Data Mining of Forensic Association Rules

James V. Hansen  
*Brigham Young University*

Paul Benjamin Lowry  
*Brigham Young University*

Rayman D. Meservy  
*Brigham Young University, rayman_Meservy@byu.edu*

Follow this and additional works at: [https://scholarsarchive.byu.edu/facpub](https://scholarsarchive.byu.edu/facpub)

Part of the *Management Information Systems Commons*

**BYU ScholarsArchive Citation**
Hansen, James V.; Lowry, Paul Benjamin; and Meservy, Rayman D., "Data Mining of Forensic Association Rules" (2003). *Faculty Publications*. 3242.  
[https://scholarsarchive.byu.edu/facpub/3242](https://scholarsarchive.byu.edu/facpub/3242)

This Peer-Reviewed Article is brought to you for free and open access by BYU ScholarsArchive. It has been accepted for inclusion in Faculty Publications by an authorized administrator of BYU ScholarsArchive. For more information, please contact scholarsarchive@byu.edu, ellen_amatangelo@byu.edu.
Data Mining of Forensic Association Rules

James V. Hansen
James_Hansen@BYU.edu

Paul Benjamin Lowry
Paul_Lowry@BYU.edu

Rayman D. Meservy
Rayman_Meservy@BYU.edu

Marriott School of Management
Brigham Young University
Provo, Utah 84602

Phone: (801) 422-2119
Fax: (801) 422-0621
Abstract

Data mining offers a potentially powerful method for analyzing the large data sets that are typically found in forensic computing (FC) investigations to discover useful and previously unknown patterns within the data. The contribution of this paper is an innovative and rigorous data mining methodology that enables effective search of large volumes of complex data to discover offender profiles. These profiles are based on association rules, which are computationally sound, flexible, easily interpreted, and provide a ready set of data for refinement via predictive models.

Methodology incorporates link analysis and creation of predictive models based on association rule input.

Key words: association rules, forensic computing, data mining, link analysis, predictive models
1. **Introduction**

Forensic computing (FC) is an evolving discipline that involves the preservation, identification, extraction, documentation, and interpretation of computer data for purposes of supporting legal actions, fraud prevention, or investigations of suspected organizational policy violations. FC methods are also finding application in accounting and intrusion forensics. The transcendent characteristic of FC is that it is primarily an after-the-fact analysis of potential and alleged computer crime, where such analysis includes these activities:

1. Acquiring evidence without altering or damaging the original data
2. Authentication that recovered evidence is the same as the originally seized data
3. Analyzing the data without modifying it
4. Investigative profiling.

The first three activities are essential, but investigative profiling is where the greatest potential exits for enhancing capability for identifying computer crime. In particular, there is a need for limiting the potentially large search space endemic to computer-based criminal behavior. This reduction is thought to be achievable through guided processing of data sources to produce patterns of suspected crime. Such patterns can provide profiles that establish the characteristics of attacks, and these profiles can help to identify the types of activities involved in the crime under investigation. (de Vel, Anderson, Corney, & Mohay, 2001)
These efforts are driven by a growing requirement for electronically-stored information that can be used as evidence legal actions, fraud prevention, and security analysis. Zaslow (2003) reports that in a survey of 1,100 U.S. companies conducted by the American Management Association and the ePolicy Institute, fourteen percent of respondents indicated during 2002 they had been required by a legal authority to produce employee e-mail. This is up nine percent from 2001. Other areas that may be sources of FC evidence include computer hardware, operating systems, application software, telephone call logs, room access logs, systems programming and programming languages, as well as systems security logs. Although this information has been traditionally used for legal actions, such as criminal prosecutions, there are other areas where investigative profiling can be particularly helpful. For example, FC evidence can also be used ex post facto to understand how to correct procedural vulnerabilities that can expose a company to fraud, security breaches, and information abuse.

More recently FC has proven useful in support of investigations into fraud and other illicit abuses of proprietary information. Recent cases requiring FC include investigations into Arthur Andersen for unlawful destruction of documents, American Home Products for indifference to human life, an Arizona ISP concerning an email worm, securities fraud using an instant messaging archive, and charges of obstruction against Credit Suisse First Boston (Volonino, 2003).

Perhaps of greatest importance, increases in cyber crimes are extending the reach of forensic methods to discover profiles that are associated with illicit intrusions. Discovered profiles can aid in devising improved intrusion detection systems. Examples of the many cyber crimes that occur on a daily basis include denial of service attacks,
Unauthorized data access, unauthorized data usage, redirection of data, spam, Trojan horses, password theft, security breaches, software piracy, data sniffing, identity theft, identity spoofing, information theft, forged digital certificates, and "salami fraud" (where participants skim small sums from many accounts to avoid detection).

Today, the volume and sophistication of daily information, security, and fraud risks necessitate the need for businesses to turn to FC for legal remedies and prevention. Fortunately, a typical business may possess a valuable mine of data that can be used to discover and prevent such abuses. It is now standard business practice to retain and archive commonly stored electronic information to support organizational needs for tracking activities such as customer and vendor transactions, performance analysis, and for retaining employee work records and email. Moreover, the storing of certain information (i.e., securities transactions, payroll, and virtually all financials), is required by external regulation. Hence, the current electronic repositories held by companies typically provide the necessary data for FC.

Even so, investigation of large volumes of data and the search for patterns and profiles that are useful is a daunting task. Data may be sourced from a complex array of computer workstations, servers, local and networked peripheral devices, network data and routers, removable hard drives, personal data assistants (PDA's), backup media, and email servers. The crucial task is to analyze these large databases effectively to find interesting and unsuspected patterns that are understandable and useful to a forensic investigator.

Data mining offers a potentially powerful method for analyzing the large data sets that are typically found in FC investigations to discover useful and previously unknown
patterns within the data. As compared to other more conventional technologies, data mining is able to produce useful results from large data sets characterized by incompleteness and noise (Ester, Kriegel, & Sander, 1996). Data mining applications currently exist for content-based and collaborative filtering-based recommendations systems, customer profiles in fraud detection, and web browsing activities.

The contribution of this paper is an innovative and rigorous data mining methodology that enables effective search of large volumes of complex data to discover offender profiles. These profiles are based on association rules, which are computationally sound, flexible, easily interpreted, and provide a ready set of data for refinement via predictive models. We posit that by profiling offenders using data mining techniques, organizations will be more effective in the use of FC to prevent fraud, information abuse, and security violations.

This paper proceeds as follows: Section 2 summarizes related work. Section 3 outlines the fundamentals of pattern discovery. Section 4 presents the association rule paradigm, which forms the core of the data mining technology that we use, and Section 5 gives the technology context in terms of behavioral profiling. Section 6 describes our database, the generalized rule induction algorithm, and experimental procedure. Section 7 suggests and demonstrates a useful extension of association rules to construction of a predictive model. Section 8 recaps the method and results. Section 9 concludes with summary remarks.

2. Related Work
Data-mining techniques for dynamic intrusion forensics have been suggested by others. Abraham and de Vel (2002) investigate the use of association rules in investigative profiling. They use an association-mining algorithm, M2IS-c, to generate rules from Linux wtmp log files. However, M2IS-c is an a priori algorithm that fails to consider a posteriori values discussed later in this paper. Abraham and de Vel (2002) use conceptual hierarchies to identify potential anomalous behavior and build the profile while concentrating on contradictions in multiple rule sets. They suggest that interestingness measures used in other data mining studies should be considered (Abraham & de Vel, 2002).

Lee, et al. (1999) describe a data mining framework for building intrusion-detection forensics. They explain three approaches to intrusion-detection rules: classification, association rules, and time related episodes. Lee and his associates also consider combining rules developed from multiple data sets and present some interesting examples of how the rules may work together.

A broader survey of intrusion detection approaches is found in Mohay et al. (2003), who examine the evolution of intrusion detection systems and the problems with modern software. They note the importance of data-mining and machine-learning techniques for dynamic compilation of new attack signatures. Although they summarize several approaches, including the data mining approach, an analysis of underlying algorithms is not presented.

3. Forensic Pattern Discovery
This section describes the analytic procedures used to identify patterns in forensic data, which patterns are then used to create profiles. A pattern describes a structure and a set of relationships that characterize a set of data records. "Interesting" patterns include those that describe a typical profile, which in turn enables an investigator to make nontrivial predictions on new data. Some example patterns are suggested by Mohay, et al. (2003):

1. 55 percent of middle managers access the financial database of a company

2. A subset of log data records reveals that certain types of users log in after hours

3. Specific log data records indicating user X is accessing system files to which the person is not authorized to access.

Patterns that profile usage patterns facilitate the identification of deviations from normal patterns. This identification is a two-phase process. The first phase is generating the patterns, which can be difficult. That is to say, if we have N attributes, each with M possible values, then we would require $M^N$ patterns; and for all but the simplest cases this is impractical. Fortunately there often exists some structure among patterns that algorithms can exploit.

In particular, some combinations of attributes may regularly occur together, and these combinations can then be used to form patterns that are termed association rules. For example, if Harry regularly logs in early in the day, then $User = Harry$ and $TimeOfDay = morning$ might be two attributes and values that frequently occur together. More complex pattern structures can be identified by robust pattern-finding algorithms. Such algorithms must be scalable, robust to high dimensionality, and resistant to over
fitting of the data. *Scalability* means that an algorithm should be able to handle both small and large databases. *Dimensionality robustness* ensures an algorithm’s effectiveness when it processes a large number of attributes. *Over fitting resistance* means that an algorithm does not fit its model so tightly to a given data set that general structures which may enable accurate forecasting are missed. Association rule mining brings these capabilities to FC as discussed in the next section.

4. **Fundamentals of Association Rules**

This section further describes association rules theoretically, and then shows how useful rules can be used for forensic profiling. Let \( I = \{ i_1, i_2, \ldots, i_n \} \) be a collection of items. Denote the task-relevant data as a set of database transactions where each transaction \( T \) is a set of items such that \( T \subseteq I \). Then we can associate each transaction with an identifier, called TID. If \( J \) is a set of items, \( T \) contains \( J \) if and only if \( J \subseteq T \). An association rule is an implication of the form

\[
J \Rightarrow K: (s, c), \text{ where } J \subseteq I, K \subseteq I, \text{ and } J \cap K = \emptyset.
\]

The intuitive meaning of such a rule is that transactions 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 \( K \), 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

\[
R: (\text{employeeType} = \text{lanAd min}) \text{ and } (\text{timeOfDay}) \text{ between } \{8\text{am and } 5\text{pm}\} \text{ and } (\text{application} = \log \text{Scan}) \Rightarrow (\text{access} = \text{valid}) \ (15\%, 70\%)
\]
produced 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 a profile inherent in the data set. Since these rules are rarely complete, the support parameter can provide useful information to the analyst: First, only data that occurs frequently enough to satisfy a predefined minimum support percentage may be allowed in the rule generation process. Second, regularities that are not included in profile (due to not satisfying the support percentage threshold) may be viewed as non-habitual and can be investigated as contrary to regular behavior.

These considerations lead naturally to the question of what constitutes an interesting rule. Consider, for example, the rule

\[ Audit \Rightarrow checkFailedLoginAttempts (8\%, 70\%) \]  

Suppose *Audit* is a parent of *transLogAudit* and about a quarter of Audit activities are *transLogAudit*, one would expect the rule

\[ transLogAudit \Rightarrow checkFailedLoginAttempts \]  

to have 2% support and 70% confidence. If the actual support and confidence for this assertion are near 2% support and 70% confidence, the rule can be assessed as redundant since it does not convey any additional information and is less general than the first rule. By this reasoning "interesting" rule can be defined as one whose support is more than *K*
times the expected value, or whose confidence is more than \( K \) times the expected value for some user-specified \( K \). In the above example, suppose the support is 5% and the confidence is 70%. If \( K \) is set at 2 then the increased level of support would make (1) an interesting rule.

5. **Investigative Profiling**

Given the above definition of interesting association rules, those rules provide a basis for creating useful offender profiles for FC. An offender profile is comprised of two elements: a factual element and the behavioral element. The factual profile is the accurate background knowledge about the offender; such as their name, employee status, computer-user name, relationships with other employees and organizations, and so forth. The behavioral profile is comprised of knowledge about an offender’s behavior.

Behavioral profile (BP) knowledge is typically constructed from sources such as log file transaction, header and body of emails, telecommunications call-record data patterns, and so forth.

A BP can be modeled in different ways, depending on the needs of an analyst. For instance, a BP could be depicted as a union of sub-profile hierarchies (PH\(_j\)) such as authorship profile, software application usage profile, log-in profile, and so forth—as denoted below:

\[
BP \leftarrow \bigcup_j^M PH_j
\]

An alternative is to model BP’s as sets of association rules:

\[
BP \leftarrow \{ R_i | i = 1, 2, ..., N \}
\]
In this case, the rule attributes can be obtained from raw data or selected from the profile hierarchy nodes. For example, the rule

\[ \text{If user U is a LAN administrator, then the application} \]
\[ Y = \text{Snort executed} \] \hspace{1cm} (7)
may be a valid rule in a LAN administrator profile; but it may not be valid for a profile of a financial analyst.

In our study, we examine user behavioral profiles derived from event data in log files. The relevant profiles are straightforwardly represented by a set of association rules

\[ BP \leftarrow \{ R_i | i = 1, 2, \ldots, N \} \] \hspace{1cm} (8)

Such rules provide an intuitive and declarative way to describe user behavior (Fawcett & Provost, 1997). For instance, recall the rule

\[ R: (\text{employeeType} = \text{lanAd min}) \text{ and } (\text{timeOfDay}) \text{ between } \{8\text{am} \text{ and } 5\text{pm}\} \text{ and } (\text{application} = \text{log Scan}) \Rightarrow (\text{access} = \text{valid}) \] \hspace{1cm} (9)
asserting that “LAN administrators have valid access to logScan applications between 8 am and 5 pm.” Such association rules are based on some correlation between a rule’s antecedent and consequent; because these are correlations, causality cannot be inferred.

The initial action in a computer forensic investigation is to look for unusual events—or exceptions to generally accepted profiles. For example, if an intruder gains root privileges to a computer system, then the intruder may use the privileges to perform actions not normally associated with administrative users. Such an intrusion event would amount to a deviation from the administrative root profile, which deviation was generated from data collected prior to the intrusion. Elsewhere (Han & Kamber, 2001) it has been
argued that there are two ways unusual events may arise in profiles generated from a target data set:

1. As data entries with not enough support to be represented as association rules. To find such entries, a mechanism must exist for investigators to query data sets for entries that do not fit a profile.

2. As association rules making up part of a profile. When this occurs, it is important to identify this portion of the profile as being different from other parts of the profile. Assigning a temporal scope to rules making up the profile could help an investigator become aware that something potentially unusual may have occurred at a certain time in the life of the system being investigated.

6. Application and Analysis

Given the preceding theoretical background, we now turn toward application and FC analysis of a real-life data set. Our approach is outlined in Figure 1, which commences with an initial set of association rules being derived from forensic data. This is followed by link analysis (Jensen & Goldberg, 1998) to determine the strongest relationships. This, in turn, facilitates refinement of the association rules, which refinement may provide sufficient basis for analysis involving the capture of digital evidence. Where intrusion forensics are involved—attacks on the computer system itself—it can be helpful to use the association rules as the basis for developing a predictive model to be used as part of an intrusion detection system.

---Insert Figure 1 approximately here---
Correspondingly, the data set we chose involves intrusion data. There is a clearly identifiable area of overlap between intrusion detection and computer forensics activities that associate the forensic investigation of attacks on a computer system (Mohay et al., 2003). Intrusion detection is the investigation of activities concerning use or access, or which threaten to use or access, a computer or computer system in a manner that is unintended or undesired for a computer system. Intrusion forensics differs from intrusion detection in that the latter refers to what is achieved by the intrusion detection system. Intrusion forensics is the gathering of evidence, which may be used to incorporate new features into an intrusion detection system, but may also be used to support evidentiary requirements for a court of law. Of growing importance is the need for intrusion forensics in support of information warfare, which is of growing concern to national security.

**Data**

We consider a set of proprietary (disguised) forensic data, which is comprised of six user (factual) variables \((U_i)\) and nine behavioral variables \((B_j)\). Although the data is disguised, the data set considered here is in the typical intrusion forensics realm, which may consider factors such as protocols used, time of connection, network service sought at destination, failed logins, errors, activities found on a 'suspicious' list, and so forth. The data set used is comprised of 1004 instances.

**Generalized Rule Induction 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 multiple concepts. 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 interestingness 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). Formally, GRI measures interestingness using a $J$-measure, which maximizes the tradeoff between simplicity and goodness-of-fit for any rule. Using both the prior and posterior probabilities, the $J$-measure facilitates the comparison and ranking of competing rules, allowing efficient pruning of the possible rule set.

**Procedure**

As a first step GRI is used to search for associations in the behavioral variable group. Using link analysis, associations having the strongest affinities for one another are then identified. If these procedures yield interesting profiles, GRI can then be applied to identify user profiles associated with the interesting behavior profiles (BPs). These procedures provide a foundation for sound forensic analysis by profiling the nature and identity of malefactors.

If we wish to extend the results of the forensics analysis to preventive identification, the GRI association rules can be used to derive a predictive model and to measure its reliability. This will be discussed in a later section.
Following the strategy outlined above, we apply the GRI algorithm initially 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 ninety-seven 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 three percent of the instances in the database. Likewise, Rule 14 asserts that when antecedent B6 occurs it will have consequent B5 with 58 percent confidence. This is supported by 29.3 percent of the instances in the database. The remaining association rules are interpreted similarly.

---Insert Table 1 approximately here---

For purposes of this paper, we set a confidence threshold of 50 percent, meaning that rules generating lesser 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 might 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. This type of visual inspection can be useful, but 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).

**Link Analysis**

The volume of data in FC, 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 the FC analyst, link analysis can indicate where to focus an investigation and confirm or deny suspicions (U.S.-Congress, 1995). Link analysis has proven its value in a number of applications, which include detecting terrorist threats, retrieving and classifying Web pages, detecting nuclear proliferation, studying transportation problems, discovering money laundering, and finding new medical knowledge.

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 now 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 can 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.

---Insert Figure 2 approximately here---

A likely 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 \{B_1, B_8\}-profile as BP1 and label our \{B_3, B_5, B_6\}-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 2.0 percent of the cases in the forensic dataset.

---Insert Table 2 approximately here---

To give this context without compromising confidentiality, consider BP1. Suppose that our in our BP1 profile (\{B_3, B_8\}) the following behaviors were represented:

B3: login belongs to a hot list
B8: host scan;

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 \( sourcehost = TTYPE2, \)
\( \text{number of root accesses} > 16.50, \) and \( \text{number of service requests} > 14.85, \) then there is
evidence that the login is on a hot list and that the accessing party is executing a host scan.

Similarly, Rule 7 would be interpreted as follows: 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.

Such information can be useful in providing evidence for legal proceedings, or for
determination of violations of corporate or human resource policies. The identification of
these final association rules can also provide useful input to predictive models, which can
extend forensic analysis into the realm of prediction and prevention. One such option is
discussed and demonstrated in the following section.

7. **Association Rules and Predictive Model Extension**

Thus far we have argued that FC can be supported in reliable ways by identifying
association rules that link user profiles with behavior modules. These rules can be
represented in textual form or as decision trees, both of which provide the investigator
with a simple way of describing the patterns relevant to an investigation (Mohay et al.,
2003)

An added benefit of having discovered relevant association rules is that we have
information that can be input into predictive models that provide rigorous measurement
of the predictive reliability of the discovered associations. Recall that association rules provide some information on likelihood in the form of confidence and support measures. However, these models are typically based on the entire data set and provide little information as to how well they may generalize to predicting behavior.

If, for example, a forensic investigation is to develop rules that can be used to update intrusion detection systems, it can be beneficial to assess how well these rules generalize to unseen instances. We use the C5.0 inductive algorithm (Witten & Frank, 2000) to generate predictive rules from the discovered association rules. C5.0 operates by splitting a training set based on the field that provides the maximum information gain. Each subsample defined by the first split is then split again (usually on a different field). This process repeats until the subsamples cannot be partitioned any further. At that point, the lowest level partitions are reevaluated; and those that do not contribute significantly to the value of the model are pruned. We note that the important association profiles generated earlier allows straightforward input to C5.0.

The predictive results for BPI are shown in Figure 3. The rule is shown in decision-tree form. Importantly, use of C5.0 has allowed use of boosting methods (Friedman, Hastie, & Tibshirani, 2000) and cross-validation to achieve a reliable estimate of generalization accuracy. The accuracy level of 94 percent suggests that this rule may be 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.

---Insert Figure 3 approximately here---
The results for BP2 are depicted in Figure 4. The accuracy on holdout sets here averages 97 percent, which again argues for the reliability of the rule, which could be used in either of the contexts mentioned above.

---Insert Figure 4 approximately here---

8. Discussion

Our objective is to develop an innovative and rigorous methodology for mining association rules in support of the investigative pattern requirement of CF. Reliable patterns can provide valuable support for computer and intrusion forensics by facilitating the search for offenders and their behaviors.

We defined an offender profile as having two components: a factual profile and a behavioral profile. In our study, BP1 and BP2 were found to be the most important behavioral profiles using link analysis. Based on these profiles, refined association rules were computed defining 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 relating behavioral profiles to user templates. These association rules enabled immediate extension to a C5.0 predictive model, which facilitated measuring generalization reliability of the resulting rules.

The GRI algorithm discovers of all association rules yielding confidence values greater than a threshold supplied by the FC analyst. High support item sets (at or above
the threshold) are transformed into association rules. It is possible that in forensic work, useful information could be lost by setting the threshold at a too-high level. While beyond the scope of this study, work has been done to mitigate this problem by effecting concept hierarchies to reduce the number of uninteresting rules, while identifying any suspicious rules (Mohay et al., 2003). This may be a fruitful area for future research.

9. Concluding Remarks

CF is the field of forensic science that deals with digital crimes, or crimes that involve the use of computers. CF is becoming increasingly important as societies find themselves living in a digital world subject to criminal behavior. Because of these threats, CF is taking center stage in both technical and legal spheres, as well as finding application in areas of corporate policy, human rights, information systems, and accounting.

Key requirements for effective CF include assurance that relevant data are accurate and have not been modified or corrupted and subsequent analysis of that data to discover evidence that affirms or disaffirms allegations. Important 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. The application used in this paper takes the context of evidence discovery pertaining to intrusion forensics. Related applications are focused on evidence of malfeasance found in logs, files, emails, discarded hard drives, graphics, and other electronic media.
Mohay, et al. (2003) point to several promising areas of future CF research. At the top of their list is the use of data mining to find useful patterns in evidence, with a focus on the use of association rules and link analysis. Other promising areas include authorship and attribution of email, stegoforensic analysis, image mining, and cryptanalysis.

The study reported here has investigated CF data mining using association rules and link analysis. We have outlined the theoretical foundations of this approach and presented a data mining investigation using actual, disguised data. Although our approach suggests promise, valuable insights and refinements may be discovered by similar studies--or by extending this methodology to other CF applications, including accounting forensics and image analysis.
Figure 1.

Use of Association Rules
In Forensic Analysis
Table 1.

Association Rules for Behaviors

<table>
<thead>
<tr>
<th>Support</th>
<th>Confidence</th>
<th>Consequent</th>
<th>Antecedent 1</th>
<th>Antecedent 2</th>
<th>Antecedent 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3.00</td>
<td>B3</td>
<td>B2</td>
<td>B5</td>
<td>B6</td>
</tr>
<tr>
<td>2</td>
<td>3.10</td>
<td>B5</td>
<td>B2</td>
<td>B3</td>
<td>B6</td>
</tr>
<tr>
<td>3</td>
<td>4.00</td>
<td>B3</td>
<td>B4</td>
<td>B5</td>
<td>B6</td>
</tr>
<tr>
<td>4</td>
<td>4.50</td>
<td>B6</td>
<td>B1</td>
<td>B3</td>
<td>B5</td>
</tr>
<tr>
<td>5</td>
<td>3.30</td>
<td>B6</td>
<td>B2</td>
<td>B3</td>
<td>B5</td>
</tr>
<tr>
<td>6</td>
<td>16.70</td>
<td>B5</td>
<td>B3</td>
<td>B6</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>4.60</td>
<td>B5</td>
<td>B1</td>
<td>B3</td>
<td>B6</td>
</tr>
<tr>
<td>8</td>
<td>17.00</td>
<td>B3</td>
<td>B5</td>
<td>B6</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>4.70</td>
<td>B3</td>
<td>B1</td>
<td>B5</td>
<td>B6</td>
</tr>
<tr>
<td>10</td>
<td>17.30</td>
<td>B6</td>
<td>B3</td>
<td>B5</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>6.00</td>
<td>B3</td>
<td>B4</td>
<td>B6</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>4.70</td>
<td>B3</td>
<td>B2</td>
<td>B6</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>7.50</td>
<td>B3</td>
<td>B4</td>
<td>B5</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>29.30</td>
<td>B5</td>
<td>B6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>8.50</td>
<td>B1</td>
<td>B6</td>
<td>B8</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>30.20</td>
<td>B3</td>
<td>B5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>30.30</td>
<td>B5</td>
<td>B3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>30.20</td>
<td>B6</td>
<td>B5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>30.30</td>
<td>B6</td>
<td>B3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>8.90</td>
<td>B8</td>
<td>B1</td>
<td>B6</td>
<td></td>
</tr>
</tbody>
</table>
Figure 2.

Weighted Graph of Associations
<table>
<thead>
<tr>
<th>Support</th>
<th>Confidence</th>
<th>Antecedent 1</th>
<th>Antecedent 2</th>
<th>Antecedent 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.50</td>
<td>BP1</td>
<td>U2 = TTYPE2</td>
<td>U6 &gt; 16.500</td>
</tr>
<tr>
<td>2</td>
<td>0.30</td>
<td>BP2</td>
<td>U2 = TTYPE2</td>
<td>U4</td>
</tr>
<tr>
<td>3</td>
<td>1.40</td>
<td>BP2</td>
<td>U5 &lt; 10.750</td>
<td>U3</td>
</tr>
<tr>
<td>4</td>
<td>2.00</td>
<td>BP2</td>
<td>U2 = TTYPE2</td>
<td>U5 &lt; 12.550</td>
</tr>
<tr>
<td>5</td>
<td>3.30</td>
<td>BP2</td>
<td>U2 = TTYPE1</td>
<td>U3</td>
</tr>
<tr>
<td>6</td>
<td>6.30</td>
<td>BP2</td>
<td>U2 = TTYPE3</td>
<td>U3</td>
</tr>
<tr>
<td>7</td>
<td>16.50</td>
<td>BP2</td>
<td>U3</td>
<td>U5 &lt; 16.950</td>
</tr>
<tr>
<td>8</td>
<td>7.50</td>
<td>BP2</td>
<td>U3</td>
<td>U4</td>
</tr>
<tr>
<td>9</td>
<td>6.90</td>
<td>BP2</td>
<td>U2 = TTYPE2</td>
<td>U3</td>
</tr>
<tr>
<td>10</td>
<td>0.90</td>
<td>BP1</td>
<td>U2 = TTYPE3</td>
<td>U3</td>
</tr>
<tr>
<td>11</td>
<td>0.70</td>
<td>BP2</td>
<td>U4</td>
<td>U5 &lt; 10.750</td>
</tr>
<tr>
<td>12</td>
<td>0.70</td>
<td>BP2</td>
<td>U2 = TTYPE3</td>
<td>U5 &lt; 10.650</td>
</tr>
<tr>
<td>13</td>
<td>2.00</td>
<td>BP1</td>
<td>U3</td>
<td>U6 &gt; 16.500</td>
</tr>
<tr>
<td>14</td>
<td>1.00</td>
<td>BP1</td>
<td>U2 = TTYPE1</td>
<td>U6 &lt; 17.500</td>
</tr>
<tr>
<td>15</td>
<td>3.80</td>
<td>BP1</td>
<td>U6 &gt; 16.500</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>2.60</td>
<td>BP2</td>
<td>U5 &lt; 10.750</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>1.30</td>
<td>BP2</td>
<td>U4</td>
<td>U5 &lt; 10.750</td>
</tr>
<tr>
<td>18</td>
<td>4.20</td>
<td>BP1</td>
<td>U2 = TTYPE2</td>
<td>U3</td>
</tr>
<tr>
<td>19</td>
<td>5.20</td>
<td>BP1</td>
<td>U2 = TTYPE3</td>
<td>U6 &lt; 19.500</td>
</tr>
<tr>
<td>20</td>
<td>3.60</td>
<td>BP2</td>
<td>U2 = TTYPE2</td>
<td>U5 &lt; 12.550</td>
</tr>
</tbody>
</table>

**Table 2.**

Rules Associating BP1 and BP2 With User Profiles
Figure 3.

BP1 Boosted Decision Tree from C5.0--
Cross-Validation Accuracy 94 Percent
Figure 4.

BP2 Boosted Decision Tree from C5.0--Cross-Validation Accuracy: 97 Percent
References


de Vel, O., Anderson, A., Corney, M., & Mohay, G. (2001). Mining e-mail content for author identification forensics. ACM SIGMOD Record, 30(4), 55-64.


