View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by El Servicio de Difusión de la Creación Intelectual

# Strengthening Intrusion Detection Techniques through Emerging Patterns

Walter M. Grandinetti

wmg@cs.uns.edu.ar

Laboratorio de Investigación y Desarrollo en Inteligencia Artificial (LIDIA)
Departamento de Ciencias e Ingeniería de la Computación
Universidad Nacional del Sur

**Key Words:** System Security, Pattern Mining, Emerging Patterns, Jumping Emerging Patterns.

## 1 Introduction

In todays world where nearly every company is dependent on the Internet to survive, it is not surprising that the role of intrusion detection has become extremely important within the last decade. Intrusion detection involves determining whether some entity has attempted to gain, or worse, it has gained unauthorized access to the system. The task of current intrusion detection systems is detect possible threats not only from insiders but also from outsiders.

Based on our current knowledge, there are two things the system administrator could do in order to keep secure his system. First, use preventive measures. Second, make use of the audit logs. Due to the sheer volume of the logs, it is required that this task be performed automatically. Data Mining field of study has help to partially automatize this process. However, the current state of art has still left too much to the administrator and sometimes it distract the administrator raising false alarms.

This work propose to apply a new technique, successfully used in others fields of knowledge as Bioinformatics and Classification Systems, in order to define more accurately user's profiles and to detect more intruders, raising a lower number of false alarms and having a precision higher than other techniques.

## 2 Intrusion Detection

### 2.1 Taxonomy of Intruders

There exist a taxonomy which helps to identify intruders into three classes, [1]:

- Masquerader: An individual who is not authorized to use the computer and who perpetrates a system's access control to exploit a legitimate user's account.
- Misfeasor: A legitimate user who is authorized for access but misuses his or her privileges.
- Clandestine User: An individual who seizes supervisory control of the system and uses this control to evade auditing and access controls or to suppress audit collection.

Nowadays a hierarchy of hackers can be found, where two major levels can be distinguished, [4]. At the higher level there are experienced users which have a very good knowledge of the underlying technology. At the lower level we find programmers with a basic understanding of how hacking tools work (e.g. those who just apply supplied cracking programs). A teamwork involving these two levels turns out to be particularly dangerous.

### 2.2 Intrusion Countermeasures

There are two complementary methods to use in order to keep secure the system. First, the usage of a complete set of preventive countermeasures which can thwart every known kind of attack. Second, exploiting the logs in order to detect possible intrusions, a technique also known as intrusion detection.

#### 2.2.1 Preventive Countermeasures

The administrator should try to thwart intrusions firstly through preventive measures. Usually, an effective way to accomplish such a task is to make use of firewalls, in order to keep outsiders restraint. To keep the system secure from insiders the administrator should establish password-based and privileges-based politics.

However, these tools just prevent intrusions which are attempted through known channels. Moreover, they do not fully solve the problem of insiders behaving as masqueraders or misfeasors.

A new preventive measure is the usage of "honeypots"[1]. A honeypot is a computer on the network with the sole purpose of looking and acting like a legitimate computer but actually configured to interact with potential hackers in such a way to capture details of their attacks. Honeypots are known also as a sacrificial lamb, decoy, or booby trap. The more realistic the interaction is, the longer the attacker will stay occupied on honeypot systems and away from the real system. The longer the hacker stays using the honeypot, the more he will disclosed about his techniques. This information can be used to identify what he is after, what his skill level is, and what tools he do use. All this information is then used to prepare better your network and host defenses.

The major drawbacks with preventive measures is that they require that the administrator to be aware of current potential threats and vulnerabilities within his computing system.

#### 2.2.2 Intrusion Detection Systems (IDS)

An important study, [1], postulate that one could distinguish between a masquerader and a legitimate user with reasonable confidence. Patterns of legitimate user behavior can be established by observing past history, and significant deviation from such patterns can be detected. This kind of detection is known as *anomaly detection*.

However, it should be noted that an anomalous sequence could be the result of three phenomena (according to [5]). First, noise (typing errors), second, concept drift (working on a different project, etc.), third, masquerader.

It is highly desirable to detect the differences between intruder's behavior and authorized user's behavior. The first thing to do is to build an accurate profile. Second, provide the IDS with the mechanism to classify either one.

In order to build an accurate user's profile a set of measures should be thoughtfully selected (log-in time, log-in location, favorite editor, sequence of actions, resources used, etc.).

Because of the sheer volume of audit data, both in the amount of audit records and in the number of system features, it is not logical to think that the manager could create and update each and every profile. Hence efficient, intelligent and automated data analysis tools are required to discover the user's behavior within the system.

[1] http://www.sans.org/resources/idfaq/honeypot2.php

Data mining techniques fit perfectly within this outline helping expert humans to develop a concise profile and at the same time showing previously unknown behavior which is usually scattered within the data that it is unlikely seen at a simple glance at the logs.

The major drawback of the dynamic anomaly detection approach is the need to accurately define the boundary between acceptable and anomalous behavior. A misplaced boundary would either alert the system managers with a false alarm or allow the entrance to an intruder and worse modify the profile creating a back door hard to be detected.

It can be seen (figure 1) that there is an overlapping area in the intruder's behavior and the profile of authorized user [1]. However is should be noted that overlapped area in the misfeasor and the authorized user's behavior is more stressed just because it is the same person.

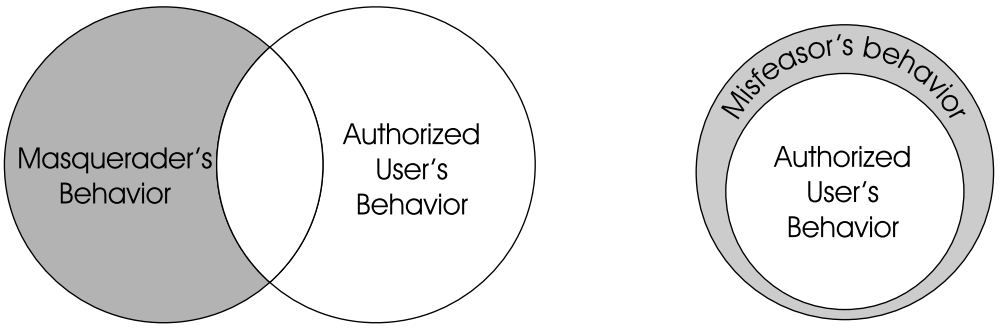


Figure 1: Overlapping of Intruders' or Misfeasors' profile and Authentic users' profile

Anderson suggests that the task of detecting a misfeasor (legitimate user performing in an unauthorized fashion) is more difficult than detecting a masquerader, in that the distinction between abnormal and normal behavior may be small. Anderson concluded that such violations would be undetectable just through the search for anomalous behavior. However, misfeasor behavior might nevertheless be detectable by intelligent definition of the class of conditions that suggest unauthorized usage. Finally, the detection of the clandestine user was felt to be beyond the scope of purely automated techniques.

It was also observed in [5] that an intruder (either masquerade or misfeasor) may perform such "anomalous" sequences repeatedly, so that their frequency becomes high enough to be detected.

As it will be intended to show later, I do not subscribe to either suggestions. I believe that misfeasors are detectable through a peruse study of the logs (however I do subscribe that this task is more difficult than detecting masqueraders) and the activities of an intruder may never become frequent enough to be detected or at the time that this happen may be to late.

## 3 Emerging Pattern Mining

The *Emerging Pattern Mining* has been proposed by Dong and Li, [2], in order to capture significant changes and differences between datasets. The outcome of this task is a set of patterns, called *emerging patterns* (EP).

Emerging Patterns are defined as itemsets whose supports increase *significantly* from one dataset to another. Namely, it look for those patterns whose frequency in one dataset, denoted $\mathcal{D}_1$, differ *significantly* from the other, denoted $\mathcal{D}_2$. A special kind of emerging pattern, called *jumping emerging pattern* (JEP), was proposed in [3]. These patterns express peculiar features occurring in one dataset but not in the other.

EPs have shown to be useful capturing trends in timestamped databases, building powerful classifiers and detecting contrasts among data classes. These application resulted in more accurate classifiers and provided a new insight of the existing data. Much of the experiments carried out using this mining technique were conducted on datasets of bioinformatics field related to gene sequences and cancer information.

From those experiment lets see the following example of an emerging pattern: "*Lung-cancer incidence rate among smokers is* 14 *times that of non-smokers*". The datasets used consist of records from smokers and non-smokers.

The set of JEPs are interesting to our proposal because of they satisfy the property of convexity, which means that JEP spaces can be bounded and concisely represented by their boundary elements. This convexity property was exploited to develop efficient maintenance algorithms to modify its boundary elements in response to changes to the data.

**Definition EP and JEP (Dong & Li)**: A set $X$ of items is called an *itemset*. A transaction $T$ contains an itemset $X$, if $X \subseteq T$. The *support* of $X$ in a dataset $\mathcal{D}$, denoted $supp_{\mathcal{D}}(X)$, is $\frac{count_{\mathcal{D}}(X)}{|\mathcal{D}|}$, where $count_{\mathcal{D}}(X)$ is the number (called *count*) of transaccions in $\mathcal{D}$ containing $X$. Given a number $\sigma > 0$, an itemset $X$ is $\sigma - frequent$ in $\mathcal{D}$ if $supp_{\mathcal{D}}(X) \geq \sigma$.

For sake of simplicity let $supp_i(X)$ denote $supp_{\mathcal{D}_i}(X)$. The *growth rate* of an itemset $X$ from a dataset $D_1$ to $D_2$ is defined as

$$GrowthRate(X) = \begin{cases} 0 & \text{if } supp_1(X) = 0 \text{ and } supp_2(X) = 0 \\ \infty & \text{if } supp_1(X) = 0 \text{ and } supp_2(X) \neq 0 \\ \frac{supp_2(X)}{supp_1(X)} & \text{otherwise} \end{cases}$$

Given a growth rate threshold $\rho > 1$, an itemset $X$ is called a $\rho-$*emerging pattern* ($\rho-$EP or simply EP) from $\mathcal{D}_1$ to $\mathcal{D}_2$ if $GrowthRate(X) \geq \rho$. EPs whose growth rate is $\infty$ are called *jumping emerging patterns* (JEPs).

## 4 Tailoring EPs to the IDS process

In order to classify the current behavior of an individual as legitimate, masquerader or misfeasor, most classifiers require being trained with data previously classified either such as legitimate $\mathcal{D}_1$ or such as intruder $\mathcal{D}_2$. The gist is that the amount of $\mathcal{D}_2$ within an IDS could be very poor or non-existent at all. Hence, the IDS should based its classification process just on the normal behavior. However, there can be used two datasets, one dataset consisting of the user's profile and the other of the recent activity.

Lets refer now to the motivation behind the usage EPs. First, classification rules obtain by traditional methods are just a subset of EPs. Second, EPs are representable concisely even when the EP space is huge. Third, statistical studies discover patterns of few variables in contrast to EPs which are fitted to manage long EPs. In [2] it was shown that it is very likely the occurrence of long EPs with much higher support than shorter EPs.

These results suggest that this technique could be of great help in any task that involve a classification problem since its accuracy its higher, it provides new insights and it is concisely represented. Therefore, in order to detect an intruder the proposal is to look for EPs using the information on the user's profile and current activities or logs. These EPs could be ranked and showed to the system administrator. From the benefits of EPs already discussed the administrator work should be assuaged.

In a previous section, it was mentioned that the usage of honeypots and more recently honeynets have helped to detect more accurately intruders. EPs could assist to the preventive measures detecting signatures and thus obtaining the patterns that belong exclusively to intruder's behavior.

Misfeasors could be detected in the same way of intruders. However, the information recollected by the miner should be canvassed to determine if it is the case of a concept drift or a potential non-authorized activity.

In addition to the already mentioned EP assets, the JEPs have the advantage that a set of Incremental Maintenance Algorithms have been developed, [3]. They take advantage of nearly repeated computations on inputs that differ slightly from one another, computing new JEP spaces incrementally.

Taking advantage of this incremental nature of the algorithms the user's profile could be updated according to more recent behavior more efficiently having a great impact on the system performance.

One of the best features of EPs is its capability to capture temporal trends. As an example of a recent discovered trend appeared in a newspaper article, [2] "*Low tuition, high standards lure U.S. students to Canada*" Dayton Daily News, 10/6/2002), concerning the emerging trends of American students studying in Canadian Universities: the enrollments of American students in Canada has been rising by about 85% in three years to a total of about 5000. This trend is an EP with low support but a large growth rate. It should be noted that these patterns with low support are not seen by miners only based on the itemset support.

This ability of EPs could be used to detect misfeasors. Keeping a record of previously discovered EPs, the administrator could notice improper behavior. For instance, lets suppose that an authorized user access to a database which either he usually do not access or he do not need to access to do his job, but he has privileges over it. This database is worthless to him but it could be valuable to an outsider, for instance a competitor. Knowing that the sysadmin analyze the user's logs, he could access interspersedly to the database. Though, the sysadmin see those EPs, he could disregard this information thinking that this spurious access could not be a flaw or an indicator. This person could perform his ants-work taking along with him all the database (having enough time). Using the EPs to detect trends, we could heed the sysadmin as early as possible of a potencial misfeasor.

## References


[1] J. Anderson, *Computer security threat monitoring and surveillance*, J. Anderson (80).

[2] Guozhu Dong and Jinyan Li, *Efficient mining of emerging patterns: Discovering trends and differences*, Knowledge Discovery and Data Mining, 1999, pp. 43–52.

[3] Jinyan Li and Kotagiri Ramamohanarao, *The space of jumping emerging patterns and its incremental maintenance algorithms*, Proc. 17th International Conf. on Machine Learning, Morgan Kaufmann, San Francisco, CA, 2000, pp. 551–558.

[4] William Stallings, *Cryptography and network security: Principles and practice*, 2 ed., Prentice Hall, 92.

[5] Mohammed J. Zaki and Karlton Sequeira, *Admit: Anomaly-base data mining for intrusions*, 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2002).