Detecting Unexplained Sequences using Top-k Algorithm

Jintu Ann John[1], Neethu Maria John [2]

***Abstract*— There are numerous techniques available to find occurrences of activities in time-stamped observation data with each occurrences having an associated probability.For the entire research community the activity recognition is considered to be a big challenge. The existing techniques cannot deal with "zero day" attacks that have never seen before .In the proposed system it is able to find the subsequenes of the observation data,called unexplained sequence,that known models are not able to "explain" with a certain confidence.Thus we have to consider a known set of activities which contains both innocuous and dangerous activities .We want to monitor and also we want to identify the unexplained subsequences in an observation sequence that are poorly explained that is,we want to identify the activities that are not present in this predefined set.Top-k algorithms are used to identify the top-k totally and partially unexplained activities.In the proposed system these algorithms are applied on the Cyber Security data sets. These algorithms are more efficient and provides faster search for identifying totally and partially unexplained sequences.**

***Index Terms*— Cyber security data set, Partially unexplained sequence, Time-stamped observation data,Top-k algorithm, Totally unexplained sequence**

I. INTRODUCTION

Data Mining is the process of extracting large amount of datas from large databases. It is the process of semiautomatically analyzing large databases to find patterns that are valid, novel, useful and understandable. One of the main task of data mining includes outlier detection. Discovering unexplained sequences is a part of outlier detection. Outlier detection means detecting datas that are not part of a particular group. Similar datas are grouped together to form clusters .The datas that does not belongs to this cluster is considered to be an outlier. Similarly in this paper, there will a predefined set of activities the activities that does not belongs to this predefined set will considered as the unexpected activities. Discovering unexplained sequences plays an important role in many applications such as fraud detection, cyber security, video surveillance etc.

In cyber security, intrusion detection can monitor network traffic for suspicious behaviour and trigger security alerts.Alert correlation methods aggregate alerts into multistep attack scenarios.

Manuscript received March, 2015.
***Jintu Ann John**, CSE, M G University Kottayam , India , 9747825486*
***Neethu Maria John,** CSE, M G University, Kottayam,India, 9947472025.*

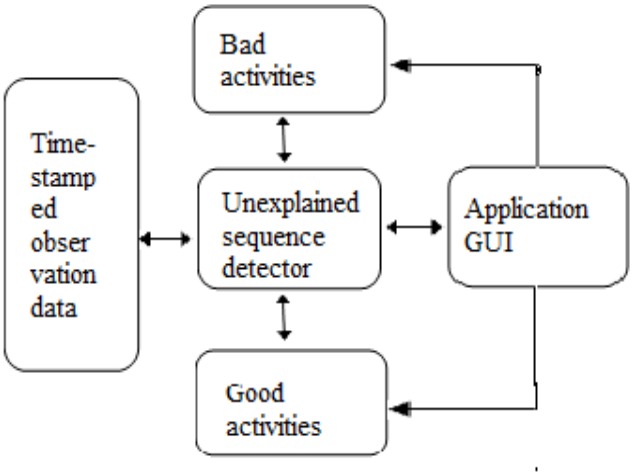

Fig.1. Overall working of unexplained sequence detector

Unexplained sequence detector is designed and implemented to discover the subsequences of the observation stream that are not sufficietly explained by the well known activity models.

Fig.1 shows how the unexplained sequence detector works.We consider a known activity model which contains both good and bad activities.Good activities are activities that are considered to be appropriate and bad acivities are activities which are inappropriate.

Unexplained sequences allow an application to identify activities that are never seen or imagined before by experts, and to add them to an increasing body of such knowledge. For instance, a zero-day attack on a computer system, may involve sequences of actions not seen before and hence not captured by past activity models.These types of observations have to be identified.

This method can be used in many applications, here we apply this method to cyber security datasets.Cyber security is to protect networks, computers, programs and data from various attacks, damage or unauthorized access.

In this paper we are considering a network ,it will contains already a predefined set of sequence of activities.If an attack occurs that is it is an activity which is not related to this network.So it will be considered as an unexpected activity and it should be identified.

In the existing approaches they rely on models encoding a priori knowledge of either normal or malicious behavior. They cannot deal with events such as "zero day" attacks that have never been seen before.This problem should be overcome by proposed approach.We introduces Top-k algorithms to find top-k totally and partially unexplained sequences.Also develops a prototype implementation on

experiments using a cyber security dataset showing that the algorithm works effectively, both from an efficiency perspective and an accuracy perspective.

The remainder of this paper is organised as follows:Related work is derscribed in the section II.Proposed system and the algorithm used is described in the section III.Section IV describes Experimental results and section V describes Conclusion.

II. RELATED WORK

Intrusion detection and alert correlation techniques provide valuable and complementary tools for identifying and monitoring security threats in complex network infrastructures.Intrusion detection systems (IDS) can monitor network traffic for suspicious behaviour and trigger security alerts accordingly. Intrusion detection is based on the assumption that an intruder's behaviour will be noticeably different from that of a legitimate user and that many unauthorized actions are detectable [5].

Intrusion detection[9]: Intrusion detection is the process of monitoring the events occurring in a computer system or network and analyzing them for signs of possible incidents, which are violations or imminent threats of violation of computer security policies, acceptable use policies, or standard security practices. If a security system fails to defend against an attack, it should be well aware of being attacked and have a mechanism to perform countermeasures in order to prevent further attacks and reduce the damage and loss resulting from the attack. That is the main aim of Intrusion Detection Systems (IDS). Intrusion prevention is the process of performing intrusion detection and attempting to stop detected incidents happening again. Intrusion detection and prevention systems (IDPS) are primarily focused on identifying possible incidents, logging information about them, attempting to stop them, and reporting them to security administrators. In addition, organizations use IDPSs for other purposes, such as documenting existing threats, identifying problems with security policies, and deterring individuals from violating them. IDPSs have become a necessary addition to the security infrastructure of nearly every organization that relies on information technology.

Intrusion detection systems[11] analyze information about the activity performed in a computer system or network, looking for evidence of malicious behaviour", that is to say an IDS is a system that detects unauthorized access or potential attacks on informatic systems through informations source available on the system (log) or on the network (network traffic). In addition, IDSs are systems used mainly for monitoring network traffic through a set of rules and flexible algorithms in order to detect attacks on the autonomous system (AS) involved.

Two types of intrusion detection systems [4]:

Signature based: An IDS is called signature based [3] if it uses a knowledge base in order to detect an attack. An IDS analyzes the collected data and compares them to a set of attack signatures to discover any anomaly actions. So, an anomaly action is discovered if it exists by a correspondence between its signature and that of a known attack, stored in a knowledge base. A signature is a pattern that corresponds to a known threat while signature-based detection is the process of comparing signatures against observed events to identify possible incidents.

Anomaly based: An anomaly based IDS [3] tries to build correct models of resources normal activity and stores them in a knowledge base. This IDS detects anomaly activities through a comparison of the stored actions with another knowledge base of known threats; quantitatively, an action is considered important if some features of actions exceed the appropriate thresholds.

Thus, Anomaly-based detection is the process of comparing definitions of what activity is considered normal against observed events to identify significant deviations. This IDS uses profiles that represent the normal behaviour of such things as users, hosts, network connections, or applications. The profiles are developed by monitoring the characteristics of typical activity over a period of time. An initial profile is generated over a period of time sometimes called a training period. Profiles for anomaly-based detection can either be static or dynamic. Once generated, a static profile is unchanged unless the IDS is specifically directed to generate a new profile. A dynamic profile is adjusted constantly as additional events are observed.

In this paper, the case where we have a set of known activities which contains both innocuous and dangerous activities and we are looking for observation sequences that cannot be explained by either (if they were, they would constitute patterns that were known a priori). These need to be flagged as they might represent "zero day" attacks that is the attacks that were never seen before and vary significantly from past known access patterns.

Correlation techniques[9,10]:It is defined as a process that contains multiple components with the purpose of analyzing alerts and providing high-level insight on the security state of the network under surveillance. Alert correlation [6] can be very beneficial especially for intrusion response. Firstly, it reduces the volume of alerts that needs to be handled. IDSs may generate thousands of alerts per day. Secondly, due to the problem of false positives, it is impossible to respond to every alert that is reported by IDS. Only those which are detected with high confidence will be considered for response action. Alert correlation provides a way to increase the detection confidence. Thirdly, correlated alerts provide a succinct, high-level view of the security state of the network under surveillance. Another important use of alert correlation is to recognize the strategies or plans of different intrusions and infer the goal of attacks. In summary, the goal of correlation is to find causal relationships between alerts in order to reconstruct attacks from isolated alerts. This goal is achieved by providing a higher level view of the actual attacks [7, 8].

III. PROPOSED SYSTEM

By using Cyber Security application it is easy to identify actions in an observation sequence .For this purpose, we have designed and developed a specific prototype implementation for cyber security domain.

The prototype consists of:

- A network sniffer
- A network Intrusion Detection System
- An Alert Aggregation module
- The UAP engine

The Sniffer captures network traffic and generates the sequence of packets, the Intrusion Detection System analyzes such traffic and generates the sequence of alerts. Then, as the number of alerts returned by the IDS may be relatively high,

the Alert Aggregation module, that takes as input the identified alerts, will aggregate multiple alerts triggered by the same event into a macro-alert, based on a set of aggregation rules. Finally, UAP Engine takes as inputs the occurrences detected during the previous step and the whole captured traffic as well and discovers the unexplained cyber activities.

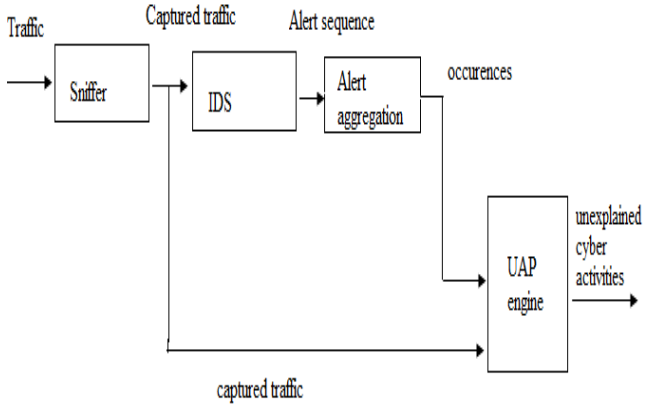

Fig. 2 The prototype architecture for cyber security context

Sniffer

The Sniffer chosen has been Wireshark. It is used to capture network traffic and generate packet sequences.

Intrusion Detection System

We have chosen Snort as Intrusion Detection System. Snort is an open source network intrusion prevention and detection system.

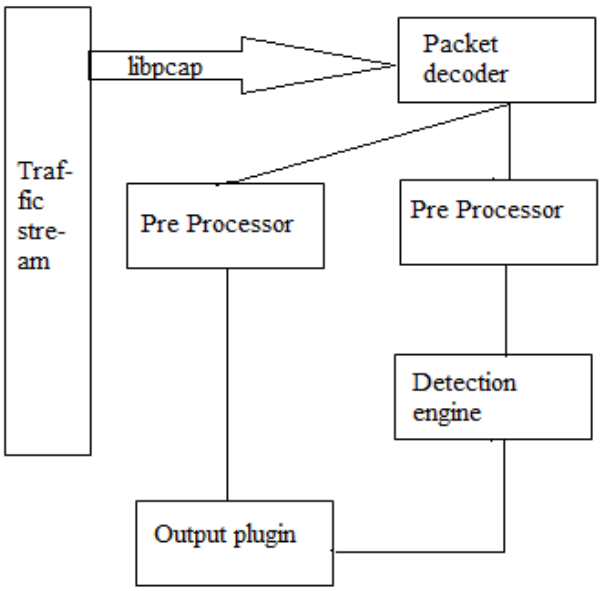

Fig:3.Intrusion Detection System

Snort is divided into 5 components:

- **Packet capturing mechanism:** to get packets into the preprocessors and then the main detection engine.
- **Packet Decoder:** as soon as packets have been gathered, Snort must decode the specific protocol elements for each packet.
- **Detection engine**: builds attack signatures by parsing Snort rules.
- **Output plugins**: It has to get intrusion data to users.

Alert aggregation

The main purpose of using alert aggregation [12] is to reduce the redundancy of alerts by grouping duplicate alerts and merging them into a single one.

UAP Engine

The UAP Engine takes as inputs the macro-alert list detected
by the Alert Aggregation Module and the whole captured traffic as well and discovers the unexplained cyber activities.

In this paper a business application is considered.So network traffic from a business network is taken.Wireshark is to capture network traffic and generate packet sequences and Snort as the set of activity models.

Top-k algorithm

1.Setting a threshold value for determining unexplained sequence
2.For each unexplained sequence
  2.1.check probability of each sequence > threshold
  2.2.then exit with no Top-k unexplained sequence
3.else
  3.1.perform binary search on probabilities of sequence using threshold value
4.return Top-k unexplained sequence

Unexplained sequences allow an application to identify activities never seen or imagined before by experts, and to add them to an increasing body of such knowledge. Based on this, users can specify a probability threshold and look for all sequences that are totally (or partially) unexplained with a probability exceeding the threshold.

Top-k algorithms[1] will returns the k results with the highest scores.The advantage of this algorithm are that it is simple to implement, requires no preprocessing, maintains a bounded buffer of size k, and is usually fast unless the repository is large.

IV. EXPERIMENTATION AND RESULTS

By applying Top-k algorithm the unexplained sequences are identified. Based on the unexplained sequences identified versus the k value graph is plotted. In the case of totally unexplained sequence the graph is represented as a straight line.Because the probability for all unexplained sequences is equal.

In the case of partially unexplained sequence the entire file is parsed and identify if any of the content is malicious.The id of that particular files are identified.The source files are identified as malicious files.The source files includes .exe, .java, .c ,etc files.If these types of files reaches it will detected as malicious. These types of files are considered as the totally unexplained sequences.

If any other type of files came the entire file is parsed and if any content is not normal then it will be considered as partially unexplained else it will be considered as a normal file.

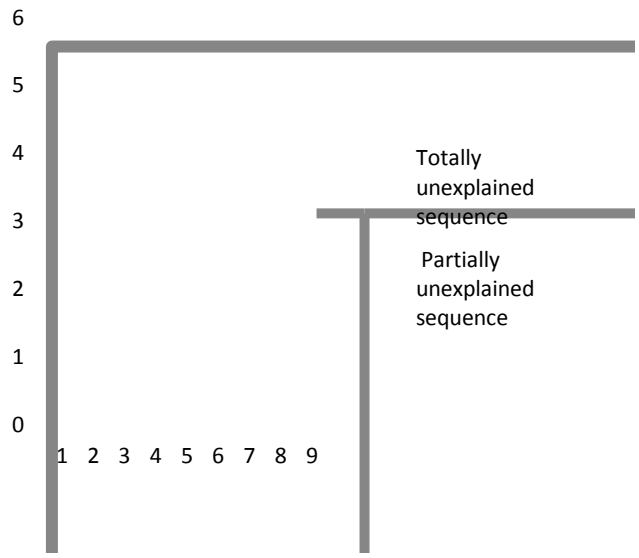

Fig: (b)
Fig. 3 Performance Evaluations

V. CONCLUSION

We have a sequence of time-stamped observation data and a set of "known" activities (normal or suspicious). This paper addresses the problem of finding subsequences of the sequence that are not "sufficiently" explained by the activities in the predefined set. We formally define what it means for a sequence to be unexplained by defining totally and partially unexplained sequences. We propose a protocol and identify interesting properties that can be leveraged to make the search for unexplained activities highly efficient. We developed the Top-K algorithms to find totally and partially unexplained activities with highest probabilities. Cyber security datasets are used as the application to find out the unexplained sequences.

ACKNOWLEDGMENT

The first author would like to thanks all those people, who guided and supported. Without their valuable guidance and support, this task was not possible and also likes to thank colleagues for their discussions and suggestions.

REFERENCES

[1] Massimiliano Albanese,Cristian Molinaro,Antonio Picariello,"Discovering the Top-k Unexplaines Sequences in Time-Stamped Observation Data",*IEEE Transactions on Knowledge and Data Engineering* ,vol.26,No.3,March 2014.

[2] P. Ning, Y. Cui, and D. S. Reeves, "Constructing attack scenarios through correlation of intrusion alerts," in CCS 2002, (Washington, DC, USA), pp. 245–254,ACM, November 2002.

[3] P.Garc´ıa-Teodoro, J. D´ıaz-Verdejo, G. aci´a-Fern´andez,and E. V´azquez,"Anomaly-based network intrusion detection: Techniques, systems and challenges,"*Computers & Security, vol. 28, pp. 18–28, February-March 2009*

[4] B. Mukherjee, L. T. Heberlein, and K. N. Levitt, "Network intrusion detection,"*IEEE Network, vol. 8, pp. 26–41, May 1994*

[5] S. Hongeng and R. Nevatia,"Multi-agent event recognition," in ICCV, pp. 84–93, 2001.

[6] S. Noel, E. Robertson, and S. Jajodia,"Correlating intrusion events and building attack scenarios through attack graph distances," in ACSAC, (Tucson, AZ, USA), pp. 350–359, December 2004.

[7] H. Debar and A.Wespi, "Aggregation and correlation of intrusion-detection alerts," in RAID (W. Lee, L. M´e, and A. Wespi, eds.), vol. 2212 of Lecture Notes in Computer Science, (Davis, CA, USA), pp. 85–103, *Springer,October 2001*.

[8] S. O. Al-Mamory and H. Zhang, "Ids alerts correlation using grammar-based approach,"*Journal of Computer Virology, vol. 5, pp. 271–282, November 2009*.

[9] X. Qin andW. Lee, "Statistical causality analysis of INFOSEC alert data," in RAID (G. Vigna, C. Kruegel, and E. Jonsson, eds.), vol. 2820 of Lecture Notes in Computer Science, (Pittsburgh, PA, USA), pp. 73–93, *Springer, September 2003.*

[10] X. Qin, "A Probabilistic-Based Framework for INFOSEC Alert Correlation". Phd thesis, Georgia Institute of Technology, August 2005.

[11] J. P. Anderson, "Computer security threat monitoring and surveillance," tech. rep.,James P. Anderson Co., Fort Washington, PA, USA, April 1980.

[12] A. Jones and S. Li, "Temporal signatures for intrusion detection," in ACSAC, (New Orleans, LA, USA), pp. 252–261, *IEEE Computer Society, December* 2001.

[13] A. J. Oliner, A. V. Kulkarni, and A. Aiken, "Community epidemic detection using time-correlated anomalies," in RAID (S. Jha, R. Sommer, and C. Kreibich, eds.), vol. 6307 of Lecture Notes in Computer Science, (Ottawa, Canada), pp. 360–381, *Springer, September 2010*

[14] A. Jones and S. Li, "Temporal signatures for intrusion detection," in ACSAC, (New Orleans, LA, USA), pp. 252–261, *IEEE Computer Society, December 2001.*

Jintu Ann John, Department of Computer Science and Engineering, Mangalam college of Engineering, Ettumanoor, Kerala, India.

.

Neethu Maria John, Associate Professor, Department of Computer Science and Engineering, Mangalam college of Engineering, Ettumanoor, Kerala, India.

