Association for Information Systems
AIS Electronic Library (AISeL)

AMCIS 2010 Proceedings

Americas Conference on Information Systems (AMCIS)

8-2010

Forensic Data Mining: Finding Intrusion Patterns in Evidentiary Data

Rayman D. Meservy
Information Systems Department Marriott School Brigham Young University, Meservy@byu.edu

James V. Hansen
Information Systems Department Marriott School Brigham Young University, James_Hansen@byu.edu

Forensic Data Mining: Finding Intrusion Patterns in Evidentiary Data

James V. Hansen
Information Systems Department
Marriott School
Brigham Young University
James_Hansen@byu.edu

Paul Benjamin Lowry
Information Systems Department
Marriott School
Brigham Young University
Paul.Lowry.PhD@gmail.com

Rayman D. Meservy*
Information Systems Department
Marriott School
Brigham Young University
Meservy@byu.edu

*Responsible for article correspondence

ABSTRACT

In The extensive growth of computing networks and tools and tricks for intruding into and attacking networks has underscored the importance of intrusion detection in network security. Yet, contemporary intrusion detection systems (IDS) are limiting in that they typically employ a misuse detection strategy, with searches for patterns of program or user behavior that match known intrusion scenarios, or signatures. Accordingly, there is a need for more robust and adaptive methods for designing and updating intrusion detection systems. One promising approach is the use of data mining methods for discovering intrusion patterns. Discovered patterns and profiles can be translated into classifiers for detecting deviations from normal usage patterns. Among promising mining methods are association rules, link analysis, and rule-induction algorithms. Our particular contribution is a unique approach to combining association rules with link analysis and a rule-induction algorithm to augment intrusion detection systems.

Keywords

Data mining, intrusion detection, pattern discovery, rule-induction algorithms, link analysis

INTRODUCTION

An intrusion can be defined as any collection of actions that threaten the integrity, confidentiality, or availability of a network resource. Internet-based global computer systems are increasingly vulnerable to malicious intrusion and related cyber threats because of the high interconnectivity among systems worldwide (Banzhaf, Nordin, Keller, & Francone, 1998; Kumar, Srivastava, & Lazarevic, 2005). Increased globalization and lack of virtual borders make it difficult to assign responsibility and accountability for intrusion prevention (de Borchgrave, Cilluffo, Cardash, & Ledgerwood, 2001; Kumar, et al., 2005). Such attacks can lead to a loss of physical or digital assets, money, consumer confidence, national security, and even life

(Hansen, Lowry, Meservy, & McDonald, 2006). The unprecedented growth in malicious intrusions worldwide is underscoring the importance of developing effective methods for forensic detection of intrusions (Han & Kamber, 2006).

According to Han and Kamber (2006), contemporary intrusion detection systems (IDS) are limiting in that they typically employ a misuse detection strategy, with searches for patterns of program or user behavior that match known intrusion scenarios, or signatures. Generally, these hand-coded signatures are provided by the laborious work of human experts and require a human security analyst to investigate the alarms and to determine what action should be taken. The development of automated or semi-automated methods to support these tasks is fundamental and is the motivation for the study reported here. The discovery of clusters via association rules as the basis for link analysis, and subsequent rule-induction in intrusion forensics, has not been studied before. We provide proof of concept by examining a set of disguised intrusion data secured from a large financial institution.

BACKGROUND

Several studies have explored the use of association rules as a standalone method for intrusion forensics. Abraham and de Vel (2002) consider the use of association rules in investigative profiling. They use an association-mining algorithm, M2IS-c, to generate rules from Linux *wtmp* log files.

Lee, Stolfo, & Mok (1999) propose a data mining framework for intrusion-detection forensics, which considers three approaches to intrusion-detection rules: classification, association rules, and time-related episodes. These authors also suggest combining rules developed from multiple data sets and present some motivating examples of how the rules can work together.

Nanopoulos, Katsaros, & Manolopoulous (2001) devise a system (MADAM ID) to mine audit data as input to automated models for intrusion detection. This method uses data mining techniques to process system audit records in order to detect patterns for defining rule sets.

Link analysis by itself has been used as an analytical tool across a broad range of applications, including detecting terrorist threats, retrieving and classifying Web pages, detecting nuclear proliferation, analyzing transportation routes, detecting money laundering, and finding previously undiscovered medical knowledge (Jensen & Goldberg, 1998). While we were unable to find documented applications of link analysis to intrusion forensics, the Financial Crimes Enforcement Network (FinCEN) (Goldberg & Senator, 1995) is a well known system that uses standalone link analysis in detecting money laundering. The complexity of this domain in disentangling multiple stages of money transfer, exposing the structure and operation of organizations, and characterizing the roles of certain entities in the network, are similar to the complexities found in intrusion forensics.

Research by Li et al. (2005) report on the use of a machine learning algorithm (C4.5) to induce rules from data generated by the Abnormal Border Gateway Protocol (BGP). Abnormal BGP events such as attacks, misconfigurations, or electricity failures can cause anomalous or pathological routing behavior and must be detected in their early stages. The rules of abnormal behavior that were learned were successfully applied to two case studies. A literature search suggests that there is ongoing interest by the research community in the use of data mining to aid in computer intrusion forensics (Qiu, Bao, & Zhu, 2009, Goel, 2009, Masud, Khan, Thuraisingham, Wang, Liu and Zhu 2009, and Weerasinghe, 2009).

While these methods used alone have value, in the case of intrusion forensics, where we must deal with entity properties as well as behavioral data, we propose that a tandem of association rules, link analysis, and machine learning adds robustness to the process. In support of this conjecture, we first introduce the fundamentals of these methods. We then show how they can provide a robust analytical platform for intrusion forensics.

Association Rule Fundamentals

Let $V = \{v_1, ...v_k\}$ be a collection of k attribute values (e.g., Attribute = timeOfDay, Attribute value= 12:00). Denote the task-relevant data as a set of database transactions where each event E is a set of events such that $E \subseteq V$. We can associate each event with an identifier TID. If *J* is a set of values, V contains *J* if and only if $J \subseteq V$. An association rule is an implication of the form:

$$J \Rightarrow K : (s, c), where\ J \subset V, K \subset V, and\ J \bigcap K = \phi$$

The intuitive meaning of such a rule is that events in the database that contain the items in *J* tend to also contain the items in *K*. The letters *s* and *c* denote support and confidence percentages for the rule, where support *s* specifies how frequently the

items in *J* and *K* occur together; and confidence *c* is the conditional probability *P(K|J)*, where the probability *P(x)* is estimated using the support percentage of the set *x*. For example, the rule

$$R:(\text{employeeType} = lanAd\ min) \text{ and } (\text{timeOfDay}) \ between\ \{8am\ \ and\ 5pm\}$$
$$and\ (\text{application} = log\ Scan) \Rightarrow (\text{access} = valid)\ (15\%,70\%)$$

asserts that when the employee type is *lanAdmin*, the time of day is *between 8 a.m. and 5 p.m.* and the application is *logScan*. Then the likelihood that access is *valid* is *70 percent*, with *15 percent* of the total records supporting this claim.

This capability allows the building of rule sets that describe behavioral data. These rules typically describe the behavior of people or the operation of specified systems. Often in computer forensic investigations, the necessary information can be found in log files on computer systems. The rule sets derived from this data can be considered to describe profiles inherent in the data set.

RESEARCH METHODOLOGY

Our research methodology is outlined in Figure 1, which commences with an initial set of association rules derived from forensic data. This is followed by a link analysis (Jensen & Goldberg, 1998) to determine the strongest relationships. This, in turn, facilitates refinement of the association rules to strengthen analysis.

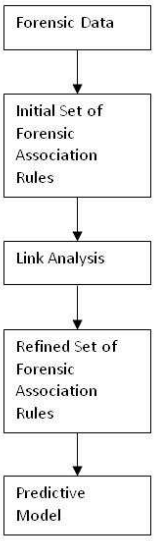

Figure 1. Use of Association Rules in Forensic Analysis

Data

Referring to Figure 1, we commenced with a disguised data set provided by a large financial institution. The dataset was comprised of 50,000 instances from extracted from audit logs of entities within their organization. The set included intrusion profiles that had actually occurred, augmented by those that were captured by honey pot strategies. The attributes of each instance included user (factual) variables (U_i) and behavioral variables (B_j).

While the data was disguised, we can offer some insight into to the type of data by describing several example attack vectors, which are associated with authentication credential theft. These vectors include

Horizontal Credential Guessing

Vertical Credential Guessing

Malware: Keystroke Logging

Social Engineering: Phishing

Social Engineering: Phishing: Man in the Middle

Ethernet Communications Intercept: CAM Table Flooding

Ethernet Communications Intercept: ARP Spoofing

Horizontal Credential Guessing is a technique for obtaining credentials to a system by attempting to authenticate to every user account using a predetermined password list, typically limited to a few commonly used passwords. Where account lockout mechanisms are implemented, the number of guesses per account will typically be set to one below the lockout threshold. Where user ids are not public, valid ids may be harvested through error messages presented during the logon or enrollment process. If user ids cannot be enumerated, a common user id list may be used in the attack. The attack may be executed manually or by using automated tools such as Hydra and Brutus.malware.

Vertical Credential Guessing is a technique for obtaining the password for a known account by repeatedly authenticating to the account using a large dictionary of passwords (dictionary guessing attack). One may also attempt to discover the password by attempting to authenticate with all possible password values (brute-force guessing attack).

Malware is malicious software designed to cause harm to or compromise a computer system. It is typically installed surreptitiously on a system through exploitation of a security hole (via worm or virus) or by deceiving the user into installing the software using social engineering techniques (Trojan software). Malware is also often installed through a dropper program, also originally installed through exploitation of a security hole or through social engineering techniques.

Phishing is a low-cost attack vector that provides a high return. The risk to the adversary in staging the attack is minimal because the attack is staged from the Internet and can be easily hosted on a compromised system, making successful pursuit of the adversary difficult. Phishing attacks are most commonly identified by vigilant users who notify the owner of the compromised brand about the malicious site. Because of the delayed identification and take-down of phishing sites, many users have likely been compromised prior to take-down.

Man in the Middle Phishing employs social engineering techniques to lure users to a proxy controlled by the adversary. The adversary lures users to his site by email or instant message. Established in between user and trusted site communications, the adversary can intercept communications and execute unauthorized transactions within the session.

Content Addressable Memory (CAM) tables reside in memory on network switches and store the MAC address, VLAN assignment, and the associated physical switch port of each system connected to the switch. In normal operation, the switch sends traffic to the physical port associated with the MAC address specified in the Ethernet frame. An adversary can attempt to intercept traffic on the switched network by filling the CAM table with unique MAC addresses using faked gratuitous ARP messages, causing the switch to broadcast all traffic to all switch ports.

ARP Spoofing is a technique that can be used to intercept communications on a switched Ethernet subnet. The adversary conducts the attack by sending spoofed gratuitous address resolution protocol messages to the target systems that advertise his own MAC address as being associated with the IP address of the system or systems for which the adversary intends to intercept communications. By doing so, all systems that received the spoofed ARP information will send information to the adversary's system rather than to the actual systems that were spoofed. Tools for facilitating this attack include Arpoison, ettercap, and parasite.

A *file stealer*, a type of malware, is malicious software that may be used by an adversary to steal files from a compromised system.

Example rules derived from this data might include these:

- If vertical credential guessing and horizontal credential guessing are evidenced with *medium to high* detection risk, then the likelihood of attack is *high* and the expected loss is *medium*.
- If a file stealer is detected, or if CAM Table Flooding and ARP Spoofing are evidenced, with *low detection* risk, then likelihood of successful attack is *low to medium* and the expected loss is *low*.
- If keystroke logging is detected with *medium to high* detection risk, then the likelihood of successful attack is *high* and the expected loss is *high*.

Association Rule Algorithm

The generalized rule induction (GRI) algorithm we use is novel in that it not only learns rules for a given concept (classification), but it concurrently learns rules relating to multiple concepts (Sarawagi, Thomas, & Agrawal, 2000). This type of learning is considerably more general than other existing algorithms. A key feature is the use of an information theoretic measure that quantifies the information content of a rule.

Importantly, the GRI algorithm can help assess the level of interest of derived rules. Some generated rules differentiate relationships in the data better than other rules. Simple rules that are easily understood may be better than complex rules that over-fit the data. Using an information-theoretic view, the confirming information is more important if an event is rare. Useful rules also tend to be very dissimilar among themselves (Smyth & Goodman, 1992). The GRI literature evaluates those differences as the interestingness of the rule (Aggelis & Christodoulakis, 2003; Chen, Han, & P., 1997; Freitas, 1999; Hilderman, Hamilton, & Barber, 1999). Formally, GRI measures interestingness using a *J*-measure (Smyth & Goodman, 1991), which maximizes the trade-off between simplicity and goodness-of-fit for any rule. Using both the a priori and a posteriori probabilities, the *J*-measure facilitates the comparison and ranking of competing rules, allowing efficient pruning of the possible rule set.

The GRI algorithm has several advantages: (a) it can work with both numeric and symbolic inputs; (b) information theoretic bounds help prune the search space; and (c) it has many desirable mathematical properties (including limiting properties).

ANALYSIS

Association Rules

Following the strategy outlined in Figure 1, we initially applied the GRI algorithm to discover associations that exist among the behavioral variables. These results are illustrated in Table 1. Inspection shows that the confidence level is highest (97 percent) for the occurrence of B2, B5, and B6 together as antecedents, with B3 as a consequent. A 97 percent confidence level indicates that when the antecedents are true in the data set, B3 occurs in 97 percent of those instances. This outcome is supported by 3 percent of the instances in the database. Likewise, rule 14 asserts that when antecedent B6 occurs, it will have consequent B5 with a 58 percent confidence. This is supported by 29.3 percent of the instances in the database. The remaining association rules are interpreted similarly.

Instance	Support	Confidence	Consequent	Antecedent 1	Antecedent 2	Antecedent 3
1	3.00	97	B3	B2	B5	B6
2	3.10	94	B5	B2	B3	B6
3	4.00	90	B3	B4	B5	B6
4	4.50	89	B6	B1	B3	B5
5	3.30	88	B6	B2	B3	B5
6	16.70	87	B5	B3	B6	
7	4.60	87	B5	B1	B3	B6
8	17.00	86	B3	B5	B6	
9	4.70	85	B3	B1	B5	B6
10	17.30	84	B6	B3	B5	
11	6.00	68	B3	B4	B6	
12	4.70	66	B3	B2	B6	
13	7.50	59	B3	B4	B5	
14	29.30	58	B5	B6		
15	8.50	58	B1	B6	B8	
16	30.20	57	B3	B5		
17	30.30	57	B5	B3		
18	30.20	56	B6	B5		
19	30.30	55	B6	B3		
20	8.90	55	B8	B1	B6	

Table 1. Association Rules for Behaviors

For the purposes of this paper, we set a confidence threshold of 50 percent, meaning that rules generating lower confidence levels were excluded from consideration. This raises the question of choosing the appropriate confidence threshold; the answer to which may be context dependent. That is to say, if a very large set of association rules is being generated, the level of confidence can be increased. If there are very few rules, this measure can be decreased.

Continuing with the analysis, notice that rule 16 shows B5 as an antecedent for the consequent B3, and rule 17 shows that B3 is an antecedent for consequent B5. Intuitively, this suggests a possible strong affinity between these two behaviors. However, even though this type of visual inspection can be useful, it can also be fraught with error and misinterpretation, particularly when there is a large set of association rules or several antecedents in most rules. A more reliable method of analyzing raw association rules, such as shown in Table 1, has been developed and termed link analysis (Jensen & Goldberg, 1998).

Refining Results with Link Analysis

The volume of data in forensic computing, which includes behaviors plus associations, can be daunting to analyze; yet the amount of relevant data may be small. Here, we use link analysis to aid this refinement. In simple terms, link analysis combines the association rules with visualization. In particular, link analysis explores associations among the behaviors and generates a graphical model of those behaviors. Strong two-way, three-way, and *n*-way relationships can be reliably identified.

For this study, link analysis results are depicted in Figure 2. Here, the bold links (weighted graph) facilitate immediate identification of the strongest affinities among user behaviors. It is straightforward to identify B1 and B8 as having a strong mutual affinity. It is also evident that there is a strong three-way affinity among B3, B5, and B6. Such information provides the foundation for refining our analysis. Having identified those associations that exhibit the strongest affinities, we use GRI to discover association rules that connect the strongly linked behavior profiles with user profiles, thereby generating forensic evidence on what type of user profile is most closely associated with those behaviors.

For reference, we will label the strong behavior profiles as BP1 and BP2, respectively. GRI includes capability for creation of nodes for behavior modules, which can then be computationally associated with user profiles.

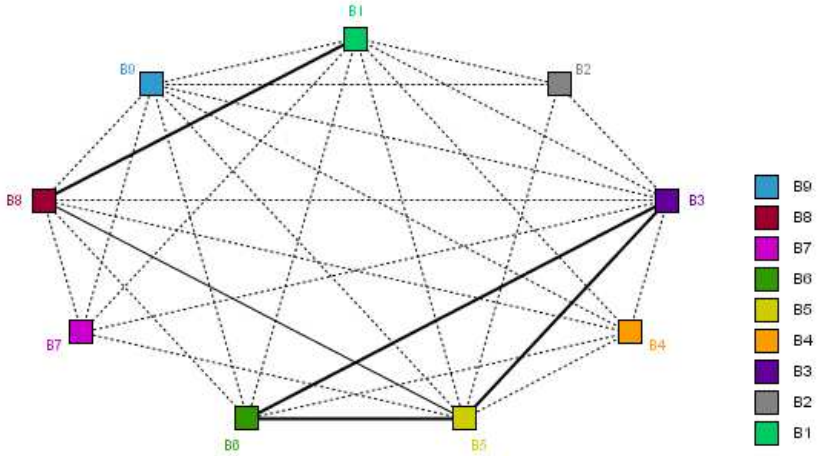

Figure 2. Weighted Graph of Associations

A question arises concerning which of the two profile types should be used as antecedents and which should be used as consequents. From a forensic perspective, the motivating interest is in the formal computation of user profiles and associated behavior profiles; thus, the direction may not be critical. Conventionally, however, behaviors are used as consequents, and user profiles are used as antecedents.

We label the {B1, B8} profile as BP1 and label our {B3, B5, B6} profile as BP2. GRI computation yields association rules between BP1, BP2, and user profiles as shown in Table 2. Observe that there are twelve rules that associate BP1 and BP2 with specific user profiles, along with their support and confidence measures. Consider, for example, rule 1. When its antecedents are satisfied, there is a likely (90 percent confidence) consequent of BP1. This result is supported by 2 percent of the cases in the forensic data set.

Instance	Support	Confidence	Consequence	Antecedent 1	Antecedent 2	Antecedent 3
1	0.50	100.00	BP1	U2 = TTYPE2	U6 > 16.500	U5 >14.850
2	0.30	100.00	BP2	U2 = TTYPE2	U4	U5 < 10.750
3	1.40	93.00	BP2	U5 < 10.750	U3	
4	2.00	90.00	BP2	U2 = TTYPE2	U5 < 12.550	U3
5	3.30	88.00	BP2	U2 = TTYPE1	U3	U5 < 16.850
6	6.30	86.00	BP2	U2 = TTYPE3	U3	U5 < 16.950
7	16.50	84.00	BP2	U3	U5 < 16.950	
8	7.50	81.00	BP2	U3	U4	U5 < 16.950
9	6.90	81.00	BP2	U2 = TTYPE2	U3	U5 < 16.650
10	0.90	78.00	BP1	U2 = TTYPE3	U3	U6 < 16.500
11	0.70	71.00	BP2	U4	U5 < 10.750	U6 < 31.500
12	0.70	71.00	BP2	U2 = TTYPE3	U5 < 10.650	
13	2.00	70.00	BP1	U3	U6 > 16.500	
14	1.00	70.00	BP1	U2 = TTYPE1	U6 < 17.500	
15	3.80	63.00	BP1	U6 > 16.500		
16	2.60	58.00	BP2	U5 < 10.750		
17	1.30	54.00	BP2	U4	U5 < 10.750	
18	4.20	52.00	BP1	U2 = TTYPE2	U3	U6 < 23.500
19	5.20	50.00	BP1	U2 = TTYPE3	U6 < 19.500	
20	3.60	50.00	BP2	U2 = TTYPE2	U5 < 12.550	

Table 2. Rules Associating BP1 and BP2 with User Profiles

To give this context, consider BP1. Suppose that in our BP1 profile {B3, B8}, the following behaviors were represented.

B3: ARM and CRM activity

B8: vertical or horizontal credential guessing;

and that the user attributes (U*i*) included

U2: source host

U5: number of root accesses

U6: number of service requests

rule 1 would then specify that when the user facts consist of source host = TTYPE2, number of root accesses > 16.50, and number of service requests > 14.85, then there is evidence that the of ARM and CRM activity and vertical or horizontal credential guessing.

Similarly, rule 7 would be interpreted as the following: When the antecedents are satisfied, there is an 84 percent confidence of affinity with BP2. This is supported by 16.5 percent of the instances in the data set.

The analysis can be enriched if these profiles are put into predictive models, which can extend forensic analysis into the realm of prediction and prevention. One such option is discussed next.

Predictive Model Extension

The procedures described above have produced information that can be input to a rule-induction algorithm, which yields models whose generalization capability can be tested. Association rules do provide some information on likelihood in the form of confidence and support measures, yet the rules are typically based on the entire data set and provide little information as to how well they may generalize to predicting behavior. In consequence, we selected a rule-induction algorithm, C5.0, whose rule-based output is well suited to representing profiles in an intrusion detection system.

We applied k-fold cross-validation methods as a means of estimating reliable prediction errors for our models. Cross validation directly estimates the out-of-sample error:

$$Err = E[L(Y, \hat{f}(X))]$$

The out-of-sample error is the generalization error when the estimator $\hat{f}(X)$ is applied to an independent test sample from the joint distribution of X and Y. We divided the training sets at random into *k* distinct segments, or folds. We then trained each model using data from *k-1* of the folds and tested performance using the remaining fold. This enables use of a high proportion (*1-1/k*) of the available data to train the models, while making use of all data instances in evaluating the cross-validation error.

Determining the appropriate value for k is not immediate. Ideally, we would use the smallest number of folds producing a stable estimate of error. In the absence of such knowledge, Hastie, et al. (2001) present evidence that five- or ten-fold cross-validation is a good compromise. Mitchell (1997) suggests using ten-fold cross validation for data sets of over *n=100* and leaving one-out (*k=n*) for sample sets from data sets of size less than 100. Accordingly, we select *k = 10* and apply the following algorithm:

Procedure: k-fold cross validation

i. Randomize the instances in the data set

ii. Divide the training set into k equal parts of size n

iii. Do i=1 to k times

> Call the ith set of n instances the validation set and set it aside. Train the system on the remaining k-1 sets; test the trained model on the validation set and record the performance. Clear memory of the learned model.

iv. Calculate the average performance over the k validation tests.

The predictive results for BP1 are shown in Figure 3. The rule is shown in decision-tree form. The accuracy level of 94 percent suggests that this rule may be a reasonably good candidate for inclusion to an update of an intrusion detection system. It can also be useful in assessing the likelihood that an act of malfeasance was committed by a particular suspect. Of course this leaves six percent as false positives, which for large-traffic systems can lead to costly investigations of false alarms. While outside the scope of our work here, if false alarms are to be avoided, the cutoff for classifying an event as an intrusion could be increased in the rule induction model, thereby reducing false alarms, but at the cost of reducing accuracy in predicting actual alarms.

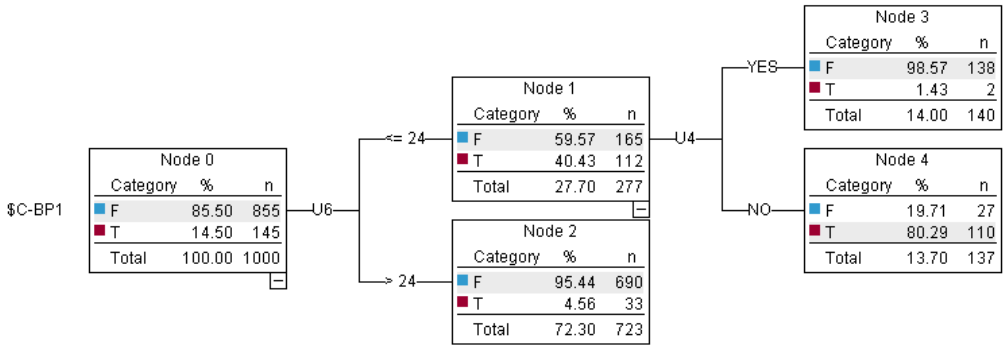

Figure 3. BP1 Boosted Decision Tree from C5.0—Cross-Validation Accuracy 94%

DISCUSSION

The objective of this research was to develop a rigorous methodology for mining association rules in support of the investigative pattern requirement of forensic computing. Clearly, reliable patterns can provide valuable support for computer and intrusion forensics by facilitating the search for offenders and their behaviors.

We denoted an offender profile as having two components: a factual profile and a behavioral profile. In our study, BP1 and BP2 were found (using link analysis) to be the most important behavioral profiles. Based on these profiles, refined association rules were computed, which defined affinities between behavioral profiles and user templates (U1 through U6).

GRI was used to discover an initial set of raw association rules, which were then refined based on the strength of affinities evaluated with link analysis. The resulting behavioral profiles were useful in computing refined association rules that related behavioral profiles to user templates. These association rules enabled immediate extension to a C5.0 predictive model, which facilitated the measuring of generalization reliability of the resulting rules. In terms of application, it is useful to note that this process requires very little user interaction.

CONCLUSION

Computer forensics is the field of forensic science that deals with digital crimes, or crimes that involve the use of computers. Critical to an effective and well-reasoned trail of evidence is the discovery of patterns or associations that are represented in the digitally recorded data. Such associations can find a variety of useful contexts.

This study investigated forensic data mining using a combination of association rules, link analysis, and rule induction. Conceptual foundations of this approach were presented and illustrated on intrusion data supplied by a large financial institution.

REFERENCES

1. Abraham, T., & de Vel, O. (2002, December 9-12). *Investigative profiling with computer forensic log data and association rules.* Paper presented at the 2002 IEEE International Conference on Data Mining (ICDM'02), Maebashi City, Japan.
2. Aggelis, V., & Christodoulakis, D. (2003). *Association rules and predictive models for e-Banking Services.* Paper presented at the 9th Panhellenic Conference in Informatics (PCI'2003), Thessaloniki, Greece.
3. Banzhaf, W., Nordin, P., Keller, R. E., & Francone, F. D. (1998). *Genetic programming: An introduction: On the Automatic Evolution of Computer Programs and its Applications*. San Francisco, California, USA: Morgan Kaufmann.
4. Chen, M., Han, J., & P., Y. (1997). Data mining: An overview from database perspective. *IEEE Transactions on Knowledge and Data Engineering, 8*(6), 866-883.

5. de Borchgrave, A., Cilluffo, F. J., Cardash, S. L., & Ledgerwood, M. M. (2001). *Cyber Threats and Information Security: Meeting the 21st Century Challenge*. Washington, District of Columbia, USA: Center for Strategic and International Studies (CSIS).
6. Freitas, A. (1999). On rule interestingness measures. *Knowledge-based Systems, 12*(5-6), 309-315.
7. Goel, S. (Editor) (2009). Digital Forensics and Cyber Crime: First International ICST Conference, ICDF2C2009, Albany, NY, USA, *Lecture Notes of the Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering,* Vol. 31.
8. Goldberg, H., & Senator, T. (1995). *Restructuring databases for knowledge discovery by consolidation and link formation.* Paper presented at the First International Conference on Knowledge Discovery and Data Mining, Menlo Park.
9. Han, J., & Kamber, M. (2006). *Data Mining: Concepts and Techniques* (2nd ed.). San Francisco: Morgan Kaufmann Publishers.
10. Hansen, J. V., Lowry, P. B., Meservy, R., & McDonald, D. (2006). Genetic programming for prevention of cyberterrorism through dynamic and evolving intrusion detection. *Decision Support Systems, 42*.
11. Hilderman, R., Hamilton, H., & Barber, B. (1999). *Ranking the interestingness of summaries from data mining systems.* Paper presented at the 12th International Florida Artificial Intelligence Research Symposium (FLAIRS'99), Orlando, Florida, USA.
12. Jensen, D., & Goldberg, H. (Eds.). (1998). *AAAI Fall Symposium on Artificial Intelligence and Link Analysis*. Orlando.
13. Kumar, V., Srivastava, J., & Lazarevic, A. (Eds.). (2005). *Managing Cyber Threats: Issues, Approaches and Challenges*: Springer.
14. Lee, W., Stolfo, S., & Mok, K. (1999, May 9-12). *A data mining framework for building intrusion detection models.* Paper presented at the 1999 IEEE Symposium on Security and Privacy, Oakland, California, USA.
15. Li, J., Dejing, D., Wu, Z., Kim, S., & Agarwal, V. (2005). An internet routing forensics framework for discovering rules of abnormal BGP events. *ACM SIGCOMM Computer Communication Review, 35*(5), 55-66.
16. Mohay, G., Anderson, A., Collie, B., de Vel, O., & McKemmish, R. (2003). *Computer and Intrusion Forensics*. Boston, Massachusetts, USA: Artech House.
17. Masud, M., Kahn, L., Thuraisingham, B., Wang, X., Liu, P. and Zhu, S. (2008). Detecting Remote Exploits Using Data Mining, *Advances in Digital Forensics IV*, 285/2008, 177-189.
18. Nanopoulos, A., Katsaros, D., and ManManolopoulos, Y. (2001, August 26). *Effective prediction of web-user accesses: A data mining approach.* Paper presented at the Workshop on Mining Log Data Across All Customer ToughPoints (WEBKDD'01), San Francisco, California, USA.
19. Qui, W., Bao, C. and Zhu, X. (2009). Computer forensic using Lazy Local bagging predictors, *Journal of Shanghai Jiaotong University (Science)*, 14, 1, 94-97.
20. Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. San Mateo: Morgan Kaufmann.
21. Sarawagi, S., Thomas, S., & Agrawal, R. (2000). Integrating association rule mining with relational database systems: Alternatives and implications. *Data Mining and Knowledge Discovery, 4*(2/3), 89-125.
22. Smyth, P., & Goodman, R. (1992). An information theoretic approach to rule induction from databases. *IEEE Transactions on Knowledge and Data Engineering, 4*(4), 301-316.
23. Smyth, P., & Goodman, R. M. (1991). Rule induction using information theory. In G. Piatetsky-Shapiro & W. J. Frawley (Eds.), *Knowledge Discovery in Databases* (pp. 159-176). Cambridge, Massachusetts, USA: MIT Press.
24. Vais, M. (2002). *Law enforcement tools and technologies for investigating cyber attacks: A national needs assessment report*. Hanover, New Hampshire, USA: Institute for Security Technology Studies.
25. Weerasinghe, D. (Editor) (2009). *Information Security and Digital Forensics: First International Conference*, ISDF 2009, London, United Kingdom, September 30 - October 2, 2009.
26. Witten, I., & Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques*. San Francisco, California, USA: Morgan-Kaufmann Publishers.