI Application Programming,* Hingham, Massachusetts, Charles River Media, INC.

LI, L., WEINBERG, C. R., DARDEN, T. A. & PEDERSEN, L. G. (2001) Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the GA/KNN method *Biostatistics,* 17**,** 1131-1142

PERNER, P. (Ed.) (2002) *Genomic Data Explosion – The Challenge for Bioinformatics,* Verlag Berlin Heidelberg, Springer.

PIATETSKY-SHAPIRO, G. & TAMAYO, P. (2003) Microarray Data Mining: Facing the Challenges. *SIGKDD Explorations,* 5**,** 5.

STEKEL, D. (2003) *Microarray Bioinformatics,* Cambridge, Cambridge University Press.

WANG, L. & FU, X. (2005) *Data Mining with Computational Intelligence*, Springer.

WENIGER, M., ENGELMANN, J. C. & SCHULTZ, J. (2007) Genome Expression Pathway Analysis Tool – Analysis and visualization of microarray gene expression data under genomic, proteomic and metabolic context. *BMC Bioinformatics,* 8.

Three-Dimensional Cellular Automation LFSR Algorithm

YONG Wang[1] , XINMIN Geng[1], YU Wang[2]
[1]Dept.of computer Science and Technology,
Shanghai University of Electric Power, Shanghai, China
E-mail: wy616@126.com

[2]China Association for International Exchange of Personnel,
Beijing, China
E-mail: wangyu_caiep@126.com

ABSTRACT
Three-dimensional cellular automation (CA) linear feedback shift register (LFSR) algorithm combines cellular automation (CA) method and linear feedback shift register LFSR method to create approximately random stream bits. There are three bit tests such as mono bit test, poker test and run test by which the algorithm feasibility can be judged. The three tests accept 20,000 bits from a random source of the algorithm according to FIPS 140-1 standard. The algorithm can pass three stream bit test. The results illustrate it is feasible which can create better random stream bit than CA or LFSR algorithm. The algorithm is less efficient and more complicated than the other algorithm.

KEYWORDS
CA; LFSR; Stream cipher; Algorithm

INTRODUCTION

Stream cipher is widely used in contemporary cryptology which can create random (not completely random) bit stream as a key. The plaintext can be changed into cipher text by XORing the plaintext with this random stream key. The plaintext is recovered by XORing the cipher text with the same random stream key. There are many ways to produce a stream key. The most common is to use a hardware device called a linear feedback shift register (LFSR).The A5 stream cipher is an irregularly clocked LFSR system.. The A5 stream cipher has been used for the protection of voice communications as part of the GSM (General System for Mobile). It involves a set of security features such as applying stream cipher A5 in encryption and decryption [1,2]. In three stream ciphers such as S1, S2 and S3 are proposed for GSM applications [3,4]. Many efforts have been made to design new stream ciphers for GSM network[5,6].A cellular automation (CA) is just an array (either one dimensional or two dimensional) of simple cells, whose value depends on the value of its neighbors and a specified rule. There dimensional cellular automation LFSR algorithm combines CA and LFSR to create random stream.

THREE-DIMENSIONAL CELLULAR AUTOMATION DEFINITION

One-dimensional cellular automation introduction
A simple one dimensional cellular automation is a 8-cell array shown as Figure 1

0	1	0	0	1	0	1	1
0	1	2	3	4	5	6	7

Figure 1: One-Dimensional Cellular Automation
The CA is initialized to 0100 1011.Each cell changes its state based on some rule. One possible rule for the examples CA could be defined by the immediate neighborhood of each cell. The cell depends on the current value in the cell and the values in the cells to its left and its right[1]. The CA is assumed to be connected in a circle, so the cell to the left of cell 0 is cell 7,and the cell to the right cell 7 is cell 0.

The CA Neighborhood is as follows:

101 010 100 001 010 101 011 110.
A specific rule for this neighborhood could be:
Neighborhood 000 001 010 011 100 101 110 111
New state: 0 1 0 1 0 1 1 0

Rules for this type of CA are identified by converting the new state bits into a decimal number. The new state for the preceding rule is 0101 0110, which in decimal is 86.

Applying this rule to the initial CA produces the following CA is 1001 0111.A CA can be used to generate random bits by selecting a rule, a CA size, an initial seed and the cell to provide to the random bit[1]. For example, we choose the 7th cell to produce the random bit. The first 5 step to produce the random bits are 10100.

Two-dimensional cellular automation introduction

A two-dimensional cellular CA offers a more powerful random number generator at the expense of additional complexity. A two-dimensional cellular CA is just an array of one-dimensional cellular CA, a cell's value is updated by some function of this current neighborhood which consist of the cells above, below, to the right, and to the left of the target cell[1]. A general rule structure can be defined as follows:

$$S_{i,j}(t+1) = Xxor\left[C \times S_{i,j}(t)\right]xor\left[N \times S_{i-1,j}(t)\right]$$
$$xor\left[W \times S_{i,j-1}(t)\right]xor\left[S \times S_{i+1,j}(t)\right]xor\left[E \times S_{i,j+1}(t)\right]$$

Where X,C,N,W,S,E are 0,1 variables. If X is 1, this is a nonlinear rule; otherwise, it is a linear rule, C,N,W,S and ,E are the center, north, west, south, and east cells, respectively. The cells that participate in updating the center cell are determined by values of these five variables. The CA is assumed to be connected in a row circle and column circle as the one-dimensional cellular. A simple two-dimensional 3*3cellular automation is like a double circle array. So the cell to the north of cell [0,0] is cell [2,0],and the cell to the west cell [0,0] is cell [0,2].

A simple two-dimensional 3*3cellular automation is shown as Figure 2:

0	North S(i-1, j) 0	1
West S(i, j-1) 0	Center S(i, j) 0	East S(i, j+1) 0
0	South S(i+1, j) 1	0

Figure 2: Two-Dimensional Cellular Automation

If N is 1,then the north cell is used to update the center cell ,The values of all six variables are used to identify each possible rule. If (X,C,N,W,S,E)=(001011), then the rule is defined as Rule 11, because 1011 is decimal 11. the rule looks like this:

$$S_{i,j}(t+1) = \left[N \times S_{i-1,j}(t)\right]$$
$$xor\left[S \times S_{i+1,j}(t)\right]xor\left[E \times S_{i,j+1}(t)\right]$$

A random stream is generated by assigning a rule to each cell, initializing the CA to a random state, and running the CA using a center cell to produce the bit stream[1], For example, the CA is initialized

to {001;000;010},each cell is assigned Rule 11.The value in the center cell is used to construct the random stream. The cells are randomly initialized, after four steps, the random-bit stream is .01101

Three dimensional cellular automation

A three-dimensional cellular CA is a cube of circle two-dimensional cellular. A cell's value is updated by some function of its neighborhood, which consists of the cells above, below, to the right, to the left, to the up and to the down of the target cell. A general rule structure can be defined as follows:

$$S_{i,j,k}(t+1) = Xxor\left[C \times S_{i,j,k}(t)\right]xor\left[N \times S_{i-1,j,k}(t)\right]$$
$$xor\left[W \times S_{i,j-1}(t)\right]xor\left[S \times S_{i+1,j}(t)\right]xor\left[E \times S_{i,j+1,k}(t)\right] S_{i,j,k}(t+1) = S_{i,j,k}(t+1)xor\left[U \times S_{i,j,k+1}(t)\right]$$
$$S_{i,j,k}(t+1) = S_{i,j,k}(t+1)xor\left[D \times S_{i,j,k-1}(t)\right]$$

A simple three-dimensional 3*3*3 cellular automation is shown as Figure 3

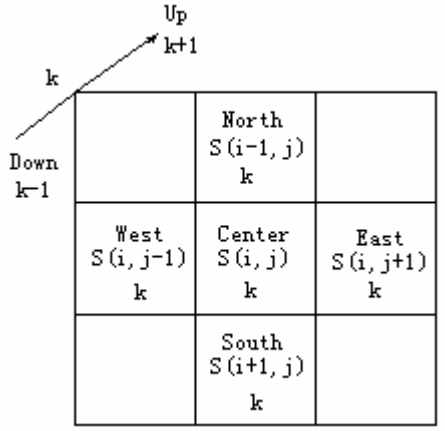


Figure 3: Three-Dimensional Cellular Automation

Where X,C,N,W,S,E,U,D are 0,1 variables. If X is 1, this is a nonlinear rule; otherwise, it is a linear rule, C,N,W,S,E,U and D are the center, north, west, south, east cells, up cells and down cells respectively.

The values of all eight variables are used to identify each possible rule. If (X,C,N,W,S,E,U,D)=(0010 1111), then the rule is defined as Rule 47, because 101111 is decimal 47. the rule looks like this:

$$S_{i,j,k}(t+1) = \left[N \times S_{i-1,j,k}(t)\right]xor\left[S \times S_{i+1,j,k}(t)\right]xor$$
$$\left[E \times S_{i,j+1}(t)\right]xor\left[U \times S_{i,j,k+1}(t)\right]xor\left[D \times S_{i,j,k-1}(t)\right]$$

A random stream is generated by assigning a rule to each cell, initializing the CA to a random state, and running the CA using a center cell to produce the bit stream.CA is initialized to three-dimensional 3*3*3 array s[i][j][k]={0}, and s[0][0][0]=1; s[0][2][1]=1; s[1][0][1]=1; s[2][1][1]=1; s[2][2][2]=1;For example, each cell is assigned Rule 47.The value s[1][1][1] in the center cell is used to construct the random stream. The cells are randomly initialized, after three steps, the random-bit stream is .01011

THREE-DIMENSIONAL CELLULAR AUTOMATION LFSR

LFSR introduction

The most common is to use a hardware device called a linear feedback shift register (LFSR). Shift register is a very useful hardware device. Because registers are in CPU, the running speed is very fast. This device saves a set of bits. The simply register is 8-bits and 32-bits register is used in P4 CPU. The bits in register can shift to right by using assembly language instruction. Shift logical right

instruction can shift each bit to the right and the leftmost bit is zero. The needed shift register's function is shifting each bits to the right, and the rightmost bit is lost, leftmost bit is replaced with the input bit. Linear feedback shift register (LFSR) choose some bits from the shift register and XOR he input shift in bits. The XOR result is the shift in bits[1].
A general LFSR can be represented by a function of its stored bits and the connections to the shift-in bits. The function is:

$$b_n = c_1 b_1 XOR c_2 b_2 XOR \cdots XOR c_n b_n$$

Where $c_i = 1$, if the bit is selected for the XOR operation; otherwise, it is 0. For example the operation of a simple 8-bit LFSR, in which the shift in is the XOR of cells 3 and 6 is shown in Figure 4.

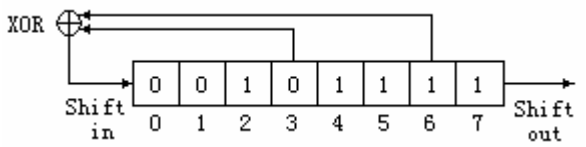


Figure 4. Linear feedback shift register (LFSR)

For example, if 8-bits LFSR is set 0010 1111,in which the shift XOR function among the Shift in bit, cell 3 and cell 6.the procedure is illustrated in Table 1.

Table 1: LFSR Shift Procedure

in	C0	C1	C2	C3	C4	C5	C6	C7	out
1	0	0	1	0	1	1	1	1	1
0	1	0	0	1	0	1	1	1	1
1	0	1	0	0	1	0	1	1	1
0	1	0	1	0	0	1	0	1	1

The LFSR shift procedure include three steps: the first step is calculating cell 3 and cell 6 by XOR to create shift in bit. The second step is shift bit including shift in bit from left to right[1]. The last step is filling the shift out cell with the cell 7.Then loop this three steps until the counter is 0;

Three-dimensional cellular automation LFSR

The three-dimensional cellular automation LFSR combine the dimensional cellular method and linear feedback shift register (LFSR) to create continues stream bits. The method can be represented by a function of its stored bits and the connections to the shift-in bits from both selected bits and CA random-bit stream. The function is

$$b_n = CArb XOR c_1 b_1 XOR c_2 b_2 XOR \cdots XOR c_n b_n$$

CArb stands for cellular automation random bit .Where $c_i = 1$, if the bit is selected for the XOR operation; otherwise, it is 0. For example the operation of a simple 8-bit LFSR, in which the shift in is the XOR of cells 3 and 6 with s[1][1][1]is shown in Figure 5.

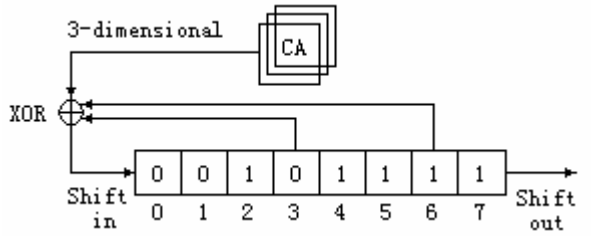


Figure 5: Three-Dimensional Cellular Automation LFSR

For example, if 8-bits three-dimensional cellular automation LFSR is set 0010 1111 same as to the CA Rule decimal 47, in which the shift XOR function among the Shift in bit, cell 3 and cell 6 with the CA random bit. The procedure is illustrated in Table 2.

Table 2: Three-Dimensional CA LFSR Shift Procedure

ca	in	c0	c1	c2	c3	c4	c5	c6	c7	out
1	0	0	0	1	0	1	1	1	1	1
0	0	0	0	0	1	0	1	1	1	1
1	0	0	0	0	0	1	0	1	1	1
1	1	0	0	0	0	0	1	0	1	1

The three-dimensional CA LFSR shift procedure include three steps: the first step is calculating cell 3 and cell 6 with CA random bit by XOR to create shift in bit. The second step is shift bit including shift in bit from left to right. The last step is filling the shift out cell with the cell 7.Then loop three steps until the counter is 0.

RANDOM BIT TESTS

LFSR random bit test

There are several ways to prove the bit stream has the characteristics expected of a random set of bits. The FIPS 140-1 test suite includes some obvious and some not so obvious tests, which are in the collection of National Institute of Standards and Technology (NIST).The FIPS 140-1 includes three tests: mono test, poker test and runs test. Three tests suite accept 20,000 bits from a random source. The first test is mono bit test, which verifies that the number of 1's and 0's are almost equal; The process counts the number of 1's:IF it is within the range 9654-10,346,then the bit stream passes the mono bit test[1].

We initial LFSR with 0010 1111 set in which the shift XOR function among the Shift in bit, cell 3 and cell 6.The mono test result is:

Cycle = 20000
Ones bits Count = 10060
Passes the One bits Count Test (Mono bit test)

The second test is poker test. Passing the mono bit test does not guarantee that a bit stream is truly random. Poker test is another specified test by FIPS 140-1. For the poker test, the 20,000 bits are divided into 4-bit segments. Each 4-bit segment represents a decimal number between 0 and 15. A truly random sequence of bits should result in a random distribution of the numbers 0-15.Let n_i be the number of occurrences of a number i. n_3 is the number of 4-bit 0110.These values are substituted into:

$$X = \frac{16}{5000}\sum_{i=0}^{15} n^2 - 5000$$

The poker test is passed if 1.03<X<57.4.The LFSR poker test result is:
Cycle = 20000
The number of occurrences of number from 0 to 15 is as follows set: {275, 317, 316, 314, 313, 315, 314, 315, 314, 316, 313, 314, 318, 316, 315, 315}
X=4.8896
1.03<X<57.4, Passes the poker test

A third randomness test is called the runs test. A run is a consecutive sequence of either 1's or 0's.In a truly random-bit stream, there should be a random distribution so maximal-length runs. If the number of each run falls within the following guidelines, then the sequence passes the test. the required interval is illustrated in Table 3.

Table 3: Runs Test Required Interval

length	1	2	3	4	5	6+
Min interval	2267	1079	502	223	90	90
Max interval	2733	1421	748	402	223	223

The LFSR runs test is as follows:
Cycle = 20000
Gaps count from 1 to 6 is set 2517, 1261, 630, 315, 157, 158; Runs count from 1 to 6 is set 2519, 1259, 630, 315, 157, 158.Passes the run gap test.

Three-dimensional CA LFSR random bit test

We initial three-dimensional cellular automation LFSR. CA is initialized to three-dimensional 3*3*3 array s[i][j][k]={0}, and s[0][0][0]=1; s[0][2][1]=1; s[1][0][1]=1; s[2][1][1]=1; s[2][2][2]=1;For example, each cell is assigned Rule 47.LFSR is 0010 1111 set in which the shift XOR function among the Shift in bit, cell 3 , cell 6 and s[1][1][1] .The mono test result is:

Cycle = 20000
One bits Count = 10054
Zero bits Count=9946
Passes the Ones Count Test (Mono bit test)

The three-dimensional CA LFSR poker test result is as follows:
Cycle = 20000
The number of occurrences of number from 0 to 15 is as follows set:{299, 303, 315, 315, 303, 314, 313, 315, 315, 317, 317, 313, 316, 318, 312, 315}
X=1.536
1.03<X<57.4, Passes the poker test

The three-dimensional CA LFSR runs test is as follows:
Cycle = 20000
Gaps count from 1 to 6 is set 2524, 1264, 630, 313, 131, 157; Runs count from 1 to 6 is set 2500, 1259, 627, 318, 158, 156.Passes the run gap test.

DISCUSSION

According to the random bits test result, three-dimensional cellular automation (CA) linear feedback shift register (LFSR) passed three stream bit test. Three tests suite accept 20,000 bits from a random source. The first test is mono bit test, which verifies that the number of 1's and 0's are almost equal, the 3-dimensional CA LFSR algorithm can create 1 bits count number is 10054 less than 6 bits compared with LFSR method.3-dimensional CA LFSR algorithm can also pass the poker test. The X value is 1.536 less then LFSR X value 4.8896. The result illustrate the algorithm can get better random stream bits than LFSR algorithm. The 3-dimensional CA LFSR algorithm can also pass the runs test, the 1 bits and 0 bits pass the run and gaps test, the result fill in required interval almost same to the LFSR run gap test. Two algorithm results are shown as Figure 6.

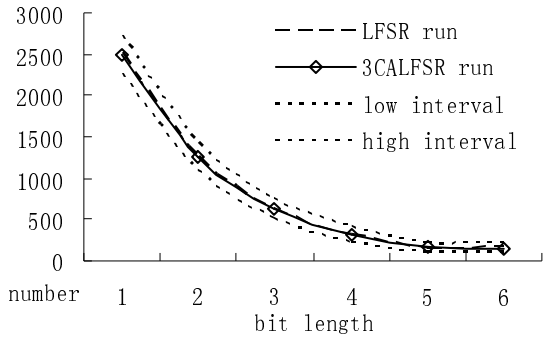


Figure 6: Three-Dimensional CA LFSR Run Gap Test

The three-dimensional cellular automation (CA) linear feedback shift register (LFSR) algorithm combined cellular automation method and LFSR method to create random stream bit, the random tests result illustrate the algorithm is feasible which can create better random stream bit than CA or LFSR algorithm. The algorithm is more inefficient and complicated than the other algorithm.

ACKNOWLEDGEMENTS

Projected supported by Shanghai Leading Academic Discipline Project, China(Grant No. P1303).
Project supported by the Shanghai Committee of Science and Technology, China (Grant No. 065115023).
Project supported by the Shanghai High Education Bureau Fund, China (Grant No. 20065302).

REFERENCES

[1] Richard J.Spillman.(2005) Classical and Contemporary Cryptology, The Tsinghua Press, China
[2] Zhang Bin, Wu Hong-Jun and Feng Deng-Guo. (2005) .On the Security of Three Stream Ciphers, Journal of Software, 16(07),pp. 1344-1351
[3] Lo CC and Chen YJ. (2001). Stream ciphers for GSM networks. Computer Communications, 24(11), pp.1090-1096
[4] Lo CC and Chen YJ. (1999). Secure communication mechanisms for GSM networks. IEEE Trans. on Consumer Electronics, 45(4), pp.1074-1080
[5] Biryukov A, Shamir A and Wagner D. (2000). Real time cryptanalysis of A5/1 on a PC. In: Schneier B, ed. Fast Software Encryption. LNCS1978, New York: Springer-Verlag, pp.2001.1-18
[6] Biham E, and Dunkelman O.(2000). Cryptanalysis of the A5/1 GSM stream cipher. In: Roy B, Okamoto E, eds. Progress in Cryptology- INDOCRYPT 2000, pp. 43-51

An Agent-Oriented Programming Based on OOP

YONG Wang[1] , XINMIN Geng[1] , YU Wang[2]
[1]Dept.of computer Science and Technology,
Shanghai University of Electric Power, Shanghai, China
E-mail: wy616@126.com

[2]China Association for International Exchange of Personnel,
Beijing, China
E-mail: wangyu_caiep@126.com

ABSTRACT
An agent-oriented programming based on C++ language is presented, which is different from object-oriented programming, but is a specialization of OOP. The research aims at defining agent components relationship architecture, and an agent has been implemented. The agent consists of components such as knowledge database, belief, learn, chat, sleep, read state and save state. Agent has public attributes: name, energy, master, init flag and so on. The most important component of an agent is belief, which can control the agent to read state, sleep, learn from master, chat with its master and save state. Without the beliefs agent can't do anything. Under the control of the belief all components can work together if the energy is enough. The energy will increase after sleeping and decrease after learning, which decides beliefs can work or not. The agent-oriented programming can run properly under the control of beliefs with the energy support.

KEYWORDS
Agent; intelligent ;OOP; programming

INTRODUCTION

Agent term is always used in artificial intelligence area, which can be implemented by JAVA or C++ object-oriented language (OOP)[1]. Agent framework is called agent-oriented programming (AOP)[2]. Since from the programmers point of view, AOP can be viewed as a specialization of the object-oriented programming (OOP). Agent consists of components such as belief, desire and intention. This agent architecture is very famous BDI model [3], and many researchers developed the model according to different special functions [4,5]. There are still many differences between AOP and OOP. Unconstrained parameters defining is valid in OOP, yet the parameters in AOP are constrained in the set beliefs, desire, intention, commitments and so on. The set must be defined according to the agent function firstly and can't change after the definition. Agent behavior in an implemented agent system is complicated [6]. Study on human-agent and agent-agent collaboration can help me understand the logic structure [7]. Intention of Agent is very import unit of all of the components [8]. So we designed an agent-oriented program based on OOP. An agent can communicate with other agents and get knowledge from the intercommunication. All agents have their public and private attributes. When the attributes are changed by environment, agent can fell the changes and transport them to the control unit. All the reaction is controlled by the agent beliefs and with the energy support.

MULTI-AGENT SYSTEM DEFINITION

Tuple definition of Multi-Agent system
Agent can't be considered independently of the environment in which they exist and through which they can interact. A Multi-Agent system is a triple, which is a set of Agents, Knowledge and master. The definition is as follows:

$$MAS =< Agents, Master, Knowledge >$$

Each agent is a triple:

$$Agents =< Agent_1, Agent_2, ..., Agent_n >$$
$$Agent_i =< State_i, Master_i, process_i >$$

State is a set of values that completely define the agent. An agent includes some public attributes such as name, energy and init Flag. Process is an executing mapping that changes the agent's state without being invoked from any outside entity. The belief is the agent central unit component that can feel the state change and react autonomously.

Master is a triple:
$$Master =< Name, Agents, process >$$
Master can talk with agents and tell them the knowledge by this means.
Knowledge is a four-tuple:
$$K =< objName, objAttr, attrVal, process >$$

Knowledge class definition for agent

The process of designing a knowledge database of agent begins with an analysis of what information agent must hold and relationships among components of that information. Agent has the learning ability when intercommunicate information with other agents. They begin to learn about the object attributes from the intercommunication. So knowledge database should have the components such as object name, attribute name, attribute value. Once agent get new knowledge, the information should be saved immediately. The save knowledge function should be declared in knowledge database class. Knowledge database definition of agent is as follows:

```
#define N 5
class Knowledge
{
        public:
        char  objectName[10];
        char  attributeName[N][10];
        int    attributeValue[N];
        void   saveKnowledge();
};
void Knowledge::saveKnowledge()
{
        FILE *fp;
        if((fp=fopen("knowledge.txt","ab+"))==NULL)
        printf("Cannot open this file\n");
        fwrite(attributeName,sizeof(Knowledge),1,fp);
        fclose(fp);
}
```

AGENT LOGIC STRUCTURE

Agent logic framework

In related past research by others in artificial intelligence, agent logic structure includes belief, desire and intention components which is the BDI agent architecture. In order to simplify the agent structure, the desire and intention components are included in belief component which control agent. Agent has the public attributes such as name, master, energy, flag, which can be loaded from and saved in database. Agent can communicate with each other, learn from master, chat with master, sleep to increase energy and get knowledge from database. All the actions of agent is under the control of its belief which is the control unit. Agent logic framework is as Figure 1.

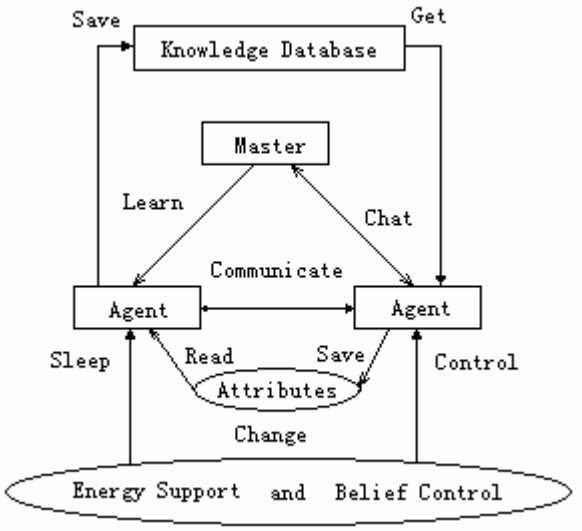


Figure 1: Agent Logic Framework

Agent Class logic structure

Chobit is used as agent class name, which comes from Japanese cartoon. In the cartoon chobit is advanced human like computers symbol and has the self-learn ability and strong feelings. I like the cartoon very much, so I use chobit as agent class name. The main components are chat, learn, sleep, read state and save state. Agent class logic structure is as follows:

```
#include <stdio.h>
#include <string.h>
#include "Knowledge.h"
#define NULL 0
class Chobit:public Knowledge
{
        public:
        int   energy;
        char name[10];
        char master[10];
        char initFlag;
        void Chobit::initChobit(char *chobit_name,char *master_name,int init_energy);
        Chobit(char *chobit_name,char*master_name,
        int init_energy)
{
        initChobit(chobit_name,
        master_name,init_energy);
        belief();
}
protected:
private:
        void readState();
        void readInitFlag();
        void saveInitFlag();
        void saveState();
        void belief();
        void sleep();
        void chat();
        void learn();
};
```

AGENT CLASS IMPLEMENTATION

After definition of agent logic structure, the programming of agents is the central topic. Agent class has some private functions such as belief, read or save state, sleep, chat, learn and so on, which can be used by agents or by themselves. Belief is a very important component, which is the control function of agent and can control all other components work together if the energy is enough. The energy will increase after sleeping and decrease after learning, which decides beliefs can work or not. The agent-oriented programming can run properly under the control of beliefs with the energy support.

Agent construction function

Agent construction function is first loaded by Agent object. The function has three parameters: chobit name, master name and init energy. Belief function is also loaded when the agent object is created. The agent construction function is as follows:

```
Chobit(char *chobit_name, char *master_name,
int init_energy)
{
        initChobit(chobit_name, master_name,init_energy);
        belief();
}
```

Agent belief

The most important component of an agent is belief, which can control the agent to read state, sleep, learn from master, chat with its master and save state. Without the beliefs agent can't do anything. Under the control of the belief all components can work together if the energy is enough. Agent belief function is as follows:

```
void Chobit::belief()
{
        readState();
        sleep();
        learn();
        chat();
        saveState();
}
```

Read and save init flag

When agent object is created, a created mark should be saved in file. Otherwise the agent program begin another time, the last information will be lost. In order to avoid missing the information, the init flag is saved in file. Read and save init flag function are as follow:

```
void Chobit::readInitFlag()
{
        FILE *fp;
        if((fp=fopen("initFlag.txt","r"))= =NULL)
        printf("Cannot open this file\n");
        initFlag=fgetc(fp);
        fclose(fp);
}
void Chobit::saveInitFlag()
{
        FILE *fp;
        if((fp=fopen("initFlag.txt","w"))= =NULL)
        printf("Cannot open this file\n");
        initFlag='n';
```

```
        fputc(initFlag,fp);
        fclose(fp);
}
```

Read and save state

Reading and saving state is different from reading and saving init flag. Init flag is used to save the agent beginning state, while when agent is running, the public attributes should be read and save in files. The reading and saving state function are as follows:

```
void Chobit::readState()
{
        FILE *fp;
        if((fp=fopen("state.txt","rb"))= =NULL)
        printf("Cannot open this file\n");
        fread(&energy,sizeof(Chobit),1,fp);
        fclose(fp);
        printf("%s belongs to %s and energy
         is %d\n",name,master,energy);
}
void Chobit::saveState()
{
        FILE *fp;
        if((fp=fopen("state.txt","wb"))= =NULL)
        printf("Cannot open this file\n");
        fwrite(&energy,sizeof(Chobit),1,fp);
        fclose(fp);
}
```

Agent sleep

Agent sleep function sounds absurdness, maybe you will think program can work without sleep. I want to simulate the real action of an agent, so I added the sleep function. Chobit can increase energy after sleeping when chobit's energy is on the decrease. The agent sleep function is as follows:

```
void Chobit::sleep()

{
        while(energy<3)
        {
                printf("%s want to sleep\n",name);
                        energy+=10;
        }
}
```

Agent chat with master

Agent chat function is interesting. You can talk with your agent and the talking records can be stored in files. But agent can't talk by self-determination. The function is only used to input and save information. Of course agent should have the self-determination ability, the ability is realized in learn function. The agent chatting function is as follows:

```
void Chobit::chat()
{
        char  language[100][50];
        char  whoSpeak[100],masterSpeak[50],
        chobitSpeak[50],index=0,j=0;
        int  byeFlag=-1;
        FILE *fp;
```

```
        while(energy>=3)
        {
                printf("%s speak: ",master);
                strcpy(whoSpeak,master);
                gets(masterSpeak);
                if((byeFlag=strcmp(masterSpeak,"bye"))= =0)
                break;
                strcat(whoSpeak,masterSpeak);
                strcpy(language[index++],whoSpeak);
                printf("%s speak: ",name);
                strcpy(whoSpeak,name);
                gets(chobitSpeak);
                if((byeFlag=strcmp(chobitSpeak,"bye"))= =0)
                break;
                strcat(whoSpeak,chobitSpeak);
                strcpy(language[index++],whoSpeak);
                --energy;
        }
        if((fp=fopen("language.txt","a"))= =NULL)
        printf("Cannot open this file\n");
        for(j=0;j<index;j++)
        {
                printf("%s\n",language[j]);
                fputs(language[j],fp);
        }
        fclose(fp);
}
```

Learn from master

Agent learning function is used to make chobit get knowledge from its master. The knowledge is saved in different files. IF agent wants to recognize an object, agent should have the knowledge about the objects. So the knowledge consists of object name and its attributes. Learning from master function is as follows:

```
void Chobit::learn()
{
        int index=0;
        char continueFlag;
        Knowledge myKnowledge;
        do{
                printf("%s asked objectName? \n",name);
                printf("%s answer: ",master);
                gets(objectName);
                printf("objectName=%s\n",objectName);
                for(index=0;index<=4;index++)
                {
                        printf("\n%s asked attributeName[%d]?
                         \n",name,index);
                        printf("%s answer: ",master);
                        gets(attributeName[index]);
                }
                do{
                        for(index=0;index<=4;index++)
                        {
                                printf("%s=",attributeName[index]);
```

```
                    scanf("%d",&attributeValue[index]);
                }
                myKnowledge.saveKnowledge();
                getchar();
                printf("intput another rule(y/n)?  ");
                continueFlag=getchar();
            }while(continueFlag= ='y');
            energy++;
            getchar();
            printf("another objectName(y/n)?  ");
            continueFlag=getchar();
            getchar();
        }while(continueFlag= ='y');
}
```

AGENT OBJECT IMPLEMENTAION

Agent object implementation is very simple. All agent class definitions are in chobit head file. So we should include the head file in front of the main function. Jin is a chobit series agent and shu is master of Jin. Jin ‘s initial energy is five. The agent implementation is as follows:

```
#include "Chobit.h"
void main()
{
        Chobit jin("Jin ","Shu ",5);
}
```

DISCUSSION

An agent-oriented programming is described, and the program can run properly in console model. The agent can learn from master, store knowledge by chatting and increase energy after sleeping. All the actions are under agent belief control with the energy support. But the program needs ameliorated in chat and knowledge components. Agent can not only store chat records but also mine knowledge from the chat records. Agent should have self-learn ability besides learning from master. Further research will be carried out in these directions and an advanced agent-oriented program with self-learn ability is the next research project. If the ability was realized, the real talking with master will come true.

ACKNOWLEDGEMENTS

Projected supported by Shanghai Leading Academic Discipline Project, China(Grant No. P1303).
Project supported by the Shanghai Committee of Science and Technology, China (Grant No. 065115023).
Project supported by the Shanghai High Education Bureau Fund, China (Grant No. 20065302).

REFERENCES

[1] G. Agha, P. Wegner and A. Yonezawa. (1993). Research Directions in Concurrent Object-Oriented Programming, The MIT Press, United States
[2] Yao Zheng and Gao Wen.(1997). Agent oriented programming, Journal of Software, The Science Press, China, pp. 824-831
[3] Hu Shan-li and Shi Chun-yi. (2000). Agent-BDI Logic, Journal of Software, The Science Press, China, pp. 1353-1360
[4] Hu Shan-li and Shi Chun-yi. (2000). An Intention Model for Agent, Journal of Software, The Science Press, China, pp. 965-970

[5] Shi Chun-yi and Zhang Wei. (2002). A Formal Semantics of Agent Organization Structure Design, Journal of Software, The Science Press, China, pp. 447-452
[6] Lam, Dung N., Barber and K. Suzanne.(2004).Debugging Agent Behavior in an Implemented Agent System, Second International Workshop on Programming Multi-Agent Systems, Springer Verlag, United States, pp. 104-125
[7] Federico Bergenti, M. Brian Blake and Giacomo Cabri. (2004). Agent-Based Computing for Enterprise Collaboration - Human-Agent and Agent-Agent Collaboration, Proceedings of the 13th IEEE International Workshops on Enabling Technologies: Infrastructure for Collaborative Enterprises, IEEE Computer Society, United States, pp. 11-12
[8] Hu Shan-li and Shi Chun-yi. (2006). An Improved Twin-Subset Semantic Model for Intention of Agent, Journal of Software, The Science Press, China, pp. 396-402