Mining Online Shopping Frauds—A Non-Data Mining Approach

Kenichi Yoshida, Kazuhiko Tsuda, Setsuya Kurahashi, and Hiroki Azuma

***Abstract*—With the growth of e-commerce, various schemes have emerged to defraud suppliers who offer services and sell goods over the Internet. The deferred payment system, which is a traditional Japanese business practice whereby customers do not pay until goods are received, facilitates online fraud. After receiving goods, fraudulent clients simply disappear and the supplier does not receive the payment. However, since the traditional deferred payment system is expected by honest customers, online shopping sites cannot eliminate this payment system, and consequently are vulnerable to this type of fraud. The conventional approach to detect online shopping fraud is the use of various data mining methods based on statistical analysis. However, outbreaks of new fraudulent clients create new samples that change the distribution of data and decrease the performance of data-mining-based fraud detection. In this study, we propose a new approach that does not rely primarily on data mining. The main characteristic of the proposed approach is the use of the nature of economic crimes. In addition, specific implementations to detect online shopping fraud are proposed. The application of the proposed approach in other areas, such as spam filtering and Internet virus detection, is also discussed.**

***Index Terms*—Online shopping, fraud detection, forged clients, data mining, economic crime.**

I. Introduction

With the growth of e-commerce, various schemes have emerged to defraud suppliers who offer services and sell goods over the Internet. In Japan, online shopping fraud has become a significant problem. The deferred payment system, which is a traditional Japanese business practice whereby customers do not pay until goods are received, facilitates online fraud. After receiving goods from the online shopping sites, fraudulent clients simply disappear and the supplier does not receive payment. However, as the traditional deferred payment system is expected by honest customers, online shopping sites cannot eliminate this payment system, and consequently are vulnerable to this type of fraud. Although online shopping sites take measures to prevent and detect fraud, sufficient outcomes have not yet been achieved.

The conventional approach to detect online shopping fraud is the use of various data mining methods based on statistical analysis. Classification, regression, clustering, prediction, outlier detection, and visualization are techniques typically used to detect fraud. Based on the statistical attributes of transaction data, these methods attempt to distinguish normal transactions from fraudulent transactions. However, fraudsters attempt to steal goods without triggering identification of a fraudulent transaction. They pose as new customers and change their address frequently. Since fraudulent clients change identity and essentially disappear before data mining methods acquire sufficient information to detect their fraudulent activities, the use of data mining methods to expose such clients cannot achieve effective outcomes.

To tackle this problem, this paper proposes a new approach that does not rely solely on data mining methods. The distinguishing feature of the proposed approach is the use of the nature of business/economic crimes. Fundamentally, perpetrators of economic crimes attempt to garner a satisfactory financial gain. This obvious characteristic of economic crimes also applies to online fraud. For example, to make a sufficient amount of money from online shopping fraud by taking advantage of the deferred payment system, fraudsters must place orders in a short period of time using a specific method. This generates particular transaction characteristics that can be analyzed to facilitate the detection and prevention of online shopping fraud.

In this paper, we explain the central concept of our approach, which we call "analyzing business/economic crime (ABC)" approach, and specific implementations to detect online shopping fraud. We also consider the potential application of the ABC approach, i.e., the analysis of the nature of business/economic crime, to other areas such as spam filtering and Internet virus detection.

The remainder of this paper is organized as follows. Section II summarizes past research on fraud detection. The ABC approach is described in Section III. Section IV describes the potential application of the ABC approach to detect spam and Internet viruses. Our findings are summarized in Section V.

II. Related Works

The ever-increasing use of Internet for e-commerce and other financial transactions has led to an inevitable increase of Internet-related crimes [1]. Since fraud detection has become an important issue, academics as well as industry participants have focused on methods that can detect various frauds. For example, [2] attempts to detect payment card fraud; [3] analyzes the relationship between malware and financial crimes. Both [2] and [3] provide practical

Manuscript received May 24, 2013; revised July 20, 2013.

Kenichi Yoshida, Kazuhiko Tsuda, and Setsuya Kurahashi are with Graduate School of Business Science, University of Tsukuba, Otsuka 3-29-1, Bunkyo, Tokyo 112-0012, Japan (e-mail: yoshida@gssm.otsuka.tsukuba.ac.jp, tsuda@gssm.otsuka.tsukuba.ac.jp, kurahashi.setsuya.gf@u.tsukuba.ac.jp).

Hiroki Azuma is with HAZS Corporation. Kojimachi6-2-6, Chiyoda-ku, Tokyo 102-0083, Japan (hirokiazuma2002@hazs.biz).

suggestions that could influence business practices.

The use of data mining techniques has been studied comprehensively. [4] surveys various research in automated fraud detection. It categorizes known frauds and describes the characteristics of data collected in affected industries. Within the business context of mining data to achieve higher cost savings, [4] presents methods and techniques and describes the associated problems. [5] presents a review of the literature on the application of data mining techniques for the detection of financial frauds, such as bank, insurance, and securities and commodities frauds. Recognizing that financial fraud detection is an important emerging topic, [5] presents a comprehensive survey. According to these surveys, various data mining techniques such as classification, regression, clustering, prediction, outlier detection, and visualization have been applied.

Since online shopping and fraud detection have attracted a great deal of attention in recent years, new approaches such as [6] have been proposed. [6] uses an unsupervised learning method based on a finite mixture model to identify pricing fraud. The importance of fraud detection and its impact on business seem to support these studies.

A common characteristic of these studies is the use of data mining techniques that analyze the statistical characteristics of fraudulent transactions. Unfortunately, outbreaks of new fraudulent clients create new samples that change the distribution of data, i.e., statistical characteristics of fraudulent transactions, and decrease the performance of such data-mining-based fraud detection methods. The effective use of data mining and statistical analysis seems to be limited.

This paper proposes a new approach that does not rely on data mining methods. The primary characteristic of the proposed approach is the use of the fundamental nature of economic crimes, i.e., perpetrators of economic crimes seek financial gain. The proposed approach tries to analyze this characteristic. Although the proposed method relies in part on data mining techniques, the analysis of the nature of economic crimes is more important.

III. Analyzing Business/Economic Crimes

The primary characteristic of the proposed ABC approach is the use of the nature of business/economic crimes. In this section, we explain a type of online shopping fraud that abuses the deferred payment system, which is common in Japan. We also describe an implementation to detect fraudulent clients who attempt to steal goods.

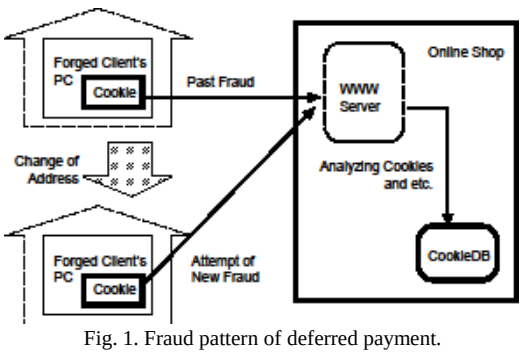

Fig. 1. Fraud pattern of deferred payment.

A. Nature of Economic Crimes

Online shopping fraud that abuses the deferred payment system in Japan is an important issue. This system, a traditional Japanese business practice, facilitates this type of crime. In this system, customers register their addresses with online shopping sites before ordering goods. After receiving the goods and the invoice for their purchase, they remit the charges to the online store's bank account.

Although credit card payment for online purchases is becoming increasingly popular, its adoption in Japan lags behind many other countries. Therefore, in Japan online stores continue to offer deferred payment. Although a credit check is conducted when a shopper registers, it is not particularly rigorous, and this lack of rigor allows fraudulent clients to register. Online stores rely on an address as the main source of information for a credit check.

Reliance on address information to establish a credit rating is problematic. Short-term apartment rentals are becoming popular in Japan. Fraudulent clients can use an apartment for a short period of time by paying rent in advance. If the value of goods they can steal is greater than the rental pay, they can profit from their illegal activities. When in a profitable position, they effectively disappear and rent another short-term apartment. Although obviously illegal, this activity constitutes a business structure for online shopping fraud, and systematic frauds are attempted.

The use of data mining techniques such as classification, regression, and outlier detection cannot detect these frauds effectively. Because the fraudsters change apartments, and therefore change their addresses, which as mentioned previously is primary data used for a credit check, these methods cannot detect relevant statistical characteristics of fraudulent behavior. Thus, relying solely on a data mining approach is insufficient to prevent online shopping fraud.

B. Fraudulent Client Detection

Fig. 1 shows a typical online shopping fraud pattern that uses deferred payment. First, fraudulent clients rent an apartment for a short period of time. Then, they order goods through WWW. After they receive the goods, they move without making payment. By simply moving to new apartments, they can continue perpetrating this fraud. Unfortunately, credit checks when renting apartments are not thorough if the rent is paid in advance. As a result, identifying and prosecuting these criminals is very difficult.

Precise Address: Ostuka 3-29, Tokyo, Japan, Tsukuba Apartment, 632
Decepted Addresses:
1: Otsuka, Tokyo, Japan, Tsukuba Apartment 632
2: Otsuka 3-29-632, Tokyo, Japan
3: Otsuka 3-29, Tokyo, Japan, Tsukuba Residence 632
4: Ostuka 3-29, Tokyo, Japan, Tsukuba Aprtment, 632
5: Otsuka 3-29, Tokyo, Japan, Apartment Tsukuba, 632
6: — Mixure of Kanji Hiragana —

Fig. 2. Address deception.

However, there exist common patterns to maximize the economic benefit. For example, fraudsters tend to continue using the same personal computers (PCs). Thus, online stores can identify PCs used in past fraudulent activities by analyzing cookies and other computer signatures.

Another important pattern in online fraud is address deception, i.e., small modifications in address strings. Fig. 2

shows examples of typical address deception. Let us suppose that the first line is the precise address. Example 1 omits numbers in the address. Example 2 omits the name of the apartment.

Example 3 changes the set phrases commonly used in apartments names, such as "residence" and "apartment." Example 4 involves intentional spelling mistakes. Examples 5 and 6 show the usage of change of the word order and character translation from Kanji to Hiragana. Although these address deceptions make the analysis by computer systems difficult, delivery services tend to find the intended location and unintentionally support fraud. By using deceptive addresses, fraudsters can place multiple orders in a short period of time to maximize the financial benefit.

C. ABC Device

Fig. 3 shows the system configuration of the online shopping fraud detection system based on the ABC approach. Although we use data mining methods to determine fraudulent clients, they are not the main components of this system. The ABC device, which preprocesses data, is of greater importance. The information supplied by preprocessing improves the final detection accuracy of the system.

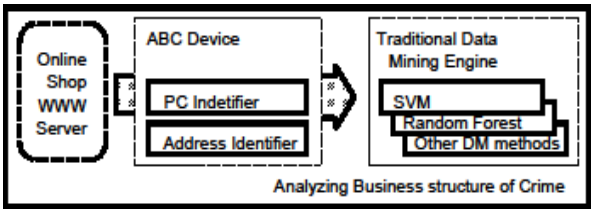

Fig. 3. System configuration.

The PC identifier and the address identifier are the main components of the ABC device. PC identifier analyzes cookies, IP addresses, ISPs, among others. Cookies are used to maintain information during multiple WWW accesses for a single purchase. For example, different WWW pages to select goods, to specify address, and to confirm purchase details share the same cookie for a single transaction. If the expiration time of cookies is sufficiently long, the client can be identified the next time they access the server.

Although cookies can be purged from PCs, such behavior is unusual for honest customers. Therefore, online shops can place higher trust on customers who do not purge cookies. In addition, the act of purging cookies yields certain information. Since IP addresses are supplied by the ISP and a fraudulent client cannot manipulate this information. It also gives variable information to PC identifier.

D. Address Identifier

Although the techniques used in the PC identifier are common within Internet marketing, the address identifier techniques require some explanation. As mention in Section 3.B, fraudulent clients modify their addresses (Fig. 2) to enable multiple simultaneous transactions. By using slightly altered addresses, fraudulent clients attempt to generate multiple identities to allow multiple transactions that exceed limitations for a single customer. However, dishonest clients cannot make significant modifications to the receiving addresses.

Fig. 4 shows the process used to detect address deception. To receive goods, deceptive addresses tend to have the correct sequence for the number component of a precise address or the correct apartment name. To detect deception, the process extracts the number sequence, apartment name, and other address components. Then, it calculates the similarity of each component. The sum of similarities is used to compare the forged addresses.

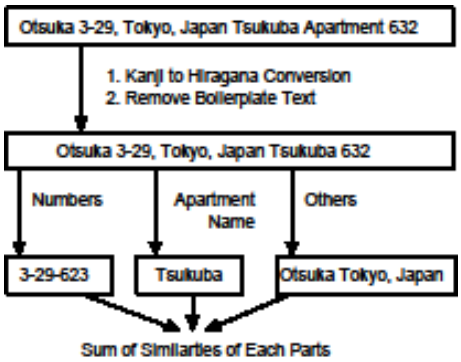

Fig. 4. Process of address identifier.

In the process, all Kanji characters are converted into Hiragana characters. Then, prefecture and city information are used to make a hash table. Although the number of addresses is too large to be retrieved by the sequential algorithm, using the hash table can speed up the process. Note that using fraudulent prefecture or city information is practically impossible. To receive goods, fraudulent clients cannot modify such critical address information. The correction of prefecture and city information using an address database is easy to implement.

The same process is used to compare the addresses of new clients with the addresses registered in the database of apartments. For example, [7] discloses a list of apartments, their addresses, and rental conditions. Since fraudulent clients tend to use short-term rentals, an online shop can restrict transactions from new customers who use such an apartment.

E. Supplemental Use of Data Mining Methods

Some outputs of the ABC device, such as cookies, can provide deterministic conclusions; however, other outputs, such as rental conditions, cannot yield a final conclusion. To analyze various combined information, we use supplementary data mining methods. However, we believe that the most significant component of our approach is the attempt of the analysis based on the nature of business/economic crimes.

IV. General Applicability of the ABC Approach

Although not evident at the time, our past applications [8, 9] can also be seen to follow the ABC approach. This section introduces our past applications to demonstrate the effectiveness of the ABC approach to detect economic crimes.

A. Spam Detection

Spam has become a serious threat, and many studies have been conducted to realize spam filters that protect e-mail

users (e.g., [10]-[12]).

Vector representation, e.g., TF IDF [13], combined with machine learning techniques [14] are commonly used to detect spams. Fundamental problem is the difficulty of the learning problem. Fig. 5 shows a sample vector representation of e-mail. Spammers today seem to have good knowledge about techniques used to detect spam. They try to make smaller information. For example, they make shorter spam with small alterations. They also add random words so that random words disturb the statistical analysis. Such tricks make learning task difficult. In other words, finding discrimination function of spam and non-spam mail on this representation alone is not a simple task.

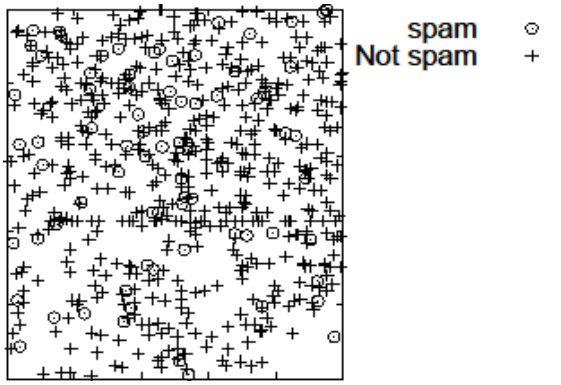

Fig. 5. Traditional SPAM representation.

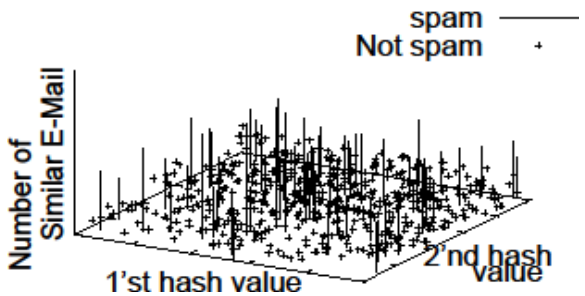

Fig. 6. Nature of SPAM as economic crime.

Although most of the conventional spam filter uses vector representation as the basic data, we use document space density [13] as the key information to distinguish spams from other e-mails. More precisely, we just count the number of similar e-mails as alternate information which we can calculate efficiently. Fig. 6 shows the histogram of e-mails shown in Fig. 5. X and Y axis are that of Fig. 5. Z axis is the number of similar e-mails. As clearly shown in the Fig., use of the histogram makes distinguishing spams from other e-mails far easier.

Spam is employed for legitimate marketing and commercial purposes as well as disreputable purposes such as phishing. To receive sufficient economic benefit, spammers must send huge volumes of unsolicited emails. By analyzing this business/economic characteristic, we have developed a spam filter [8]. The experimental results showed that a simple threshold was sufficient to distinguish spam from a legitimate email. Breakeven performance of the resulting spam filter was approximately 100%. This spam filter analyzes business characteristics of spam; the results show the effectiveness of the ABC approach.

B. Virus Detection

Viruses and malware also have identifiable characteristics. To maximize results, they attempt to infect as many PCs as possible. Thus, once a PC is infected by a virus or malware, the infected PC tries to spread the infection to many other PCs. This produces many packets that contain the source IP addresses of the infected PCs. Mori et al. [15] have reported that the number of distinct elements in TCP/IP headers is important when capturing such behavior. They refer to the number of distinct elements in TCP/IP headers as "cardinality." When a malicious client attempts to locate vulnerable servers, it contacts many servers (scanning). Therefore, malicious clients can be identified by analyzing the number of distinct servers that a single client attempts to access [15]. This can be performed by counting the number of distinct destination IP addresses (Dst_IPs) contained within packets that share the same source IP address (Src_IP).

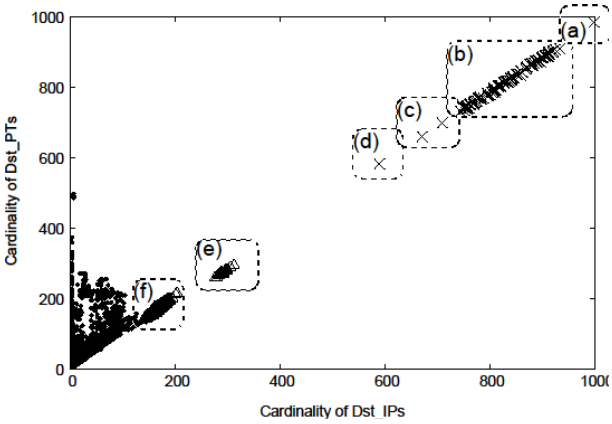

Fig. 7. Nature of internet virus.

On the basis of this concept, we developed an Internet virus detector that analyzes the cardinality of network data. Fig. 7 shows examples of results from the detector. Each point in the figure represents the reported number of distinct Dst_IPs (horizontal axis) and destination ports (Dst_PTs, vertical axis) for specific Src_IP-source port (Src_PT) combinations. These numbers are reported in units of 1000 packets for each Src_IP-Src_PT combination, and the same Src_IP-Src_PT combinations generate multiple points. For example, item (d) in the figure indicates that the number of distinct Dst_PTs for packets from a single port of a single host is approximately 600, as well as the number of distinct Dst_IPs for the same packets are approximately 600.

The figure shows the statistics collected for an Src_IP-Src_PT combination that mainly reflects server output behavior. These statistics can be classified into three groups according to the number of different Dst_IPs.

More than 500 different Dst_IPs: Four combinations of the same Src_IP and Src_PT were found in the data set ((a), (b), (c), and (d) in Fig. 7. These four hosts were either under attack or were trying to find vulnerable hosts. Host (a) sent only packets with RST and ACK flags, and it always used Src_PT 7000. Host (b) sent only packets with SYN and ACK flags, and it also used Src_PT 7000. Hosts (c) and (d) sent only packets with SYN and ACK flags, and they used Src_PT 80.

150-500 different Dst_IPs: Two combinations of the same Src_IP and Src_PT were found in this category (see (e) and (f)). They also used a combination of the same TCP ports and UDP ports. These two hosts appeared to be hub hosts of an overlay network. They tried to keep sessions active with a large number of hosts in order to maintain the overlay network.

Less than 150 different Dst_IPs: These are flows for which no conclusion could be made from this analysis.

Clearly, groups of packets with more than 150 different Dst_IPs tend to be suspect. The analysis of cardinality can identify such problematic traffic caused by a virus or malware. The ABC approach, which focuses on common and identifiable characteristics, is effective in finding viruses and malware.

V. Conclusion

The detection of online shopping fraud by dishonest clients is a significant issue that must be addressed. However, traditional methods that use data mining techniques to detect fraud cannot achieve sufficient performance.

To tackle this problem, we have proposed the "Analysis of Business/economic Characteristics (ABC)" approach. The key ideas of the ABC approach are as follows:

- Business/economic crimes, such as online shopping fraud, share common characteristics to maximize benefits.
- The use of the deferred payment system and short-term apartment rentals are common identifiable characteristics among perpetrators of online shopping fraud in Japan that enable the detection of such fraudulent activity.
- The ABC approach analyzes these common characteristics, such as apartment address and web browser cookie information, to detect fraudulent online shopping.

Our past works, which employed the ABC approach, i.e., a spam filter and Internet virus detector, were introduced to show the general advantages of the ABC approach.

Acknowledgment

Part of this study was supported by JSPS Grants-in-Aid for Scientific Research 25280114.

References

[1] The Ministry of Justice Japan. (2012). White paper on crime (in Japanese). [Online]. Available: http://hakusyo1.moj.go.jp/jp/59/nfm/mokuji.html

[2] K. Brown, D. Pariseau, and D. Chatelain, "Automated payment card fraud detection and location," US Patent 7,543,739, 2009.

[3] BITS: The Financial Services Roundtable. (2011). Malware risks and mitigation report. [Online]. Available: http://www.nist.gov/itl/upload/BITS- Malware-Report-Jun2011.pdf

[4] C. Phua, V. C. S. Lee, K. Smith-Miles, and R. W. Gayler, "A compre-hensive survey of data mining-based fraud detection research," *CoRR*, vol. abs/1009.6119, 2010.

[5] E. Ngaia, Y. Hub, Y. Wonga, Y. Chenb, and X. Sunb, "The application of data mining techniques in financial fraud detection: A classification frame- work and an academic review of literature," *Decision Support Systems*, vol. 50, pp. 559-569, 2011.

[6] K. Kim, Y. Choi, and J. Park, "Pricing fraud detection in online shopping malls using a finite mixture model," *Electronic Commerce Research and Applications*, vol. In Press, 2013.

[7] NEXT Co. Ltd. [Online]. Available: http://www.homes.co.jp/

[8] K. Yoshida, F. Adachi, T. Washio, H. Momotda, T. Homma, A. Nakashima, H. Fujikawa, and K. Yamazaki, "Density-based spam detector," in *Proc. Proceed- ings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2004, pp. 486-493.

[9] Y. Shomura, Y. Watanabe, and K. Yoshida, "Analyzing the number of variety in frequent found flows," *Transaction of the Institute of Electronics, Information and Communication Engineers*, vol. E91-B, no. 6, pp. 1896- 1905, 2008.

[10] P. Graham. (2003). Better bayesian filtering. *The 2003 Spam Conference*. [Online]. Available: http://paulgraham.com/better.html

[11] G. Robinson. (2002). Spam detection [Online]. Available: http://radio.weblogs .com/0101454/stories/2002/09/16/spamdetection. html

[12] SpamAssassin. (2003). [Online]. Available: http://useast.spamassassin.org/

[13] G. Salton and M. J. McGill, *Introduction to modern information retrieval*, McGraw Hill, 1983.

[14] T. Mitchell, *Machine Learning*, McGraw Hill, 1997.

[15] T. Mori, R. Kawahara, N. Kamiyama, K. Ishibashi, and T. Abe, "Detection of worm-infected hosts by communication pattern analysis," in P*roc. Technical Report of the Institute of Electronics, Information and Communication En- gineers*, 2005, pp. 1-6.

Kenichi Yoshida received his Ph.D. from Osaka University in 1992. In 1980, he joined Hitachi Ltd., and is working for University of Tsukuba from 2002. His current research inter- est includes application of Internet and application of machine learning techniques. He is a member of IEEE, ACM, AAAI, IPSJ, and IEICE.

Kazuhiko Tsuda received his Ph.D. from Tokushima Univer- sity in 1994. He joined Mitsubishi electric Ltd. in 1986-1990, Sumitomo Metal Ltd. in 1991-1998, and and is working for University of Tsukuba from 1998. His current research interest includes information retrieval and natural Lan- guage processing. He is a member of IEEE, IPSJ, and IEICE.

Setsuya Kurahashi received his B.S. from The Open University in Japan, M.S., and Ph.D. degrees from University of Tsukuba, Japan, in 1994, 1997, and 2001, respectively. In 2006, he joined University of Tsukuba, where he is currently an Associate Professor of Graduate School of Business Sciences. His research interests include social simulation and social networks. He is a member of IEEE, IPSJ, SICE, JSAI, JASMIN, CSSSA.

Hiroki Azuma is the president of HAZS Corporation. He received his B.S. from Osaka University of Economics in 1987. He joined Royal co., Ltd. in 1987-1989, APLUS Co., Ltd. in 1990-2006, and is working for HAZS Corporation from 2007. His current research interest includes claim information control and small-scale credit risk management.