The Early Bird Gets the Buzz: Detecting Anomalies and Emerging Trends in Information Networks*

Brian Thompson
Rutgers University
110 Frelinghuysen Road
Piscataway, NJ 08854
bthom@cs.rutgers.edu

ABSTRACT

Anomaly detection has a wide range of real-world applications, including: monitoring computer network usage, virus detection (computer or human), credit card fraud detection, and natural disaster prediction. However, unprecedented growth in the capability to collect massive amounts of data has introduced new challenges in efficiency and scalability. Furthermore, communication data is highly dynamic, so a comprehensive solution should exploit temporal as well as relational aspects of network communication.

In this work we propose a novel approach to anomaly detection in streaming communication data that is able to leverage the wealth of temporal and relational information inherent in the data. We first build a stochastic model for the system based on temporal communication patterns across each edge, which we call the REWARDS (REneWal theory Approach for Real-time Data Streams) model. We then define a measure of anomaly for an arbitrary subgraph based on the likelihood of its recent activity given past behavior. Finally, we develop an algorithm to efficiently identify subgraphs with the most anomalous activity. Experiments on a variety of real-world data show the effectiveness and scalability of our approach.

Although our work has until now focused on the cybersecurity domain, the model we present is more broadly applicable to information retrieval in data streams and information networks.

Categories and Subject Descriptors

E.1 [**Data Structures**]: Graphs and networks; G.2.2 [**Graph Theory**]: Graph algorithms; G.3 [**Probability and Statistics**]: Time series analysis; G.3 [**Probability and Statistics**]: Renewal theory

1. INTRODUCTION

1.1 Motivation

Anomaly detection has attracted attention in recent years, motivated by its applicability to a variety of existing and emerging domains, particularly cybersecurity. Despite the importance of network monitoring, the standard toolbox used is fairly primitive, including simple procedures for measuring total network traffic, activity on the major hubs, and hosts with the most traffic. While often sufficient for identifying malicious behavior such as port scanning and denial-of-service attacks, these techniques are limited in their ability to capture the complexities of network dynamics and therefore are not conducive to more sophisticated anomaly detection methods or other data mining applications.

Our contributions can be summarized as follows:

- The REWARDS (REneWal theory Approach for Real-time Data Streams) model, a streaming model for analyzing the dynamics of communication networks
- A formalization of the notion of recency for temporal data represented as a sequence of time-stamped events
- A statistical approach to anomaly detection in communication networks that incorporates both temporal patterns and knowledge of network structure
- An efficient algorithm (linear in number of edges) that identifies subgraphs of a network exhibiting anomalous behavior

2. RELATED WORK

Anomaly detection has been studied extensively in a variety of domains, from detecting credit card fraud to monitoring activity in computer networks. In this work, we focus on applications to the domain of time-evolving networks. For an overview of the many different models and approaches used for anomaly detection, we refer the reader to a recent survey by Chandola et. al. [6]. In their classification, our approach can be described as using statistical methods for identifying collective, contextual anomalies. That means that we define an anomaly to be a set of objects whose collective behavior is rare or improbable within the context of their environment. We now take a look at several recent approaches to anomaly detection in time-evolving networks.

2.1 Static Graph Analysis

One popular approach begins by segmenting time into blocks, and constructing a graph (weighted or unweighted) to represent the communication aggregated over each time window, called a *summary graph*. Then, static graph algorithms such as clustering, spectral, neighborhood, and centrality analysis are applied. Anomalies are defined as nodes or substructures that are statistical outliers in the graph based on some pre-defined measure.

Noble at. al. identify subgraphs that appear infrequently, using a data mining tool called Subdue and a variant of the Minimum Description Length (MDL) Principle [13]. Sun et. al. identify outlier nodes in bipartite graphs based on properties of their neighborhood [18]. Akoglu et. al. propose OddBall, which takes a similar approach for finding outlier nodes in a weighted summary graph [2].

*This is an abridged version of joint work with James Abello.

WSDM '12 Seattle, Washington USA

2.2 Temporal Node Comparison

A second approach evaluates each node based on its past history and uses this information to compare different nodes across the network at a single time. Priebe et. al. represent the network as a sequence of summary graphs with 1-week time blocks, examining properties of each node's neighborhood over time [15]. Candia et. al. study anomaly detection in spatio-temporal phone data, analyzing daily call volume from each cell tower and comparing to the mean call volume for that tower [4]. They use ideas from percolation theory to identify times and spatial regions of high activity.

2.3 Change Detection

Change detection looks on how summary graphs change over time. Sun et. al. take an approach based on the MDL Principle, where clustering is performed to minimize the number of bits required to represent the graph, and *change points* occur when there is significant change in the representation [17]. Henderson et. al. measure properties of each summary graph, performing more detailed analysis of community structure and individual node behavior only when there are significant changes in the global metrics [9].

2.4 Time Series Analysis

Another approach models communication data as a continuous time series, and uses tools from signal processing to analyze patterns of communication for edges, nodes, or the whole graph. Ihler et. al. use a hidden Markov model to understand temporal patterns in network traffic volume and distinguish between normal and abnormal behavior [10]. Cao et. al. learn a B-spline model, identifying both short-term deviations and long-term trends [5].

Lakhina et. al. do statistical outlier detection on the time series of traffic volume across origin-destination flows, in terms of # of bytes, # of packets, and # of IP-flows [12]. Their approach is to use PCA to find the most prevalent trends across all flows (top-k eigenflows), and then mark a flow as anomalous at a particular time based on how well it matches the eigenflow prediction. Abello et. al. also use a time series model, but consider the activity of each node or edge in the context of the overall network behavior [1].

2.5 Temporal Path Tracing

Temporal path tracing considers paths composed of time-ordered edges. Tang et. al. use temporal path length to study the rate of information diffusion across a network [19]. Xie et. al. propose an approach for identifying the origin of viruses in a network using *moonwalks*, random walks on the graph backward in time [21].

2.6 Time-scale Bias

Almost all of the above approaches are susceptible to a phenomenon we call *time-scale bias*. There is no unified way of defining time blocks that will capture deviations in graph structure at every time scale. Too short a block length may result in high variance and sample data that is not representative of the underlying graph structure; too long a block length may have a smoothing effect that hides shorter-term deviations in behavior. Furthermore, different time granularities may be appropriate for different nodes or edges within the same network. To the best of our knowledge, the existing literature does not address this fundamental issue.

In addition, most existing approaches are designed for off-line analysis. Our goal is to develop efficient and scalable methods for analyzing streaming data in information networks.

3. MODEL AND APPROACH

To address these challenges, we propose a novel statistical approach for anomaly detection in communication networks. First, we introduce the REWARDS (REneWal theory Approach for Real-time Data Streams) model for understanding communication patterns across each edge. We then perform statistical analysis to quantitatively measure correlation of communication across edges in an arbitrary subgraph. Finally, we develop a graph algorithm to efficiently identify subgraphs with the most anomalous behavior.

3.1 The REWARDS Model

We consider data in the form of a stream of time-stamped events. In this paper we define an event to consist of undirected communication between two entities. We use $\langle (v,u);t \rangle$ to denote communication between entities v and u at time t, and refer to the pair as a *network edge*. For each network edge $e = (v,u)$, let $T_{e,t}$ denote the sequence of timestamps $t_i \leq t$ during which there was communication between v and u.

While $T_{e,t}$ includes much temporal information, it may not be feasible to store the entire history of communication across all edges in a large network. We address this issue by studying the *distribution of inter-arrival times* between consecutive events. For this, we appeal to the field of renewal theory. A *renewal process* is a continuous-time Markov process where new events occur with inter-arrival times (IATs) sampled independently from a fixed distribution. For a more thorough introduction to renewal theory, we refer the reader to [16].

We model communication across a network edge $e = (v,u)$ as a renewal process with IATs sampled from a probability distribution μ_e. In the absence of prior knowledge of the true distribution, we approximate μ_e based on $T_{e,t}$, the sequence of IATs previously seen. For details on our implementation, see Section 4.

In the real world, two entities may, for various reasons, sever communication permanently. To model this phenomenon, we say network edge $e = (v,u)$ is *alive* at time t if the elapsed time since their last communication is less than $iat_{e,t}^{max}$, the maximum inter-arrival time previously seen for that pair, and *dead* otherwise. We now define the graph $G_t = (V_t, E_t)$, where $V_t = \{\text{entities}\}$ and $E_t \subseteq V_t \times V_t$ consists of all network edges that are alive at time t.

The IAT distribution summarizes the long-term behavior of a network edge. We now introduce the second part of our REWARDS model, the notion of *recency*, which assigns greater relevance to communication that happened more recently.

3.2 Recency

In renewal theory, the *age* of a renewal process is the time elapsed since the last event. Adapting this terminology, we define $Age(e,t) = t - t_e^{last}$, the time elapsed since the last communication across edge e. Consider the cumulative distribution function for values of $Age(e,t)$, sampled uniformly over all times at which e is alive:

$$CDF_e^{Age}(\tau) = \mathbf{P}(Age(e,t) \leq \tau).$$

We now define the recency of an edge e at time t:

$$Rec(e,t) = 1 - CDF_e^{Age}(Age(e,t)).$$

We observe that Rec is a decreasing function on the age of a network edge, having value 1 at the time of a new event, 0 when iat_e^{max} time has elapsed, and taking values uniformly in $[0,1]$ when sampled over times at which e is alive. (In fact, it is the unique such function.) Formally, the uniformity property states that for any edge e and time t, $Pr(Rec(e,t) \leq z) = z$ for all $z \in [0,1]$. Recency thus gives us a simple, unbiased way to measure how much time has passed since the last time-stamped event, and lays the foundation for applying statistical analysis. However, we have not yet made use of the graph structure. In our next step, we consider the collective behavior of a subset of edges.

3.3 Divergence

Given a graph $G = (V, E)$ with weight function $w : E \to [0, 1]$, $E' \subseteq E$, and threshold $\theta \in [0, 1]$, let $X_{E',\theta} = |\{e \in E' : w(e) \geq \theta\}|$. We define the θ-divergence of E' as follows:

$$Div_\theta(E') = \frac{1}{\mathbf{P}(X \geq X_{E',\theta})} = \left(\sum_{i=X_{E',\theta}}^{|E'|} \binom{|E'|}{i} (1-\theta)^i \theta^{|E'|-i} \right)^{-1}$$

where $X \sim Bin(|E'|, 1-\theta)$ is a Binomially distributed random variable representing the sum of Bernoulli trials, one for each edge in E'.

However, a challenge still remains in choosing the threshold value. In fact, we claim that a high-divergence edge set at any threshold is meaningful. We address this by introducing the concept of *max-divergence*, which is defined as follows:

$$Div_{max}(E') = \max_{\theta \in [0,1]} Div_\theta(E').$$

We further define the max-divergence of a vertex $v \in V$ as $Div_{max}(v) = Div_{max}(E(v))$, where $E(v)$ is the set of edges incident to v, and the *maximum vertex divergence* of a graph as $MVD(G) = \max_{v \in V(G)} Div_{max}(v)$.

Given a weighted graph $G = (V, E)$ with weight function $w : E \to [0, 1]$ and a threshold $\theta \in [0, 1]$, a *maximal θ-component* of G is a connected subgraph $C = (V', E')$, $V' \subseteq V$, $E' \subseteq E$, for which the following conditions hold: (1) $w(e) \geq \theta$, $\forall e \in E'$; and (2) $w(e) < \theta$, $\forall e \notin E'$ with at least one endpoint in V'. Throughout the paper, when θ is not specified, the term *θ-component* refers to a maximal θ-component for any θ.

We define the maximal θ-component around $v \in V$, denoted as $C_\theta(v)$, to be the maximal θ-component containing v. Let $\mathcal{C}_\theta(G)$ denote the set of all nonempty maximal θ-components of G. Note that $\mathcal{C}_\theta(G)$ naturally corresponds to a partition of V.

Just as we considered the divergence for a set of edges or a vertex, we now define the θ-divergence of a maximal θ-component C:

$$Div_\theta(C) = \frac{1}{\mathbf{P}(X \geq |E(C)|)}$$

with $X \sim Bin(n, 1-\theta)$ and $n = \sum_{v \in V(C)} deg(v) - |E(C)|$, the total number of edges incident to vertices in C. When considering a maximal θ-component C, we are typically only concerned with the divergence value at threshold θ. Consequently, we simply refer to this quantity as the divergence of C, denoted $Div(C)$. We now define the *maximum component divergence* of a graph:

$$MCD(G) = \max_{C \subseteq \mathcal{C}(G)} Div(C)$$

where $\mathcal{C}(G)$ is the set of all θ-components of G.

Divergence gives us a way to quantitatively measure the degree of anomaly of a fixed subgraph. However, what if the "suspect set" of subgraphs to monitor is not known? The number of possible subgraphs grows exponentially in the size of the graph, so monitoring even a constant fraction of them is not computationally feasible on large networks. This leads us to our last step, the MCD Algorithm, which takes as input a weighted graph $G = (V, E)$ with weight function $w : E \to [0, 1]$, and outputs a list of disjoint θ-components that partition $V(G)$, in order of decreasing divergence (see Figure 1). This is accomplished by running a variant of the Union-Find Algorithm [20], incrementally adding edges in order of decreasing weight. The details of the MCD Algorithm can be found in the full version of the paper.

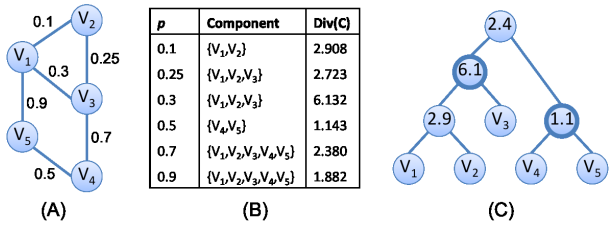

Figure 1: (a) A weighted graph, (b) the step-by-step progression of the MCD Algorithm, and (c) the MCD Tree with resulting partition highlighted.

4. EXPERIMENTS

Our experiments are designed to demonstrate the effectiveness of our approach in two ways. First, they show that modeling network edges as renewal processes captures essential temporal properties of real-world datasets. Second, they verify that the anomalous nodes and connected components identified by divergence analysis yield meaningful results that agree with human intuition, by comparing with hand-labeled anomalous activity and performing validation experiments on simulated data. Table 1 lists the datasets used in our experiments.

4.1 Modeling Inter-Arrival Times

To apply the REWARDS model, we must estimate the distribution of inter-arrival times across each network edge. Barabási et. al. suggest that inter-arrival times for communication follow a power law [3]. We found that this claim holds across all of our datasets, regardless of the communication medium. We then compared several distribution models and parameter estimation methods, and found that the Bounded Pareto Distribution, a truncated version of the common power-law distribution, using a Maximum-Likelihood approach to estimate the model parameters, consistently out-performed the other models, further corroborating Barabási's claim. (Details of experiments omitted due to space constraints.) Therefore, we use the Bounded Pareto with Maximum-Likelihood Estimation for all further experiments.

4.2 Understanding the Results

In Figure 2 we examine plots of MCD and MVD over time for the TWITTER and BLUETOOTH networks, and compare them with the total network activity (number of time-stamped events). In both datasets, the number of TSEs follows a clear weekly and daily pattern. For TWITTER, the MCD and MVD values are almost always identical, indicating that the dataset is composed of disjoint stars, so that the divergence of each component is equal to the divergence of the vertex at the center of the star. Manual inspection of the data confirms that this is indeed the case (the top MCD values correspond to spammers), and similarly for the ENRON and LBNL datasets (corresponding to multiple-recipient emails and scanning activity, respectively). On the other hand, MCD values do not seem to be heavily dependent on the number of TSEs for TWITTER or LBNL. This is likely due to the large number of TSEs in the dataset, and highlights the ability of our approach to find local regions of heightened recent activity in large networks.

Due to the nature of the BLUETOOTH dataset (measuring physical proximity), denser subgraphs are more common, so heightened activity within a component can not necessarily be attributed to a single individual. This is most noticable on Day 136, where the MCD reaches its peak, yet the MVD and number of TSEs at that time are not significantly greater than on other days or weeks.

Dataset	Description	Timespan	# of nodes	# of edges	# of TSEs
ENRON [11]	email network	2 years	1141	2017	4847
BLUETOOTH [8]	proximity of mobile devices	9 months	101	2815	102563
LBNL [14]	IP traffic	1 hour	3317	9637	9258309
TWITTER [7]	user-messages-user	1 year	262932	307816	1134722

Table 1: Datasets used in our experiments

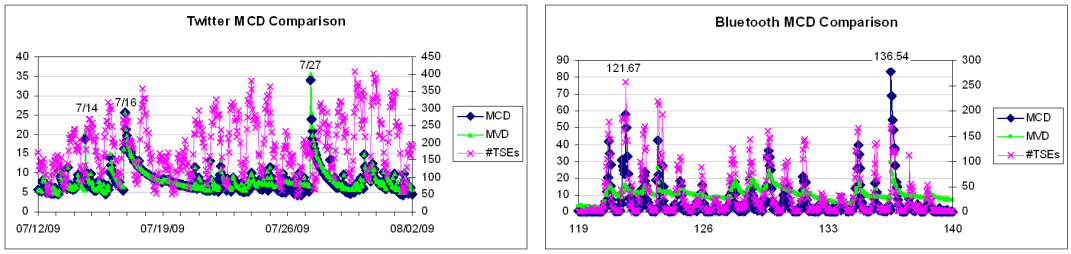

Figure 2: Plots of $\log_{10}$(MCD), $\log_{10}$(MVD), and # of TSEs over time for TWITTER and BLUETOOTH networks.

4.3 Validation of Results

Consider the following experiment on a communication network: Let G_t be the unweighted graph of edges that are alive at time t. Assign each edge e a weight $w(e)$ chosen independently and uniformly at random from [0,1]. Run the MCD Algorithm on the resulting weighted graph (G_t, w).

For each network and every time t, we ran 1000 independent trials of the above experiment, and calculated the average and standard deviation of the MCD across all trials. Figure 3 compares the results of running the MCD Algorithm on the LBNL and ENRON datasets with the results from the validation experiments. We observe that some spikes in each dataset fall many standard deviations above the mean, indicating times at which a particular subset of nodes demonstrated a strong correlation in their communication activity, far beyond what is likely to occur if edges are acting independently. Furthermore, for LBNL the MCD frequently fell well within three standard deviations of the mean, indicating that IP traffic between different pairs of nodes is typically independent, whereas the ENRON dataset shows greater dependency. This is reflective of the burstiness of human behavior [3].

4.4 LBNL Case Study

The LBNL dataset was collected by Pang et. al. [14] by monitoring network flows for IP traffic on a large enterprise network. The authors then identified several sources of "scanning activity," that is, sources contacting more than 50 distinct IP addresses in ascending or descending order, as well as two known internal scanners.

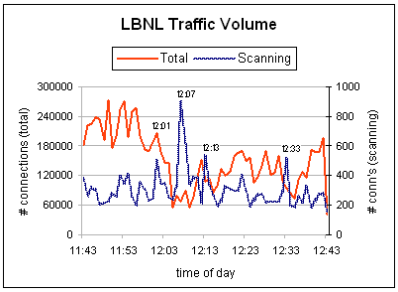

Figure 4: Volume of IP traffic per minute in the LBNL dataset. Note the different scale for volume of scanning activity.

Figure 4 shows the traffic volume per minute, and how much of that was due to scanning activity. We note that spikes in scanning activity do not necessarily coincide with spikes in total volume, and that scanning activity accounts for less than 2% of all traffic even at its peak. Next, we compare this with the MCD plot in Figure 3 , and the network visualization in Figure 5. The times of greatest MCD in the LBNL dataset are 11:58, 12:13, 12:26, 12:33, and 12:40. Times 12:13 and 12:33 stand out in terms of scanning activity, but the other times of high MCD, including the largest peak at time 12:26, were not discovered by the LBNL researchers because it did not fit their criteria for scanning activity. This demonstrates the flexibility of our approach for detecting anomalous activity based on both temporal and relational factors.

5. COMPLEXITY ANALYSIS

Our REWARDS model requires only $O(m)$ space, with constant-time updates (per communication event). The MCD Algorithm runs in $O(n + m \log m)$ time, and $O(n + m)$ time in practice. Detailed analysis can be found in the full version of the paper.

6. QUESTIONS AND FUTURE DIRECTIONS

Until recently our work has centered on anomaly detection in the communication domain, where traffic or message contents may be encrypted or unavailable. Our current interest is to apply our techniques in the context of web search in information networks, where semantic information is readily available. Namely, how do we identify emerging trends and patterns of information diffusion in a streaming fashion? How do we find information that is both time-relevant and of likely interest to an individual? We suggest that our REWARDS model can be a good first step in this direction.

7. REFERENCES

[1] J. Abello, T. Eliassi-Rad, and N. Devanur. Detecting novel discrepancies in communication networks. In *10th IEEE International Conference on Data Mining*, pages 8–17, 2010.

[2] L. Akoglu, M. McGlohon, and C. Faloutsos. Oddball: Spotting anomalies in weighted graphs. In M. Zaki, J. Yu, B. Ravindran, and V. Pudi, editors, *Adv. in Knowledge Disc. and Data Mining*, volume 6119 of *Lecture Notes in Comp. Sci.*, pages 410–421. Springer Berlin / Heidelberg, 2010.

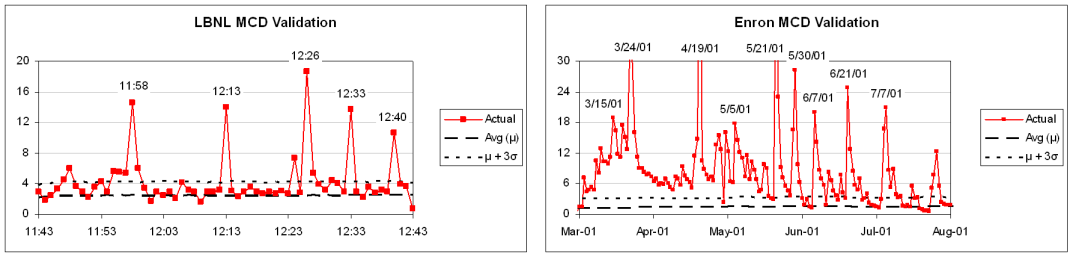

Figure 3: Plots of $\log_{10}$(MCD) over time for LBNL and ENRON networks, along with validation results.

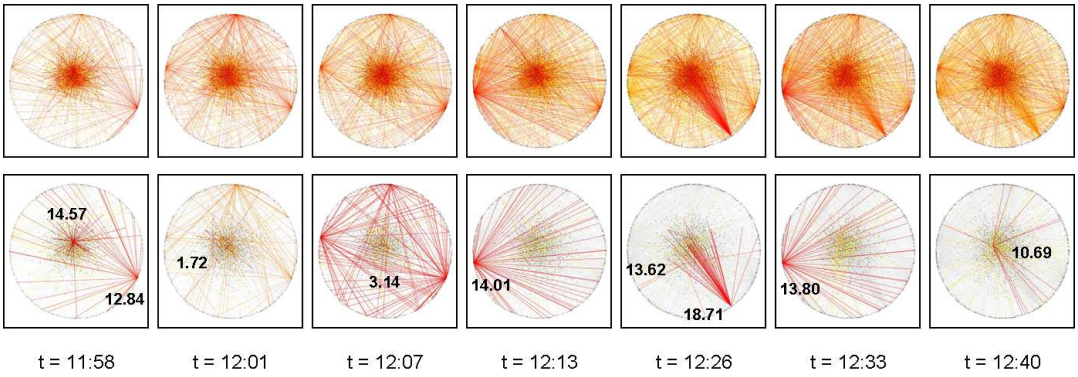

Figure 5: Graph of live edges in LBNL network (top); θ-components with highest divergence output by MCD Algorithm (bottom).

[3] A.-L. Barabási. The origin of bursts and heavy tails in human dynamics. *Nature*, 435(7039):207–211, May 2005.

[4] J. Candia, M. C. González, P. Wang, T. Schoenharl, G. Madey, and A.-L. Barabási. Uncovering individual and collective human dynamics from mobile phone records. *J. Physics A: Mathematical and Theoretical*, 41, June 2008.

[5] J. Cao, A. Chen, T. Bu, and A. Buvaneswari. Monitoring time-varying network streams using state-space models. In *INFOCOM 2009, IEEE*, pages 2721 –2725, 2009.

[6] V. Chandola, A. Banerjee, and V. Kumar. Anomaly detection: A survey. *ACM Comput. Surv.*, 41(3), 2009.

[7] M. D. Choudhury. How birds of a feather flock together on online social spaces. In *Grace Hopper Celebration of Women in Computing*, Atlanta, GA, USA, 2010.

[8] N. Eagle, A. S. Pentland, and D. Lazer. Inferring friendship network structure by using mobile phone data. *Proc. of the Nat'l Academy of Sciences*, 106:15274–15278, Aug. 2009.

[9] K. Henderson, T. Eliassi-Rad, C. Faloutsos, L. Akoglu, L. Li, K. Maruhashi, B. A. Prakash, and H. Tong. Metric forensics: a multi-level approach for mining volatile graphs. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '10, pages 163–172, New York, NY, USA, 2010. ACM.

[10] A. Ihler, J. Hutchins, and P. Smyth. Adaptive event detection with time-varying poisson processes. In *Proc. of 12th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining*, KDD '06, pages 207–216, 2006.

[11] B. Klimt and Y. Yang. The enron corpus: A new dataset for email classification research. In J.-F. Boulicaut, F. Esposito, F. Giannotti, and D. Pedreschi, editors, *Machine Learning: ECML 2004*, volume 3201 of *Lecture Notes in Computer Science*, pages 217–226. Springer Berlin / Heidelberg, 2004.

[12] A. Lakhina, M. Crovella, and C. Diot. Characterization of network-wide anomalies in traffic flows. In *Proc. of the 4th ACM SIGCOMM Conf. on Internet Measurement*, IMC '04, pages 201–206, 2004.

[13] C. C. Noble and D. J. Cook. Graph-based anomaly detection. In *Proc. of the 9th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining*, KDD '03, pages 631–636, New York, NY, USA, 2003. ACM.

[14] R. Pang, M. Allman, V. Paxson, and J. Lee. The devil and packet trace anonymization. *SIGCOMM Comput. Commun. Rev.*, 36:29–38, January 2006.

[15] C. E. Priebe, J. M. Conroy, D. J. Marchette, and Y. Park. Scan statistics on enron graphs. *Computational and Mathematical Organization Theory*, 11:229–247, Oct. 2005.

[16] S. M. Ross. *Renewal Theory and Its Apps.* Acad. Press, 2010.

[17] J. Sun, C. Faloutsos, S. Papadimitriou, and P. S. Yu. Graphscope: parameter-free mining of large time-evolving graphs. In *Proceedings of the 13th ACM SIGKDD Int'l Conference on Knowledge Discovery and Data Mining*, KDD '07, pages 687–696, 2007.

[18] J. Sun, H. Qu, D. Chakrabarti, and C. Faloutsos. Neighborhood formation and anomaly detection in bipartite graphs. In *5th IEEE Int'l Conf. on Data Mining*, pages 418–425, 2005.

[19] J. Tang, M. Musolesi, C. Mascolo, and V. Latora. Temporal distance metrics for social network analysis. In *Proceedings of the 2nd ACM workshop on Online social networks*, WOSN '09, pages 31–36, New York, NY, USA, 2009. ACM.

[20] R. E. Tarjan. Efficiency of a good but not linear set union algorithm. *J. ACM*, 22:215–225, April 1975.

[21] Y. Xie, V. Sekar, D. A. Maltz, M. K. Reiter, and H. Zhang. Worm origin identification using random moonwalks. In *Proc. of the 2005 IEEE Symposium on Security and Privacy*, pages 242–256, 2005.