Diffusion and Graph Spectral Methods for Network Forensic Analysis

Wei Wang and Thomas E. Daniels
Department of Electrical and Computer Engineering
Iowa State University
Ames, IA 50010
{weiwang, daniels}@iastate.edu

Abstract

In this paper we propose the new paradigm of applying diffusion and graph spectral methods for network forensic analysis. Based on an evidence graph model built from collected evidence, graph spectral methods show potential in identifying key components and patterns of attack by extracting important graph structures. We also present the novel view that the propagation of suspicion in an attack scene could be modelled in analogy with heat diffusion in physics systems. In this paradigm, the evidence graph becomes the basis for a physical construct, which derives its properties such as conductivity and heat generation from evidence features. We argue that diffusion and graph spectral methods not only provide a mathematically well grounded approach to network forensic analysis, but also open up the opportunity for applying structured parameter refinement and high performance computation methods to forensic analysis field.

Categories and Subject Descriptors

K.6.5 [**Management of Computing and Information Systems**]: Security and Protection

General Terms

Security,Management

Keywords

Network forensics, graph spectrum, diffusion model

NSPW 2006, September 19–22, 2006, Schloss Dagstuhl, Germany.

1. Introduction

Networks today are plagued by the increasing scale and impact of cyber attacks. In addition to detection and prevention of intrusions, it is important to have post hoc investigation mechanisms that hold attackers responsible for their malicious actions. Similar to its physical world counterpart, network forensic analysis aims to identify suspicious entities in the scene of attack and reconstruct stepwise actions of the attacker by reasoning with evidence captured from networked environments. In contrast to sophisticated multi-stage attacks and huge amount of available sensor data, current practices in network forensic analysis are still mainly done by manual ad-hoc methods, a time-consuming and error prone process[6]. The analysis process remains challenging due to the lack of scientifically well-founded and systematic methods.

In this paper we present a novel view to address the limitations of current ad-hoc analysis methods. Based on our previous work in forensic analysis with the evidence graph model, we propose a graph theoretic approach with diffusion and spectral methods. Graph-spectral and related kernel based methods show potential in attack scenario extraction and attack case profiling with their ability to efficiently extract structure characteristics of the evidence graph. The propagation of suspicion in the attack scene is modelled in analogy with heat diffusion in physics terms. In addition, our approach provides well a structured framework for parameter refinement and high performance computation.

The remainder of the paper is organized as follows. A brief review of our background work in evidence graph model and hierarchical reasoning framework is presented in section 2. Section 3 describes the supporting theoretical background for proposed diffusion and graph-spectral methods and their applications in forensic analysis. Section 4 discusses practical considerations for model refinement and high performance computation. Finally, section 5 presents related work and section 6 concludes this paper.

2. Preliminaries

In this section, we give a brief introduction to the evidence graph model and hierarchical reasoning framework as the basis of our proposed methods. In our previous work [18], an extensible graph model has been developed to integrate collected evidence from heterogenous sources. The resulting graph structure captures entities, events and functional states in the attack scenario for analysis.

2.1 The Evidence Graph Model

An evidence graph is a quadruple $G = (N, E, L_N, L_E)$, where N is the set of nodes, E is the set of edges, L_N is the set of labels for attributes of nodes and L_E is the set of labels for attributes of edges.

In our host-centric evidence graph, each node n_i represents a host level entity of forensic investigation interest and each edge e_i represents a piece of observed forensic evidence. Each node is characterized by a set of fuzzy functional states for attack scenario analysis. For example, a limited set of fuzzy states could be S=$\{$*Attacker*, *Victim*, *Stepping Stone*, *Affiliated*$\}$, which describes possible roles of hosts in the attack scenario. This set of functional classification could clearly be refined, but we believe it is adequate as an illustrative start.

Each edge in the evidence graph is characterized by a set of numerical attributes *weight*, *relevancy* and *context importance*. These attributes are instantiated based on domain knowledge. *Weight* of the edge is a fuzzy value $h \in [0, 1]$ that represents impact of the attack represented with higher value indicates more serious impact. *Relevancy* $r \in [0, 1]$ represents the belief that the underlying attack indicated by the evidence would successfully achieve its expected impact. Currently we apply a static evaluation approach that compares the prerequisites of an attack with target host's configuration. If all prerequisites are completely satisfied, its relevancy value is assigned as 1. If one or more contradicting configurations are found, its relevancy value is assigned as 0; otherwise we are unable to determine the relevancy value is assigned as 0.5. *Context importance* $h \in [0, 1]$ is used to relate significance of evidence with value of the hosts involved, which is predefined from site specific knowledge of the network under investigation. Finally, we calculate *priority score* for an edge as the product of its weight, relevancy and context importance to indicate overall priority of the evidence.

2.2 Building the Evidence Graph

With the evidence graph model, we transform forensic evidence from heterogeneous information sources into a weighted digraph which may include multi-edges. This graph structure lays the basis for our proposed diffusion and spectral analysis methods. To construct the evidence graph, the sequence of intrusion evidence is processed in time order, starting from the first evidence record and moving towards the latest evidence record. Evidence with interval time stamps is added to the graph in order of the start time in their interval and ties are broken arbitrarily.

```
input  : Stream of evidence records in time order
output: Evidence graph $G$
begin
    $G \leftarrow \phi$;
    foreach evidence E in stream do
        foreach host V as subject or object of E do
            if V does not exist in G then
                CreateNode $(G, V)$;
            end
        end
        CreateEdge $(G, E)$;
        foreach host V as subject or object of E do
            UpdateNode $(E, V)$;
        end
    end
end
```

Algorithm 1: Building an evidence graph

As shown in algorithm 1, we evaluate its subject and object node of each evidence record in the first step. The *CreateNode* and *CreateEdge* functions add nodes and edges to the evidence graph. The *UpdateNode* function performs fuzzy inference to determine states of the subject and object nodes, which corresponds to the local reasoning procedure in our hierarchical reasoning framework.

2.3 Hierarchical Reasoning Framework

We perform forensic analysis with a hierarchical reasoning framework of two levels: local reasoning and global reasoning. The objective of local reasoning is to evaluate functional states for hosts from local observations. In the evidence graph context, "local" means reasoning is solely based on the node's incident edges and states of its neighbors. A fuzzy inference approach based on Rule-Based Fuzzy Cognitive Maps (RBFCM) has been developed to model the set of node states S described in model definition. The local reasoning process is regarded as part of the model generation stage described in the next section. The fuzzy states inferred could be used to evaluate suspicion generation in the diffusion model.

The global reasoning process aims to extract the set of entities tightly connected in the foreground attack scene and infer their relationships. Based on the evidence graph, our former approach to the global reasoning process was as a group detection problem that works in two different phases: (1) creating new attack groups by generating seeds for a group and (2) extending the existing group by discovering more hidden members.

As described in previous work [18], we start with a seed host empirically chosen from its context or network centrality metrics. Following that a greedy iterative algorithm is used to extend the attack group by adding nodes with correlation strength above a predefined threshold as new seeds. The attack scenario is reconstructed from the extracted subgraph of correlated nodes and corresponding functional states from local reasoning results.

The limitation of this approach is that the reasoning process remains ad-hoc and is intractable for large scale analysis. Metrics for seed generation and thresholds in group expansion process are determined to a large degree by analyst expertise. The iterative group expanding process is computationally expensive for massive graphs. Diffusion and graph spectral methods will provide more systematic and efficient solutions for both phases of the global reasoning process.

- The seed generation phase aims to discover important suspicious entities as initial seeds of attack group. In the evidence graph space, the problem can be transformed into identifying clusters of certain important structure characteristics. In recent graph theory [13, 19] and link analysis [7] work, graph spectral and kernel methods have been well studied to evaluate structure of large complex graphs.
- The group expansion phase is based on the invariant that entities belong to the attack scenario of interest should be strongly correlated through certain suspicious relations. In essence, group expansion can be regarded as tracking the flow of suspicion across the evidence graph, which could be explored in analogy with the diffusion model in classical physics [8].

3. Forensic Analysis with Diffusion and Graph-Spectral Methods

Based on the evidence graph model, we foresee the potential of diffusion and graph spectral methods in two major applications of network forensics analysis: attack scenario extraction and attack case profiling.

- *Attack Scenario Extraction* is the process of inferring the set of entities and events associated with the attacker. In our approach, scenario extraction can be seen as identifying suspicious nodes, extracting clusters in which the constituent nodes are tightly correlated and tracking the flow of suspicion among them.
- *Attack Case Profiling* is based on the promise that attackers tend to repeat certain behavioral patterns or strategies. Given a new extracted scenario, the natural question to ask is whether it is similar to patterns seen in the past or elsewhere. Knowing the recurrence of attack scenarios would help the investigator form appropriate responses. As our evidence graph model presents a graph based description of attack patterns, this is similar to the problem of subgraph isomorphism, a known NP-complete decision problem [5]. Our goal is to develop efficient graph spectra characterizations for encoding and matching of attack cases.

In the following, we will describe the supportive background of diffusion and graph spectral methods as well as their potentials in the above forensic analysis applications.

3.1 Analysis Model Overview

Figure 1 shows the major stages of our network forensics analysis process.

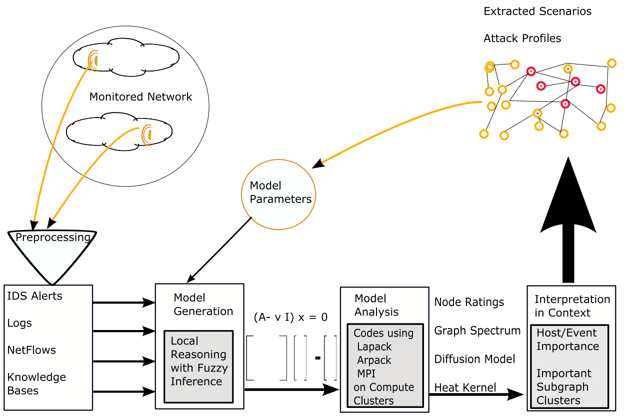


Figure 1. Overview of network forensic analysis process

- In *Evidence Preprocessing* phase, intrusion evidence from heterogeneous information sources such as IDS alerts, flow records and host logs are normalized into a unified conceivable format. Abstraction and aggregation are performed to reduce the redundancy in raw evidence.
- In *Model Generation* phase, the preprocessed forensic evidence are transformed into evidence graphs. In this stage attributes of the evidence graph are instantiated based on information retrieved from evidence depository and knowledge bases. Fuzzy based local reasoning is also performed to determine each node's functional states.
- In *Model Analysis* phase, we apply diffusion and graph spectral methods to extract information such as cluster importance and flow of suspicion from the evidence graph structure, which corresponds to global reasoning in our hierarchical reasoning framework.
- In *Scenario Interpretation* phase, results of diffusion and graph spectral methods are interpreted and filtered with domain specific knowledge to produce final analysis report.

3.2 The Laplacian Spectrum and Kernel Based Methods

Our first approach is to use the Laplacian spectrum of evidence graphs for scenario extraction. It is known that many principal properties of a graph are closely related to its graph spectrum [13, 20]. To characterize the properties of a graph and extract information from its structure, we compute the graph spectrum using its Laplacian matrix representation. Let the graph denoted by $G = (V, E)$ where

V is the set of nodes and $E \subseteq V \times V$ is the set of edges. The square adjacency matrix A of G is defined as:

$$A(u,v) = \begin{cases} 1, & \text{if (u,v)} \in E \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

where u and v are nodes in the graph. The diagonal degree matrix D is constructed as $D(u,u) = \Sigma_{v \in V} A(u,v)$. Then the Laplacian matrix representation of G is given by $L = D - A$. The normalized Laplacian is defined as $\hat{L} = D^{-\frac{1}{2}} L D^{-\frac{1}{2}}$. For a weighted multigraph like our evidence graph, we can easily deduce its normalized Laplacian as follows:

$$\hat{L}(u,v) = \begin{cases} 1 - \frac{w(u,u)}{d_u}, & \text{if } u = v \text{ and } d_u \neq 0 \\ -\frac{w(u,v)}{\sqrt{d_u d_v}}, & \text{if } u \text{ and } v \text{ are adjacent} \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

To simplify eigendecomposition, we ignore the directness of edges in the evidence graph G to avoid complex eigenvalues for nonsymmetric matrix. In equation 2, $w(u,v)$ represents the sum of priority scores for all edges between node u and v in the evidence graph. Degree of node u is defined as $d_u = \sum_v w(u,v)$. Rows and columns of the Laplacian matrix $\hat{L}(G)$ are indexed by vertices of the evidence graph G.

Given the Laplacian representation $\hat{L}$, spectrum of the evidence graph is obtained by the eigendecomposition $\hat{L} = \Phi \Lambda \Phi^T$ where $\Lambda = diag(\lambda_1, \lambda_2, ..., \lambda_{|V|})$ is a diagonal matrix of eigenvalues and $\Phi = (\phi_1, \phi_2, ..., \phi_{|V|})$ is the matrix composed with eigenvectors as columns. The Laplacian spectrum of graph G refers to the set of eigen values $(\lambda_1, \lambda_2, ..., \lambda_{|V|})$.

We observe that structure of the evidence graph offers a first approximation of attack patterns, though much irrelevant noise is included. Therefore we are currently investigating several graph spectral measures to extract the attack scene from large complex evidence graphs. The Laplacian spectrum has been extensively explored in graph theory to characterize graph level structure properties such as connectivity, diameter and path length distribution [13, 1]. Specifically we are investigating spectral metrics that extract two types of information from the evidence graph: (1) Spectral features that represent important individual nodes, i.e. the ”key player” in the attack scenario (2) natural clusters of highly correlated suspicious nodes, i.e. the extended attack group in the scenario.

As an initial experiment, we perform graph spectrum analysis on the Lincoln lab LLDOS 2.0 dataset. The LLDOS 2.0 dataset contains a multi stage attack scenario that include the following sessions:

1. The attacker probes the target network using a valid DNS HINFO query;
2. The attacker compromises the DNS server by exploiting the Solaris Sadmind vulnerability.
3. The attacker uploads exploit scripts and mstream DDoS agent/master to the compromised DNS server by FTP.
4. The attacker telnet to the compromised DNS server and repeats the probing and Sadmind attack process towards hosts in the same domain. After successful attack against a Solaris host, the attacker uploads mstream agents by FTP.
5. The attacker access the compromised hosts by telnet and initiates DDoS attack towards an external web server.

We us Snort as the IDS sensor to detect intrusions in the traffic dump. Flow records are also extracted and stored into a MySQL database. In the preprocessing phase, raw alerts are aggregated to reduce redundancy. File transfer(ftp) and remote access(ssh,rlogin,telnet) flows associated with hosts that have $Attacker$, $Victim$ or *Stepping Stone* states activated in selected time frames are incorporated to build the evidence graph.

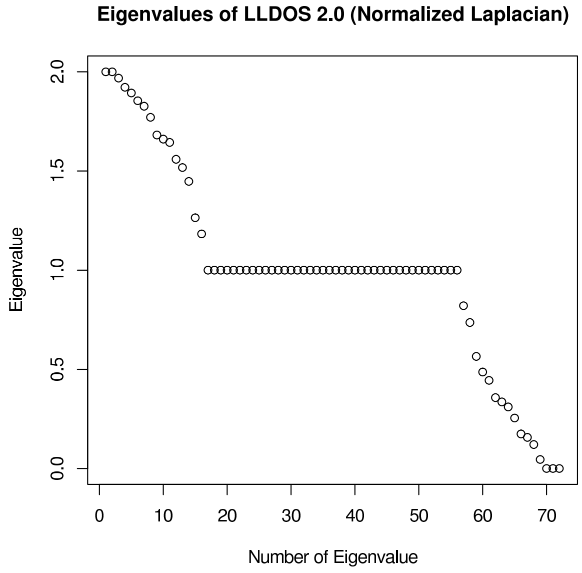


Figure 2. Graph Laplacian spectrum for LLDOS 2.0 dataset

In the next step, we perform spectral analysis on the evidence graph. Figure 2 shows eigenvalues of the graph's normalized Laplacian. We truncate the graph spectrum by picking leading eigenvectors corresponding to the n largest eigenvalues $v_1, v_2, ..., v_n$ as the largest eigenvalues are more informative of graph structure. Distribution of eigenvectors v_1, v_2, v_3, v_4 is shown in figure 3. In each eigenvector we identify the significant components and find the corresponding subgraph in the evidence graph for examination.

- For principal eigenvector v_1, the corresponding subgraph is shown in figure 4(a). The largest magnitude component in $v_(1)$ locates host 131.84.1.31, which is target of the DDOS attack. An isolated cluster caused by background attack is also extracted.

- Figure 4(b) shows the subgraph corresponding to significant components in eigenvector v_2. We observe that it represents a subset of the components represented in dominant eigenvector v_1.

- As can be seen in figure 4(c), components of significant magnitude in eigenvector v_3 map to a cluster of routers and service hosts in the DMZ, which are brought up by background traffic artifacts.

- In eigenvector v_4, we observe that components of most significant magnitudes maps to the two DDOS agent hosts 172.16.115.20 and 172.16.112.50. The other nodes shown in figure 4(d) are caused by background artifacts.

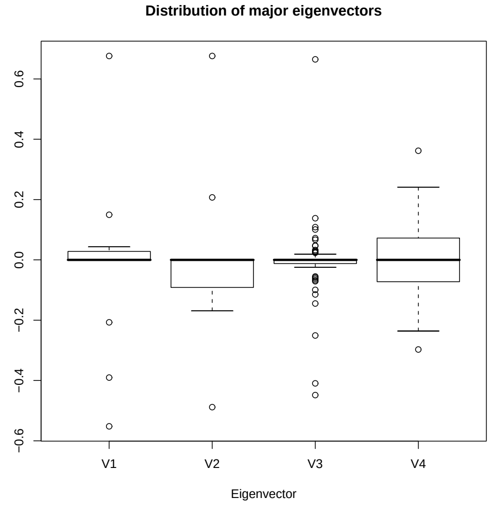


Figure 3. Distribution of major eigenvectors from LLDOS 2.0 graph

It is interesting to see that we can extract important components in the attack scenario and identify clustering patterns from the leading eigenvectors. Eigenvectors with lower magnitude of eigenvalues exhibit a random pattern and mostly correspond to trivial entities in the attack scenario. To reduce the amount of irrelevant clusters in the extracted subgraph, we could build a history profile for filtering purpose. The failed break-in attempt against inside host 172.16.112.207 is not extracted, which indicates that better spectra metrics and model refinement need to be investigated. Moreover, due to the deficiency in background traffic and attack generation of LLDOS 2.0 dataset [11, 12], possible correlation between the scale of background traffic and attack activity may lead to biased results.

In addition to Laplacian spectrum analysis, we are also investigating closed related kernel based methods to analyze the evidence graph. In machine learning works, kernel methods have been used to capture correlations between data points represented in a graph structure. In essence kernel methods can be viewed as an implicit mapping from data space to some feature space that better captures inherent structure of the data.

Specifically, we are interested in a class of kernels denoted as *heat kernels* for their suitability in discrete graph space. The heat kernel is defined by the heat equation associated with the graph Laplacian, i.e. $\frac{\partial h_t}{\partial t} = -\hat{L}h_t$, where h_t is the heat kernel and t is time. The solution is found by exponentiating the Laplacian spectrum, i.e. $h_t = \Phi e^{-\Lambda t}\Phi^T$. The resulting heat kernel for the graph is a $|V| \times |V|$ matrix, where $h_t(u,v) = \sum_{i=1}^{n} e^{-\lambda_i t}\phi_i(u)\phi_i(v)$.

We believe kernel based methods have potential in our graph based network forensic analysis. It has been shown that short time behavior of the heat kernel is determined by local topology of the graph while its long time behavior is determined by the global structure of the graph [19]. Thus by varying t we are able to extract attack patterns at multiple scales from the underlying evidence graph structure. Consequently the heat kernel has been explored to characterize and compare graphs [20, 19], which leads to application in attack case profiling. As attack patterns are represented by extracted graph structure and the kernel function maps graph structure into a vector space, we can evaluate similarity between attack patterns by making comparisons between corresponding point distributions. Here a major challenge is to develop appropriate kernel parameters that minimize the cospectrality effect of graphs and provide an effective spectral characterization. Adequate abstraction of the evidence graph might be needed prior to encoding and matching. Moreover, information such as functional states of nodes need to be considered together with the spectral signature for profile building.

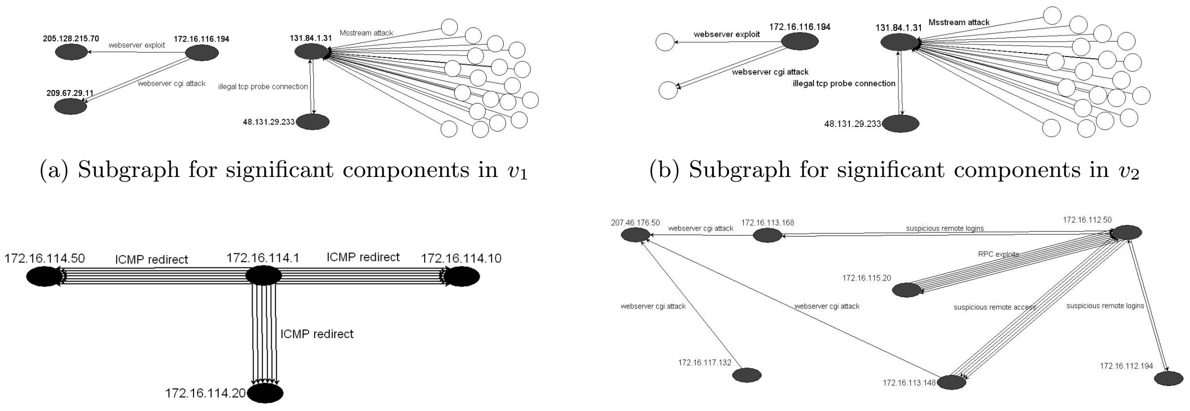
3.3 Diffusion Model

Our second approach is to transform evidence graph analysis into approximations of steady state energy diffusion problems. In classical physics, the heat diffusion equation is used to describe the diffusion of heat through continuous media:

$$\nabla^2 T = \frac{\partial^2 T}{\partial x^2} + \frac{\partial^2 T}{\partial y^2} + \frac{\partial^2 T}{\partial z^2} = \frac{c}{k}\frac{\partial T}{\partial t} - Q(x,y,z) \qquad (3)$$

where ∇ is the continuous Laplacian operator. In the equation, Q represents the effects of an internal source of heat while c and k are constants representing the heat capacity and conductivity of the material.

The diffusion model is closely related to the spectral methods described above. For a graph $G = (V, E)$, if we consider each node as $|V|$ independent physical items



(a) Subgraph for significant components in v_1

(b) Subgraph for significant components in v_2

(c) Subgraph for significant components in v_3

(d) Subgraph for significant components in v_4

Figure 4. Sub evidence graphs extracted from eigenanalysis

where diffusion of heat can occur across the edges in E, then $\frac{\partial^2 T(u)}{\partial v^2}$ is analogous to $\hat{L}(u,v)\vec{T}$ where $\vec{T}$ is a vector of temperatures defined at all V.

As described in the previous section, the heat kernel of a graph can be exponentiated to generate time step solutions to a diffusion problem. However, extracting all eigenpairs by eigendecomposition may be impractical for large directed graphs. On the other hand, the steady state diffusion model can be solved by more direct numerical methods as they do not require extraction of eigenvalues and eigenvectors. With the steady state diffusion model we are able to handle massive directed evidence graphs for more effective scenario extraction.

Diffusion problems in the steady state are often solved by means of a finite element analysis (FEA) as well as the closely related finite difference analysis. FEA attempts to approximate the solutions of a set of partial differential equations at a finite set of elements subject to a set of boundary conditions. The general approach in FEA is to make a first order Taylor series approximation of the differential equation at each point (element) in the material based on its neighbor elements and the associated boundary conditions. These approximations typically lead to n equations in n unknowns which can then be solved using straightforward Gaussian elimination, iterative methods, or even multi-resolution techniques [3].

Our fundamental motivation for exploring the diffusion model is that the attack process can be regarded as the propagation of suspicion in the network, which could be modelled in analogy with heat diffusion in the evidence graph. To model properties of the diffusion model, we explicitly utilize the concept of suspicion in that each entity and piece of evidence are regarded as the container or carrier of certain amount of suspicion. In this paradigm, the evidence graph becomes the basis for a physical construct. Heat represents suspicion and temperature indicates the level of suspicion. Based on local reasoning results in the model generation stage, the inferred fuzzy states are used to emulate suspicion injected into the graph. Each node with certain activated states propagates suspicion to its neighborhood. On the other hand, each edge acts as an insulated conductor connecting nodes. Conductivity of the edge could be modelled based on features of evidence, such as duration, classification and traffic rate. Additional parameters include the boundary conditions that models suspicion radiating and entering nodes at some rate. The resulting suspicion flow map would present a quantitative measure of attack propagation and help the investigator focus on "hot" areas for further investigation. The flow map would also help us identify hosts and connections that are involved in the attack scenario but are not directly associated with security alerts.

4. Practical Issues

In this section we discuss important issues in the practical application of our proposed diffusion and graph spectral methods. We argue that our approach has significant advantages in model refinement and computation performance.

4.1 Model Refinement

As shown in figure 1, effectiveness of network forensic analysis is affected by parameters assigned in various stages. In the model generation stage, appropriate weights should be decided for different types of evidence. In the model analysis phase, it is important to find appropriate parameters for kernels and diffusion models such as decay factor and conductivity. In current forensic analysis process, parameters are just chosen by expert knowledge, which inevitability lead to the question that these values are somewhat arbitrary. Moreover, it is difficult to evaluate the quality of parameters and their impact on analysis results.

One significant advantage of our proposed diffusion and graph spectral methods is that it provides a mathematically well-grounded approach to refine models by reverse solving analysis results. The initial arbitrary designed model is refined by algebraically and then numerically back solving for parameters in an iterative process. First, analysis is performed on the labelled training datasets with a parameter set initialized manually. Second, the analysis results are evaluated with the ground truth. We go back to the model and attempt to find an assignment of parameters that achieves the most desirable results. An intuitive approach is to create an over determined system that we will solve using simplex or gradient descent methods. Finally, the revised set of parameters will then be used against new scenarios in a forward analysis to evaluate its performance. Here a big challenge is to develop training datasets that are representative of real world attack scenarios.

4.2 Computation for Large Scale Analysis

The increasing scale of cyber attacks and huge amount of evidence necessitates efficient computational forensic analysis. To our knowledge, most work in intrusion analysis field such as alert correlation methods follow a serial computation model and little has been done in using high performance computing for security data analysis.

One advantage of our approach is that we can utilize well established computational methods and software packages for large scale network forensic analysis. The computational building blocks of diffusion and graph spectral methods including eigenanalysis and solution of linear systems have been studied for over 60 years in electronic digital computers. This has lead to a wealth of computational methods and software that accommodates huge problem sets in high performance, parallel and distributed computing environments.

Eigenanalysis, the process of solving problems by finding eigenvalues and eigenvectors, has been explored in a variety of fields. Efficient algorithms and software for eigenanalysis of systems with 10^6 variables have been developed [10]. This would easily handle an evidence graph containing the same number of hosts. From the perspective of available solutions for linear system equations supporting finite element simulations and eigenanalysis, we foresee several exciting possibilities. The first is to use computer clusters for analysis of network datasets larger than any considered feasible before. Second, by using problem decomposition and multilevel solution techniques [3], it is possible to perform online intrusion analysis with cooperating agents distributed across networks. For example, each monitor would utilize iterative methods derive solutions from its own view and share solutions between each other. The performance is related to iterative solution techniques and the level of coupling between evidence observed by the distributed monitors.

5. Related Work

Our approach is stemmed from the evidence graph model and reasoning framework proposed in [18]. To our knowledge, no similar techniques have been used in the network forensic analysis field. Most past work such as alert correlation techniques [17, 2, 15] focus on a specific type of evidence instead of forming big picture incident reconstructions. Network forensic tools like eTrust [4] and NetDetector[14] have been widely used to capture evidence and investigate security breaches, however the analysis procedures are mostly ad-hoc or based on hard coded knowledge.

Important properties of the Laplacian graph spectrum are summarized in [13, 1]. In [20], a wide variety of spectrums based on different graph representations are explored as metrics for graph matching. It has been shown that the heat kernel performs well in characterizing structure of graphs [19]. In computer vision work, spectral characterizations has been developed to capture hierarchical graph structures into a low-dimensional vector space [16].

Heat kernels on discrete graph space are first defined in [8]. There are studies that explore the application of kernel methods to link analysis [7, 9]. In [7], kernels based on the graph Laplacian are used to yield link analysis measures such as importance and relatedness. Property of the kernel based measures is evaluated with a network of bibliographic citations. These kernel measures could be extended to deal with more complex relationships represented by our evidence graph model.

6. Conclusion

In this paper we have presented a novel paradigm that applies diffusion and graph spectral methods to network forensic analysis. Our approach has potential to transform the current ad-hoc forensic investigation process into a systematic framework with well grounded mathematical methods. We show that diffusion and graph spectral methods have promise for attack scenario extraction and attack case profiling.

This work is the starting point towards an entirely new paradigm of network forensic analysis. We are currently in the process of investigating appropriate graph spectral measures and steady state diffusion model. An important concern is how our proposed approach would perform in a realistic noisy network environment where most of the evidence are false positives or irrelevant background attacks. With the development of diffusion and graph spectral measures, more experiments will be conducted to evaluate their performance under different scenarios and signal-noise ratios.

7. References

[1] F. R. K. Chung. *Spectral Graph Theory*. American Mathematical Society, 1997.

[2] H. Debar and A. Wespi. Aggregation and Correlation of Intrusion-Detection Alerts. In *Proceedings of the 4th International Symposium on Recent Advances in Intrusion Detection(RAID)*, October 2001.

[3] C. C. Douglas. Multigrid methods in science and engineering. *IEEE Computer Science Engineering*, 3:55–68, 1997.

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

[5] M. R. Garey and D. S. Johnson. *Computers and Intractability: A Guide to the Theory of NP-Completeness*. W.H. Freeman, 1979.

[6] Institute for Security Technology Studies. Law enforcement tools and technologies for investigating cyber attacks: Gap analysis report. February 2004.

[7] T. Ito, M. Shimbo, T. Kudo, and Y. Matsumoto. Application of kernels to link analysis. In *KDD '05: Proceeding of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 586–592, New York, NY, USA, 2005.

[8] R. I. Kondor and J. D. Lafferty. Diffusion kernels on graphs and other discrete input spaces. In *ICML '02: Proceedings of the Nineteenth International Conference on Machine Learning*, pages 315–322, 2002.

[9] J. Lafferty and G. Lebanon. Information diffusion kernels. 2002.

[10] R. Lehoucq, D. Sorensen, and C. Yang. Arpack users' guide: Solution of large scale eigenvalue problems with implicitly restarted arnoldi methods. Technical Report from http://www.caam.rice.edu/software/ARPACK/, Computational and Applied Mathematics, Rice University, October 1997., 1997.

[11] J. McHugh. The 1998 lincoln laboratory ids evaluation. In *Recent Advances in Intrusion Detection*, pages 145–161, 2000.

[12] J. McHugh. Testing intrusion detection systems: a critique of the 1998 and 1999 darpa intrusion detection system evaluations as performed by lincoln laboratory. *ACM Trans. Inf. Syst. Secur.*, 3(4):262–294, 2000.

[13] B. Mohar. The laplacian spectrum of graphs. *Graph Theory, Combinatorics, and Applications*, (2):871–898, 1991.

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

[15] P. Ning, Y. Cui, and D. S. Reeves. Constructing attack scenarios through correlation of intrusion alerts. In *9th ACM Conference on Computer and Communicaitons Security*, November 2002.

[16] A. Shokoufandeh, D. Macrini, S. Dickinson, K. Siddiqi, and S. W. Zucker. Indexing hierarchical structures using graph spectra. *IEEE Trans. Pattern Anal. Mach. Intell.*, 27(7), 2005.

[17] A. Valdes and K. Skinner. Probablistic alert correlation. In *Proceedings of the 4th International Symposium on Recent Advances in Intrusion Detection(RAID)*, October 2001.

[18] W. Wang and T. E. Daniels. Building evidence graphs for network forensics analysis. In *Proceedings of the 21st Annual Computer Security Applications Conference(ACSAC)*, Tucson, Arizona, December 2005.

[19] B. Xiao, R. C. Wilson, and E. R. Hancock. Characterising Graphs using the Heat Kernel. In *Proceedings of the 16th British Machine Vision Conference*, Oxford Brookes University, Oxford, September 2005.

[20] P. Zhu and R. C. Wilson. A study of graph spectra for comparing graphs. In *The 16th British Machine Vision Conference*, Oxford Brookes University, Oxford, September 2005.