Information Sciences 231 (2013) 1–3

Contents lists available at SciVerse ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Editorial

Guest editorial: Special issue on data mining for information security

1. Introduction

Computer and communication systems are subject to repeated security attacks. Given the variety of new vulnerabilities discovered every day, the introduction of new attack schemes, and the ever-expanding use of the Internet, it is not surprising that the field of computer and network security has grown and evolved significantly in recent years. Attacks are so pervasive nowadays that many firms, especially large financial institutions, spend over 10% of their total information and communication technology budget directly on computer and network security. Changes in the type of attacks, such as the use of botnets and the identification of new vulnerabilities, have resulted in a highly dynamic threat landscape that is unamenable to traditional security approaches.

Data mining techniques which incorporate induction algorithms that explore data in order to discover hidden patterns and develop predictive models, have proved to be effective in tackling the aforementioned information security challenges. In recent years classification, associations rules, and clustering mechanisms, have all been used to discover and generalize attack patterns in order to develop powerful solutions for coping with the latest threats such as: distributed denial of service (DDoS) attacks, host-based intrusions [15,17], data leakage, SPAM and malicious code including Trojan, Worms and computer viruses [8,9,13,14,16,20].

2. The special issue

The papers in this special issue are clustered into four groups. The first group focuses on employing data mining techniques for coping with intrusion detection. The second group deals with using classification techniques to identify malicious code. The third group mainly addresses privacy preserving data mining. Finally, the fourth group presents new techniques for the detection of the presence of embedded secret messages (Steganalysis) using machine learning techniques.

In the following subsections, we introduce all the papers included in this special issue.

2.1. Intrusion detection systems

Intrusion detection systems aim at detecting and, in many cases, preventing attacks against an information technology resource such as a network router, computer server, etc. Many intrusion detection systems are based on machine learning techniques.

Song et al. [19] propose a new anomaly detection method that can automatically tune and optimize the values of its parameters without predefining them. Song et al. [19] emphasize that in recent studies, many machine learning techniques were employed by intrusion detection systems (IDSs) designers. Unsupervised anomaly detection techniques draw major attention since they can be constructed without using a labeled dataset which, in many cases, does not exist. Despite their advantages, deployment of an anomaly detection system in a real environment is not an easy task since it requires careful settings of values to many parameters that control the system operation. Song et al. [19] propose a new anomaly detection method by which its operators can automatically tune and optimize the values of its parameters without predefining them. The newly proposed method was evaluated using real traffic data collected from Kyoto University honeypots. The evaluation results suggest that the newly proposed method outperforms previous ones.

SQL injections attacks are one of the most widely spread threats for modern databases. In this type of attack the attacker tries to inject a value that appears to be legitimate, but actually modifies the SQL statement that is sent to the database management system. Pinzón et al. [10] present a multi-agent architecture aimed at detecting SQL injection attacks. The proposed method is based on a hierarchical and distributed strategy where different tasks are structured on layers. The agents in each one of the layers focus on specific tasks, such as data gathering, data classification, and visualization. The proposed method is based on two key agents under hybrid architecture: a classifier agent that incorporates a Case-Based Reasoning engine

0020-0255/$ - see front matter © 2013 Published by Elsevier Inc.
http://dx.doi.org/10.1016/j.ins.2013.01.027

employing advanced algorithms in the reasoning cycle stages, and a visualizer agent that integrates several techniques to facilitate the visual analysis of suspicious queries. The former agent incorporates a new classification model based on a mixture of a neural networks and a Support Vector Machine in order to classify SQL queries. The latter agent combines clustering and neural projection techniques to support the visual analysis and identification of the attacks. The proposed detection method was evaluated using real data and the experimental results confirmed the effectiveness of the proposed method.

Protecting sensor network nodes by adding security mechanisms is considered a challenge because wireless sensor nodes are congenitally limited by insufficient hardware resources such as memory size and battery life. Huang et al. [3] propose a new intrusion detection system called the Markovian which aims at protecting sensor network nodes from malicious attacks. The proposed system incorporates game theory with anomaly and misuse detection to determine the best defense strategies. It also employs a Markov decision processes with an attack-pattern-mining algorithm to predict future attack patterns and implement appropriate measures.

2.2. Detecting malicious code

Users download a wide variety of content from the internet including new computer software, some of which can be malicious software. Malwares usually display a hostile behavior which aims to disrupt normal operations or gain unauthorized access to system resources or sensitive information. The volume of malware is growing faster every year and poses a serious global security threat. Consequently, malware detection is becoming a key issue in information security. Presently, signature-based detection methods are the most common methods used in commercial anti-virus. Despite the widespread use of this approach, it fails to detect new malware because, generally, it can only detect malicious executable after it has manifested itself in some way (usually by causing a damage).

Shahzad et al. [18] present a novel concept of genetic footprint which is capable to detect malicious processes at run time. The footprints are discovered by mining the information in the kernel Process Control Blocks (PCBs) of the suspicious process. The genetic footprint consists of selected parameters, maintained inside the PCB of a kernel for each running process, that define the semantics and behavior of an executing process. A systematic forensic study of the execution traces of benign and malware processes is performed to identify discriminatory parameters of a PCB. The new method meets the following characteristics: (1) high detection accuracy, (2) low false alarm rate, (3) short detection time, and (4) resilience to run-time evasion attempts.

Santos et al. [12] propose a new method to detect unknown malware families by analyzing the frequency of the appearance of opcode (operation code of the machine language instruction) sequences. In particular, Santos et al. [12] describe a technique to mine the relevance of each opcode and evaluate how often each opcode sequence appears.

2.3. Privacy and anonymity

Many data mining applications involve mining data that includes private and sensitive information about users. To avoid such situations, privacy regulations were promulgated in many countries (e.g., privacy regulation as part of HIPAA in the USA). The data owner is required to omit identifying data so that to ensure, with a high probability, that private information about individuals cannot be inferred from the dataset released for analysis or sent to another data owner. At the same time, omitting important fields from datasets, such as age in a medical domain, might reduce the accuracy of the model derived from the data by the DM process. Privacy-preserving data mining (PPDM) deals with the trade-off between the effectiveness of the mining process and privacy of the subjects, aiming at minimizing the privacy exposure with a minimal effect on mining results [6,7].

Privacy protection has become increasingly important as Web 2.0 applications are swiftly gaining popularity. Additionally, online social networks accumulate important user data such as age, occupation and role, personality, and more to create a rich identity. Adversaries may extradite sensitive information to unauthorized agents. The pervasiveness of location-aware devices has been the focus of extensive research in trajectory data mining, resulting in many influential applications. Yet, the privacy issue in sharing trajectory data creates an obstacle for effective data mining. For example, based on geo-tagged photos that users share over the internet, one can analyze the trajectories of those users [5]. Chen et al. [1] study the challenges of anonymizing trajectory data; high dimensionality, sparseness, and sequentiality. Employing traditional privacy models and anonymization methods often leads to low data utility in the resulting data and ineffective data mining. As an illustration, Chen et al. [1] aim at preserving both instances of location-time doublets and frequent sequences in a trajectory database, both being the foundation of many trajectory data mining tasks.

Authorship analysis is a forensic tool which aims to expose the author of online textual documents. Authorship analysis is the statistical study of linguistic and computational characteristics of written documents of individuals. Iqbal et al. [4] present a unified data mining solution to address authorship analysis problems based on the concept of frequent pattern-based writeprint.

Privacy-preserving set operations, such as set union and set intersection on distributed sets, are widely used in data mining for which the preservation of privacy is of the utmost concern. Chun et al. [2] extended privacy-preserving set operations and considered privacy-preserving disjunctive normal form (DNF) operations on distributed sets.

2.4. Steganalysis

Steganography is a form of security through obscurity. The idea of steganography is to deliver hidden information in such a way that no one, apart from the sender and intended recipient, may know even suspect them. The advantage of steganography over cryptography is that it does not arouse attention. For example, the well-known MP3 audio format can be used to carry the audio steganography on the Internet. The aim of statistical steganalysis schemes is to detect the existence of secret information embedded by steganography. In their paper, Qiao et al. [11] indicate that there are few steganalysis methods proposed for audio steganography, and especially the scarcity of steganalysis methods of the information-hiding behavior in MP3 audio. Thus, Qiao et al. [11] propose a comprehensive approach for steganalysis of MP3 audio by deriving new features.

Acknowledgements

We would like to thank all the authors who submitted papers for consideration to the special issue. We have received 55 papers from which we could include in the special issue 9 papers. We would especially like to thank the reviewers for their time and detailed reviews that helped us to decide which papers to include in the special issue. Finally, we would like to thank the Editor-in-Chief, Prof. Witold Pedrycz, and Prof. Paul P. Wang, Special Issue Editor, for their valuable guidance and encouragement and the editorial staff for their support in the production of this special issue.

References

[1] R. Chen, B.C.M. Funga, N. Mohammed, B.C. Desai, K. Wang, Privacy-preserving trajectory data publishing by local suppression, Information Sciences 227 (2013) 82–96.
[2] J.Y. Chun, D. Hong, I.R. Jeong, D.H. Lee, Privacy-preserving disjunctive normal form operations on distributed sets, Information Sciences 227 (2013) 112–121.
[3] J.Y. Huang, I.E. Liao. Y.F. Chung, K.T. Chen, Shielding wireless sensor network using Markovian intrusion detection system with attack pattern mining, Information Sciences 227 (2013) 31–43.
[4] F. Iqbal, H. Binsalleeh, B.C.M. Fung, M. Debbabi, A unified data mining solution for authorship analysis in anonymous textual communications, Information Sciences 227 (2013) 97–111.
[5] S. Kisilevich, D. Keim, L. Rokach, A novel approach to mining travel sequences using collections of geotagged photos, in: Proceedings of the Thirteenth International Conference on Geographic Information Science, Springer-Verlag, Berlin, 2010, pp. 163–182.
[6] S. Kisilevich, L. Rokach, Y. Elovici, B. Shapira, A multi-dimensional suppression for K-anonymity, IEEE Transactions on Knowledge and Data Engineering 22 (3) (2010) 334–347.
[7] N. Matatov, L. Rokach, O. Maimon, Privacy-preserving data mining: a feature set partitioning approach, Information Sciences 180 (14) (2010) 2696–2720.
[8] E. Menahem, A. Shabtai, L. Rokach, Y. Elovici, Improving malware detection by applying multi-inducer ensemble, Computational Statistics and Data Analysis 53 (4) (2009) 1483–1494.
[9] R. Moskovitch, Y. Elovici, L. Rokach, Detection of unknown computer worms based on behavioral classification of the host, Computational Statistics and Data Analysis 52 (9) (2008) 4544–4566.
[10] C.I. Pinzón, J.F. De Paz, A. Herrero, E. Corchado, J. Bajo, J.M. Corchado, idMAS-SQL: intrusion detection based on MAS to detect and block SQL injection through data mining, Information Sciences 227 (2013) 14–30.
[11] M. Qiao, A.H. Sung, Q. Liu, MP3 audio steganalysis, Information Sciences 227 (2013) 122–133.
[12] I. Santos, F. Brezo, X. Ugarte-Pedrero, P.G. Bringas, Opcode sequences as representation of executables for data-mining-based unknown malware detection, Information Sciences 227 (2013) 63–81.
[13] A. Shabtai, E. Menahem, Y. Elovici, F-Sign: automatic, function-based signature generation for malware, IEEE Transactions on Systems, Man and Cybernetics: Part C 41 (4) (2011) 494–508.
[14] A. Shabtai, R. Moskovitch, Y. Elovici, C. Glezer, Detection of malicious code by applying machine learning classifiers on static features – a state-of-the-art survey, Information Security Technical Report 14 (1) (2009) 16–29.
[15] A. Shabtai, U. Kanonov, Y. Elovici, Intrusion detection for mobile devices using the knowledge based temporal-abstraction method, Journal of Systems and Software 83 (8) (2010) 1524–1537.
[16] A. Shabtai, Y. Fledel, D. Potachnik, R. Moskovitch, Y. Elovici, Monitoring, analysis and filtering system for purifying network traffic of known and unknown malicious content, Security and Communication Networks 4 (8) (2011) 947–965.
[17] A. Shabtai, Y. Wiess, U. Kanonov, Y. Elovici, C. Glezer, Andromaly: an anomaly detection framework for android devices, Journal of Intelligent Information Systems, submitted for publication.
[18] F. Shahzad, M. Shahzad, M. Farooq, In-execution dynamic malware analysis and detection by mining information in process control blocks of Linux OS, Information Sciences 227 (2013) 44–62.
[19] J. Song, H. Takakura, Y. Okabe, K. Nakao, Towards a more practical unsupervised anomaly detection system, Information Sciences 227 (2013) 3–13.
[20] D. Stopel, Z. Boger, R. Moskovitch, Y. Shahar, Y. Elovici, Using artificial neural networks to detect unknown computer worms, Neural Computing and Applications 18 (7) (2009) 663–674.

Yuval Elovici
Lior Rokach
Department of Information Systems Engineering and the Deutsche Telekom Laboratories at Ben-Gurion University, Ben-Gurion University of the Negev, Israel
E-mail addresses: elovici@bgu.ac.il (Y. Elovici), liorrk@bgu.ac.il (L. Rokach)

Sahin Albayrak
DAILAB, Technische Universität Berlin, Germany
E-mail address: Sahin.Albayrak@dai-labor.de