Network-Aware Behavior Clustering of Internet End Hosts

Kuai Xu, Feng Wang
Arizona State University
Email: {kuai.xu, fwang25}@asu.edu

Lin Gu
Hong Kong University of Science and Technology
Email: lingu@cse.ust.hk

***Abstract*—This paper explores the behavior similarity of Internet end hosts in the same network prefixes. We use bipartite graphs to model network traffic, and then construct one-mode projection graphs for capturing social-behavior similarity of end hosts. By applying a simple and efficient spectral clustering algorithm, we perform network-aware clustering of end hosts in the same prefixes into different behavior clusters. Based on information-theoretical measures, we find that the clusters exhibit distinct traffic characteristics which provides improved interpretations of the separated traffic compared with the aggregated traffic of the prefixes. Finally, we demonstrate the applications of exploring behavior similarity in profiling network behaviors and detecting anomalous behaviors through synthetic traffic that combines Internet backbone traffic and packet traces from real scenarios of worm propagations and denial of service attacks.**

I. Introduction

As Internet devices and applications continue to grow, it becomes increasingly important to understand network behavior for efficient network management and security monitoring. Understanding traffic behavior of backbone networks or large enterprise networks is a challenging task because of massive traffic data and the wide diversity of end hosts. While there exist a number of work on traffic behavior profiling [1], [2], [3], [4], these prior work mostly focus on traffic behavior at host-level, application-level and network-level. The studies on profiling host behaviors [1], [2] shed light on the behavior patterns of individual end hosts, however the increasingly large number of end hosts poses significant challenges for such fine-granularity analysis for backbone networks or enterprise networks in real-time environments. The application-level and network-level traffic profiling [3], [4] show aggregated communication patterns for a given application or a certain network, however the aggregated or coarse-granularity traffic profiling often requires further in-depth analysis to uncover the distinct behaviors of individual end hosts.

To fill the gap between host-level and network-level traffic behavior profiling, this paper proposes a new perspective of profiling network traffic by identifying and analyzing *behavior clusters of end hosts in the same network prefixes*. Focusing on clustered traffic behavior in network prefixes not only reduces the number of behavior profiles for analysis compared with host-level traffic profiling, but also reveals detailed patterns for a group of end hosts sharing similar behavior compared with network-level traffic profiling.

Towards this end, we introduce a novel approach, namely *network-aware behavior clustering*, which contains multiple steps to discover and make sense of inherent behavior clusters in networks prefixes. First, we use bipartite graphs to model *social-behavior* of Internet end hosts [5], [6]. The *social-behavior* of an end host is represented by *whom does the host communicate with?* Subsequently, we derive one-mode projection graphs to capture the social-behavior similarity of host communications through edges between source (or destination) hosts that talk to the same destinations (or sources). Third, we apply a simple spectral clustering algorithm to discover the inherent behavior clusters within the same network prefixes. Each cluster consists of end hosts that communicate with similar sets of severs, clients or peers, thus sharing strong social-behavior similarity.

For each traffic cluster of network prefixes, we use relative uncertainty concepts in information theory to characterize and interpret its behavior patterns based on traffic features such as source ports, destination ports and IP addresses. The results show that the clustering algorithm extracts behavior clusters with distinct traffic characteristics from the aggregated traffic of network prefixes. The distinct patterns of traffic clusters provide improved interpretations of network traffic, and illustrate the practical values of our proposed network-aware behavior clustering of Internet end hosts. In addition, our analysis also finds that the majority of end hosts stay in the same clusters over time, which suggests the high temporal stability of behavior clusters within the same network prefixes.

Finally we use case studies to demonstrate the applications of behavior clusters in discovering traffic patterns for traffic engineering or access control list constructions, and in detecting anomalous behavior such as scanning activities, worms and denial of service (DoS) attacks through synthetic traffic traces. The synthetic traffic combines backbone network traffic and real scenarios of worm propagations and denial of service attacks.

The contributions of this paper are summarized as follows:

- We use bipartite graphs to represent communication patterns between source and destination addresses, and construct one-mode projection graphs to capture the social-behavior similarity of host communications;
- We explore behavior similarity of Internet end hosts in the same network prefixes using a simple spectral clustering algorithm and discover the distinct behavior clusters of network prefixes;

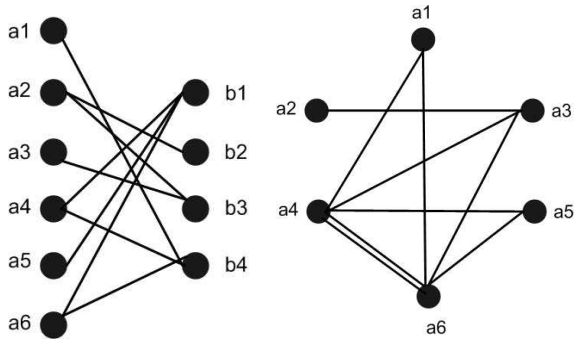
- We use relative uncertainty concepts to characterize and interpret traffic patterns of behavior clusters based on the distributions of traffic features such as source ports, destination ports and IP addresses;
- We demonstrate the applications of exploring behavior similarity in profiling network prefixes and detecting anomalous traffic patterns such as scanning activities, worms or denial of service attacks through synthetic traffic traces.

This paper is organized as follows. Section II discusses bipartite graphs for modeling data communication in network traffic and the one-mode projection for capturing social-behavior similarity of end hosts in the same prefixes. Section III is devoted to the spectral clustering algorithm for discovering behavior similarity of end hosts in the same prefixes. Section IV describes the datasets used in the study and presents the distinct characteristics of behavior clusters and temporal stability of the clusters. Section V demonstrates the applications of behavior similarity in discovering traffic patterns in network prefixes and detecting anomalous behaviors. Section VI discusses the related work, and Section VII concludes this paper and outlines the future work.

II. Modeling Host Communications with Bipartite Graphs and One-mode Projections

In this section, we first use bipartite graphs to represent hosts communications in network traffic, and subsequently use the one-mode projection of bipartite graphs to capture the social-behavior similarity of end hosts in the same network prefixes.

A. Bipartite graphs of host communications



(a) Bipartite graph (b) One-mode projection

Fig. 1. (a) A example of bipartite graphs based on host communications between the source hosts (a_1 - a_6) and the destination hosts (b_1 - b_4); (b) The one-mode projection on the vertex set of the left-side nodes, i.e., the source hosts (a_1 - a_6)

The host communications observed in network traffic of Internet backbone links or Internet-facing links of border routers for enterprise networks could be naturally modeled with a bipartite graph $\mathcal{G} = (\mathcal{A}, \mathcal{B}, \mathcal{E})$, where $\mathcal{A}$ and $\mathcal{B}$ are two disjoint vertex sets, and $\mathcal{E} \subseteq \mathcal{A} \times \mathcal{B}$ is the edge set [7]. Specifically, all the `source` IP addresses form the vertex set $\mathcal{A}$, while the vertex set $\mathcal{B}$ consists of all the `destination` addresses. Each of the edges, e_k in $\mathcal{G}$ connects one vertex $a_i \in \mathcal{A}$ and another vertex $b_j \in \mathcal{B}$.

To analyze the traffic behavior for network prefixes which include end hosts with the same network-bits in their IP addresses, we further decompose the bipartite graph of all the traffic into a set of smaller disjoint bipartite subgraphs such that each bipartite subgraph captures the host communications for a single source or destination IP prefix, e.g., $\mathcal{G}_{\mathcal{P}} = (\mathcal{A}_{\mathcal{P}}, \mathcal{B}, \mathcal{E}_{\mathcal{P}})$ and $\mathcal{G}_{\mathcal{Q}} = (\mathcal{A}, \mathcal{B}_{\mathcal{Q}}, \mathcal{E}_{\mathcal{Q}})$ representing the bipartite subgraphs of host communications for the source IP prefix $\mathcal{P}$ and the destination IP prefix $\mathcal{Q}$, respectively.

B. One-mode projection of bipartite graphs

To study the behavior similarity of end hosts in the same network prefixes, we leverage one-mode projection graphs of bipartite graphs that are used to extract hidden information or relationships between nodes within the same vertex sets [7]. Figure 1[a] shows a simple bipartite graph that is generated based on host communications between the six source hosts (a_1 - a_6) and the four destination hosts (b_1 - b_4), while Figure 1[b] illustrates the one-mode projection of the bipartite graph on the vertex set of the six left-side nodes, i.e., the source hosts (a_1 - a_6). An edge connects two nodes in the one-mode projection if and only if both nodes have connections to at least one same node in the bipartite graph. One could easily draw the one-mode projection graph for the vertex set of the right-side nodes using the same methodology.

The one-mode projection of the bipartite graphs uses edges between end hosts in the same prefixes to quantify the similarity of their network connection patterns. For example, in Figure 1[b] the edge between a_2 and a_3 reflects the observation that both a_2 and a_3 talk with the same destination host b_3 in the bipartite graph (Figure 1[a]), and the double edges between a_4 and a_6 capture the observation that both a_4 and a_6 talk with the destinations b_1 and b_4. Therefore, given a bipartite graph $\mathcal{G}_{\mathcal{P}} = (\mathcal{A}_{\mathcal{P}}, \mathcal{B}, \mathcal{E}_{\mathcal{P}})$ for a source prefix $\mathcal{P}$, we could construct the one-mode projection graph, $\mathcal{G}'_{\mathcal{A}_{\mathcal{P}}} = (\mathcal{A}_{\mathcal{P}}, \mathcal{E}'_{\mathcal{A}_{\mathcal{P}}})$, where $\mathcal{A}_{\mathcal{P}}$ consists of all source end hosts observed in $\mathcal{P}$ and $\{p_i, p_j\} \in \mathcal{E}'_{\mathcal{A}_{\mathcal{P}}}$ if and only if two hosts p_i and p_j talk with at least one same destination host. The similar process could generate the one-mode projection for any destination prefix $\mathcal{Q}$ as well.

To retain the information on the count of all the shared destinations, we use the normalized weight of the edges in the one-mode projection graph. Let $\mathcal{N}_{p_i}$ and $\mathcal{N}_{p_j}$ represent the numbers of destination hosts for the two hosts p_i and p_j in the prefix $\mathcal{P}$, respectively. We then use $w_{\{}p_i, p_j\}$ to denotes the weigh for the edges between two hosts p_i and p_j in the one-mode projection,

$$w_{\{}p_i, p_j\} = \frac{|\mathcal{N}_{p_i} \bigcap \mathcal{N}_{p_j}|}{|\mathcal{N}_{p_i} \bigcup \mathcal{N}_{p_j}|}, \tag{1}$$

where $|\mathcal{N}_{p_i} \bigcap \mathcal{N}_{p_j}|$ denotes the total number of the shared destinations in the bipartite graph between the two hosts p_i and p_j, and $|\mathcal{N}_{p_i} \bigcup \mathcal{N}_{p_j}|$ denotes the total number of the

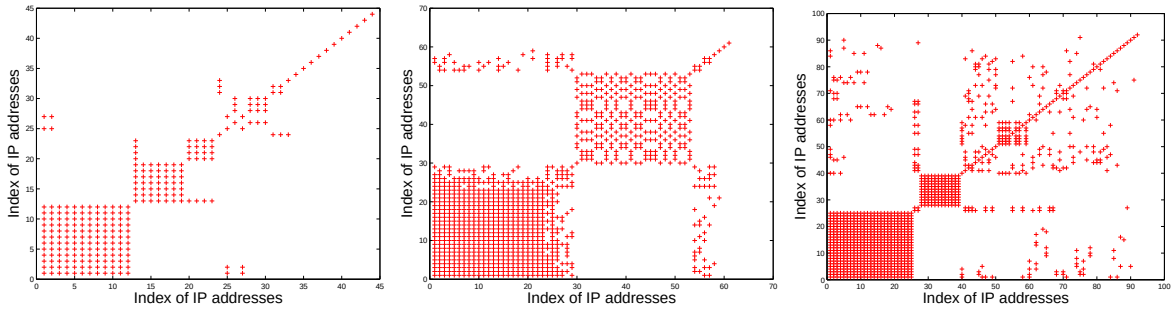


Fig. 2. Visualization of the adjacency matrix for one-mode projections of bipartite graphs for three network prefixes.

unique destinations of p_i and p_j. Note that $w_{\{p_i,p_i\}} = 1$. The weighted adjacency matrix of the one-mode projection graph for the network prefix $\mathcal{P}$ then becomes $\mathcal{M}_{\mathcal{P}} = (m_{i,j})_{|\mathcal{P}|\times|\mathcal{P}|}$, $m(i,j) = w_{\{p_i,p_j\}}$. The similar process could lead to the weighted adjacency matrix $\mathcal{M}_{\mathcal{Q}}$ of the one-mode projection graph for the destination prefix $\mathcal{Q}$.

One interesting observation of the one-mode projection graphs for host communications lies in the clustered patterns in the weighted adjacency matrix. The scatter plots in Figure 2 visualize the adjacency matrices of the one-mode projection graphs for three different network prefixes with 44, 61, and 92 end hosts, respectively. For each prefix, we sort the IP addresses based on the hosts' degree (number of neighbors in the one-mode project graph) in non-increasing order. Both x-axis and y-axis represent the indices of IP addresses in the same prefix, and each "+" point (i,j) in the plots denotes an edge with a positive weight between two sorted hosts p_i and p_j in the one-mode projection graph, i.e., $m(i,j) = w_{\{p_i,p_j\}} > 0$. As shown in Figure 2, each prefix has a few well-separated blocks that divide end hosts into different clusters. This observation on the adjacency matrix motivates us to further explore spectral clustering techniques and graph partitioning algorithm [8], [9], [10] to uncover these behavior clusters of end hosts in the same network prefixes.

III. Discovering Behavior Clusters with Spectral Clustering Algorithms

In this section, we describe a spectral clustering algorithm for automatic discovery of behavior clusters in network prefixes based on host communications. Figure 3 illustrates the schematic process of the algorithm from constructing bipartite graphs based on IP packets to analyzing behavior clusters of network prefixes. The previous section has discussed the construction of bipartite graphs and one-mode projection graphs, and this section is devoted to building similarity matrix and using spectral clustering algorithms to discover behavior clusters of network traffic. The next section will analyze behavior clusters and shed light on the characteristics of these clusters for in-depth understanding of network traffic.

Clustering algorithms have recently been used to analyze and profile hosts in campus networks in [1], where an agglomerative clustering algorithm is applied to characterize end hosts based on traffic features in IP packet headers, such as the

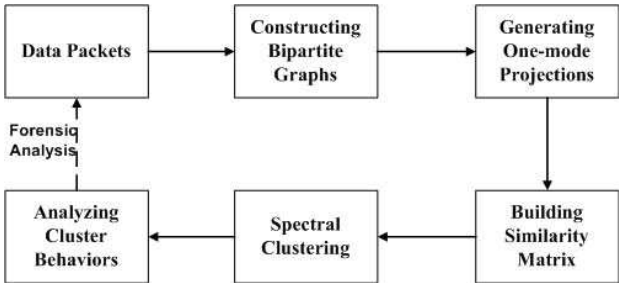


Fig. 3. The schematic process of network-aware behavior clustering algorithm for discovering behavior clusters of network prefixes

number of distinct destination IP addresses, the daily count of network traffic volumes in bytes, average TTL (time-to-live) value, etc. In this study, we focus on the social-behavior of end hosts in data communications through bipartite graphs and one-mode projection graphs, and are interested in exploring the social-behavior similarity of end hosts to discover inherent traffic clusters in the same network prefixes.

An important starting point of a clustering algorithm is to define the appropriate similarity metrics between data points. In this paper we use the weighted edge between two hosts u and v of the same prefix in the one-mode projection graph as the similarity measure $s_{u,v}$ between u and v, because the weighted edges capture and quantify the social-behavior similarity of host communications in network traffic. Therefore, the weighted adjacency matrix of the one-mode projection graphs for the prefix $\mathcal{M}_{\mathcal{P}}$ essentially becomes the similarity matrix $\mathcal{S}_{\mathcal{P}}$ which will be used as an input to the spectral clustering algorithm outlined below.

This study applies a simple spectral clustering algorithm developed in [10] due to its wide applications in graph partitioning and its small running time. The original spectral clustering algorithm [10] requires an explicit input of k as the expected number of clusters. Given the infeasibility of predicting the optimal number of behavior clusters in network prefixes without analyzing the traffic data, we therefore augment the algorithm by adding a step of automatically selecting an appropriate value of k as the desired number of the clusters based on the eigenvalue distribution. The detail of this step is explained in the following algorithm.

Algorithm 1 outlines the major steps of the spectral clus-

Algorithm 1 Algorithm of discovering behavior clusters using an augmented spectral clustering algorithm

Input: IP packet traces during a given time window and a source or destination prefix $\mathcal{P}$;

1: Construct bipartite graphs of host communications from data packets;
2: Generate the one-mode projection of bipartite graphs and its weighted adjacency matrix $\mathcal{M}_\mathcal{P}$ for end hosts in the prefix $\mathcal{P}$, and then obtain the similarity matrix $\mathcal{S}_\mathcal{P} \in \mathbb{R}^{n\times n}$ for the prefix $\mathcal{P}$;
3: Let A be the diagonal matrix with $A(i,i) = \sum_{j=1}^{n} s_{i,j}$, where $i = 1, \ldots, n$;
4: Compute the Laplacian matrix $L = A^{-1/2} S A^{-1/2}$;
5: Find the largest k eigenvalues, λ_1, λ_2, $\cdots$, λ_k such that $\sum_{i=1}^{k} \lambda_i \geq \alpha \times \sum_{j=1}^{n} \lambda_n$ and $(\lambda_k - \lambda_{k+1}) \geq \beta \times (\lambda_{k-1} - \lambda_k)$;
6: Use the corresponding k eigenvectors (e_1, e_2, $\cdots$, e_k) as columns to construct the matrix $E = [e_1 e_2 \cdots e_k] \in \mathbb{R}^{n\times k}$;
7: Construct the matrix $\mathcal{Z}$ through renormalizing E such that each row has a unit length, and consider each row as a point;
8: Run k-means clustering algorithm to cluster the points of $\mathcal{Z}$ into k clusters ($\mathcal{Y}_1$, $\mathcal{Y}_2$, $\cdots$, $\mathcal{Y}_k$)
9: Assign the original IP address p_i to the cluster $\mathcal{C}_j$ if the row i of $\mathcal{Z}$ is assigned to the cluster $\mathcal{Y}_j$.

Output: clusters $\mathcal{C}_1$, $\mathcal{C}_2$, $\cdots$, $\mathcal{C}_k$, where $\mathcal{C}_i = \{p_j | z_j \in \mathcal{Y}_j\}$.

tering algorithm with the augmented change of automatically selecting k clusters based on the traffic patterns. The input of this algorithm is IP data packet traces during a given time window and a source or destination prefix $\mathcal{P}$. The first step is to preprocess the IP data packets by constructing bipartite graphs of host communications from data packets, while the second step is to generate the one-mode projection of bipartite graphs and its weighted adjacency matrix $\mathcal{M}_\mathcal{P}$ for end hosts in the prefix $\mathcal{P}$, and then to obtain the similarity matrix $\mathcal{S}_\mathcal{P} \in \mathbb{R}^{n\times n}$.

Next, we compute the Laplacian matrix $L = A^{-1/2} S A^{-1/2}$, where A is the diagonal matrix with $A(i,i) = \sum_{j=1}^{n} s_{i,j}$ and $i = 1, \ldots, n$. Then in the augmented step we search for the largest k eigenvalues, λ_1, λ_2, $\cdots$, λ_k such that $\sum_{i=1}^{k} \lambda_i \geq \alpha \times \sum_{j=1}^{n} \lambda_n$ and $(\lambda_k - \lambda_{k+1}) \geq \beta \times (\lambda_{k-1} - \lambda_k)$. In other words, the augmented step searches an appropriate value for k by finding the largest k eigenvalues that account for at least α of the total variances and stopping at the eigenvalue λ_k where the distribution of eigenvalues exhibits a sharp slope change[1]. Figure 4 illustrates an example of the distribution of the eigenvalues for an IP prefix during a one-minute time window. It is very clear to observe a sharp elbow in the plot which indicates there exist a few eigenvectors that account for the majority of the variances in the similarity matrix. In other words, it shows that there indeed exist a few traffic clusters that could group end hosts in the same prefixes together based on their similar social-behavior patterns.

We use the corresponding k eigenvectors (e_1, e_2, $\cdots$, e_k) as columns to construct the matrix $E = [e_1 e_2 \cdots e_k] \in \mathbb{R}^{n\times k}$, and subsequently construct the matrix $\mathcal{Z}$ through re-normalizing E such that each row has a unit length.

[1]In our experiments, we use 0.9 and 2 for the α and β, respectively.

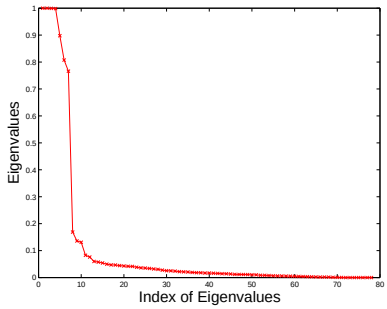


Fig. 4. Eigenvalue distribution of the similarity matrix for a network prefix

Considering each row as a point, the final step of the algorithm is to run a k-means clustering algorithm to cluster the points of $\mathcal{Z}$ into k clusters ($\mathcal{Y}_1$, $\mathcal{Y}_2$, $\cdots$, $\mathcal{Y}_k$), and then assign the original IP address p_i to the cluster $\mathcal{C}_j$ if the row i of $\mathcal{Z}$ is assigned to the cluster $\mathcal{Y}_j$.

The output of this algorithm is a set of k clusters ($\mathcal{C}_1$, $\mathcal{C}_2$, $\cdots$, $\mathcal{C}_k$), each of which includes a group of end hosts sharing similar social-behavior patterns in network traffic. In the following two sections, we will study traffic characteristics of behavior clusters discovered by the spectral clustering algorithm, and then demonstrate the applications of behavior clusters for discovering traffic patterns and detecting anomalous behaviors.

IV. Characteristics of Behavior Clusters

In this section we first briefly describe the datasets used in this study, and then discuss traffic characteristics of behavior clusters in network prefixes and cluster stability of end hosts.

A. Datasets

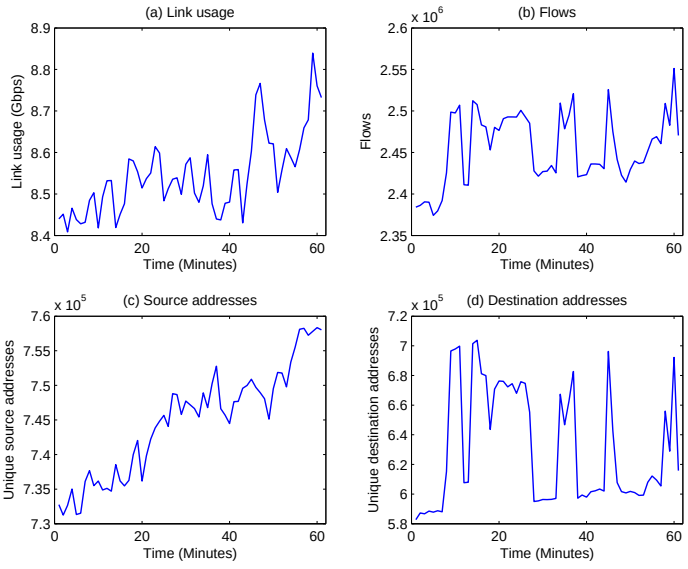


Fig. 5. Statistics of CAIDA Internet traffic trace

The datasets used in our analysis are collected from CAIDA's equinix-chicago and equinix-sanjose network mon-

itors [11] on bidirectional OC-192 Internet backbone links of a large Internet service provider during December 17, 2009. The CAIDA Internet traffic traces are anonymized using CryptoPAn *prefix-preserving* anonymization [12] for privacy reasons, however such *prefix-preserving* process does not affect our analysis that explores behavior similarity of end hosts within the same *network prefixes*.

Similar to the observations in previous studies [2], Internet backbone links carry large volumes of network traffic, which poses a challenging problem for real-time or near real-time traffic analysis. The total size of the compressed dataset used in this study is over 200GB. Figure 5[a] shows an average of 8.6 Gbps link usage during a one-hour duration, and Figure 5[b] illustrates millions of network flows for every minute during this period. In addition, Figures 5[c][d] show the total numbers of unique source and destination IP addresses, respectively. Such a large number of unique IP addresses in the packet traces makes it very challenging to analyze traffic behavior at host-level [1], therefore the focus on behavior clusters of network prefixes becomes an intuitive alternative for scalable analysis on Internet backbone traffic.

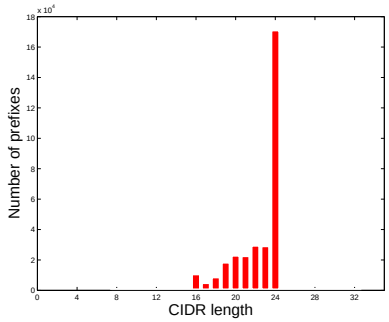


Fig. 6. Block size distribution of BGP prefixes

In our analysis we use $/24$ block as the network prefix granularity for analysis for two reasons. First, $/24$ is a common block size of BGP routing prefixes based on the observations on BGP routing tables. Figure 6 illustrates the block size distribution of BGP prefixes in a recent snapshot of BGP routing table from the RouteView project [13]. The $/24$ blocks account for over 50% of all the total prefixes on the Internet. Secondly, the prefix-preserving anonymization process makes it impractical to map anonymized IP addresses in the packet traces to the real BGP prefixes, although it is ideal to use BGP prefixes in the analysis. On the other hand, our methodology could be applied to BGP prefixes if data packets are not anonymized in other datasets. Our analysis is applied to both source and destination prefixes, since the bipartite graphs and one-mode projection graphs in the previous sections could be established for both sides. We will be presenting the results from both source and destination prefixes throughout the rest of this paper.

B. Distinct traffic characteristics of behavior clusters

The network-aware behavior clustering of end hosts shifts traffic analysis from host-level to prefix-level clusters, and

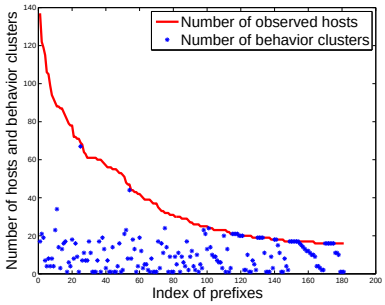


Fig. 7. The number of observed hosts and behavior clusters in all the prefixes with at least 16 hosts during one-minute time window

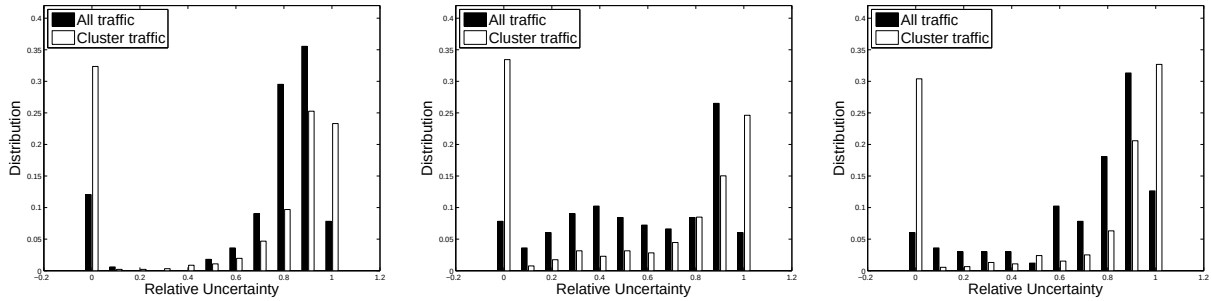
increases the granularity of traffic analysis compared with host-level traffic profiling, thus could successfully reduce the number of behavior profiles for analysis. Figure 7 illustrates the size of the prefixes with at least 16 end hosts and the number of their clusters during a one-minute time window. As we can see, the number of clusters is much smaller than the size of prefixes, as each behavior cluster groups many end hosts together due to their common social-behavior patterns. This observation holds for other time windows as well.

After obtaining separate behavior clusters, the next question we ask is *do the clusters indeed exhibit distinct behaviors?* Towards answering this question, we study the distributions of traffic features in each of behavior clusters, and then compare them with the aggregated traffic of the prefixes. We use an information-theoretic measure, relative uncertainty (RU) introduced in [2] to analyze the traffic features in individual clusters and the aggregated traffic. Given a variable X with a probability distribution, $p(x_i) = n_i/n, x_i \in X, i = 1, 2, \cdots, m$, where n_i is the number of times X is observed with the value x_i, the relative uncertainty (RU) on the variable X is defined as follows:

$$RU(X) = \frac{H(X)}{H_{max}(X)} = \frac{H(X)}{\log m}, \tag{2}$$

where the entropy, $H(X)$, is defined as $H(X) = -\sum_{x_i \in X} p(x_i) \log p(x_i)$, and $H(X)$ measures the variety or diversity in the observed values of X [14]. The relative uncertainty value $RU(X)$ quantifies the randomness or uniqueness of the observed values. In general, $RU(X)$ being 1 or approximately to 1 shows that the observed values of X are closer to being uniformly distributed, while $RU(X)$ being 0 or approximately to 0 indicates that the values of X are concentrated to one or a few frequently observed values [2].

Our results show that behavior clusters separate different traffic patterns of the same prefixes for improved understanding and interpretation. Figure 8 shows the distribution of relative uncertainty on destination IP addresses, source ports, and destination ports, respectively for all the source prefixes and their behavior clusters during a one-minute time window. Compared with relative uncertainty values for network prefixes, the behavior clusters have much larger percentages of relative uncertainty values on all of these features being



(a) Destination IP addresses (b) Source Ports (c) Destination Ports

Fig. 8. Histograms of relative uncertainty distributions for behavior clusters and the aggregated traffic

0 and 1 or approximately being 0 and 1, which reveal concentrated patterns on a unique port and IP address, or random patterns on ports and addresses. This result shows that the clustering algorithm extracts behavior clusters with distinct traffic characteristics from the aggregated traffic in the network prefixes, thus significantly improve the understanding of the traffic patterns with detailed and meaningful interpretations.

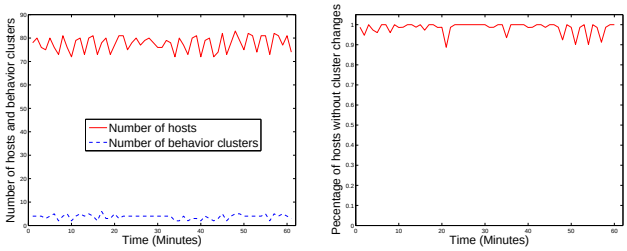
C. Temporal stability of behavior clusters

The second question on the characteristics of behavior clusters we ask is *are the clusters stable over time?* In other words, *do end hosts of network prefixes change clusters over time?* To address this question, we study the temporal stability of behavior clusters and the dynamics of cluster changes for end hosts over time. Figure 9[a] illustrates the high temporal stability of behavior clusters for one IP prefix during the one-hour time window. As shown by the top line in Figure 9[a], the number of end hosts in the prefix fluctuates slightly over time, since some hosts do not continuously send or receive traffic. More importantly, the number of behavior clusters, illustrated by the bottom line in Figure 9[a], also exhibits slight fluctuations over time. Similar observations hold for other prefixes.

In addition, we find the majority of end hosts stay in the same behavior cluster over time. Figure 9[b] shows the high percentage of end hosts in the network prefix in Figure 9[a] without changing clusters over consecutive time windows. In average, 71.8% of all the end hosts in the traffic traces do not change clusters during the one-hour time period. These observations confirm that network-aware behavior clustering separates end hosts of network prefixes into distinct and stable behavior clusters.

V. Applications

In this section we will demonstrate the applications of network-aware behavior clustering in understanding traffic patterns at the network prefix level and detecting anomalous behavior. We use traffic traces collected from Internet backbone links of a large ISP and real traces of worm propagations and denial of service attacks to generate synthetic traffic for the study.



(a) Number of hosts and behavior clusters over time (b) Percentage of end hosts without changing clusters

Fig. 9. Temporal stability of behavior clusters in a network prefix

TABLE I
Traffic clusters of an example destination prefix.

Cluster ID	Size	Flows	Patterns
1	20	422	(sip [87], spt *, dip [20], dpt 9050)
2	8	15	(sip [15], spt *, dip [8], dpt 80)
3	8	79	(sip [38], spt 80, dip [8], dpt *)
4	33	33	(sip [1], spt *, dip [33], dpt 445)

A. Discovering traffic patterns in network prefixes

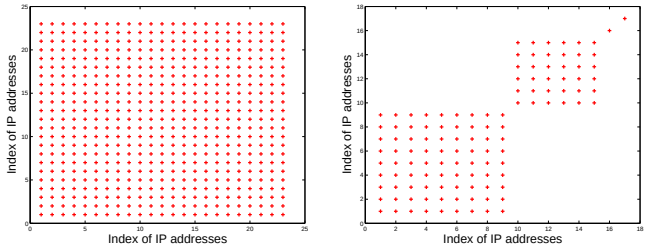
One major motivation of exploring behavior similarity is to gain a deep understanding of Internet traffic in backbone networks or large enterprise networks. Therefore, we first demonstrate the applications of network-aware behavior clustering on discovering traffic patterns. The traffic clusters discovered in each prefix reveal groups or clusters of traffic activities in the same prefixes, and understanding these patterns could be used for fine-grained traffic engineering and access control list (ACL) constructions.

Behavior clusters provide an improved understanding of traffic patterns in network prefixes compared with the aggregated traffic of network prefixes. For example, Table I lists four traffic clusters for one destination prefix with 69 active end hosts during one time window. The first cluster consists of 20 destination hosts (*dip [20]*) to which 87 source hosts (*sip [87]*) talk on destination port 9050 (*dpt [9050]*) with random source ports (*spt* *), while the second cluster consists of 8 hosts to which 15 source hosts talk on destination port 80. In the third cluster, 38 source hosts talk to 8 hosts using source port 80. Finally, the last cluster consists of 33

hosts to which a single source host talks on the destination port 445 that is associated with well-known vulnerabilities. In other words, the last cluster is very likely corresponding to a scanning activity towards these hosts. If the traffic of this prefix is mixed together for analysis, it becomes very difficult to interpret and understand since there are multiple behavior patterns simultaneously occurring within the same prefix. However, by separating the traffic into different clusters, the behavior of each cluster becomes much easier to interpret. More importantly, the patterns discovered by certain clusters could provide sufficient information for creating appropriate ACL rules for blocking malicious traffic without affecting other hosts. For example, we could build a simple ACL rule, <deny, tcp, $a.b.c.d$, any, any, eq 445> based on the last cluster in Table I, where $a.b.c.d$ denotes the scanner, for filtering further unwanted scanning traffic.

B. Detecting anomalous behavior

1) Detecting scanning activities with behavior clusters of destination prefixes: One interesting finding on the behavior clusters of network prefixes is that many prefixes with tens of end hosts have only a single cluster, i.e., all hosts in each of these prefixes talk with the same set of hosts. For example, Figure 10[a] shows one case of such activity towards one prefix with 23 end hosts in one time window. Upon close examination, we find that in this case one particular source IP scans all 23 IP addresses, thus explaining the single traffic cluster of the network prefix.



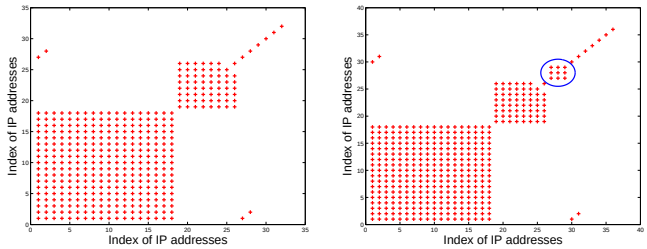
(a) Scanning Case 1 (b) Scanning Case 2

Fig. 10. Behavior clusters formed by scanning activities towards end hosts in the same prefixes

Detecting such simple scanning scenarios is not surprising, since many other existing approaches could reveal these patterns. However, the behavior clusters of destination prefixes are also able to reveal more challenging scanning cases from the massive traffic data. For instance, Figure 10[b] shows four behavior clusters of an IP prefix. The first cluster includes nine end hosts, while the second includes six hosts. Each of the last two clusters includes a single host since they do not share any social-behavior similarity with other hosts.

By studying network traffic in each cluster, we find that the first two clusters are corresponding to two independent scanning behaviors at the same time. The first cluster is due to one scanner targeting nine different hosts, while the second cluster is caused by a different scanner targeting six other hosts. It is very interesting to note that in terms of packet counts, the last two small-sized clusters account for 99.76% of network traffic (6655 out of 6671 data packets), while the first two clusters, having only nine and seven packets respectively, accounting for a very small percentage of the traffic. If traffic analysis is simply focused on the entire prefix, such low-volume anomalous patterns could simply be missed. Therefore, this suggests that behavior analysis on host communication patterns is complementary to existing volume-based techniques for detecting scanning behavior patterns.

2) Detecting worm behavior in its early phases: To demonstrate the application of network-aware behavior clustering in detecting worm behavior, we use real traces of Witty worm collected by CAIDA [15] and combine it with the backbone network traffic into synthetic traffic. The behavior clustering is able to detect a new cluster in one of the prefixes during the very beginning of worm propagations. Figures 11[a][b] show behavior clusters of this prefix before and after worm propagations, respectively. An emerging small cluster consisting of three end hosts marked by the circle in Figure 11[b] is actually triggered by data packets containing the Witty worms.



(a) Behavior clusters of a network prefix before Witty worm propagations (b) Behavior clusters of the prefix during the first few minute of Witty worm propagations

Fig. 11. Emerging behavior clusters formed by worm propagations

Such emerging behavior clusters of a network prefix triggered by worm propagation events or other suspicious activities serve as strong alarm signals to network operators for immediate response and in-depth analysis. More importantly, this approach does not require a large number of packets for early detections, since the new clusters could be formed by a very small number of end hosts that share the similarity of traffic communications with the same set of hosts. This above example shows that analyzing unusual or emerging behavior clusters could help detect worm propagations or other anomalies in their early phases. Such early-detection ability is very important for attack responses.

3) Detecting DDoS attacks: Detecting and mitigating DDoS attacks is one of the challenging tasks facing network operators or security analysts at edge networks due to the nature of these attacks in saturating network links. However, we argue that pushing the detection from edge networks to backbone networks is beneficial, since backbone networks have sufficient bandwidth and diverse routing paths compared with edge networks. By combining backbone traffic from

a large ISP and real cases of DDoS attacks identified in the previous work [16], we demonstrate the application of behavior similarity in detecting DDoS attacks in Internet backbone networks.

Figures 12[a-d] illustrate the behavior clusters of two source IP prefixes before and during DDoS attacks based on the synthetic traffic traces. The spectral clustering reveals emerging clusters or cluster changes during DDoS attacks for both source prefixes (Figure 12[b] and Figure 12[d]). For the first prefix, 39 end hosts form an emerging cluster in Figure 12[b], while in Figure 12[d] the existing cluster of 25 end hosts of the second prefix in Figure 12[c] is expanded to a much larger cluster with 52 end hosts. The reason for the abnormal expansion of the cluster in the second prefix is that the existing 25 hosts join other 27 hosts in the same prefix in launching the DDoS attacks while sending normal data traffic as well. Compared with other methods of detecting DDoS attacks, the advantage of behavior clusters is to leverage the small emerging clusters and the dynamics of existing clusters for capturing interesting events, such that the attacks could be detected before the traffic arrives at edge networks and saturates network links connecting edge networks to the Internet.

VI. Related Work

Most of the prior work have focused on profiling network behavior of individual end hosts [5], [2], [17], [18], or classifying the roles and communities of end hosts based on their traffic patterns [19], [20]. In [5], the authors study the host behavior at the social, functional and application levels for classifying traffic flows, while [2] builds behavior profiles of end hosts using traffic communication patterns and [18] merges packet header data into clusters based on the similarity of network traffic features. Through the information available on the web, Trestian et al. develop a traffic classification technique based on Google search engine [17]. [19] implements two algorithms to group hosts of enterprise networks into different roles based on their observed connection patterns. Similarly, [20] studies the communities of interest (COI) for hosts communications using traffic data collected from a large enterprise network. In contrast to these works, our goal in this paper is to study the social-behavior similarity of end hosts in the same network prefixes. Similar to our work, in [1] Wei at el. apply agglomerative algorithms to cluster end hosts into clusters based on host profiles that consist of a number of traffic features including daily destination number, daily byte number, average TTL, TCP and UDP ports, and aggregated statistics for each destination. Our study is different from [1] in that we build bipartite graphs based on host communications, and apply spectral clustering techniques to uncover groups of end hosts in the same prefixes that share common social-level behavior patterns.

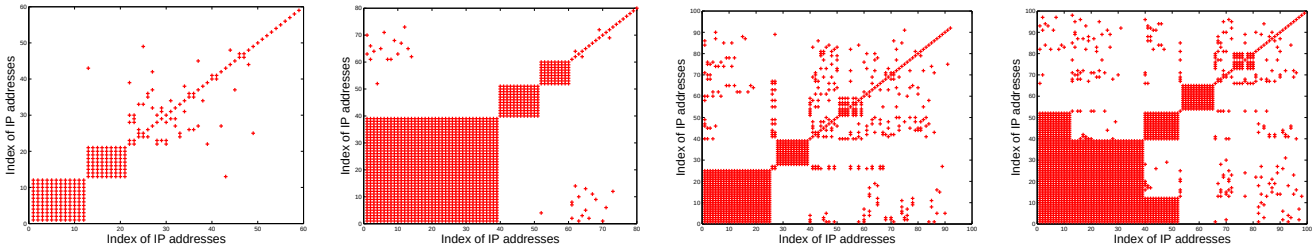
In addition, several work have studied the aggregated traffic behavior on the link-level, prefix-level or network-level [21], [3], [22], [23], [24], [25], [26]. In [21] Soule et al. develop a histogram-based method for classification of BGP-level prefix flows and use histograms to capture the entire distribution properties of highly aggregated flows, while [3] creates a traffic profile for each network prefix through behavior analysis of aggregated traffic at the network prefix level. In [22] Bhattacharyya et al. study traffic dynamics at POP-level and access-link-level in a major POP (Point-Of-Presence) in a commercial Tier-1 IP backbone. Lakhina et al. use PCA (Principal Component Analysis) to detect, identify, and quantify network-wide traffic anomalies from the normal background traffic [23]. [24] and [25] study the network-level behavior of spammers, while [26] proposes a network-aware clustering method to identify Web client clusters using BGP routing information. Different from these work on the aggregated behavior, our paper focuses on studying the similarity of host behavior in the same prefixes for network management and security monitoring.

Recent years have witnessed an increasing interest in IP address dynamics [27] and block level address usage in the visible Internet [28]. [27] develops an automated algorithm for identifying dynamic IP addresses and computing IP volatility based on Hotmail user-login data, while [28] identifies how Internet IP addresses are used from active probing, and finds that contiguous addresses are often used in a similar manner. Our work extends this work by studying traffic communication graphs to classify end hosts in the same IP prefixes into different clusters based on their traffic patterns.

Graph analysis has been widely used in monitoring and visualizing network traffic [6], studying network traffic behavior [4], discovering shared-interest relationships based on email communication history [29], and localizing botnets members based on the communication patterns used for command and control [30]. In [6], Iliofotou et al. use traffic dispersion graphs to model the social-behavior of hosts where the edges represent the interactions between source and destination hosts, while [4] uses traffic activity graphs to capture the interactions among hosts engaging in specific types of communications. The early work [29] constructs email communication graphs and employs interest-clustering algorithms for discovering email users with particular interests or expertise. [30] develops an inference algorithm to search botnet communication structures from the background communication graphs constructed from the collected network traffic. Inspired by these studies, our work also uses graph analysis to construct the bipartite graphs from host communication and then to generate the one-mode projection graphs for uncovering the social-behavior similarity among end hosts.

VII. Conclusions and Future Work

This paper explores network-aware clustering of Internet end hosts in the same IP prefixes. We use bipartite graphs and one-mode projection graphs to model host communication patterns observed on Internet backbone links. By applying spectral clustering algorithms on the one-mode projection, we find the clustered behavior of end hosts in the same network prefixes. We use relative uncertainty concepts to study the distinct traffic characteristics of behavior clusters, and analyze



(a) $Prefix_1$ before the attack (b) $Prefix_1$ during the attack (c) $Prefix_2$ before the attack (d) $Prefix_2$ during the attack

Fig. 12. Emerging behavior clusters of two independent source prefixes formed during DDoS attacks

the temporal stability of behavior clusters by examining the cluster changes of end hosts. Finally, we demonstrate the applications of behavior similarity in discovering traffic patterns in IP prefixes and detecting anomalous behaviors through synthetic traffic that combines backbone network traffic and real scenarios of worm propagations and denial of service attacks. Our future work includes developing a prototype system to evaluate the operational feasibility of network-aware behavior clustering of Internet end hosts in real-time. We are also interested in applying behavior clustering in other networks such as online social networks.

ACKNOWLEDGMENT

We would like to thank CAIDA for providing Internet traces dataset in this research. This work was supported in part by Arizona State University New College SRCA grants, and in part by HKUST grants DAG08/09.EG11 and REC09/10.EG06.

REFERENCES

[1] S. Wei, J. Mirkovic, and E. Kissel, "Profiling and Clustering Internet Hosts," in *Proceedings of the International Conference on Data Mining*, June 2006.

[2] K. Xu, Z.-L. Zhang, and S. Bhattacharyya, "Profiling Internet Backbone Traffic: Behavior Models and Applications," in *Proceedings of ACM SIGCOMM*, August 2005.

[3] H. Jiang, Z. Ge, S. Jin, and J. Wang, "Network Prefix-level Traffic Profiling: Characterizing, Modeling, and Evaluation," *Computer Networks*, 2010.

[4] Y. Jin, E. Sharafuddin, and Z.-L. Zhang, "Unveiling core network-wide communication patterns through application traffic activity graph decomposition," in *Proceedings of ACM SIGMETRICS*, June 2009.

[5] T. Karagiannis, K. Papagiannaki, and M. Faloutsos, "BLINC: Multilevel Traffic Classification in the Dark," in *Proceedings of ACM SIGCOMM*, August 2005.

[6] M. Iliofotou, P. Pappu, M. Faloutsos, M. Mitzenmacher, S. Singh, and G. Varghese, "Network monitoring using traffic dispersion graphs," in *Proceedings of ACM SIGCOMM Internet Measurement Conference*, 2007.

[7] J.-L. Guillaume, and M. Latapy, "Bipartite graphs as models of complex networks," *Physica A: Statistical and Theoretical Physics*, vol. 371, no. 2, pp. 795 – 813, 2006.

[8] Y. Weiss, "Segmentation Using Eigenvectors: A Unifying View," in *Proceedings of the International Conference on Computer Vision*, 1999.

[9] M. Maila and J. Shi, "Learning Segmentation with Random Walk," in *Proceedings of Neural Information Processing Systems (NIPS) Conference*, 2001.

[10] A. Ng, M. Jordan, and Y. Weiss, "On Spectral Clustering: Analysis and an Algorithm," in *Proceedings of Neural Information Processing Systems (NIPS) Conference*, 2001.

[11] Cooperative Association for Internet Data Analysis (CAIDA), "Internet Traces," http://www.caida.org/data/passive/passive_2009_dataset.xml.

[12] J. Fan, J. Xu, M. Ammar, and S. Moon, "Prefix-preserving IP address anonymization: measurement-based security evaluation and a new cryptography-based scheme," *Computer Networks*, vol. 46, no. 2, pp. 253–272, October 2004.

[13] University of Oregon, "Route Views Project," http://www.routeviews.org/.

[14] T. Cover and J. Thomas, *Elements of Information Theory*. Wiley Series in Telecommunications, 1991.

[15] C. Shannon and D. Moore, "The spread of the witty worm," *IEEE Security and Privacy*, vol. 2, no. 4, pp. 46 – 50, 2004.

[16] A. Hussain, J. Heidemann, and C. Papadopoulos, "A Framework for Classifying Denial of Service Attacks," in *Proceedings of ACM SIGCOMM*, August 2003.

[17] I. Trestian, S. Ranjan, A. Kuzmanovi, and A. Nucci, "Unconstrained endpoint profiling (googling the internet)," in *Proceedings of ACM SIGCOMM*, August 2008.

[18] K. Theriault, D. Vukelich, W. Farrell, D. Kong, and J. Lowry, "Network Traffic Analysis Using Behavior-Based Clustering," BBN Technologies Technical Report, 2010.

[19] G. Tan, M. Poletto, J. Guttag, and F. Kaashoek, "Role Classification of Hosts within Enterprise Networks Based on Connection Patterns," in *Proceedings of USENIX Annual Technical Conference*, June 2003.

[20] W. Aiello, C. Kalmanek, P. McDaniel, S. Sen, O. Spatscheck, and J. Merwe, "Analysis of Communities of Interest in Data Networks," in *Proceedings of Passive and Active Measurments Workshop*, March 2005.

[21] A. Soule, K. Salamatia, N. Taft, R. Emilion, and K. Papagiannaki, "Flow classification by histograms: or how to go on safari in the internet," in *Proceedings of ACM SIGMETRICS*, June 2004.

[22] S. Bhattacharyya, C. Diot, J. Jetcheva, and N. Taft, "Pop-level and access-link-level traffic dynamics in a tier-1 POP," in *Proceedings of ACM SIGCOMM Workshop on Internet Measurement*, San Francisco, CA, 2001.

[23] A. Lakhina, M. Crovella, and C. Diot, "Diagnosing Network-Wide Traffic Anomalies," in *Proceedings of ACM SIGCOMM*, 2004.

[24] A. Ramachandran and N. Feamster, "Understanding the Network-Level Behavior of Spammers," in *Proceedings of ACM SIGCOMM*, September 2006.

[25] Z. Qian, Z. Mao, Y. Xie, and F. Yu, "On Network-level Clusters for Spam Detection," in *Proceedings of Network and Distributed System Security Symposium (NDSS)*, March 2010.

[26] B. Krishnamurthy, and J. Wang, "On network-aware clustering of Web clients," in *Proceedings of ACM SIGCOMM*, August 2000.

[27] Y. Xie, F. Yu, K. Achan, E. Gillum, M. Goldszmidt, and T. Wobber, "How dynamic are IP addresses?" in *Proceedings of ACM SIGCOMM*, August 2007.

[28] X. Cai, and J. Heidemann, "Understanding Block-level Address Usage in the Visible Internet," in *Proceedings of ACM SIGCOMM*, August 2010.

[29] M. Schwartz, and D. Wood, "Discovering shared interests using graph analysis," *Communications of the ACM*, vol. 36, no. 8, pp. 78 – 89, August 1993.

[30] S. Nagaraja, P. Mittal, C.-Y. Hong, M. Caesar, and N. Borisov, "BotGrep: Finding P2P Bots with Structured Graph Analysis," in *Proceedings of USENIX Security Symposium*, August 2010.