International Journal of Data Warehousing and Mining
Volume 15 • Issue 2 • April-June 2019

Change Detection in Large Evolving Networks

Josephine M. Namayanja, University of Massachusetts Boston, Boston, USA
Vandana P. Janeja, University of Maryland Baltimore County, Baltimore, USA

ABSTRACT

This article presents a novel technique for the detection of change in massive evolving communication networks. This approach utilizes a novel hybrid sampling methodology to select central nodes and key subgraphs from networks over time. The objective is to select and utilize a much smaller targeted sample of the network, represented as a graph, without loss of any knowledge derived from graph properties as compared to the entire massive graph. This article uses the targeted samples to detect micro- and macro-level changes in the network. This approach can be potentially useful in the domain of cybersecurity where this article highlights the importance of graph sampling and multi-level change detection in identifying network changes that may be difficult to detect on a larger scale. This article therefore presents a means to audit large networks to establish continuous awareness of network behavior.

KEYWORDS

Big Data, Central Nodes, Cyber Threats, Hybrid Sampling, Key Subgraphs, Macro-Level Changes, Micro-Level Changes, Temporal Binning

INTRODUCTION

A network is described as a set of interconnected nodes, such as a computer network, social network, communication network (Sun, Faloutsos, Papadimitriou & Yu, 2007), to mention a few. This study focuses on modelling the relationship between the network points represented as a graph of the communication on a computer network over a period of time in terms of connectivity, unlike network traffic measurement in the form of packets or their size (bytes). Given that data in computer networks consists of billions of nodes, which are communicating with each other as indicated by the edges, makes it massive. When we consider such data over a period of time, this data becomes even more massive. As a result, the massive size of such network structures makes them vulnerable to various cyberattacks such as Advanced Persistent Threats (APTs) that can be entrenched into the network and stay undetected for long periods of time. Our study therefore leverages from the context of attacks that prevent the normal use of a computer and may cause it and any affiliated resources from being reliable, available or accessible. Furthermore, in real-world networks, large volumes of network traffic pose a challenge in efficiently monitoring them for attack detection and thus creates a need to characterize network behavior to create awareness in order to determine potential vulnerabilities.

The process of change detection in such large evolving networks can be used to establish awareness on a network by monitoring the network to detect shifts (McCulloh, 2009). While changes do occur as graphs evolve over time, certain changes are more significant than others, creating the task to make sense of changes that take place. In the case of massive computer networks that are vulnerable

DOI: 10.4018/IJDWM.2019040104

to various cyberattacks, change detection can be used as a trigger for further investigation into the network to determine if changes in the network are associated to a previous, present or potential cyber threat. Therefore, our objective is to conduct a selective analysis of massive network structures in order to mine this massive data efficiently and to accurately make sense out of the knowledge discovered.

Our approach is multifaceted whereby we apply graph sampling to identify and select representative subsets of the network. We consider this as a strategic sampling approach in which we utilize a hybrid methodology that combines sampling, clustering and stratified binning to select key nodes, namely central nodes and key subgraphs associated to the central nodes from a network over time as presented in our preliminary work in (Namayanja & Janeja, 2013, 2014, 2015). We consider a central node to represent a key (critical) node on the network such as a server. Such central nodes are considered important points on a network based on their role in the network that can be defined in terms of their centrality such as degree centrality and betweenness centrality (Freeman, 1979), eigenvector centrality (Bonacich, 1987), PageRank (Page, Brin, Motwani, Winograd, 1999) among others. Determining the role of a node in a network can be useful in threat detection (Scripps, Tan & Esfahanian, 2007) and according to (Shen, Nguyen, Xuan & Thai, 2012), an assessment of network vulnerabilities indicates that an attacker is likely to exploit the weak points such as critical nodes whose corruption greatly affects network performance.

Based on this, we utilize our targeted samples to detect multi-level changes in the network over time. First, we detect micro-level changes which are shifts in the presence and centrality of central nodes over time to determine Consistent and Inconsistent (CoIn) central nodes as well as times when changes in node behavior are significant referred to as CoIn Time Periods of Change (CoIn-TPC). While micro-level changes can be used to characterize the behavior of such critical nodes on a network, they do not relay information about the bigger picture in the overall network. Therefore, we also detect macro-level changes, where we detect shifts in graph level properties in key subgraphs to determine Network Level CoIn (NL-CoIn) Change Points. These macro-level changes are times when the fundamental structural or network level properties significantly change as a result of changes in the behavior of central nodes. Our multi-level change detection approach also aims to determine similarities and correlations between micro- and macro-level changes in the network over time. Specifically, this paper makes the following contributions:

- We present a novel method that utilizes hybrid sampling to select top central nodes and key subgraphs from graphs representing large evolving networks. This method aims to illustrate the importance of utilizing strategic sampling in large evolving networks;
- We illustrate empirically that the key subgraphs selected follow the properties highlighted in the literature for full graphs;
- We present extensive experimental and comparative results in temporally evolving networks using real-world internet traces from Center for Applied Internet Data Analysis (CAIDA) (Anonymized Internet Traces 2008, 2009, 2010; Network Telescope "Three Days Of Conficker"). This method aims to identify correlations between micro- and macro-level changes in large evolving networks.

We thus extend our approach presented in our preliminary work to multiple real-world network traffic datasets, evaluation of our strategic sampling method, analyzing both directed and undirected networks, testing multiple centrality measures as well as comparing our approach to existing change detection approaches. While this study has been conducted extensively, we limit the discussion of our findings in this paper due to space constraints and to avoid redundancy. The rest of this paper is organized as follows: In section 2, we discuss related work. In section 3, we present our methodology. In section 4, we present experimental results and discussion of findings. Lastly, in section 5, we present our conclusions and future work.

RELATED WORK

Change Detection in Network Structures: The process of change detection in network analysis have been widely explored in relation to the use of Statistical Process Control (SPC) charts specifically, Sequential Probability Ratio Test (SPRT), Cumulative Sum (CUSUM), Exponentially Weighted Moving Average (EWMA), and the Shiryaev–Roberts (SR) procedure (McCulloh, 2009; Tartakovsky, Polunchenko & Sokolov, 2013). However, SPC techniques assume that the data is sequential or time-ordered (Slavin, 2006) and are therefore not suitable for analysis of network behavior from other dimensions such as detecting shifts in node behavior whose evaluation in comparison to other nodes maybe based on a non-sequential process. Additionally, change detection approaches in networks have been limited to detect changes at the structural level by observing network level properties like density, diameter or average node centrality (Namayanja & Janeja, 2015; McCulloh, 2009; Gaston, Kraetzl & Wallis, 2006; Akoglu & Faloutsos, 2013; Opsahl, Agneessens, Skvoretz, 2010). On the otherhand, (Namayanja & Janeja, 2013, 2014) study changes in the network by evaluating the behavior of key nodes in the network in terms of changes node centrality. In order to evaluate the overall structure of the network, these node level centrality measures can also be translated into network level metrics by averaging them out (McCulloh, 2009). This study on the otherhand extends (Namayanja & Janeja, 2013, 2014, 2015) by evaluating the level of change in node behavior over time and correlating changes at the node and network level.

Application of Sampling Techniques in Large Graphs: (Krishnamurthy et al., 2005) and (Leskovec & Faloutsos, 2006) propose sampling algorithms associated to random node selection, random edge selection, and random deletions. (Vitter, 1985) also proposes a reservoir approach to select random samples where it is not possible to determine the actual size of the data. However, the utilization of certain random sampling techniques such as random deletions may result into the elimination of key network components or may not capture sub components that are representative of the full graph. Interestingly, (Namayanja & Janeja, 2013, 2014, 2015) propose a strategic sampling method that focuses on key nodes and subgraphs in the network to detect changes in the network over time. However, these studies do not justify the reliability of the samples selected with respect to the overall graph as discussed in (Leskovec & Faloutsos, 2006; Leskovec, Kleinberg & Faloutsos, 2007). (Namayanja & Janeja, 2013, 2014, 2015) are also limited to analyzing samples of a network and thus do not justify the importance of the targeted samples. This study extends (Namayanja & Janeja, 2013, 2014, 2015) and aims to clearly demonstrate the reliability of samples selected by evaluating fundamental graph properties related to growth of the network over time. Additionally, this study also justifies the need for sampling large graphs in the process of change detection by comparing findings to a non-sampled approach. Next, we discuss our approach.

METHODOLOGY

Our approach utilizes graph sampling to select targeted samples from a network represented as a graph. We then use the selected samples to detect changes in the network as described in our detailed approach that follows.

Graph Sampling

G*iven a graph* $G = (N, E)$, $N = \{n_1 \ldots n_v\}$ *is a set of nodes and* $E = \{e_1 \ldots e_w\}$ *is a set of edges, such that* (n_i, n_j) *is an edge between nodes* n_i *and* n_j. A graph can be directed or undirected where the former identifies the direction between edges and the latter as otherwise.

First, we take random temporal network snapshots of selected times $t_a \ldots t_s$. Each temporal snapshot represents a sample size of 25% per time slice. Interestingly taking as low as 15% of a large graph in its current state can be used to capture the smaller structure of the same graph during its previous state (Leskovec & Faloutsos, 2006). However, in this study, we increase the sample size to capture

a larger portion of the network. Because of the dynamic nature of graphs, the 25% sample size may vary for different temporal snapshots. In the future, we plan to test our approach on varying sample sizes to determine the optimal minimum sample size for such large networks and to support efficient processing at this phase. We also want to determine if there are any similarities and/or variations in the changes detected across varying sample sizes.

We then utilize k-means clustering to group nodes from each snapshot into clusters. We apply k-means clustering on the centrality values associated to each node using Ideal k where k is the number of clusters (Namayanja & Janeja, 2011). Based on our multiple set of experiments, it is determined that Ideal k ranges from 2 – 3 clusters due to high disparity in the centrality values of nodes which results into skewness of clustering. Hence, we use clustering to identify such a small grouping of high centrality values and then match the nodes that correspond to those values. Such a cluster consisting of the identified nodes is referred to as a target cluster. Next we select the clusters of nodes with the highest centrality level. We then combine the nodes from each targeted cluster which we refer to as a set of central nodes $C = \{c1 \ldots co\}$ where o is the total number of central nodes. Thus, a central node is a key node with a high centrality on the network. For this study we consider the following centrality measures, degree centrality, betweenness centrality, eigenvector centrality and PageRank respectively, although other centrality measures may be used. We apply different centrality measures to broaden our selection pool for selecting central nodes. Interestingly, the top nodes discovered across centrality measures overlap, which can be attributed to the correlation between different centrality measures as discussed in (Friedkin, 1991; Lee, 2006). Overlaps also occur across temporal snapshots with respect to the datasets utilized in this study.

Then using equal frequency binning, we select a set of temporal bins $B = \{b_1 \ldots b_x\}$, where each temporal bin b_i consists of z continuous time periods. In this study, we use network traffic provided by CAIDA where we consider a temporal bin as a month and for each month we select 10 minutes mainly because of the massive volume of the network traffic. In the future, we plan to test our approach on varying number of minutes to detect variations in changes based on a sliding-scale window approach. For each temporal bin b_i, we select o central nodes and their adjacent edges to form a set of subgraphs which we refer to as key subgraphs or hybrid samples. This is a guided approach to stratified sampling where we select the key subgraphs associated to central nodes only. The result is a set of (x * z) key subgraphs. It should be noted that in each temporal bin, we select the same central nodes. However, the edges to these central nodes may vary for each temporal bin since the network organically grows or shrinks over time which thus accounts for the dynamic structure of the network over time. Such variations are key to determining the changes in the network over time.

As described in our graph sampling approach, sampling of a graph means taking subsets of it. While sampling of large graphs is essential according to (Krishnamurthy et al., 2005; Leskovec et al., 2007), the goal is to maintain the structural properties of the graph, such as the densification power law to be considered representative of the full graph. For instance, (Leskovec et al., 2007) clearly demonstrate that graphs obey the densification power law where edges grow faster than nodes. First, the graph over time maintains a power law degree distribution with a constant exponent γ. If $\gamma < 2$ and is constant over time, then the graph is said to densify. Therefore, using these properties, we would also like to illustrate how our subgraphs are representative of the full graph by showing the behavior B of samples; 1) if the samples exhibit the power law degree distribution exponent γ and 2) if the densification exponent α is reflected as a function of γ. Effectively if 1 and 2 are satisfied, then the behavior H of central nodes emulates behavior H of the graph, that is $H[C] \approx H[N]$ where H[C] is the behavior of central nodes and H[N] is behavior of the graph. Based on the degree and densification exponents γ and α respectively, our goal is to determine if the key subgraphs are representative of the entire graph and can be useful in the process of change detection. In this paper, we demonstrate the importance and validity of strategic sampling in large evolving networks using the graph properties defined. An evaluation of our sampling approach is provided in our results section.

Multi-Level Change Detection

Using the selected central nodes and key subgraphs, we follow up their behavior to detect; i) changes with respect to structural characteristics in the behavior of central nodes (micro-level) in the graph over time and ii) changes with respect to structural characteristics in the behavior of fundamental graph properties (macro- level) in the network over time. We refer to this framework as multi-level change detection.

Micro-Level Change Detection

For micro-level changes in the graph we observe the presence and centrality of each central node respectively and in comparison to other central nodes across temporal bins to identify steady or changing nodes. We refer to these as Consistent or Inconsistent (CoIn) central nodes respectively where a Consistent central node is one whose presence or centrality meets a user defined threshold and is thus stable over time, while an Inconsistent central node is one whose presence or centrality does not meet a user defined threshold and thus fluctuates or changes over time. Additionally, we identify those temporal bins or times significantly associated to inconsistency or change in presence or centrality respectively of central nodes. We refer to these as CoIn Time Periods of Change (CoIn-TPC). We evaluate CoIn-TPC using ground truth based on examples of times when real-world cyber events occurred as described in table 2 and 3 in the results section. Our objective is to determine if an association exists between changes in key network components and big cyber events. It should be noted that our findings are only used to indicate an association of changes in the network to cyber events and does not justify changes discovered as a cause of the cyber events. We also compare CoIn-TPC against CUSUM and EWMA (based on 0.2 and 0.3 weighting factors for EWMA respectively). Both techniques have built-in change detection capabilities that support time-dependent data.

Next, we discuss how we detect changes at the network level based on changes in CoIn central nodes.

Macro-Level Change Detection

Now, given CoIn central nodes, we follow-up on the behavior at the network level by evaluating network level properties particularly density, average degree of all nodes and diameter to detect change points at the network level which we refer to as Network Level CoIn (NL-CoIn) change points and thus macro-level change detection. Essentially, each key subgraph becomes a CoIn Subgraph. Therefore, for each CoIn subgraph we determine the following; density as NL-CoIn-DEN, average degree of all nodes as NL-CoIn-DEG and diameter as NL-CoIn-DIAM respectively. Hence, like CoIn-TPC, we identify time points where graph properties deviate from other time points and we evaluate NL-CoIn change points. We also evaluate using ground truth and compare our approach against CUSUM and EWMA. Next we present our experimental results.

EXPERIMENTAL RESULTS

Dataset Description

For our experiments, we use real-world internet traffic datasets from the Center for Applied Internet Data Analysis (CAIDA) (Anonymized Internet Traces 2008, 2009, 2010, Network Telescope "Three Days Of Conficker"). Here we select internet traffic associated to different time periods and scenarios. According to this study, the datasets used are as follows:

- **CAIDA 2008:** Anonymized Internet Traces 2008 (Anonymized Internet Traces 2008);
- **CAIDA 2008-2010:** A combination of Anonymized Internet Traces 2008, Anonymized Internet Traces 2009 and Anonymized Internet Traces 2010 (Anonymized Internet Traces 2008, 2009, 2010);

- **CAIDA Conficker - Telescope:** Three Days of Conficker (Network Telescope "Three Days Of Conficker").

We also selected these specific datasets as an example for this study because they represent large amounts of network traffic data captured over an extensive period of time, but most importantly they are also aligned with network traffic captured during an on-going cyberattack, specifically the Conficker Worm that started in 2008 until 2009 [12]. Hence, we feel that this data was ideal for our experiments in change detection with respect to an example related to big cyber events. However, for future studies, we plan to compare our approach using more recent datasets that capture large network traffic and cyber events that occur over an extended period of time such as the Conficker Worm. Table 1 provides a summary of each dataset as used in this study.

It should be noted that the top CoIn central nodes overlap across the directed and undirected network structures.

Experimental Setup

The experiments were conducted on a machine with 3.4GHz Intel Core i7-3770, 4 cores, 16GB RAM, 3401 MHz, and running Microsoft Windows Professional version 6.1.7601 Service Pack 1 Build 7601. Given that our study deals with large datasets, our experiments were also conducted using a high-performance cluster, specifically the Maya cluster which is managed by the UMBC High Performance Computing Facility (HPCF) (High Performance Computing Facility - UMBC). The Maya cluster is a heterogeneous cluster with equipment acq uired between 2009 and 2013. During the time of this study, the cluster contained a total of 324 nodes, 38 GPUs and 38Intel Phi co-processors, and over 8 TB of main memory. All nodes run Red Hat Enterprise Linux 6.4. It should be noted that the Maya cluster has since been updated. In the selected parallel processing environment, we utilize MapReduce (Elser & Montresor, 2013) to support our methods in graph sampling and change detection that require intensive processing. Algorithms for graph analysis were implemented using Java version 6. Additionally, for clustering, we used Weka 3.6.10 (Hall et al., 2009).

For our analysis, we present and discuss results on the following:

1. Micro-Level Change Detection;
2. Macro-Level Change Detection;

Table 1. Summary of datasets

	Dataset Time Period	Total Number of Edges	Number of Temporal Bins	Number of Central Nodes – Directed Network	Number of Central Nodes – Undirected Network
CAIDA 2008: Anonymized Internet Traces 2008	**March 2008 – December 2008**	**55,634,738**	**10**	**15**	**21**
CAIDA 2008-2010: Anonymized Internet Traces 2008, 2009, 2010	**December 2008 – January 2010**	**67,383,262**	**14**	**12**	**15**
CAIDA Conficker: Network Telescope: Three Days of Conficker	**November 2008 – January 2009**	**42,680,485**	**3**	**7**	**7**

3. Similarity Analysis for Micro-Level and Macro-Level changes;
4. Correlation Analysis for Micro-Level and Macro-Level changes;
5. Evaluation of Hybrid Sampling.

Micro-Level Change Detection

For our micro-level analysis, we present findings on Consistent and Inconsistent (CoIn) central nodes and Time Periods of Change due to CoIn Central Nodes (CoIn-TPC).

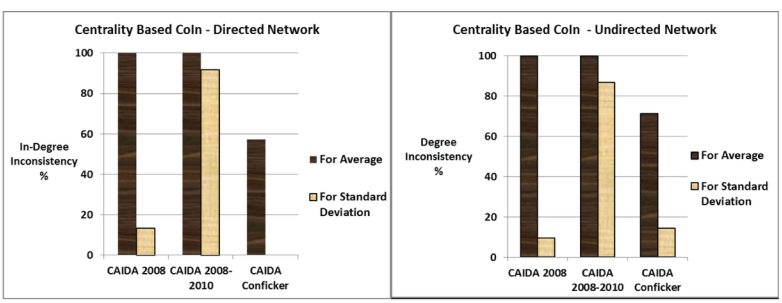
Evaluation of Consistent and Inconsistent (CoIn) Central Nodes Using CoIn Threshold Approaches

We focus on inconsistency because it represents change in behavior of central nodes over time. Although it should be noted that the inconsistency of central nodes is the inverse of consistency levels.

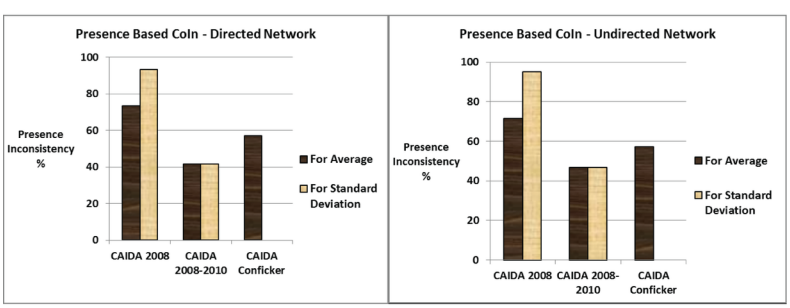
For centrally-based CoIn, Figure 1(a) show higher inconsistency levels for average compared to standard deviation. Similar findings were also observed in other centrality properties, although we neglect to discuss those here to avoid redundancy. With the average behavior, there is an expected level of how central nodes should behave over time such that anything below that expected level signifies a change in behavior. For example, top firms such as banks or telecom companies whose operations run on critical systems are expected to have their systems efficiently available except in planned or routine cases.

Alternatively, in 1(b) changes in presence-based CoIn vary for the different thresholds across datasets. Thus, centrality of central nodes is a better indicator of change in the behavior of central nodes over time. Our findings confirm that detecting change in the behavior of central nodes over time benefits from analyzing multiple perspectives, such as node availability and node connectivity metrics in order to capture changes.

Figure 1. (a) Inconsistency levels for centrality based CoIn; (b) Inconsistency levels for presence based CoIn



(a)



(b)

Assessment of CoIn Levels Based on Varying Thresholds

We assess the level of consistency and inconsistency ranging from 25% to 100% respectively for central nodes over time. Only a fraction of our results is presented to avoid redundancy.

Figure 2(a) and (b), indicate that as the threshold level reduces, the level of inconsistency of central nodes either stabilizes or decreases at 75% of the time. In other words, atleast 75% of the time the behavior of central nodes in the network remains stable or meets its expected level in this case. This can be used heuristically to state at what point in time, changes in the behavior of central nodes should be alarming.

It should be taken into account that we do not have any ground truth to evaluate the actual node level changes with respect to these datasets. Therefore, we evaluate the time periods associated to changes in node behavior in association to examples of cyberattacks that have already taken place.

Ground Truth Evaluation of CoIn Time Periods of Change (CoIn-TPC)

In Tables 2 and 3, we provide a summary of real-world events relevant to datasets CAIDA 2008-2010 and CAIDA Conficker respectively although a summary of events for CAIDA 2008 are also available but not presented here. Each summary of events is derived from internet security reports by Akamai Technologies (State of the Internet Connectivity Reports – 2008, 2009, 2010) which interestingly are also associated to the CAIDA internet traffic analysis. Evaluation is based on accuracy, precision and recall. Any blank value in the graphs indicates zero accuracy, precision and recall respectively. The datasets used here are not labelled, however, labels are created based on real-world events, which is helpful in real-world networks whose state is never definite.

Figure 3 indicate that CoIn-TPC generally performs better than CUSUM and EWMA. Particularly, we are able to detect time periods of change associated to the onset of big attacks such as Conficker Worm that started in November 2008. We are also able to detect when the attack intensifies during February 2009, especially when it is not masked by another non-attack event such as the Inauguration of President Barack Obama that occurred in January 2009. Similar findings are also reflected in directed networks.

Macro-Level Change Detection

Here we identify and evaluate change points associated to densification and diameter of the network (NL-CoIn) as a result of changes in the behavior of CoIn Central Nodes.

Ground Truth Evaluation of Network Level CoIn (NL-CoIn) Change Points

In order to avoid redundancy, we only present findings based on the analysis of undirected networks per dataset.

Figure 2. (a) Inconsistency levels for Centrality based CoIn in undirected networks using threshold adjustment; (b) Inconsistency levels for presence based CoIn in undirected networks using threshold adjustment

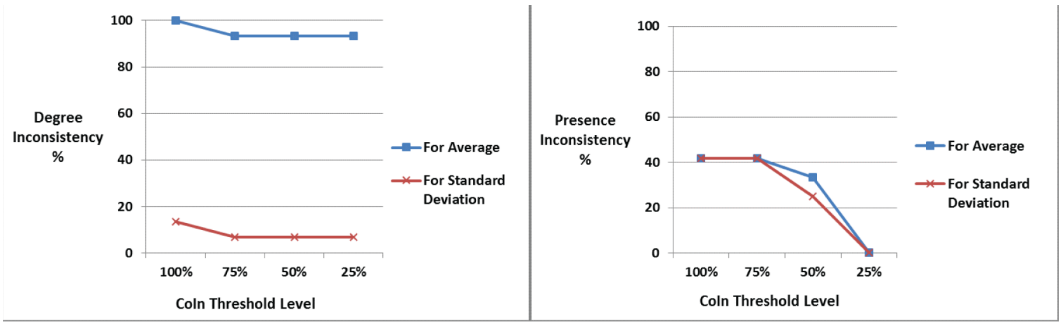


Table 2. Summary of real-world cyber events for CAIDA 2008-2010

MONTH	EVENTS
DECEMBER, 2008	- High attack traffic due to Conficker Worm. - High traffic levels and application problems combined to take various U.S. Postal Service Web-based services offline.
JANUARY, 2009	- Monthly peaks in observed attack traffic roughly correlated with the publication of Microsoft Security Bulletins. - Inauguration of US President Barack Obama increased levels of traffic, but at the same time impacted usage of online retail, messaging and search sites. - Infected machines with the Conficker Worm grew from 2.4 million to over 8.9 million in just four days. - Large-scale attacks on DNS servers that impacted thousands of websites. - DDOS attack on servers that affected Telecommunication, Media & Production, Data Centers, Gmail Services.
FEBRUARY, 2009	- Monthly peaks in observed attack traffic roughly correlated with the publication of Microsoft Security Bulletins. - Time Warner Cable's DNS servers were targeted by a DDoS attack. - DDOS attack on servers that affected Telecommunication, Media & Production, Data Centers, Gmail Services.
MARCH, 2009	- Monthly peaks in observed attack traffic roughly correlated with the publication of Microsoft Security Bulletins. - Decrease in attack traffic related to Conficker Worm because the authors of the Conficker Worm were trying to protect already infected machines from anti-virus software instead of trying to infect additional system machines. Also attack traffic on port 445 decreased. - DDOS attack on servers that affected Telecommunication, Media & Production, Data Centers, Gmail Services.
APRIL, 2009	- Attack traffic peaks in April before and after Microsoft Security Bulletin. - Conficker Worm updated itself on already infected machines. - Network outages at AT &T in California. - Ddos Attacks on Twitter, Yahoo, Coca-Cola, Microsoft, and Google sites
MAY, 2009	- Attack traffic peaks before and after Microsoft Security Bulletin. - Website outages for Google, Gmail, e.t.c.
JUNE, 2009	- Complete Outage of Internet Traffic from Iran.
JULY, 2009	
AUGUST, 2009	
SEPTEMBER, 2009	
OCTOBER, 2009	
NOVEMBER, 2009	
DECEMBER, 2009	
JANUARY, 2010	

(Key Time Periods of Changes detected: December 2008, February 2009)

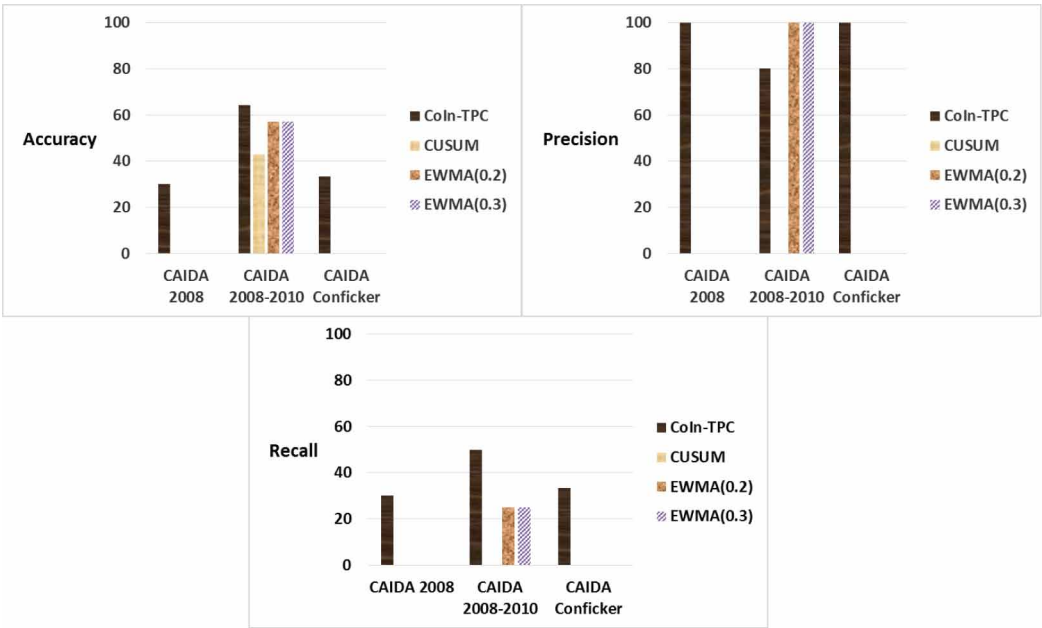
Table 3. Summary of real-world cyber events for CAIDA Conficker

MONTH	EVENTS
NOVEMBER, 2008	- Conficker Worm Onset which resulted into High attack traffic. - Unavailability of Dr. Pepper's website, due to a surge of traffic related to a one-day marketing promotion that promised everyone in America a free bottle of soda if the new Guns n' Roses album was released during 2008. - A "catastrophic" UPS failure caused a power outage at a Santa Clara data center operated by Quality Technology Services, triggering days of performance problems for the social network Friendster.
DECEMBER, 2008	- High attack traffic due to Conficker Worm. - Equipment failure caused a complete loss of Internet connectivity for two of the three ISPs in Haiti that connect to the Internet through the ARCOS submarine cable. - Over a million subscribers of Time Warner Cable saw their Internet connectivity impaired for approximately two-and-a-half hours due to the failure of Time Warner's DNS servers. - High traffic levels and application problems combined to take various U.S. Postal Service Web-based services offline.
JANUARY, 2009	- Monthly peaks in observed attack traffic roughly correlated with the publication of Microsoft Security Bulletins. - Inauguration of US President Barack Obama increased levels of traffic, but at the same time impacted usage of online retail, messaging and search sites. - Infected machines with the Conficker Worm grew from 2.4 million to over 8.9 million in just four days. - Large-scale attacks on DNS servers that impacted thousands of websites. - DDOS attack on servers that affected Telecommunication, Media & Production, Data Centers, Gmail Services.

(Key Time Periods of Changes detected: November 2008)

(Note: For the period of July 2009 to January 2010, no significant cyber threats were reported by (State of the Internet Connectivity Reports – 2008, 2009, 2010)).

Figure 3. Accuracy, precision and recall for CoIn-TPC, CUSUM and EWMA in undirected networks



Like CoIn-TPC, figure 4(a), 4(b) and 4(c) indicate that NL-CoIn generally performs better than CUSUM and EWMA for NL-CoIn-DIAM and NL-CoIn-DEG. This signifies that our method can be potentially useful in detecting critical cyber threats that may not be captured by existing change detection approaches.

Similarity Analysis for Micro-Level and Macro-Level Changes

We compare our findings for CoIn-TPC to NL-CoIn using jaccard coefficient. Jaccard Coefficient is ideal because it identifies change points that overlap across the micro-level and macro-level structure of the network.

Figures 5(a) and 5(b) indicate similarities between change points detected in CoIn-TPC and NL-CoIn, particularly in relation to diameter (NL-CoIn-DIAM) and average degree (NL-CoIn-DEG) whose findings are presented here. In this case, changes in centrality and presence of central nodes impact the network's overall connectivity.

We can also confidently state that if critical points on a network are compromised then the entire network is in jeopardy. More so, network diameter can be used to determine how the absence of key points that act as mediator points in the network impacts overall communication within the network. It can also be used as a good predictor for analyzing network level behavior over time. On the other hand, because our findings for density (NL-CoIn-DEN) do not indicate any significant similarities to CoIn-TPC, we neglect to portray them. Here micro-level changes in the network are not reflected in the density of the network.

Significant levels of similarity are observed between CoIn-TPC and NL-CoIn for centrality-based CoIn in CAIDA Conficker which captures a long-term cyberattack. Our findings confirm that analyzing change in network behavior from multiple perspectives, is beneficial in order to capture changes that otherwise cannot be detected.

Figure 4. (a) Accuracy, precision and recall for NL-Coln-DEN, CUSUM and EWMA in undirected networks; (b) Accuracy, precision and recall for NL-Coln-DIAM, CUSUM and EWMA in undirected networks; (c) Accuracy, precision and recall for NL-Coln-DEG, CUSUM and EWMA in undirected networks

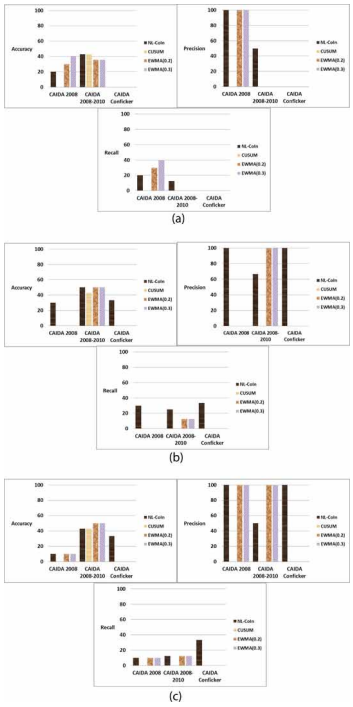
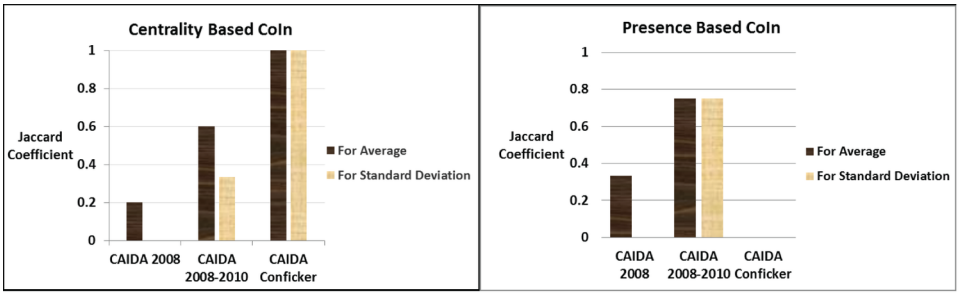
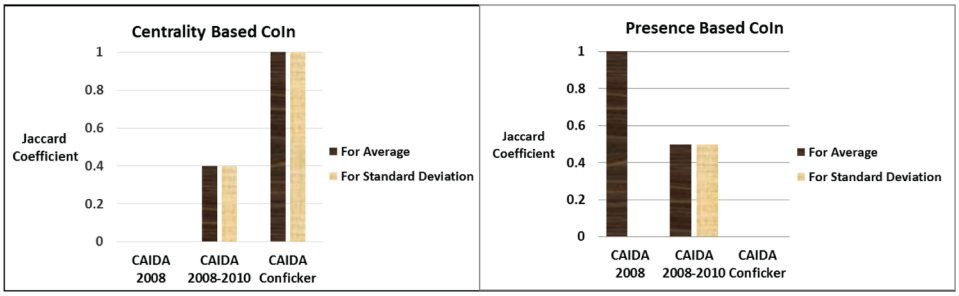


Figure 5. (a) Jaccard similarity between NL-Coln-DIAM and Coln-TPC in undirected networks; (b) Jaccard similarity between NL-Coln-DEG and Coln-TPC in undirected networks



(a)



(b)

Correlation Analysis for Micro-Level and Macro-Level Changes

Using Pearson's correlation we observe the relationship between the node inconsistency and the network level properties. Similar findings are noted in directed network structures.

Figure 6 shows that there is a high positive correlation between macro level and micro level changes particularly for diameter (NL-CoIn-DIAM) whereby an increase in inconsistency of central nodes results into an increase in the diameter, while a decrease in inconsistency of central nodes results into a decrease in the diameter. Similarly, a positive correlation is observed between average degree of nodes (NL-CoIn- DEG) and the inconsistency of central nodes.

On the other hand, we observe both minimal to zero correlation between the inconsistency of central nodes and changes in network density (NL-CoIn-DEN) which confirms our findings of minimal to zero levels of similarity between CoIn-TPC and NL-CoIn-DEN.

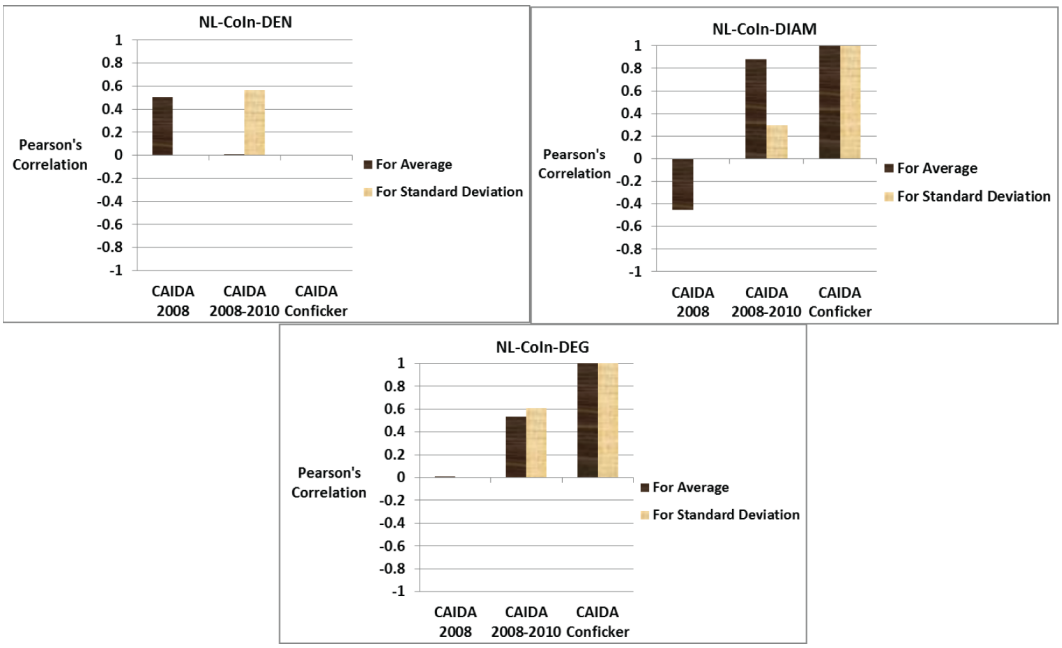
Evaluation of Hybrid Sampling

Given that the focus of this study lies on strategic sampling of large evolving networks, we also evaluate our hybrid sampling approach.

Comparison of Network Level CoIn (NL-CoIn) Change Point Detection for CoIn Subgraphs and Full Graphs

Here we extend our analysis from CoIn subgraphs to the full graphs by not applying any sampling on the graph. Our findings in figure 7 portray that the detection of changes in CoIn subgraphs indicates higher accuracy, precision and recall compared to full graphs and is thus presents a potentially useful way to drill down into the network to gather a more focused perspective of the network to detect changes that may otherwise not be captured on a large scale.

Figure 6. Pearson's correlation between NL-CoIn and centrality based CoIn central nodes in undirected networks



Empirical Validation of Densification, Degree Distribution and Diameter Over Time for Hybrid Samples

Here we validate our sampling approach to determine if the hybrid samples selected and utilized in micro- and macro-level change detection follow fundamental network properties.

Analysis of Densification and Degree Distribution Over Time

Figure 8 indicates that the degree distribution in CoIn subgraphs (hybrid samples) has a long-tailed distribution and thus follows a power law distribution. Additionally, we also see that the degree exponent is less than 1 and constant overtime. Hence, our CoIn subgraphs show significant densification over time mainly because they capture the core part of the network. Thus, our hybrid samples are representative of the full graph.

These findings are reflected in both directed and undirected networks across datasets as summarized in Table 4.

In Table 4, our findings indicate DPL exponent $\alpha > 2$ in all cases which is an indicator of significant densification of CoIn subgraphs over time. Thus, the behavior H of the hybrid samples comprised of central nodes, is such that H[C] => a) $\gamma < 1$ and b) $\alpha >= 2$. Hence, the CoIn subgraphs selected during hybrid sampling obey the densification power law and can be representative to detect changes in the network over time. Additionally, preserving these properties in our hybrid samples also signifies that changes manifested at the network level can be traced back to the node level and vice versa mainly because the selected samples are key in the network structure, which is also observed in our similarity and correlation analysis.

Figure 7. (a) Accuracy, precision and recall for network level change point detection in CoIn subgraphs and full graphs in CAIDA 2008; (b) Accuracy, precision and recall for network level change point detection in CoIn subgraphs and full graphs in CAIDA Conficker

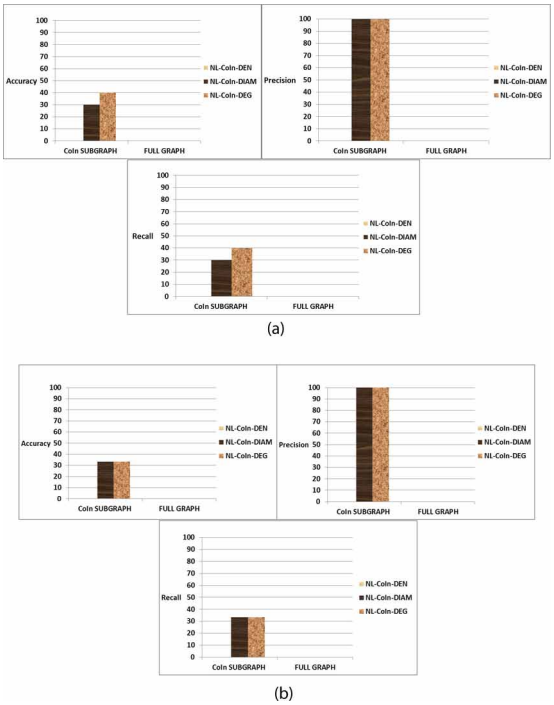


Figure 8. Degree distribution and degree exponent over time

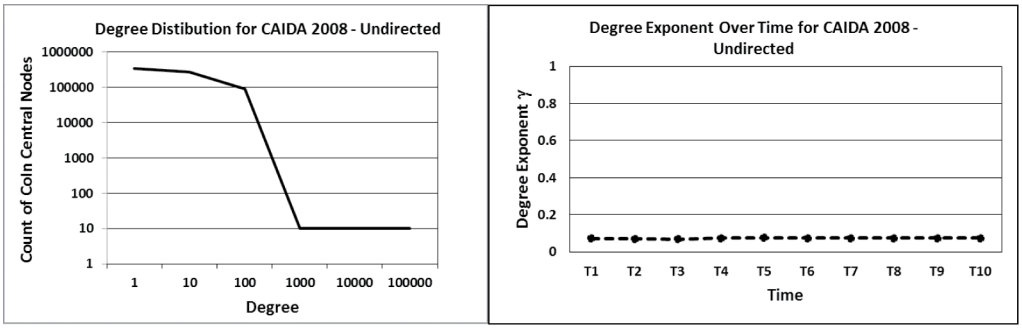


Table 4. Evaluation of densification power law and degree distribution over time

Type of Network		CoIn Central Nodes	Edges	Degree Exponent (γ)	DPL Exponent(α)
Directed	CAIDA 2008: In Degree	14	134623	0.06	3.69
	CAIDA 2008: Out Degree	15	13544	0.06	2.72
	CAIDA 2008-2010: In Degree	10	441532	0.05	4.69
	CAIDA 2008-2010: Out Degree	12	1667	0.14	2.06
Undirected	CAIDA 2008: Degree	20	184968	0.07	3.38
	CAIDA 2008-2010: Degree	15	599238	0.05	4.1

Analysis of Diameter Over Time

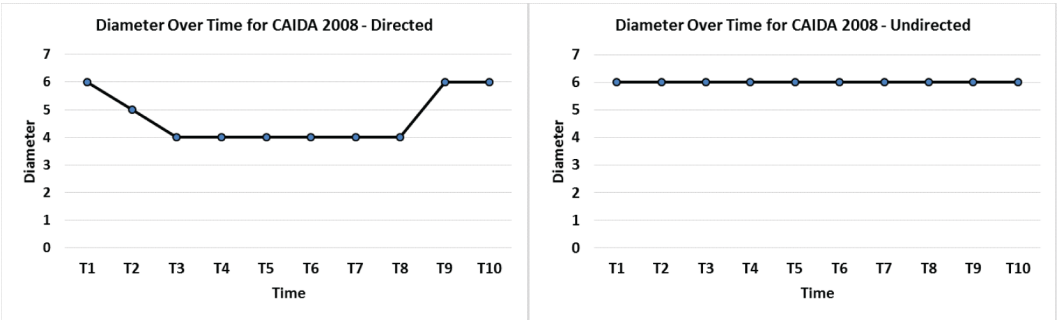
In Figure 9, our findings indicate that the diameter remains constant most of the time. In some cases, the diameter increases at certain time points in both the directed and undirected network structures where such time points have been identified as change points in the network as discussed in NL-CoIn. This is also observed in other datasets and in both directed and undirected networks.

Discussion of Findings

We highlight some key findings here:

- Our findings indicate that changes are apparent when detected based on an expected or fixed level in network behavior. In turn, this requires fundamental knowledge about the network and its key elements to determine when changes are occurring especially if such changes are associated to vulnerabilities in the network;
- Our findings indicate that by observing various reference points for network behavior, this can be used heuristically to state at what point in time changes in the network such as the behavior of key network points like central nodes should be considered alarming;
- Our findings indicate that an assessment of multiple factors in a network such as the level of connectivity (centrality of nodes) in the network can be used to detect changes that may otherwise be undetected;

Figure 9. Diameter over time



- Additionally, strategically sampling the network can be useful to detect network activity that may be undetected on a larger scale;
- Our findings indicate that the observation of network activity on a micro-level and macro-level scale can be used to evaluate the vulnerability of a network by assessing the spread or impact of a cyber threat on the network.

Practical Implications

Our approach also has several practical implications; (1) The analysis of key elements in the network such as central nodes and key subgraphs can be utilized in forensic analysis to monitor and analyze computer networks to support intrusion detection. (2) Monitoring such key network components can also be utilized in identifying the occurrence of persistent big cyberattacks or events such as Advanced Persistent Threats (APTs) that occur over long periods of time. (3) Changes in the structure of central nodes and key subgraphs can be used to determine training sets that can play a role in predicting the behavior of other key network components as well as predict the occurrence and evolution of Advanced Persistent Threats. (4) The analysis of key subgraphs can also be utilized in offensive cyber-operations to identify nodes to attack or segments of the network that are vulnerable. (5) Identifying central nodes can be used as a strategic approach to mitigate network downtime by establishing redundancy on the network.

Clarifications/ Assumptions

In this study, we have focused on changes in computer networks where we have analyzed internet traffic over periods of time. Here, we take into consideration devices such as personal computers and servers in the network where data is exchanged in the form of packets. Additionally, our findings are applicable to both directed and undirected graphs. We also assume that the networks are time evolving such that addition and deletion of nodes and edges takes place over time. More so, our findings clearly indicate that the degree distribution of nodes follows a densification power law distribution which further portray that the graphs densify over time. Next, we present our conclusions and future research directions.

CONCLUSION AND FUTURE WORK

This paper presents a combination of methods that can be applied in the discovery of changes that may occur in massive temporally evolving network structures. Particularly our methods are summarized into: i) graph sampling by node selection and ii) multi-level change detection in evolving networks. We also present extensive experimental results for our methods using real-world internet traffic data from the Center for Applied Internet Data Analysis (CAIDA). Our results show that the process of

change detection in evolving computer networks should be explored from diverse perspectives to identify persistent or re-occurring changes. Additionally, our results indicate that changes detected at the micro-level of the network described by inconsistency in presence or centrality of central nodes over time, does affect the structural characteristics defined at macro-level. The major contribution of this paper is that we are able to identify key feature vectors for overall network analysis as well as change detection in other domains. Overall, our approach can also be adapted to the process of threat detection based on the suspicious behavior of nodes in the network. In the future, we plan to extend our approach to determine optimal sampling rates for large networks. More so, we plan to extend our research to utilize predictive modelling for both micro- and macro- level characteristics in the network. We would also like to utilize networks with more contextual information in regards to the role of nodes in the network to further support our overall approach and analysis. Lastly, our current framework can be adapted to the space paradigm particularly by identifying spatial regions on the network that are associated with changes or shifts in the overall network behavior.

REFERENCES

Akoglu, L., & Faloutsos, C. (2013). Anomaly, Event, and Fraud Detection in Large Network Datasets. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining* (pp. 773-774). doi:10.1145/2433396.2433496

Bonacich, P. (1987). Power and Centrality: A Family of Measures. *American Journal of Sociology*, 1170–1182.

Elser, B., & Montresor, A. (2013). An Evaluation Study of Big Data Frameworks for Graph Processing. In *2013 IEEE International Conference on Big Data*. doi:10.1109/BigData.2013.6691555

Freeman, L. C. (1979). Centrality in Social Networks: Conceptual Clarification. *Social Networks*, *1*(3), 215–239. doi:10.1016/0378-8733(78)90021-7

Friedkin, N. E. (1991). Theoretical Foundations for Centrality Measures. *American Journal of Sociology*, *9*(16).

Gaston, M., Kraetzl, M., & Wallis, W. (2006). Using Graph Diameter for Change Detection in Dynamic Networks. *Australasian Journal of Combinatorics*, *35*, 299–311.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA Data Mining Software: An Update. *SIGKDD Explorations, 11*(1). Retrieved from https://hpcf.umbc.edu/

Krishnamurthy, V., Faloutsos, M., Chrobak, M., Lao, L., Cui, J. H., & Percus, A. G. (2005, May). Reducing large internet topologies for faster simulations. In *International Conference on Research in Networking* (pp. 328-341). Springer.

Lee, C. (2006). Correlations Among Centrality Measures in Complex Networks.

Leskovec, J., & Faloutsos, C. (2006). Sampling from large graphs. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*. doi:10.1145/1150402.1150479

Leskovec, J., Kleinberg, J., & Faloutsos, C. (2007). Graph evolution: Densification and shrinking diameters. ACM Transactions on Knowledge Discovery from Data.

McCulloh, I. (2009). *Detecting Changes in a Dynamic Social Network* [Thesis]. Institute for Software Research School of Computer Science Carnegie Mellon University.

Namayanja, J. M., & Janeja, V. P. (2011). Subspace Discovery for Disease Management: A Case Study in Metabolic Syndrome. *International Journal of Computational Models and Algorithms in Medicine*, *2*(1), 38–59. doi:10.4018/jcmam.2011010103

Namayanja, J. M., & Janeja, V. P. (2013). Discovery of persistent threat structures through temporal and geo-spatial characterization in evolving networks. In *IEEE International Conference on Intelligence and Security Informatics* (pp. 191-196). doi:10.1109/ISI.2013.6578817

Namayanja, J. M., & Janeja, V. P. (2014). Change Detection in Temporally Evolving Computer Networks: A Big Data Framework. In *First International Workshop on High Performance Big Graph Data Management, Analysis, and Mining, co-located with the IEEE Big Data 2014* doi:10.1109/BigData.2014.7004372

Namayanja, J. M., & Janeja, V. P. (2015). Change Detection in Temporally Evolving Computer Networks: Changes in Densification and Diameter Over Time. In *IEEE International Conference on Intelligence and Security Informatics*.

Opsahl, T., Agneessens, F., & Skvoretz, J. (2010). Node Centrality in Weighted Networks: Generalizing Degree and Shortest Paths. *Social Networks*, *32*(3), 245–251. doi:10.1016/j.socnet.2010.03.006

Page, L., Brin, S., Motwani, R. & Winograd, T. (1999). The PageRank Citation Ranking: Bringing Order to the Web.

Scripps, J., Tan, P., & Esfahanian, A. (2007). Node Roles and Community Structure Node Roles and Community Structure in Networks. In *Proceedings of the 9th WebKDD and 1st SNA-KDD. Workshop on Web Mining and Social Network Analysis. ACM.*

Shen, Y., Nguyen, N. P., Xuan, Y., & Thai, M. T. (2012). On the Discovery of Critical Links and Nodes for Assessing Network Vulnerability. *IEEE/ACM Transactions on Networking*.

Slavin, V. (2006). Improper Use of Control Charts: Traps to Avoid. Retrieved from https://www.sei.cmu.edu/library/assets/slavin.pdf

State of the Internet Connectivity Reports. (2008). Retrieved from https://www.akamai.com/us/en/about/our-thinking/state-of-the-internet-report/archives/state-of-the-internet-connectivity-reports-2008.jsp

State of the Internet Connectivity Reports. (2009). Retrieved from https://www.akamai.com/us/en/about/our-thinking/state-of-the-internet-report/archives/state-of-the-internet-connectivity-reports-2009.jsp

State of the Internet Connectivity Reports. (2010). Retrieved from https://www.akamai.com/us/en/about/our-thinking/state-of-the-internet-report/archives/state-of-the-internet-connectivity-reports-2010.jsp

Sun, J., Faloutsos, C., Papadimitriou, S., & Yu, P. S. (2007). Graphscope: Parameter-Free Mining of Large Time-Evolving Graphs. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 687-696). doi:10.1145/1281192.1281266

Tartakovsky, A. G., Polunchenko, A. S., & Sokolov, G. (2013). Efficient Computer Network Anomaly Detection by ChangePoint Detection Methods. *IEEE Journal of Selected Topics in Signal Processing*, *7*(1), 7–11.

The CAIDA UCSD Anonymized Internet Traces. (2008). Retrieved from http://www.caida.org/data/passive/passive_2008_dataset.xml

The CAIDA UCSD Anonymized Internet Traces. (2009). Retrieved from http://www.caida.org/data/passive/passive_2009_dataset.xml

The CAIDA UCSD Anonymized Internet Traces. (2010). Retrieved from http://www.caida.org/data/passive/passive_2010_dataset.xml

The CAIDA UCSD Network Telescope. (n.d.). Three Days Of Conficker. Retrieved from http://www.caida.org/data/passive/telescope-3days-conficker_dataset.xml

Vitter, J. S. (1985). Random Sampling with a Reservoir. In ACM Transactions on Mathematical Software, 11, 35-57. doi:10.1145/3147.3165

Josephine Namayanja is an Assistant Professor of Management Science and Information Systems at University of Massachusetts Boston. She completed her Ph.D in Information Systems at University of Maryland, Baltimore County in May 2015 where she also received her M.S in Information Systems in May 2010. She received her B.S in Information Technology from Makerere University Kampala in Uganda in 2007. Her general area of research is data mining with a broad interest in various domains. Her research focuses on the application of data mining to detect persistent patterns associated to change in network behavior. This research poses potential benefits in the domain of cyber security. She has also worked on research in healthcare where she has proposed methods that can be utilized in disease management and patient care, particularly to address needs for individuals with metabolic syndrome and Type II diabetes, respectively. Her research has been published in prestigious venues such as IEEE Big Data, IEEE Intelligence and Security Informatics, and Journal of Medical Systems. Her research has also been recommended by principle scientists in the domain and voted among the top articles in the domain. She is currently working on various research-related projects on Risk Management for IT Projects, Risk Identification for Post-Traumatic Stress Disorder (PTSD) and Cyber Security Analytics for Massive Communication Graphs.

Vandana Janeja is an Associate Professor at the Information Systems department at UMBC. Her research is in the area of Big Data Analytics with a focus on data mining and anomaly detection across multiple application areas. She has published in various refereed conferences such as ACM SIGKDD, SIAM Data Mining, IEEE ICDM, National Conference on Digital Government Research, IEEE ISI and journals such as IEEE TKDE, DMKD, KAIS and IDA. She holds a Ph.D. in Information Technology from Rutgers University. She completed her MBA from Rutgers University and MS in Computer Science from New Jersey Institute of Technology. Her research is funded through federal, state and private organizations including NSF, U.S. Army Corps of Engineers, MD State Highway Administration.