Community Detection in large-scale IP networks by Observing Traffic at Network Boundary

Ahmad Jakalan, Jian Gong, Qi Su, Xiaoyan Hu

Abstract— **Internet communications are becoming more and more complex due to the exponential growth in Internet applications which created a new challenging task to accurately and efficiently monitor and manage the huge and vast network traffic. Community detection in large-scale IP networks is an important and challenging research topic. This paper proposes a methodology of unsupervised clustering of IP addresses within a managed network domain (e.g., campus network) based on inter-IP communication structure. We propose a novel approach and an efficient algorithm to discover communities based on bipartite networks and one mode projection and the basis of graph partitioning of the similarity graph. Bipartite networks were built using a NetFlow dataset collected from a boundary router in an actual environment, and then a one-mode projection has been applied over the outside IP nodes to build a social similarity graph of the inside IP addresses. We extract communities based on graph partitioning into sub-graphs (communities). Experimental results demonstrate that our approach can discover communities from real managed domain networks and obtain high quality of partitioning communities.**

Index Terms—**Computer networks, networks security, host clustering, IP relationship discovery, Profiling IP networks.**

I. Introduction

Discovering communities in networks is one of the important and challenging research topics of network management and network security in addition to the researches in the social network analysis. With the continuous growing in the number and diversity of internet hosts and applications, it's becoming more increasingly important to understand traffic patterns of end-hosts and network applications to achieve a more efficient network management and security monitoring. Many researchers have focused on analyzing traffic behavior of individual hosts and applications. However, an increasingly large number of end-hosts, a wide diversity of applications, and massive traffic data pose significant challenges for such fine-granularity analysis for backbone networks, large enterprise networks, and Internet service providers IPS. These challenges make it difficult for researchers to study traffic patterns of end hosts independently, so it's important to discover groups of hosts that share similar behaviors. Different researchers tried to discover such clusters based on traffic patterns of end hosts such as in [1] where we have applied unsupervised machine learning techniques to detect clusters of similar traffic behaviors based on traffic patterns of individual hosts. In this research we are going in another direction for clustering IP hosts by detecting groups of similar social behavior; these groups are called communities of interests. A community of interest is a collection of hosts that share a common goal or environment or a collection of interacting hosts[2]. In complex networks, communities are defined as groups of densely interconnected nodes that are only sparsely connected with the rest of the network. Community detection in computer networks has different purposes such as detecting network traffic anomalies[3] and behavior analysis of internet traffic and application identification[4, 5]. The difficulties that face researchers when they study the problem of community detection in complex networks include the expected number of communities they are going to detect, because in most of the cases, the number of communities that the network should be partitioned into and the numbers of elements in each community are both unknown in advance before clustering, so it's important to know at which level of cutting edges of the graph should be applied on the graph to deduce a well and an efficient graph-partitioning. Many research approaches adopted the ''minimum cut'' for graph-partitioning which requires to know the minimum number of edges needed to disconnect a graph. However, the community structure problem differs crucially from graph partitioning in that the sizes of the communities are not normally known in advance. Community detection methods operate under the intuition that intra-community connections are more common than inter-communities connections. This paper proposes a methodology of unsupervised clustering of IP addresses within a managed network domain (e.g., campus network) based on inter-IP communication structure. Fig. 1 illustrates the overall scenario of the problem. Our focus is to find groups of hosts that communicate with the same external IP addresses (have similar social relationship with the outside network). The key idea of

Manuscript received July 09, 2015; revised July 23, 2015. This work was conducted under the support of Jiangsu Key Laboratory of Computer Networking Technology and the Key Laboratory of Computer Network and Information Integration (Southeast University), Ministry of Education. This work was sponsored by the National Grand Fundamental Research 973 program of China under Grant No. 2009CB320505, the National Nature Science Foundation of China under Grant No. 60973123, and the Technology Support Program (Industry) of Jiangsu under Grant No. BE2011173. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of those sponsors.

Ahmad Jakalan(email: ahmad@njnet.edu.cn, phone: 0086-15366165651), Jian Gong(jgong@njnet.edu.cn), Qi Su(qsu@njnet.edu.cn), Xiaoyan Hu(xyhu@njnet.edu.cn) are with the School of Computer Science & Engineering, Southeast University, Nanjing 210096, China. And with the Jiangsu Key Laboratory of Computer Network Technology, Nanjing 210096, China.

ISBN: 978-988-19253-6-7
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

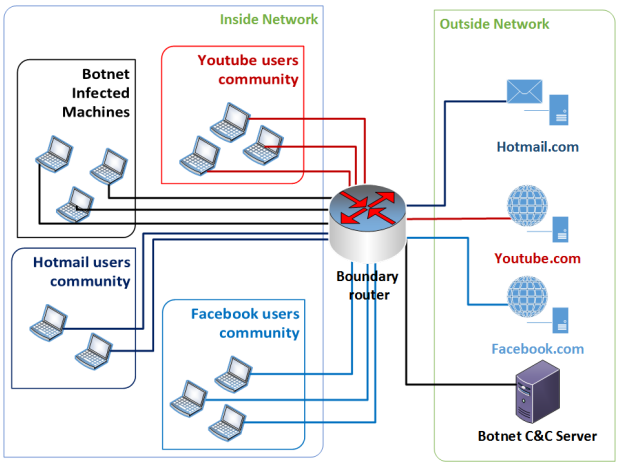


Fig. 1. The overall scenario of the problem: community detection within the managed network by observing their traffic at network boundary.

the proposed method is to explicitly add location information (internal/external) for IP clustering, different from previous works which focus only on sources and destinations, we split the entire IP address space into Internal (inside the managed domain) and External (outside) ones; the clustering method is to group a set of Internal IP addresses that communicate with common external IP addresses (i.e., the similarity measure of two internal IP addresses is the unique number of the common external IP addresses). The primary aim of this methodology is to find good quality of clusters, which is evaluated mainly on the basis of graph modularity. This approach was applied using a NetFlow dataset obtained from a boarder router in an actual environment and could be applied using any other types of datasets. The contributions of this paper include:

- We present an intuitive methodology based on global communication structure, i.e., inside-outside communication pattern represented as a bipartite graph.
- We adopt an efficient clustering algorithm to discover clusters with similar social behavior IP addresses.
- This methodology is based only on IP addresses and does not require information about TCP/UDP port numbers (which are occasionally obfuscated) or packet payloads (which are sometimes encrypted or unavailable from aggregated flow records), the use of an actual measured dataset is also the strength of this paper.
- We demonstrate practical benefits of exploring social behavior similarity of Internet hosts in understanding application usage, users' behavior, finding malicious users, and/or finding users of prohibited applications.

The outline of this paper is as follows: In section 2 we discuss others' works in the field of community detection related to our work. Section 3 describes in details the implementation of our methodology, then section 4 presents the experimental results of our approach, and in section 5 we discussed the results and evaluate the proposed algorithm, and in section 6 we present interpretation of results in terms of IP networking. Finally in section 7 we present our conclusion and future work.

II. Related Works

Different researches appeared to analyze Internet end host behavior [1, 5-17]. Unsupervised classification of internet hosts based on the communication patterns of Internet hosts in a space of traffic features are proposed in [1, 8]. Illiofotou et al. [18] uses IP communication graph and information about some applications used by few IP-hosts for the purpose of profiling Internet backbone traffic. Bipartite graphs have been widely used to analyze complex networks[19], Internet traffic [4], and social networks [20]. Kuai Xu et al. [4, 5] used graph analysis to construct the bipartite graphs from host communication and then to generate the one-mode projection graphs for uncovering the communication patterns behavior similarity among the end hosts within the same network prefix. Bipartite networks are graphs with two parties with links connecting vertices between different parties, and not possible to have links between two nodes from the same part. In [4, 5] the two sides of the bipartite graph are the source IP addresses and the destination IP addresses and so the study period should be as short as possible, because actually we can't say that source IP addresses and destination IP addresses could be a fully separated groups if we want to build the bipartite graph over a long period. In our work, we construct the bipartite graphs from hosts' communications provided by NetFlow records of the boundary router, the two "fully" separated groups of entities are IP addresses from the two sides of the Internet, one we called the managed domain or the "Internal" IP addresses, and the other is the "External" IP addresses. Since the managed domain IP addresses are known and could be mapped, so any other IP address is considered as an External IP address. Our focus is to detect communities from the total managed domain which may contain tens or hundreds of thousands of IP nodes, not only detecting communities from the hosts within the same network prefix as in[5], so it's important to adopt a new and a robust algorithm which can perform the clustering in an efficient manner. We apply one mode projection on the bipartite graph over the out-side nodes, the result of one-mode projection is the social similarity graph, each two nodes have an edge connecting them if both IP addresses have common external IP address, and the weight of the link is the number of common external IP addresses. We call the adjacency matrix of this graph as the similarity matrix. Then we apply our clustering algorithm based on the a concept we call it affiliation factor which measures the degree of affiliation of each node to a group of nodes, so we add each IP address with other IP addresses which have the same social behavior to the same group.

The problem of community detection from graph has been discussed by researchers from different disciplines where systems are often represented as graphs like sociology, biology and computer science. This problem is very hard and not yet satisfactorily solved. Huge effort of a large interdisciplinary community of scientists has been spent on it over the past few years. Community detection has different applications, Krishnamurthy et al. [21] introduces clustering Web clients who have similar interests and are close together topologically and likely to be under common administrative control to improve the performance of services provided on the World Wide Web. Krishna et al. [22] Identify clusters of customers with similar interests in the network of purchase relationships

ISBN: 978-988-19253-6-7
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

between customers and products of online retailers which enables to set up recommendation systems that guides customers and enhances the business opportunities. The problem has a long tradition and it has appeared in various forms in several disciplines. Newman et al. [23] proposed a new algorithm, aiming at the identification of edges lying between communities and their successive removal, a procedure that after some iterations leads to the isolation of the communities, intercommunity edges are detected based on the importance of the role of the edges in processes where signals are transmitted across the graph following paths of minimal length. Newman [24] has examined the problem of detecting community structure in networks as an optimization task to find the maximal value of the quantity called as modularity over possible divisions of a network. Modularity[25] is one measure of the structure of networks or graphs. It was designed to measure the strength of division of a network into modules (also called groups, clusters or communities). Networks with high modularity have dense connections between the nodes within modules but sparse connections between nodes from different modules. Modularity is easy to compute and widely applicable. However, modularity optimization methods suffer from a "resolution limit" problem that depends on the size and connectivity of the network[26]. Spectral and min-cut techniques have been applied, but exhibit a bias such that aggressive maximization of certain community score functions can destroy intuitive notions of cluster quality [10]. The proposed algorithm is a heuristic approximation algorithm; it is better than the previous algorithms from the theoretical viewpoint and useful for the actual problem instances.

III. Methodologies

We study the community detection based on NetFlow records collected from the boundary routers, but at the same time the methodology could be applied on any type of datasets that provide the trace of IP activities on the border routers. The work is not limited to the managed domain, and it could be more general. The main focus is to be able to setup a model to detect the social behavior communities of IP addresses in one side of the Internet based on its social relationship with the IP addresses on the other side, each IP address is considered as an entity, we consider IP addresses as individual nodes. Fig. 1 shows the overall scenario of the problem, community structure detection within the managed network by observing their traffic at network boundary. The strength of this approach is that it is based only on IP relationship, not based on other traffic contents which are sometimes obfuscated. As in the Fig. 3, our intention

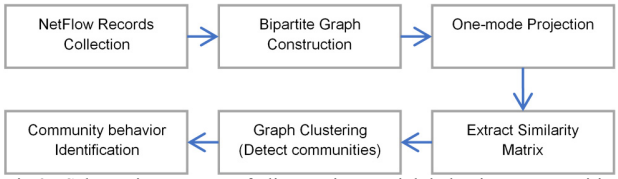


Fig.2. Schematic process of discovering social behavior communities within the managed domain network.

is to group inside IP addresses that are connected to the same IP address from the outside network in the same group, this will be useful to be able to have a better understanding of what services are requested or provided to the outside network, at the same time it will be helpful to identify some closed user groups such as botnets. Fig. 2 shows the schematic process of our methodology.

This methodology is defined in the following steps:

A. Construction of the bipartite graph

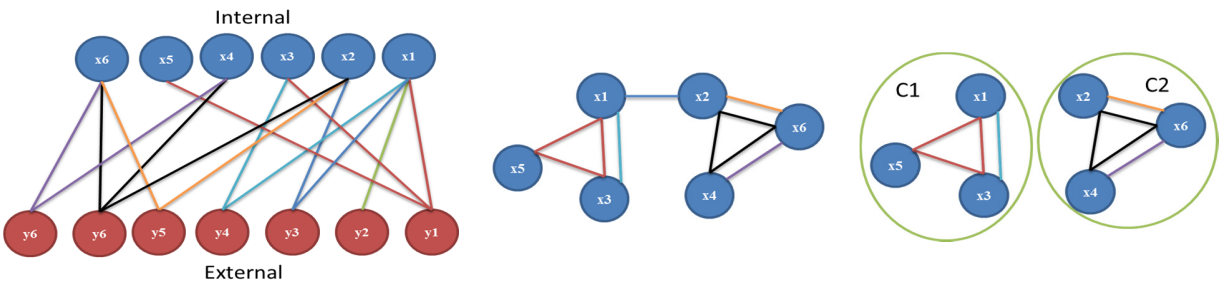
First we start with building the bipartite graph from the flow records captured on the border router between the managed network and the outside network, bipartite graph is a graph whose vertices can be divided into two disjoint sets (independent sets) such that every edge connects two vertices each of them belongs to one of the two independent sets, no edges can exist within one of these groups. The IP addresses that appear in the flow records are separated into two groups, the internal IP addresses and the external IP addresses. As we have mentioned, we are going to detect communities of social behavior within the managed domain so we first separate the monitored IP addresses as inside vertices X (the IP addresses which belong to our managed domain) and external vertices Y (the IP addresses which are not belonging to our managed domain). The bipartite graph is represented with its adjacency matrix. Let n=|X| is the number of inside IP addresses (internal vertices), p=|Y| is the number of External IP addresses (external vertices), and then:

G=(X, Y, E), where X is the group of internal IP addresses and Y is the group of external IP addresses. For a vertex, the number of adjacent vertices is called the degree of the vertex and is denoted deg(v). The degree sum formula for a bipartite graph states that

$$\sum_{x \in X} \mathbf{deg}(x) = \sum_{y \in Y} \mathbf{deg}(y) = |E| \qquad (1)$$

The adjacency matrix of the bipartite graph is defined as the following:

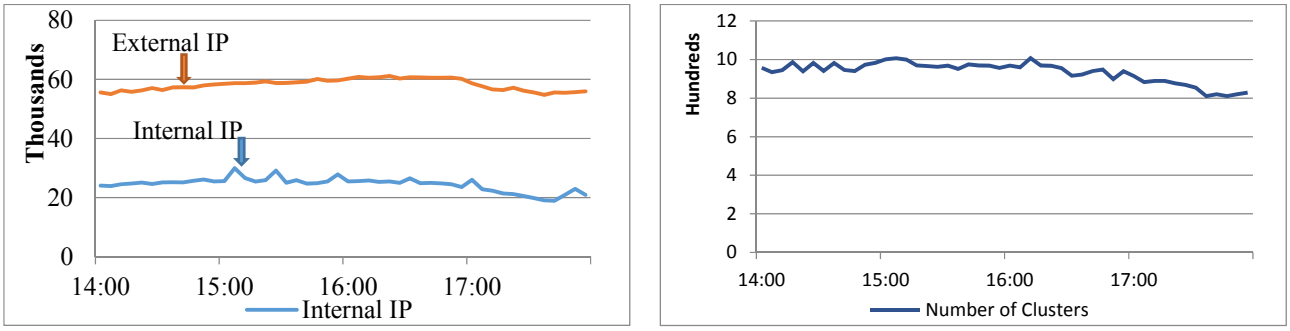
$$B_{n*p} = \begin{cases} 1 \; if \; there \; exists \; at \; least \; one \; flow \; between \; i \; and \; j \\ 0 \; if \; there \; is \; no \; flows \; between \; the \; nodes \; i \; and \; j \end{cases}$$



(a) (b) (c)

Fig. 3. (a) Bipartite network generated from the internal and external network, (b) The one-mode projection of the internal nodes, (c) The network after applying the partitioning algorithm divided into two communities.

ISBN: 978-988-19253-6-7
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)



(a) (b)

Fig.4. The experimental results of clustering algorithm over 4 hours, (a) shows the number of the monitored internal and external IP addresses captured by the Netflow records. (b) Shows the number of clusters generated by the algorithm in the same periods.

B. One-mode projection

A one mode projection over the external IP nodes is performed; in a one mode projection of a bipartite graph; an edge connects two nodes from the same side of the bipartite graph if and only if both nodes have connections to at least one same node in the other side of the bipartite graph. Fig. 2 (b) Illustrates one mode projection of the internal nodes over the external nodes of bipartite graph in Fig. 2 (b). A new graph is built, we call it the social behavior similarity graph, its vertices are internal IP addresses of the managed domain, and an edge appears between two nodes if they have a common external IP address they communicate with, the weights of the edges represent the number of distinct common external IP addresses between the two nodes. We call the adjacency matrix of the one mode projection as the similarity matrix S which represents the similarity in social behavior between IP addresses. Similarity matrix is a symmetric matrix; all entities on the main diagonal are zeros $s_{ii} = 0$. $S_{n*n} = [s_{ij}]$ Where s_{ij} is the number of common external IP addresses between *i* and *j*.

C. Graph partitioning algorithm

Communities are defined as groups of densely interconnected nodes that are only sparsely connected with the rest of the network. Our Clustering algorithm is based on the principle that one IP address should appear only in one community, and removed from other communities where it shares with them less number of common external nodes. When we think about the problem from the view point of graph theory, each line in the adjacency matrix represents a group containing the element at the row index with each column index element where the value of that cell in the matrix is larger than 0 is a member in the initial cluster that is identified by an id which is the row index, in other words, we consider each node with all of its neighbors as an initial group. We consider each internal IP address with other IP addresses who have similar behavior (share common external IPs with it) as a new group, so each line from the similarity matrix is first considered as a new cluster. For example the i^{th} element in the similarity matrix and all elements in the same line where they have a common external IP node where $s_{ij}>0$ are considered as one cluster. So, the initial maximum number of clusters is n. It's true that the similarity matrix is a symmetric matrix, but we have found that taking the entire matrix lines as initial clusters gives better results even it costs more processing time. As we have mentioned, the problem here is that neither the number of communities nor the number of elements in each community are known. For that, we remove elements from clusters based on an Affiliation Factor (AF). AF is defined as the degree of affiliation for each node x_i to a cluster C_k it belongs to by the following equation:

$$AF(x_i, C_k) = \sum_{j \in C_k} S_{ij} \quad (2)$$

We set the minimum number of elements in one cluster is 2 elements; otherwise the cluster will be deleted. We start the partitioning job based on the affiliation factor AF to keep elements in clusters they belong to them more than others, so for each element x_i from cluster $C_k$we check if it exists in another cluster C_l then we calculate its affiliation to both clusters, we have three situations:

- $AF(x_i, C_k) > AF(x_i, C_l)$ then x_i will be removed from C_l ;
- $AF(x_i, C_k) < AF(x_i, C_l)$ then x_i will be removed fromC_k;
- $AF(x_i, C_k) = AF(x_i, C_l)$ then x_i will be removed from the cluster with least number of elements.

It's important to check the number of elements in the clusters after each removal to confirm that the minimum number of cluster elements is 2; otherwise the cluster will be deleted.

IV. EXPERIMENTAL RESULTS

Our study is based on China Education and Research Network (CERNET) backbone data. We use IP Flow data collected from Netflow of border routers generated over different periods of time. The collected data is stored in files of a limited period of 5-minutes to be used later for analysis. Fig. 4. shows the experimental results of clustering algorithm over a time of 4 hours, (a) shows the number of the monitored internal and external IP addresses captured by the Netflow records, (b) shows the number of clusters (communities) detected by the algorithm over the same duration of time. It is clear that the number of clusters is stable on time based on the number of internal IP addresses.

V. DISCUSSION AND EVALUATION

A. Modularity

Modularity was proposed by Newman et al. [24] and then it has been used as a standard to measure the strength of division of a network into modules or the quality of community detection algorithms [27]. Modularity is defined as the fraction of the edges that fall within the given groups minus the expected such fraction if edges were distributed at random. It compares the number of edges inside a cluster with the expected number of

ISBN: 978-988-19253-6-7
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

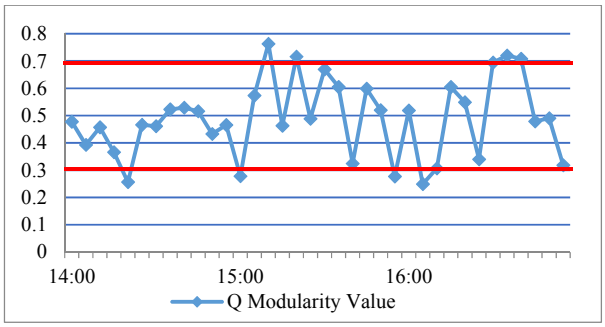


Fig. 5. Modularity value for clustering over a duration of three hours each period is 5 minutes, we calculate the modularity after each clustering, and we notice that most of the modularity values lie in between 0.3 and 0.7 which means that the modularity value of the clustering is good enough.

edges that one would find in the cluster if the network were a random network with the same number of nodes and where each node keeps its degree, but edges are otherwise randomly attached. It is positive if the number of edges within groups exceeds the number expected on the basis of chance. The value of the modularity lies in the range [−1/2, 1). Modularity reflects the concentration of edges within modules compared with random distribution of links between all nodes regardless of modules. Networks with high modularity have dense connections between the nodes in the same cluster but sparse connections between nodes from different clusters. The main consideration of modularity is the degree of distribution of the nodes in the network. In our network G, the adjacency matrix is given by the Similarity matrix S; the network contains a total of n nodes (vertices) and m edges, and d_i, d_j are the degrees of nodes i and j respectively. For any node, differences between the actual interactions and the expected numbers of connections can be obtained by calculating $S_{ij} - \frac{d_i d_j}{2m}$, so for a community C, the strength of community effect can be defined as:

$$\sum_{i \in C, j \in C} S_{ij} - \frac{d_i d_j}{2m} \quad (3)$$

So for the network G, it has been divided into k communities, and its modularity can calculated by the following equation:

$$Q = \frac{1}{2m} \sum_{l=1}^{k} \sum_{i \in C, j \in C} S_{ij} - \frac{d_i d_j}{2m} \quad (4)$$

The division on 2m is to regulate the Q value between -1 to 1. Practical implementation of this measurement by different researches confirm that a division of a network is considered a good division if the Q value lies between 0.3 and 0.7. We have applied the concept of modularity on the results to evaluate our algorithm, and it showed good results with most of the Q values lie in between 0.3 and 0.7 as shown in Fig. 5.

B. Internal and External Links

Another method to evaluate our clustering algorithm results is to compare the number of internal edges (with weights) within the same community with the number of edges connecting nodes from different communities. Fig. 6 shows a color scaled matrix of the sum of edges between elements of 100 communities. It's clear that the total number of edges between nodes from the same cluster represented in the main diagonal is much bigger than the number of edges connecting nodes from different clusters. And based on the definition of communities which are groups of densely interconnected nodes that are only sparsely connected with the rest of the network, we can prove

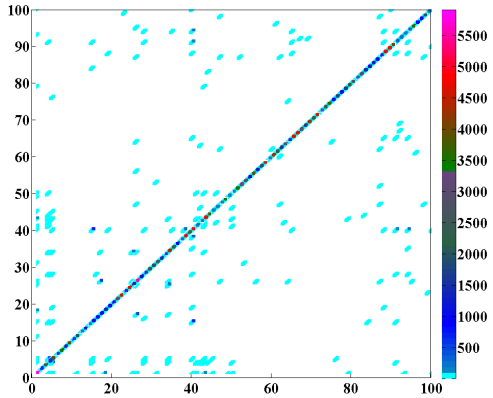


Fig. 6. A color scaled matrix of the sum of number of links between nodes from 100 communities, the main diagonal represents sum of links inside the communities

that the Clustering algorithm is giving good results. It's true that there are some clusters that have some or many edges connecting between nodes from different clusters, but they still are much less than the total number of edges connecting nodes from the same cluster.

VI. Interpretation of Results in Terms of IP Networking

To evaluate our approach practically, we have selected some clusters (communities) from the clustering results of one time period from 15:00 to 15:05 and applied further inspection on the flow records where the elements of these clusters are part of these flows (source or destination IP addresses), it was very clear that there is a dominant behavior of IP addresses in the same cluster like most frequently used application, most accessed website, or join the same botnet. Also we found that there is one or some outside IP addresses talking with all or most of the cluster's IP addresses. As we notice from Table I, some clusters are very big, and the common service they are using is very common like in cluster (0) where the most common service accessed by cluster members is searching the web using baidu.com, or using other services provided by the giant website in China. An abnormal behavior could be noticed We compared the clustering results of this period with previous and successive periods to find that the same IP address remain connecting with the same cluster members, and these IP addresses provide http service only to this outside IP address.

VII. Conclusion

In this paper we have presented an approach to detect IP networking social behavior communities from the managed domain network based on their social behavior with the outside network. Experimental results demonstrate that our approach can discover communities from real managed domain networks and obtain high quality of partitioning communities. We have discussed and evaluated our approach and shoed how good results we have obtained. Our experiments demonstrate that clustering quality is very good with a good value of modularity and the algorithm runs very efficiently even with the big data analyzed. For our knowledge, this is the first work to discover social communities of IP networks by splitting network into inside and outside networks and detect communities of similar social behavior of inside hosts based on their relationship with the outside Internet hosts. We have identified applications for

ISBN: 978-988-19253-6-7
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

TABLE I
DOMINANT BEHAVIORS OF SOME SELECTED CLUSTERS TO EVALUATE RESULTS IN TERMS OF IP NETWORKING FOR A SINGLE TIME PERIOD (5 MINUTES)

Cluster ID	Cluster Size	Dominant Protocol	Dominant Application	Notes about most frequent common external IP addresses
0	2623	TCP: 92.8% UDP: 6.69%	web: 90% P2P: 8.5%	All IP addresses in this cluster accessed baidu.com website using http
3	762	UDP: 99.61% TCP: 0.24%	service: 99.74% web: 0.22%	All IP addresses in this cluster have UDP connections with 112.124.*.*:53
13	253	UDP: 99.9%	P2P: 88.53% service: 11.19%	All IP addresses in this cluster connect with 69.22.*.* using different src/dst ports
18	168	TCP: 98.30% UDP: 1.55%	web: 73.47% P2P: 25.41%	All IP addresses in this cluster accessed google.com
90	33	TCP: 100%	P2P: 84.23% web: 14.35%	All IP addresses in this cluster have P2P connections with 202.119.*.* Internal port 3389, External port: random
96	31	TCP: 64.08% UDP: 33.70%	P2P: 54.14% web: 40.88%	P2P with 60.28.*.* Internal port 6000, External port 25607 Port 6000 with TCP protocol refers to X11 but here the connection is done over UDP
107	27	TCP: 100%	web: 96.77% P2P: 3.22%	Internal Port is 80, 97 %, with a single external IP. This is an abnormal behavior: a single external IP address with random port numbers communicating over TCP (Web service) with 27 internal IP address as if they are all web service providing service on port 80, this is a suspicious botnet behavior.
131	22	TCP: 95.30% UDP: 4.69%	web: 95.09% service: 4.90%	Internal Port 80, 95.09%. Two common external IP addresses accessing web service provided by internal IP addresses
205	13	TCP: 100%	P2P: 94.30% web: 2.98%	All IP addresses in this cluster have connections to * TCP, P2P fixed local port 3389 (officially registered as Windows Based Terminal (WBT)) and random external port
209	13	TCP: 100%	P2P: 81.30% web: 17.75%	All IP addresses in this cluster are simultaneously connected to both IP addresses *, * Using TCP, P2P, local port 3389, external port random

some selected discovered communities for the purpose of evaluation. Further work includes implementing this approach to identify applications on a large scale to provide a better understanding of network behavior and users' behaviors, the detection of closed user groups, and also implementing current work to setup a model for anomaly detection.

REFERENCES

[1] A. Jakalan, G. Jian, W. Zhang et al., "Clustering and Profiling IP Hosts Based on Traffic Behavior," Journal of Networks, vol. 10, no. 2, pp. 99-107, 2015-03-03, 2015.

[2] W. Aiello, C. Kalmanek, P. McDaniel et al., "Analysis of communities of interest in data networks," in Proceedings of the 6th international conference on Passive and Active Network Measurement, Boston, MA, 2005, pp. 83-96.

[3] W. X. Liu, and J. Cai, "A New Method of Detecting Network Traffic Anomalies," Applied Mechanics and Materials, vol. 347, pp. 912-916, 2013.

[4] K. Xu, F. Wang, and L. Gu, "Network-aware behavior clustering of Internet end hosts." pp. 2078-2086.

[5] K. Xu, F. Wang, and L. Gu, "Behavior Analysis of Internet Traffic via Bipartite Graphs and One-Mode Projections," IEEE-ACM Transactions on Networking, vol. 22, no. 3, pp. 931-942, Jun, 2014.

[6] Y. Himura, K. Fukuda, K. Cho et al., "Synoptic Graphlet: Bridging the Gap Between Supervised and Unsupervised Profiling of Host-Level Network Traffic," Ieee-Acm Transactions on Networking, vol. 21, no. 4, pp. 1284-1297, Aug, 2013.

[7] L. Bin, L. Chuang, Q. Jian et al., "A NetFlow based flow analysis and monitoring system in enterprise networks," Computer Networks, vol. 52, no. 5, pp. 1074-1092, Apr 10, 2008.

[8] G. Dewaele, Y. Himura, P. Borgnat et al., "Unsupervised host behavior classification from connection patterns," International Journal of Network Management, vol. 20, no. 5, pp. 317-337, Sep-Oct, 2010.

[9] T. Karagiannis, K. Papagiannaki, N. Taft et al., "Profiling the end host," Passive and Active Network Measurement, pp. 186-196: Springer, 2007.

[10] S. Wei, J. Mirkovic, and E. Kissel, "Profiling and Clustering Internet Hosts," DMIN, vol. 6, pp. 269-75, 2006.

[11] X. Kuai, Z. Zhi-Li, and S. Bhattacharyya, "Internet Traffic Behavior Profiling for Network Security Monitoring," IEEE/ACM Transactions on Networking, vol. 16, no. 6, pp. 1241-1252, 2008.

[12] B. Li, M. H. Gunes, G. Bebis et al., "A supervised machine learning approach to classify host roles on line using sFlow," in Proceedings of the first edition workshop on High performance and programmable networking, New York, New York, USA, 2013, pp. 53-60.

[13] H. Qiao, J. Peng, C. Feng et al., "Behavior Analysis-Based Learning Framework for Host Level Intrusion Detection," in Proceedings of the 14th Annual IEEE International Conference and Workshops on the Engineering of Computer-Based Systems, 2007, pp. 441-447.

[14] K. Xu, Z.-L. Zhang, and S. Bhattacharyya, "Profiling internet backbone traffic: Behavior models and applications," Computer Communication Review. pp. 169-180.

[15] Z. Zhang, B.-Q. Wang, H.-C. Chen et al., "Internet traffic classification based on host connection graph," Dianzi Yu Xinxi Xuebao(Journal of Electronics and Information Technology), vol. 35, no. 4, pp. 958-964, 2013.

[16] T. Karagiannis, K. Papagiannaki, and M. Faloutsos, "BLINC: multilevel traffic classification in the dark." pp. 229-240.

[17] S. Bhattacharyya, K. Xu, and Z.-L. Zhang, "Identifying significant behaviors within network traffic," US Patent 8,204,974, 2012.

[18] M. Iliofotou, B. Gallagher, T. Eliassi-Rad et al., "Profiling-By-Association: a resilient traffic profiling solution for the internet backbone," in Proceedings of the 6th International Conference, Philadelphia, Pennsylvania, 2010, pp. 1-12.

[19] J. L. Liu, and J. Cai, "Complex Network Community Structure of User Behaviors and Its Statistical Characteristics," in Proceedings of the 2011 Third Intl. Conference on Multimedia Information Networking and Security, 2011, pp. 366-370.

[20] E. A. Horvat, and K. A. Zweig, "One-mode Projection of Multiplex Bipartite Graphs." pp. 599-606.

[21] B. Krishnamurthy, and J. Wang, "On network-aware clustering of Web clients," SIGCOMM Comput. Commun. Rev., vol. 30, no. 4, pp. 97-110, 2000.

[22] P. Krishna Reddy, M. Kitsuregawa, P. Sreekanth et al., "A Graph Based Approach to Extract a Neighborhood Customer Community for Collaborative Filtering," Databases in Networked Information Systems, Lecture Notes in Computer Science S. Bhalla, ed., pp. 188-200: Springer Berlin Heidelberg, 2002.

[23] M. Girvan, and M. E. J. Newman, "Community structure in social and biological networks," Proceedings of the National Academy of Sciences, vol. 99, no. 12, pp. 7821-7826, June 11, 2002, 2002.

[24] M. E. Newman, "Modularity and community structure in networks," Proc Natl Acad Sci U S A, vol. 103, no. 23, pp. 8577-82, Jun 6, 2006.

[25] "Modularity (networks) - Wikipedia, the free encyclopedia," 2015-03-11, 2015; http://en.wikipedia.org/wiki/Modularity_(networks).

[26] S. Fortunato, and M. Barthelemy, "Resolution limit in community detection," Proc Natl Acad Sci U S A, vol. 104, no. 1, pp. 36-41, Jan 2, 2007.

[27] M. E. Newman, and M. Girvan, "Finding and evaluating community structure in networks," Phys Rev E Stat Nonlin Soft Matter Phys, vol. 69, no. 2 Pt 2, pp. 026113, Feb, 2004.

ISBN: 978-988-19253-6-7
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)