Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online): 0493-2137
E-Publication: Online Open Access
Vol:54 Issue:10:2021
DOI: 10.17605/OSF.IO/23B98

GRAPH THEORY BASED PROBE FOR FORENSIC NETWORKS AND ANOMALY FERRET OUT

YEGNANARAYANAN VENKATARAMAN[1] and GEORGE BARNABAS J[2]

[1&2] Department of Mathematics and Department of Information Technology Kalasalingam Academy of Research and Education, Deemed to be University, Anand Nagar, Krishnankoil-626126. Tamilnadu. India.
E-mail: prof.yegna@gmail.com

Abstract:
Forensic network analyzes intrusion evidence obtained to find out suspicious members and initiate step by step actions in an attack scenario. The evidence graph model serve as collected evidence. Depending on it one can form a framework that is based on hierarchical reasoning. Fuzzy inference comes in handy to comprehend host's functional states from local observations. Graph structure analysis can be done through global reasoning to determine the potential attackers. We evaluate various techniques through obtrusion ferreting out datasets and trial and error results and establish that evidence graph model is compelling to detect multi-stage attacks. Then, for fraud ferret out problems, the data evolves continually from the system under consideration. Moreover, the underlying concept changes from time to time dynamically and is understood as concept drift. Mostly the frauds are rarely observed compared to the normal behavior of the system. It is very difficult or expensive to simulate fraudulent behavior from the system. Data mining warrants robust, dependable anomaly ferreting out systems. It is a fact that research so far happened has not focused much on graph-based data. Suppose that a real graph with weighted edges is known in advance and we are interested to find a method to classify vertices as strange? Answering this is quite important for applications such as: obtrusion ferreting out mechanisms while facing the fraud happening in credit/debit/calling cards and many others. We probe further on this here.

Keywords: Forensic network, Forensic evidence, Graph based data, Anomaly Ferret out

Acronyms
FN = Forensic Network, FE = Forensic Evidence, OE = Obtrusion Evidence, OFS = Obtrusion ferret out Systems, AFO = Anomaly ferret out

1. Introduction

Elimination of cyber-attack threats requires prevention and ferret out mechanisms but this alone is not sufficient. We also need post-incident probe mechanisms to catch attackers found for their malevolent behaviors. FN's realm is built around this. FN strives to locate dubious units in the attack scene and rebuild the attacker's action sequence by comparing with evidence of intrusion collected from environments that are networked. FN possesses a huge problem space that comprises documentation, conservation, analysis and extending. FN analyzers are dumped with low quality evidence in plenty. Evidence extracted from OFS alerts are loaded with a lot of noise that emanates from false positives and irrelevant attacks. From a forensic expert's view point, it is pertinent to wisely correlate each piece of evidence to spell an engrossed view at macro level of what happened in the attack. It is irredundant to develop a viable strategy that is suitable to current sensor alarms for security with least dependence on existing knowledge models. Present practices in FN analysis hovers around manual methods that are error prone, non-scalable and time-consuming. The need of the hour is effective and automated techniques that are extensible. Graphs are Ubiquitous**:** computer networks, social networks, www to list a few. Suppose that a real graph is known in advance, with edges allotted weights, how to ferret out strange vertices? Finding answer for this is pertinent for applications such as: debit/credit/calling card frauds and many others. Anomaly obtrusion is vital for figuring rare events and data purification. It is connected with unearthing its pattern and underlying properties. Care should be taken to ensure that majority of vertices that exhibits power law pattern. This is so because, only then we can consider as outliers those vertices which tend to behave erratically.

In the last ten years, data mining has evolved as a hot topic of research, probing gripping research matters and summons real-life applications. To start with the data formats which are objective were limited to relational tables. Executions where every instance is denoted by a row in a table. But, the probes carried out in the last ten years or so attempt to transform the data to a semi-structured ones like XML or HTML texts, connections, ordered trees, symbolic sequences are denoted by well-developed logics. Graph mining has effective appeal with the aforementioned data mining that is multi-relational. But, the goal of graph mining is to develop new rules and step by step strategies to exploit substructures that are topological and imbibed in graph data, however the prioritized goal of data mining that is multi-relational is to develop rules to exploit the patterns that are relational, denoted by logical languages that are expressive. The first said is more inclined to geometry and the one discussed second is bent towards logic and relation based. For more see [9] and the references therein.

AD process usually happens in two stages namely training and ferret out. The former constructs suitable silhouette to replicate the deportment of the network being scrutinized while the latter alerts when deviation occurs beyond a fixed threshold. AD is able to recognize attacks as it is not dependent on pre-existing knowhow of particular attacks. But it is tough to identify normal silhouettes as the similarity of benevolent activities have huge deviations and anomalous deportment is witnessed in lot of cases. Also attackers endeavor to up skill the AD process to deem intrusive behavior as normal. So AD process are prone to both false negatives and false positives.

2. Main Problem Statement

a) How to productively scrutinize forensic evidence (FE) from assorted provenances to spot outfits and haps apt to varied level strafe schema in a methodical and modular outlook ?

The above main task can be broken into various subtasks such as how to a) productively tackle a plenty of IE and colander out unwanted noise for a studious scrutiny? b) do after-the-incident obtrusion probe in a modular and computationally effective set up? c) Rebuilt furtive varied level attack schema with minimum dependence on adept mastery? d) Coalesce data from assorted provenances to spot implicit baleful pursuits of attackers?

b) There is a growing concern for dependable AFO process in Data mining. Although investigations done are huge not much is done on graph-based data. Suppose we are provided with a real graph whose edges carry weights, the challenge lies in declaring vertices as strange. Given a massive network, what characteristics we should use to identify a neighborhood? How a neighborhood that is normal will appear in such a massive network? Answering this is pertinent for applications such as: debit/credit/calling card frauds and many others.

3. Problem Background

In 2001, a Research Workshop organized by digital forensic group declared Forensics network as “The use of scientifically proved techniques to collect, fuse, identify, examine, correlate, analyze, and document digital evidence from multiple, actively processing and transmitting digital sources for the purpose of uncovering facts related to the planned intent, or measured success of unauthorized activities meant to disrupt, corrupt, and or compromise system components as well as providing information to assist in response to or recovery from these activities [10].” Investigation tools in forensics akin to EnCase [11] and Safeback [12] concentrate on encapsulate and scrutiny of evidence from storage media on a particular host. Other software tools like tcp trace [13], tcpflow [14], flowtools [15] and netcat [16] backs capture of network traffic and analysis of sessions, Tools that are commercial such as eTrust network forensics tool [17] and NetDetector [18] collects plain data and probe deviations inside cooperative networks. However the method is hands on. ForNet[19] converts raw data into a brief write-up that can be preserved for years to carry out the forensic scrutiny.

The authors in [1-9] discussed some techniques adopted for graph-based AFO by exploiting the system named Subdue. Ferret out of anomalous substructure is the first technique that searches for sub edifices lying in a graph. Then the next technique endeavors anomalous subgraph ferret out is concerned with partitioning the graph into subgraphs with non-intersecting sets of vertices, and they are subjected to testing one another strange patterns. Further they also introduce some measures for regularity of graphs through borrow concept from theory of information. Substructure entropy measure elucidates bit requirement for fixed size substructure that is arbitrary. Conditional substructure entropy measure elucidates bit requirement that narrates a substructure's surroundings. They reported real life results that are deducted as network obtrusion dataset during the proceedings of KDD Cup in1999. It also includes data produced artificially. Some major types of anomalous vertices one can sport are: a) Near cliques and stars b) Heavy Vicinities c) Dominant heavy links. Then Outlier ferret out methods lead to two types called non-parametric and parametric. Parametric types deem the presence of distribution of the observations that yields best date fit. Non parametric types admits distance/density based data mining techniques. Feature bagging technique can sort out issues for high dimensionality, where features are split into multiple sets of different sizes randomly and outlier ferret out step by step procedures are carried out on every non intersecting set and then combine the scores. At last clustering step by step techniques reveal outliers as a by-product.

4. Objectives

a) To create evidence graph model with vertices pointing to network identities and links at host level and forensic evidence extracted from multiple sources for both abstraction and scalability.

b) To create effective step by step procedures for the spectral clustering and Page rank strategies to support analysis on large networks.

c) To determine the pattern and laws to be obeyed by the huge graphs modeling big data

d) To determine the 'features' to be extracted from the vertices of the above said graph

e) Some major types of anomalous vertices one can sport are: a) Near cliques and stars b) Heavy Vicinities c) Dominant heavy links

5. Related work

Widely explored techniques outside the context of security are Pagerank algorithm and Graph clustering ([20, 21]) that are multi-stage attack based. The authors in [22] adopted the recursive clustering to reduce the dependence for parameter tuning and pre awareness of the clusters. The Pagerank algorithm [23] is meant for ranking web pages. Mehta et al. [24] suggested to apply it for grading various security states in the case of attack graphs. Similarly one can adopt the Pagerank algorithm personalized [25] on graphs that are evidence based for focused probes and to identify secrete-attacker.

Outlier ferret out techniques are conceived as parametric and non-parametric classes. Parametric choice deem the presence of a distribution of the observations that are underlying and standard to fit the data. Non-parametric choices are both distance and density based techniques of data mining. Feature bagging method handles dimensions of higher value, where attributes are classified as groups randomly and becomes multiple sets of various sizes and outlier ferret out step by step procedures are done on each distinct set to level the scores. Then clustering algorithms reveal outliers as a by-product. For more see [26-35].

6. Methodology

• To create evidence collection module to gather digital evidence from different category of sensors put in place to monitor networks and hosts under probe.

• To create evidence pre-processing module to transform gathered evidence into standard format and eliminate repetitions in raw evidence through proper grouping and edit operations.

• Make use of Attack/Assets knowledge base for graph construction and attack reasoning.

• To do reasoning that is hierarchical depending on the graph that is evidence based to pinpoint intruders in more than one attack attempts and employ inference that are fuzzy based, methods, that are clustering/Pagerank based, evaluation of clusters etc, for reconstruction scenario.

To adopt ferret out of anomalies one can look for following attributes.

a) Cardinality of the ego neighbor set?
b) Cardinality of the egonet edge set? (by “egonet” we mean the 1-step neighbor of a vertex)
c) Cumulative weight of egonet?
d) Egonet’s weighted adjacency matrix’s principal eigen value?
e) Radius, diameter, connected components and PageRank.

7. Evidence based Graph Paradigm

Evidence based graph paradigm is the base of the process of forensic analysis. Performances of the evidence graph paradigm include:

1. Gives an intentive impersonate of gathered FE.

2. Forms a basis for functional and structural probe at local and global streams in attack scenario

3. Gives a user friendly platform that permits the researcher to set hypotheses and transform into reasoning mode the out-of-band counsel

A typical graph G, that is evidence based can be deemed as a quadruple (V, A, LV, LA), where V/ A/LV/LA stands for set of vertices/directed arcs/ labels of vertices/edges as the case may be. Here, each vertex v_i denotes a host-level entity and each edge e_i denotes a portion of pre-mediated FE. The purpose of confined logical thinking is to detect the states that are functional of an entity from its restricted observations. Here confined means that the ferret out is only dependent on counsel of the vertex itself and its adjacent vertices. Following closely the states of the host has pertinence in forensics probe. Attacker’s actions are denoted by events that are beyond doubt when analyzed case-by-case bereft of context. So, states’ host give the context to find events that are concealed for further probes. The complexity associated with host types and cyber-attacks makes it complicated to arrive at a conclusion regarding states of the host. So, a fuzzy approach is suggested as it is apt in figuring out the process related to decision making that lay emphasis on quality factor.

8. Conclusion

To conclude, we have considered pertinent problems namely anomaly ferret out and forensic analysis and discussed how graph theory approach could be helpful. We hope to revert more on this elsewhere. Graph theory concepts resolve intrinsic conflicts. For instance, an exact graph of an incident is known, then another incident that yields an isomorphic graph could very well be related. This graph isomorphism concept when applied to serial killers, then the graphs of previously happened crimes could be matched with graphs of new crimes to find whether they were done by the same killer. Also the concept of partial graph isomorphism is deemed as the percentage to which they are isomorphic. To determine the degree of partial isomorphism between two graphs, percentage of same nodes added to the percentage of same links and that sum divided by two, results in the percentage of isomorphism among them.

Acknowledgement

The author acknowledge the National Board of Higher Mathematics, Department of Atomic Energy, Government of India, Mumbai for financial support by their grant no. 02011/10/21NBHM-(R.P)/R&D-II/8007/Date:13-07-2021

References

[1]V. Barnett and T.Lewis, Outliers in Statistical Data, John Wiley and Sons, Chichester, New York, 1994.

[2]Stephen Bay, Krishna Kumarasamy, Markus G. Anderlie, Rohit Kumar, and David M. Steier, Large scale ferret out of irregularities in accounting data, In ICDM, 2006.

[3]Deepayan chakrabarti and Christos Faloutsos, Graph mining : Laws, generators and algorithms, ACM. Comput. Surv, 38(1),2006.

[4]William Eberle and Lawrence B.Holder, Discovering structural anomalies in graph-based data, In ICDM Workshops, 393-398,2007.

[5]Mary McGlohan, Leman Akoglu, and Christos Faloutsos, Weighted graphs and disconnected components: Patterns and a model, In ACM SIG-KDD, Las Vegas, Nev., USA, August 2008.

[6]U. Kang, Hanghang Tong and Jimeng Sun, Fast random walk Kernel, In SIAM Int. conf. on Data mining (SDM) 2012.

[7]U. Kang, charalampas E, Tsourakakis, and Christos Faloutsos, Pegasus: mining peta-scale graphs, Knowledge and Information systems. (KAIS), 27(2), 303-325, 2011.

[8]Caleb C.Noble and Diane J Cook., Graph Based Anomaly Ferret out, SIGKDD '03, August 24- 27, 2003, Washington, DC, USA.

[9]Amit Kr. Mishra, Pradeep Gupta, Ashutosh Bhatt, Jainendra Singh Rana, Innovative Study to the Graph-based Data Mining: Application of the Data Mining, International Journal of Engineering and Innovative Technology (IJEIT) Volume 2, Issue 2, August 2012.

[10]G.Palmer. A roadmap for digital forensics research. In Proceedings of the first Digital Forensic Research Workshop, Utica, New York, USA, 2001.

[11]EnCase Forensic Tool. Available at http://www.guidancesoftware.com.

[12]SafeBack Bit Stream Backup Software. Available at http://www.forensics-intl.com/safeback.html.

[13]tcptrace. http://www.tcptrace.org/.

[14]tcpflow. http://www.circlemud.org/ jelson/software/tcpflow/.

[15]flow-tools. http://www.splintered.net/sw/flow-tools/.

[16]netcat. http://netcat.sourceforge.net/.

[17]eTrust Network Forensics Solution. Available at http://www3.ca.com/.

[18]NetDetector. Available at http://www.niksun.com/Products-NetDetector.htm.

[19]Kulesh Shanmugasundaram, Nasir Memon, Anubhav Savant, and Herve Bronnimann. ForNet: A Distributed Forensics Network. In Proceedings of the Second International Workshop on Mathematical Methods, Models and Architectures for

Computer Networks Security, 2003.
[20]Fan R. K. Chung. Spectral Graph Theory. American Mathematical Society, 1997.
[21]Amy N. Langville and Carl D. Meyer. Deeper inside pagerank. Internet Mathematics, 1(3):335–380, 2004.
[22]ChrisDing. A tutorial on spectral clustering. Talk presented at ICML2004.Slides available at http://crd.lbl.gov/ cding/Spectral/.
[23]Amy N. Langville and Carl D. Meyer. Deeper inside pagerank. Internet Mathematics, 1(3):335–380, 2004.
[24]V. Mehta, C. Bartzis, H. Zhu, E.M. Clarke, and J.M. Wing. Ranking Attack Graphs. In Proceedings of Recent Advances in Intrusion Ferret out 2006(RAID'06), Hamburg, Ger- many, September 2006.
[25]Scott White and Padhraic Smyth. Algorithms for estimating relative importance in net- works. In KDD '03: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, pages 266–275, 2003.
[26]W. Geamsakul, T. Matsuda, T. Yoshida, H. Motoda, and T. Washio. Classifier construction by graph-based induction for graph-structured data. In PAKDD'03: Proc. of 7th Pacific-Asia Conference on Knowledge Discovery and Data Mining, LNAI2637, pages 52{62, 2003.
[27]A. Inokuchi, T.Washio, and H. Motoda. Complete mining of frequent patterns from graphs: Mining graph data. Machine Learning, 50:321{354, 2003.
[28]Kuramochi and G. Karypis. Frequent subgraph discovery. In ICDM'01: 1st IEEE Conf. Data Mining, pages 313{320}, 2001.
[29]X. Yan and J. Han. gspan: Graph-based substructure pattern mining. In ICDM'02: 2nd IEEE Conf. Data Mining, pages 721-724, 2002.
[30]K. Yoshida, H. Motoda, and N. Indurkhya. Graph- based induction as a unified learning framework. J. of Applied Intel, 4:297{328, 199
[31] Cook, D.J. and Holder, L.B. Graph-Based Data Mining. IEEE Intelligent Systems, 15(2), pages 32-41, 2000.
[32]Gonzalez, J., Holder, L.B., and Cook, D.J. Graph-Based Concept Learning. Proceedings of the Seventeenth National Conference on Artificial Intelligence, 2000.
[33]Jonyer, I., Holder, L.B., and Cook, D.J. Discovery and Evaluation of Graph-Based Hierarchical Conceptual Clusters. Journal of Machine Learning Research, 2.
[34]Lee, W. and Xiang, D. Information-Theoretic Measures for Anomaly Ferret out. Proceedings of The 2001 IEEE Symposium on Security and Privacy**,** Oakland, CA, May 2001.
[35]Maxion, R.A. and Tan, K.M.C. Benchmarking Anomaly-Based Ferret out Systems. International Conference on Dependable Systems and Networks, pages 623-630, New York, New York; 25-28 June 2000.
[36]http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html