A Study of Privacy Protection from Risk of Hijacking Data

Kyong-jin Kim[1*], Seng-phil Hong[1**] and Joon Young Kim[2]

[1]*Sungshin Women's University, Korea*
[2]*Kudos Financial Research Institute, Korea*
{kyongjin, philhong}@sungshin.ac.kr, jkim@kudos.co.kr
** First Author, ** Corresponding Author*

Abstract

The emergence of new technologies is permitting malicious users to collect big data in unexpected way. In particular, privacy accidents related to invasion of privacy and breach of confidentiality happen for many reasons including mere curiosity, and deliberate. However, the hackers today are no longer amateur teenagers who try to break into some data base just to prove they can do it. We are to deal with more organized professional cyber criminals who have enough economic incentive to steal large scale data. Large scale data hijacking affects privacy information breach problem to a large number of people at once. Further it can pull back the big data and related industry to grow. To solve the privacy problem with big data, we considered it necessary to develop the architecture for privacy protection. In order to protect the stored data more efficiently, we need to design the system where we can link the protection algorithm closely with the data protection law and regulation.

Our proposed architecture helps to prevent the data loss or hijacking especially in a large scale.

Keywords: *Privacy Protection, Big data, Personal Information Security*

1. Introduction

We get to witness many cases of personal identity theft, financial data security breach almost every day. Previously, it was more like a personal ID theft of amateur hackers, but the hackers are now more interested in large data hijacking then just somebody else's credit card information. They are no longer small time inefficient child play. As data protection becomes more secure and developed, the attackers gets more organized and professionally equipped with focused intention. Recently, AntiSec hackers claimed that they had been able to breach into FBI servers and gained access to a list of 12 million Apple UDID[1] number. The hackers leaked a portion of hijacked UDIDs online, claiming them to be a part of the list [3].

How this group actually hacked into 'untouchable' fortress of FBI database network is not a main question we have to ask [6]. The concern that we have is how to remedy the problem of large scale data breach and protecting privacy information from now. The problem with large scale data breach is the individual users who had their information hijacked don't even realize what kind of harm or damage it will be for them. Large scale data can provide important implications on many things from upcoming fashion trend and product

[1] "Unique Device Identifier" is simply a specific serial number for your iDevice. It can be considered as social security number for your smart device

development ideas to predicting the election outcome by analyzing people's behavior and information. Therefore, protecting the sensitive information in personal level is very important, but securing the large scale collective data becomes more important than before.

According to the latest findings from Experian's "Life in a Box" experiment [4], which has found that 19.7 million pieces of information were bought and sold illegally in the first half of 2012, which is more incidents than in the whole of 2011, when 19.04 million records were traded. The danger of personal identity breach and the magnitude of its damage are more serious when the data is hijacked in a large scale.

The single largest data breach in Korea was SK Comms data breach in 2011. It affected an estimated 35 million users of SK Comms in South Korea. SK Comms is a subsidiary of SK Telecom providing one of the largest services in Korea that offers on-line and mobile social networking, and instant-messaging (IM) services. The breach affected user accounts of Nate portal and Cyworld, both under SK Comms.

Large scale data breach is not only a problem of private information breach, but it should also be considered 'big data' hijacking in industry level. The "big data" technologies [1, 2] is considered as one of the most valuable assets in recent years, which can analyze huge volumes of data to look for patterns and trends, are now becoming targets to hijack. In a big data, there are personal data that the users already agreed to open and even being used by the third party, but multiple data approved by the users can cumulatively expose sensitive information of the user that he or she didn't want to let others know.

According to McKinsey's recent study, big data market is expected to grow into $53 billion market in 5 years time. Of course the size of the market relies on the optimal combination of hardware, software, and services and applications using the data. If there's weakest link among any of the key success factors, it will negatively affect the size of the market. The incident such as the aforementioned Cyworld data breach case can not only damage at the firm level, but has butterfly effect at the industry level.

Company like IBM is investing heavily on big data. Its Business Analytics and Optimization strategy lead to a series of large investment in the area. IBM acquired SPSS at $1.2 billion in 2010 and Netezza, a data warehousing company at $1.7 billion in 2010 as well. While some companies are spending phenomenal amount of money to gain access to the large data and ability to analyze it, some group or people with malicious intention are hijacking it. The large scale data theft has a very high social cost.

Therefore, the data collected from the users should be stored without personal identification information and treated as a collective statistical data. People agree to open their personal information to some degree to subscribe many digital services based on 'trust' that their information will be stored and used safely. As much as there are huge values from the big data, more caution in handling the data is very important. To address this issue, we present protection architecture to prevent the privacy data loss in a large scale.

The protection of the big data to prevent the individual clients' collective personal information has to be closely related to the government's regulatory efforts. To achieve this, we can suggest imbedded automatic matching and controlling system between the data protection technology and the related law and regulations.

However, there should be enough incentive for service providers to put extra efforts to protect the large scale data. Because preventing data loss and providing better security may incur higher cost. As Akerlof [5] described in his well known Market for Lemon problem, when the provider of lower security quality charges lower price, the high quality service provider may be crowded out if consumers choose cheaper service. Lenard and Rubin [7, 8] suggest that notifying data loss or breach to customer may not be cost effective to the service providers and firms do have less incentive to provide if keeping up with the privacy act and

regulation is too costly. Their thoughts may be justified from the pure economic incentive theory, but they somewhat ignored the potential damage of large scale data loss in their benefit-cost conjecture.

2. Problem Statements

The explosion of big data continues as it brings to picture a wealth of information possessed by many industries such as credit card information, personal security details, medical procedures, diagnosis codes, insurance claims and more. Big data companies and others who would love to get hands on the data try to find the best way to leverage the overwhelming information to achieve tangible business benefits and in the process, head start in the competition. Given the vast amount of personal, financial and behavioral data, the possibilities presented by big data are substantial and set to significantly transform the industry. This possibility also increases the risk of data hijacking in a large scale.

So what could this mean? There can be four areas of risk in data loss.

Table 1. Four Areas of Risk in Data Loss

Items	Details
Loss of Personal ID Info	The loss of personally identifiable information such as date of birth, driver's license number, security number etc.
Loss of Financial Data	With banks and individuals getting more proactive about protecting their financial information, many industries such as online shopping malls have become targets for hackers. Credit card details are increasingly being stolen from purchasing records for making unauthorized purchases.
Loss of Behavioral Data	The location based and customer behavioral data is being fed by GPS tracking, internet site visits, social media, purchasing habits, exercise activity, and self-reporting. Therefore, cyber thieves are targeting for the behavioral data as it can help conjecture asymptotically accurate representations of consumers' traits and behaviors. This kind of data is in huge demand among marketing companies and also others. With the increasing usage of various smart devices, behavioral data is becoming more vulnerable to theft.
Loss of Corporate Data	It is one the most dangerous threats to the corporate security because It is difficult to notice and control. Loss of corporate data can also harm its reputation and the customers may lose their faith. More problems with the loss of corporate data is, when the nature of the lost data is collected data from the general public and contains sensitive personal information, the individuals who entrusted the company with their data wouldn't even know their data is breached until there's some damage. And collectively leads to the deterrence of the related industry in the long run.

According to CS Computer Crime and Security Survey, the satisfactory level is about 2.5 on a scale of 1 to 5 where 1 meaning 'not at all satisfied, 3 meaning 'satisfied' and a rating of 5 meaning 'exceptionally satisfied'. Therefore, respondents were mildly satisfied, but not exceptionally satisfied. On a flip side, this result means that we really don't have reliable solutions for the latest generation of threats and new investments for solution will need to be made.

As more technologies enable us to make our lives easier, the more complications we face in our lives. Our personal information can be in other people's hands so easily when malicious hackers are trying to steal our data. We not only have to try secure our personal data individual level but also in corporate level of big data to make sure that sensitive data cannot be breached and hijacked. We do need to devise some mechanism and protocol to prevent the data loss or hijacking especially in a large scale.

There are many industries that can be affected by the data breach problem, but the followings are more recognized than others;

a. Medical/Healthcare Sector: In the United States, over 20% of the data breaches were in the medical and healthcare sector since 2005. The patients' medical record can be very useful to insurance companies, pharmaceutical companies, and others

b. Financial Sector: About 14% of the data breaches were in the financial field in the United States since 2005. The large scale data loss in collective personal financial data can do direct damage to the exposed individuals.

c. Internet Portals and e-Commerce: massive data is collected and stored with the company.

d. Of course there are other areas of industry such as Education, defense, *etc.*, but we are focusing more on protecting private information in a large scale.

In order to protect the stored data more efficiently, we need to design the system where we can link the protection algorithm closely with the data protection law and regulation. Our solution, Prevent Loss of Privacy Data Architecture (PLPD) is designed to provide cost and time efficient way of data security to minimize aforementioned problem.

3. Prevent Loss of Privacy Data Architecture

Big data companies and others have collected the large scale data in order to achieve their purposes. But in fact the collect of data have led to dangers involving privacy problem with big data. For this reason, we considered it necessary to develop the architecture to prevent the privacy data loss. Our proposed the PLPD architecture aimed to protect and prevent the privacy information in vulnerable network environment (see Figure 1).

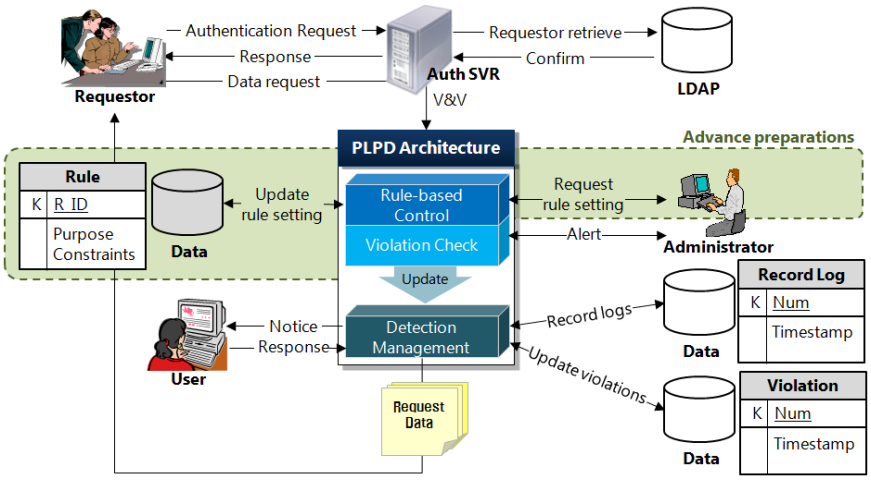

Figure 1. Overall View for Protecting the Privacy Data in Vulnerable Network Environment

PLPD consists of four distinct parts: a Unique ID Authentication Mechanism (UAM), a Rule-based Control Mechanism (RCM), a Violation Check Mechanism (VCM) and a Detection Management Mechanism (DMM). An overall view of the architecture is shown in Figure 2. In order to understand the use of the PLPD architecture, the following flow is considered: (1) an authentication allows a user to access this system. (2) It is managed when using the privacy data by providing the function that automatically compares the privacy rule (3-5), and (6) it informs requestor about the status (permit or deny) of access control in accordance with the rule and constraints. (7) It is important to record the log and violation in all processing.

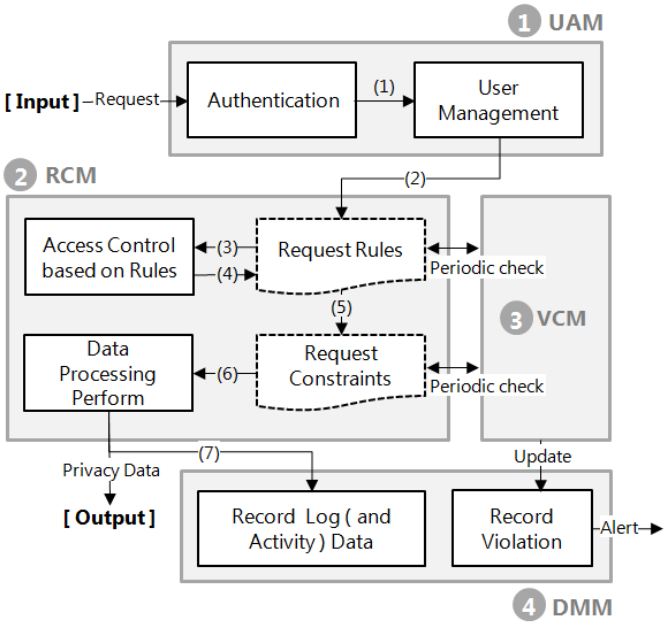

Figure 2. The PLPD Architecture

In this architecture, **1) the UAM** may involve verifying whether a user is suitable as a prerequisite to allowing access to personal information on the network.

2) The RCM is restricted access according to that of the authenticated user through the UAM. Access control in this mechanism is an important function to detect any abnormalities that check the number of frequencies. The access to private data including unique identification and sensitive information can be checked. And then must be check by conditions such as the consent of the data subject prior to the processing of personal identification information. And it performs the obligations after an access action is executed. In order to perform these rules, this mechanism includes rules to ensure compliance with privacy laws and regulations.

All the process can be recorded in **3) the DMM**, and it is to support accountability efforts. The DMM used the log file to record who logged in, when logout, and which devices they use. The log file must be retained for accountability purposes, and also have information about the performed basis by mapping access control rules. It is important to understand condition and obligations based regulations because users have the right to protect own personal information.

4) The VCM is performed when violations occur within system. It is automatically analyzed suspicious activity. The violation includes that an inquiry does not meet user's authorized limit and the quantity of inquiries exceeds the specific standards. In this case, our proposed architecture is able to ensure that the log file records violation in all processing to prevent the data loss or hijacking in a large scale. When the quantity of violation exceeds the specific standards, this mechanism can block user by specific IP address from accessing this system, and then sends an alert with detailed problems to a system administrator when such violation occurs.

4. Implementation and Performance

4.1. Algorithm

Our proposed architecture can play an important role for privacy protection in a large scale. Based on the proposed architecture, we present an algorithm to protect and prevent the privacy data. The key point of our algorithm is that a mechanism protects big data using the control rules closely with the data protection law and regulation.

Algorithm

(1) **if** IsValidating(u_{id}, u_{pwd}, u_{cert}) **then**
(2) *u.valid* ← VerifyPermission(*u*);
(3) *role* ← RoleAssign (*u.valid*);
(4) **if** *u* requests *s.data* **then**
(5) *rule-result* ← RulebasedAccessControl (*role, s.data*);
(6) **switch** (*rule-result*)
(7) **case** PERMIT:
(8) gives all permissions to *u;*
(9) RecordLog (*u*, ActionType, Timestamp);
(10) **break;**
(11) **case** RESTRICT:
(12) **if** *constraints* are not met **then**
(13) gives denied permissions to *u;*
(14) RecordLog (*u*, ActionType, Timestamp);
(15) **else**

```
(16)                    checkResult ← AlertAdministrator(u, s.data);
(17)                    if checkResult is confirmed then
(18)                        gives limited permissions to u;
(19)                        SendNoticetoSubject(u, consent:response);
(20)                    end if
(21)                    RecordLog (u, ActionType, Timestamp);
(22)                end if
(23)                break;
(24)            case DENY:
(25)                gives denied permissions to u;
(26)                RecordLog (u, ActionType, Timestamp);
(27)                break;
(28)        end switch
(29)        if verified u then
(30)            data_size ← CheckSize(s.data);
(31)            if data_size > StandardSize then
(32)                if u.perm ← CheckViolation(u, ActionType) then
(33)                    RecordViolation(u.perm, s.data, violation:reason, Timestamp);
(34)                end if
(35)                RecordLog (u, ActionType, Timestamp);
(36)            end if
(37)            if HasSensitiveData(s.data) then
(38)                RecordDetection (u, ActionType, Timestamp);
(39)                response ← SendAlertMsg (administrator, detect:reason);
(40)                if response is not null then
(41)                    SendNoticetoSubject(u, detect:reason);
(42)                end if
(43)            end if
(44)        end if
(45)    else
(45)        SendNoticetoUser(u, authenticate_failed:reason);
(46)        RecordLog(u, ActionType, Timestamp);
(47)    end if
(48)    if violation is generated then
(49)        result ← CheckAlertEvent(u, ActionType, violation:reason);
(50)        if result is checked then
(51)            SendAlerttoSubject(warning:reason, u, event);
(52)            PerformAudit(s, event, Timestamp);
(53)        end if
(54)    end if
```

4.2. Prototyping

To demonstrate the feasibility of our architecture, we implement a prototype that prevents to the privacy data loss in a large scale. The main composition module is that a mechanism protects personal information based on access control rules.

Figure 3 shows that an administrator can control rules including laws, regulations, and guidelines. The security rule information about the privacy data loss in a large scale located on the top, and it shows personal information type including id, name, etc. And the information about the performed basis by mapping the data protection law and regulation can be verified. Using this screen for the rule setting, administrator can select the privacy information appropriate for request access and make a rule control. At the time, the rule is transmitted as XACML policy document based on XML.

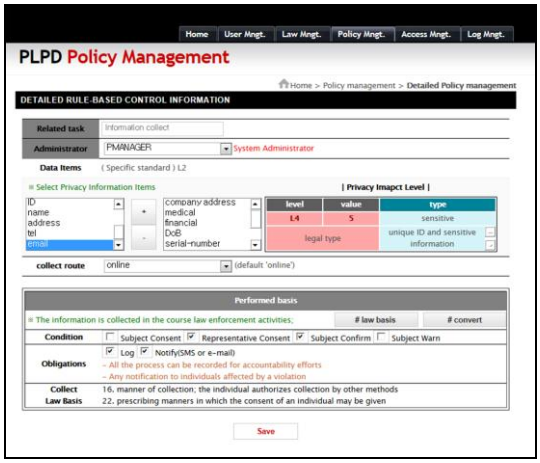

Figure 3. Setting a Rule based on Access Control

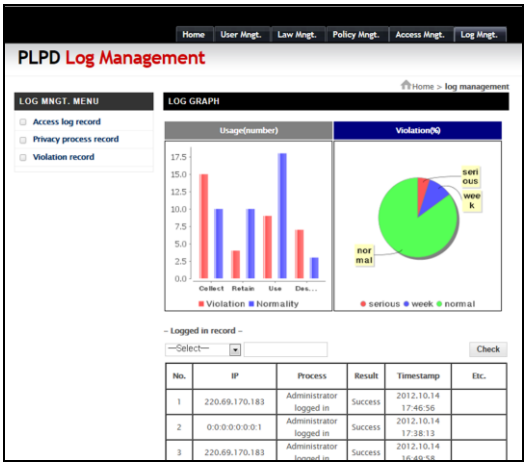

Figure 4. Graph and Record of Violation

Figure 4 presents the risk analysis based rules to ensure compliance with privacy laws and regulations. It shows by who, when, and the steps of processing what data, and it enabled the administrator to recognize the violation rates in graph. If violation occurs, the system is displayed with detail problem information. And it sends an alert with detailed problems to a subject, shown in Figure 5. Also, an administrator receives notice of warning message before using or disclosing when making an information request to a user that want to use the big data, she or he receives notice. Thus, the log data about the all processing records can be detected, and it enables to respond immediately to threats in large scale data.

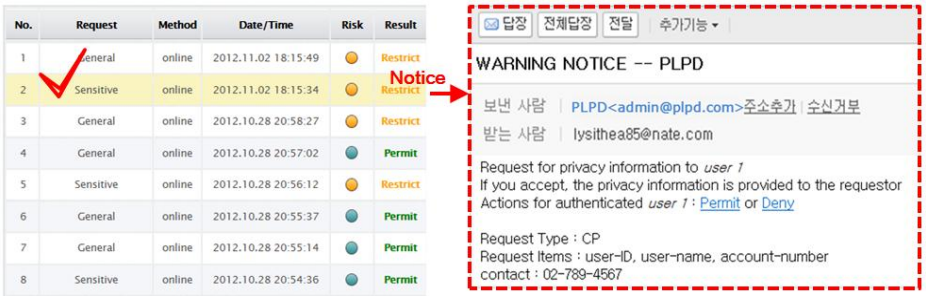

Figure 5. Notice of Alert Message

4.3. Simulation

Based on the prototype, we design a program to simulate its performance. The simulations were performed on Intel Core 2 Quad CPU 2.4GHz with 3.25 GB RAM running on Windows 7 and Apache Tomcat. This result was measured in milliseconds and computed based on the average over simulated runs. As shown in Figure 6, it is compared and analyzed against others with regard to privacy information protection. The blue, red and green lines include our proposed system (shown as a blue line), the existing system with XACML (red line in middle graph) , and the existing system without XACML. Access control rules using XACML affect the response time perceived by the user. Also, compare to the existing system, we can see the PLPD shows better performance in a large scale data.

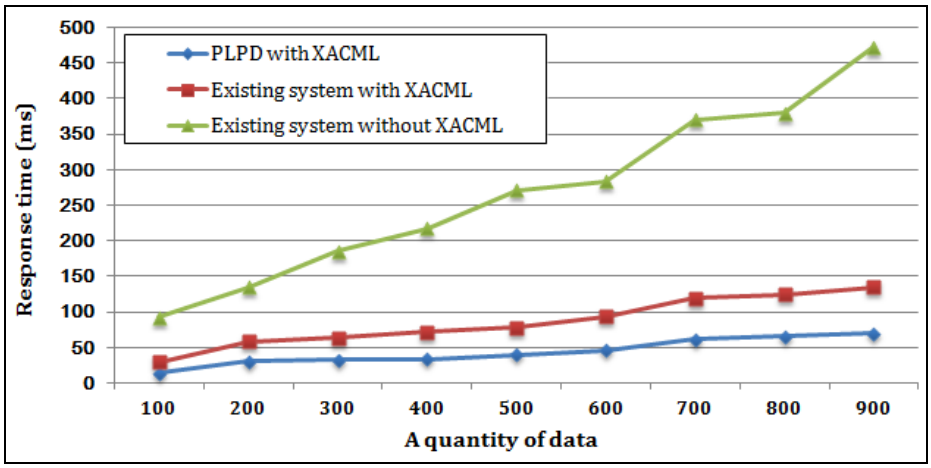

Figure 6. Performance Evaluation

5. Conclusion and Future Work

The progress in the area of information society has increased the risk of invasion of privacy due to the unfair or excessive privacy data collection. In particular, the emergence of new technologies is more interested in large data hijacking. Also, multiple data about private approved by the users can cumulatively expose sensitive information of the user that he or she didn't want to let others know. The problem with large scale data breach is considered as one of the most valuable assets in recent years, are now becoming targets to hijack. As data protection becomes more secure and developed, the attackers gets more organized and professionally equipped with focused intention. For this reason, we do need to devise the protection module to prevent the large scale data loss.

In this paper, we introduce the PLPD to protect and prevent the revelation for privacy information based on access control rules in order to solve the privacy problem with big data. It designed to provide cost and time efficient way of data security to minimize aforementioned problem. Further, it ensures that trusted collect and use of privacy data by applying the relevant laws and rules, and the Internet users or the subjects can be protected. Future studies will continue to focus on the development of our proposed architecture for the practical applicability in order to establish the infrastructure of trusted system in large scale data.

Acknowledgements

This work was supported by the Sungshin Women's University Research Grant of 2012.

References

[1] V. Anil, "NextGen Infrastructure for Big Data", Big Data industry Report, **(2011)**.
[2] Big Data - extracting value from your digital landfills, AIIM- The Global Community of Information Professionals, **(2012)**, www.aiim.org.
[3] J. Ong, Editor, "AntiSec hackers leak 1,000,001 Apple device IDs allegedly obtained from FBI breach", **(2012)**, www.thenextweb.com.
[4] A. Petrou, Editor, "Nearly 20 million pieces of private information illegally traded online", **(2012)**, http://news.techeye.net.
[5] G. Akerlof, "The Market for Lemons: Quality Uncertainty and the Market Mechanism", Quarterly Journal of Economics, vol. 84, **(1970)**, pp. 488-500.

[6] Computer Security Institute 2010/2011 CSI/FBI Computer Crime and Security Survey, **(2011)**, https://cours.etsmtl.ca/log619/documents/divers/CSIsurvey2010.pdf.
[7] T. Lenard and P. Rubin, "Much Ado About Notification", Regulation, vol. 29, no. 1, **(2006)**, pp. 44-50.
[8] T. Lenard and P. Rubin, "In Defense of Data: Information and the Costs of Privacy", Policy & Internet, vol. 2, **(2010)**, pp. 149-183.

Authors

Kyong-jin Kim

Kyong-jin Kim received her B.S. degree and M.S. degree in Computer Science from Sungshin Women's University. Currently she is studying for her Ph.D. course at Sungshin Women's University, and she is majoring in Information Protection. She research interests include access control, privacy protection, and security framework.

Seng-phil Hong

Professor Seng-phil Hong received his BS degree in Computer Science from Indiana State University, and MS degree in Computer Science from Ball State University at Indiana, USA. He researched the information security for PhD at Illinois Institute of Technology from 1994 to 1997, He joined the Research and Development Center in LG-CNS Systems, Inc since 1997, and he received Ph.D. degree in computer science from KAIST University in Korea. He is actively involved in teach and research in information security at Sungshin Women's University, Korea. His research interests include access control, security architecture, Privacy, and e-business security.

Joon Young Kim

Dr. Joon Young Kim received his BA degree in Economics from Yonsei University, and BEc degree in Finance from Macquarie University, Sydney Australia. Dr. Kim received MA degree in Economics from Yonsei University. He received Ph.D. in Economics at Claremont Graduate School, Claremont California in 2006. His major is Industrial Organization and Public Choice in Microeconomics. He was a faculty at University of Southern California as a research assistant professor for the Center for Communications Law and Policy and Center for Asian Pacific Leaderships. He had hosted two series of 'Symposium for Telecommunications Regulations' during his tenure at USC having most of former FCC Chief economists and prominent scholars such as Jerry Hausman, Simon Wilkie, Nicholas Economedies, Philip Weiser, and Michael Riorden. He served as a senior economist at SK Research Institute and actively involved various research societies for telecommunications and broadcast media. Currently, he is the Managing director at Kudos financial research institute.