THESIS FOR THE DEGREE OF LICENTIATE OF ENGINEERING

Towards Understanding the Social Structure of Email and Spam Traffic

FARNAZ MORADI

Networks and Systems
Department of Computer Science and Engineering
CHALMERS UNIVERSITY OF TECHNOLOGY
Göteborg, Sweden 2012

Towards Understanding the Social Structure of Email and Spam Traffic
Farnaz Moradi

Technical report 98L
ISSN 1652-876X
Department of Computer Science and Engineering
Networks and Systems
Chalmers University of Technology
SE-412 96 GÖTEBORG, Sweden
Phone: +46 (0)31-772 10 00

Author e-mail: `moradi@chalmers.se`

Printed by Chalmers Reproservice
GÖTEBORG, Sweden 2012

Towards Understanding the Social Structure of Email and Spam Traffic

Farnaz Moradi

Networks and Systems, Chalmers University of Technology

ABSTRACT

Email is a pervasive means of communication on the Internet. Email exchanges between individuals can be seen as social interactions between email sender(s) and receiver(s), thus can be represented as a *network*. Networks of human interactions such as friendship relations, research collaborations, and phone calls have been widely studied before to allow understanding of the characteristics, as well as the structure and dynamics of such social interactions. In this thesis, we look into the social network properties of *email networks* generated from real traffic, and investigate how a vast amount of unsolicited email traffic (*spam*) affect these properties.

Current advances in Internet data collection and processing has facilitated the study of the characteristics of email traffic observed on the Internet. In our study, we have collected large-scale email datasets from traffic traversing a high-speed Internet backbone link and have generated email networks from the observed communications to analyze the structure and dynamics of these social interactions. Moreover, we aim at unveiling the distinguishing characteristics of legitimate and unsolicited email communications.

We show that the networks of legitimate email traffic has the same structural and temporal properties that other social networks exhibit, and therefore can be modeled as small-world scale-free networks. However, the unsolicited email communications cause deviations and anomalies in the structure of email networks, and this deviation from the expected social structural properties can be used to find the sources of spam email.

We also show that email networks, similar to other social networks, have a community structure which can be found using different community detection algorithms. However, not all community detection algorithms can identify structural communities that coincide with the true logical communities of email networks, i.e., distinct communities of legitimate and unsolicited email. Our study shows that a link-based community detection algorithm is more suitable for this purpose than more widely used node-based algorithms.

The possibility of merely using the social structure of email traffic to identify the source of spam and separate the unsolicited email from legitimate email, can potentially be used to improve the protection against spam and other types of malicious activities on the Internet.

Keywords: Internet Backbone Traffic, Email Networks, Social Network Analysis, Spam, Community Detection, Anomaly Detection

Acknowledgments

First and foremost, I would like to express my profoundest gratitude to my supervisors, Prof. Philippas Tsigas and Associate Prof. Tomas Olovsson, for their encouragement, understanding, and constant guidance. They have always inspired me by showing excitement for any result I have presented during our meetings and cheering me up anytime I was disappointed. I am also very much in their intellectual debt.

I would also like to acknowledge .SE – The Internet Infrastructure Foundation for funding the anti-spam project and SUNET for supporting our research by providing the access to the network traffic data. We have also received funding from the European Union Seventh Framework Programme (FP7/ 2007-2013) under grant agreement no. 257007.

I am also grateful to the current and former members of the networks and systems devision for a friendly and productive environment. Thanks Ali, Andreas, Asrin, Bapi, Daniel, Elad, Erland, Georgios, Ioannis, Laleh, Magnus, Marina, Negin, Nhan, Olaf, Pierre, Valentin, Vilhelm, Zhang, and especially Wolfgang who encouraged me to use the Internet backbone data in the first place and helped me in the collection and processing of the data.

Last but certainly not least, I would like to thank my family, my beloved father, mother, and brothers. Words cannot express my feelings, I am always grateful to you for your unwavering love, support, and encouragement over the years. And thanks to my dearest husband, Mohammad Reza, who supported me at each step of the way with his love and patience.

Farnaz Moradi
Göteborg, September 2012

List of Appended Papers

I **Farnaz Moradi**, Magnus Almgren, Wolfgang John, Tomas Olovsson, Philippas Tsigas, "*On Collection of Large-Scale Multi-Purpose Datasets on Internet Backbone Links,*" in *Proceedings of the First Workshop on Building Analysis Datasets and Gathering Experience Returns for Security (BADGERS)*, pp. 60 - 67, ACM, Salzburg, Austria, April 10, 2011.

II **Farnaz Moradi**, Tomas Olovsson, Philippas Tsigas, "*Towards Modeling Legitimate and Unsolicited Email Traffic Using Social Network Properties,*" in *Proceedings of the 5th Workshop on Social Network Systems (SNS'12)*, pp. 9:1 - 9:6, ACM, Bern, Switzerland, April 10, 2012.

III **Farnaz Moradi**, Tomas Olovsson, Philippas Tsigas, "*Mining Network-Level Communication Patterns of Email Traffic for Spotting Unsolicited Email,*" Technical Report no. 2012-12, ISSN: 1652-926X, Chalmers University of Technology, 2012.

IV **Farnaz Moradi**, Tomas Olovsson, Philippas Tsigas, "*An Evaluation of Community Detection Algorithms on Large-Scale Email Traffic,*" in *Proceedings of the 11th International Conference on Experimental Algorithms (SEA'12)*, Lecture Notes in Computer Science Vol.: 7276, pp. 283 - 294, Springer-Verlag, Bordeaux, France, June 7-9, 2012.

Contents

Part I

INTRODUCTION

1 Introduction

Email is one of the most common services on the Internet with everyday business and personal communications depending on it. Unfortunately, the vast amount of unsolicited email (*spam*) consumes network and mail server resources, imposes security threats, and costs businesses significant amounts of money. Spam can also be exploited for phishing and scam and it can carry Trojans, worms, or viruses, making email unreliable.

It is known that a large fraction of spam originates from *botnets* [29, 44]. A botnet is a collection of compromised hosts (*bots*) where each bot contributes to conducting malicious activities or attacks such as distributed denial of service (DDoS), scanning, click frauds, and sending spam. Therefore, identifying the source of spam can lead to the detection of the source of other malicious activities on the Internet.

Numerous attempts to fight spam have led to implementation of anti-spam tools that are quite successful in hiding the spam from users' mailboxes. Most of the conventional approaches inspect email contents at the receiving mail servers, and are very resource-intensive. Although such *content-based filters* are effective in learning what the content of spam looks like, the spammers are very agile in obfuscating email contents and encapsulating their messages in other formats such as images to bypass these filters.

As a complement to content-based filters, *pre-filtering* strategies are widely used to stop spam before the email content is received and examined by the mail servers. A commonly used pre-filtering method is *IP blacklisting*. The receiving mail servers can consult IP blacklists to decide whether to accept or reject an incoming email transaction. Early rejection of spam can dramatically decrease the workload on mail servers and reduce the cost. However, IP addresses are not persistent, they can be obtained from dynamic pools of addresses and they can be stolen [12, 44]. In addition, bots usually send spam at a low rate to each individual domain and do not reuse the IP addresses that have become blacklisted.

In addition to the above mentioned anti-spam strategies, numerous other spam detection and prevention techniques have been introduced. Approaches such as enforcing laws and regulations, requesting proof-of-work (e.g., processing time) [2], mail quota enforcement [54], port blocking, and user monitoring are proposed to stop spam at the sender side. Greylisting [21], reputation-based approaches, sender authentication, and domain verification are approaches that can be used on the receiver side before accepting email contents. Replacing SMTP with a new protocol or deploying overlay authentication protocols, are some of the ideas proposed to stop spam during transit.

Despite the considerable advances in spam detection and prevention methods, there is a constant battle between spammers and anti-spam strategies. Therefore, better understanding of the behavior of spam is crucial in order to find methods that can stop spam as close to its source as possible. Recently, approaches that focus on the network-level behavior of spam have gained attention. These approaches are concerned about email sending behavior of the spammers, which is expected to be more difficult for them to change than for instance the content of the email [8, 20, 45]. In order to improve and come up with more such methods, there is a need to understand the network-level characteristics of spam and how it differs from legitimate email (*ham*) traffic.

The study of the characteristics of email and spam can be conducted using different types of email data. A number of studies have used SMTP log files from mail servers [12, 18, 19, 51, 57]. Although such datasets are limited to communications to/from a single domain, they contain detailed information about each email and the statistical summaries of accepted and rejected email communications, which allows the comparison of the behavior of spam, ham, and the rejected traffic. The spam captured in honeypots or relay sinkholes have also been used to study the characteristics of spam [43, 45]. The honeypots only attract spammers, therefore they do not allow the comparison of different characteristics and communication patterns of spam and ham. Flow-level data collected on access routers have also been used to study the properties of spam and rejected traffic [49]. These flows only contain packet headers, and although they are not limited to a single domain, they do not carry enough information to allow distinguishing spam from ham to study their distinct characteristics. Another type of data that has been used to understand the sending behavior of spam was collected from inside spam campaigns [23, 28, 29]. The data collected at these campaigns has the view point of spammers and makes it possible to closely investigate how spam is sent.

In our studies, we have used yet another type of email data. Our dataset enables us to study the behavior of legitimate and unsolicited traffic from the perspective of a network device which monitors the traffic traversing a backbone link. Recent advances in large-scale data collection and processing have enabled us to collect SMTP traffic on high-speed Internet backbone links. The collected email traffic is not limited to a single organization or domain and allows us to classify the observed email into ham, spam, and rejected communications to compare their characteristics. The collection and processing of such data, however, is not trivial. The challenges involved are both of general and technical nature. Getting access to the link, handling the huge amount of traffic on the link, privacy concerns, pre-processing and processing of the large-scale dataset are just a number of challenges that need to be addressed before the study of the characteristics of ham and spam traffic becomes possible.

After collecting the email dataset, the next step has been to look for the network-level characteristics of spam that are distinct from ham. It is known that spam is sent automatically, therefore it is expected that it does not exhibit the *social network properties* of human-generated communications [10, 18, 30, 51]. The social network properties of email communications can be studied by analyzing the structure of *email networks* generated from email traffic. An email network is an implicit social network in which each node represents an email address and each edge represents an email. It has been shown that email networks have the same structural properties that other social and interaction networks have [13, 27, 33].

This thesis is concerned with the study of social structure of email networks generated from real traffic which contain both legitimate and unsolicited email communications. The goal is to find out how spam emails affect the structural and temporal properties of email networks and propose methods to separate them from a mixture of email traffic and spot the sources of spam based on their antisocial behavior rather than on what they contain.

1.1 Social Network Properties

1.1.1 Network Structure

An extensive amount of work has aimed at understanding the structure and dynamics of network systems such as the Internet router structure [16], online social networks [37], the World Wide Web [11], phone call and SMS graphs [40], and email networks [13]. Numerous studies have focused on characterizing, modeling, and analyzing such networks to shed light on the behavior of the system as a whole. Understanding the structure and dynamics of such networks have also found many applications such as identifying *Sybil* identities in a network [53, 59], spam detection [10, 18, 30, 51], stopping unwanted communications [38], traffic classification [22, 25], identifying botnets [39], understanding the behavioral patterns of email usage [26], personalized email prioritization [58], and anomaly and fraud detection [4, 35].

Traditionally, network data was studied as random graphs [14]. However, empirical studies on different type of real network data have revealed interesting properties such as the “small world phenomenon” [55], also known as “six degrees of separation” [36], and the scale-free behavior of networks [7, 16]. These properties show that social and interaction networks are fundamentally different from other types of networks such as random networks [41]. A review of the structural properties of these networks can be found in [5].

Many real networks have been modeled as *small-world* and *scale-free* networks. In a small-world network, the distance between any pair of nodes is relatively short. The distance between two nodes is measured as the number of edges on the shortest path connecting them. In addition to short average path length, small-world networks tend to be highly clustered which can be quantified using the average *clustering coefficient* [55]. Another robust measure of the structure of networks is their *degree distribution*. The degree distribution of a network characterizes the spread in the node degrees. It has been shown that for social and interaction networks the degree distribution has a power law tail [16]. Such networks are known as scale-free networks [7].

Numerous attempts to model the structure of social networks have also taken other graph properties of these networks into account: the distribution of the size of the connected components of the network, the presence of a giant connected component (GCC), and the community structure of the networks. The study of the changes of different graph properties over time have also revealed fascinating properties of network evolution such as *shrinking diameter* and *densification power law* [32]. As social networks grow over time, they become more connected, the size of their GCC increases, their diameter shrinks, while their average clustering coefficient value stays constant.

The first study of the structure of email networks was performed by Ebel et al. [13]. They studied an email network generated from log files of the mail server of their university (Kiel University) and showed that this email network is scale free and exhibits the properties of small-world networks. Studies on the evolution of email networks have shown that email networks, similar to other social networks, densify over time and their diameter shrink [33]. It was also observed that the power law degree distribution shape and exponent of email networks remain relatively constant over time [27, 33].

1.1.2 Community Structure

Another excessively studied structural property of social and interaction networks is their community structure. A *community*, also known as a *cluster*, is

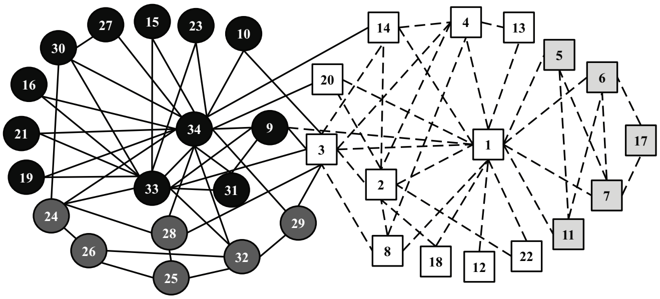

Figure 1.1: *The communities in the Zachary karate club network found by applying the fast modularity optimization algorithm by Blondel et al. [9] which coincides with the actual partitioning of the network into two groups (square and round nodes). A link-based community detection algorithm can also reveal the true partitioning of the network (solid and dashed edges), in which nodes with both type of edges are overlapping between the two communities [15].*

usually considered to be a set of nodes that are densely connected to each other and have less connections to the rest of the network. This property has been observed in many real networks, and particularly in social networks. A wide variety of *community detection* algorithms, also known as *clustering* algorithms have been proposed to extract the communities in a network merely based on the structure of the network. Figure 1.1 shows an example of the communities identified by a community detection algorithm on a real network, Zachary's network of karate club members [60], which coincide with the true partitioning of the network. A review of different type of community detection algorithms can be found in [17, 48, 56].

Community detection has also found many applications such as finding users with similar interests in a social network in order to provide recommendations to them, clustering users or clients that are geographically close or communicate a lot with each other in a system to improve the performance of the service provided to them [24], and identifying the community of malicious users in order to mitigate Sybil attacks [53]. In our study, we also show that a community detection algorithm can be used to separate unsolicited and legitimate email into distinct communities.

Community detection algorithms typically aim at partitioning the nodes of a network into clusters so that a number of structural properties are satisfied. Different community detection algorithms use different approaches and yield different communities. In order to find out which algorithm yields the best clustering, it is required to use a quantitative measure to assess the quality of different clusterings. The most widely used structural *quality function* is *modularity* [42] which is also widely used as an *objective function* to be optimized by the community detection algorithms to create communities with high structural quality. It has been shown that there is no single perfect quality function for the comparison of the communities detected by different algorithms [6]. Therefore, other well-known quality functions such as coverage, expansion, performance, density, and conductance can also be investigated to allow the selection of the most suitable algorithm for the network data at hand. In addition to the above structural quality functions, the logical quality of the clustering can also be determined based on how homogeneous the edges inside the communities are.

Recently, new community detection approaches have emerged that are based on partitioning the edges of a network into communities rather than partitioning the nodes [3, 15]. Figure 1.1 shows that both node-based and edge-based community detection algorithms identify the true partitioning of the network. The solid lines and the dashed lines represent the two communities identified by an edge-based community detection method. The nodes with both type of edges are overlapping between these communities. Overlap can naturally exists in many social and interaction networks [56].

The high diversity of community detection algorithms have made it necessary to perform experimental evaluation of the algorithms on specific types of networks to find the most suitable method for that type of network data. Email communications can be either seen as flow of data or as pairwise relations between people, therefore both flow-based community detection approaches such as *markov clustering* by Dongent [52] and *maps of random walks* by Rosvall et al. [46, 47], and topological approaches such as *fast modularity maximization* by Blondel at al. [9], and *link community detection* by Ahn et al. [3], can be suitable for clustering email networks. Leskovec et al. [34] have empirically compared different clustering algorithms on different real networks including email networks of single organizations. Lancichinetti et al. [31] have also compared different community detection algorithms and have shown that the structure of the communities in an email network of an organization is similar to the structure of communities in other communication networks.

1.2 Our Approach

Previous studies on the structure of real networks including email networks have revealed interesting properties that make these networks fundamentally different from random networks. In this section, we present our approach towards understanding the structure and dynamics of email networks generated from real email traffic captured on a high-speed Internet backbone link. We also present the distinguishing characteristics of legitimate and unsolicited email communications, as well as how these differences can lead to identification of the antisocial nodes in the network that are sending spam.

1.2.1 Data Collection and Processing

In order to study the characteristics of email traffic, we have collected two large-scale email datasets by passively capturing traffic on a 10 Gbps backbone link of SUNET (the Swedish University Network) [1]. Each dataset was collected over 14 consecutive days with roughly a year time span between them. The email traffic was collected by filtering packets to port 25 in both directions of the link. The collected packets which belonged to the same flow were aggregated and the email data was extracted from the flows. Then, each email communication was classified as either *rejected*, *spam*, or *ham* using a well-trained anti-spam tool to provide the ground truth for our study. Finally, the email contents were discarded and the IP addresses and the email addresses were anonymized so that all the information about the original senders, receivers, and the content of the emails are lost.

Overall, our email datasets which have the perspective of a network device monitoring traffic on a high-speed backbone link, provide us with means to characterize, model, and analyze the social network properties of email and spam traffic. In this thesis, we present the challenges involved in the collection and processing of large-scale traffic datasets, particularly the datasets used for the study of email and spam characteristics.

1.2.2 Structural and Temporal Analysis

In order to understand the characteristics of unsolicited email traffic and how it differs from legitimate traffic, we have performed a social network analysis of the captured email traffic. We have generated *email networks* from the observed email communications in which each node represents an email address and each edge represents an observed email communication between a pair of nodes. It is important to note that the aim of our study has not been to generate and model a complete social network of email communications, rather our goal is to highlight the differences in the social network properties of the legitimate and unsolicited traffic passing through a backbone link.

Based on our ground truth, we have generated a number of ham, spam, rejected, and complete email networks, and have studied and compared their structural and temporal properties. We have looked into the (in-/out-)degree distribution, average shortest path length, average clustering coefficient, distribution of the size of the connected components, the percentage of total nodes in the giant connected component, as well as how these properties change over time as the networks grow. Although the collection duration of 14 consecutive days is not long enough to study the growth and the evolution of the email networks, they still provide us with the possibility to perform a temporal analysis of the structural properties of the email networks and gain some evidence on how different properties of these networks change over time. Our study reveals the similarities and differences in the structural and temporal properties of email networks of ham and spam, and shows that the antisocial behavior of spam and rejected traffic are not hidden in the structure of complete email networks.

The analysis has also revealed that the unsolicited traffic causes deviation from the normal network-level behavior of email traffic. We have shown that this anomalous behavior can be detected by applying an anomaly detection method. Since spam is always present in the email traffic, it is needed to deploy an anomaly detection method that can point out the nodes that are the source of spam. In our study, we have used a distribution-based anomaly detection approach to spot the nodes that cause anomalies in the distributions of the structural properties of email networks. The anomalies are identified by comparing the feature distributions (i.e., degree distribution, community size distribution, and *egonet* size distribution) of current email traffic against baseline distributions generated from previously observed legitimate email traffic in our datasets.

Overall, we show how the social network analysis of our email datasets can reveal the differences in the network-level characteristics of spam and ham traffic, and show that a number of spam sending nodes can be identified based on their anomalous behavior.

1.2.3 Evaluation of Community Detection Algorithms

Our study of the social network properties of email traffic has also revealed that email networks exhibit a community structure. Despite the excessive number of studies on the quality of algorithms for community detection, there is still no consensus on which algorithm to use for which type of network. Therefore, we have conducted an empirical study to compare and evaluate a variety of community detection algorithms based on a set of structural and logical quality functions on our email networks. Our aim is to find the most suitable approach that can separate ham and spam emails from the mixture of traffic into distinct communities by merely analyzing the structure of email networks.

1.3 Contributions

1.3.1 PAPER I

Collection and processing of large-scale datasets can be very challenging. In this paper, we have described the data collection procedure and the challenges we have faced when dealing with high-speed data collection on an Internet backbone link. In particular, we have discussed the process of collecting and analyzing SMTP traffic in order to study the network-level characteristics of legitimate and unsolicited email traffic.

1.3.2 PAPER II

Social network analysis of email traffic allows us to understand the differences in the network-level behavior of legitimate and unsolicited email traffic. In this paper, we have shown that the collected legitimate email traffic exhibit similar structural properties to other social and interaction networks. Therefore, a ham network can be modeled as a scale-free small-world network. We have also shown the similarities and the differences in the structure of spam networks and how they change over time compared to ham networks and other social networks. We have also revealed that the antisocial behavior of spam is not hidden in a mixture of email traffic and causes anomalies (outliers) in the structure of email networks.

1.3.3 PAPER III

In this paper we have taken advantage of our observations that unsolicited email traffic deviate from legitimate email traffic to unveil a number of spam sending nodes in a network. We have deployed an anomaly detection technique which computes the divergence between the social network properties of observed email traffic and the properties of legitimate traffic to identify the communication patterns that do not conform to the expected normal behavior. We have used a time-series of email networks generated from traffic collected during time periods of fixed-length and have used the anomaly detection method to point out the anomalous nodes in each network. Our experiments have shown that the length of the period does not affect the performance of the anomaly detection; the percentage of spam sent by the identified anomalous nodes is highly correlated to the percentage of total spam in the network; and there is a trade-off in the number of false positives and the percentage of the total spammers that can be detected by this method.

1.3.4 PAPER IV

In this paper, we have shown that both ham and spam networks, as well as networks containing a mixture of both, exhibit a community structure, and that different community detection algorithms can be used to unfold the communities of these networks. However there is a trade-off in creating high structural quality and high logical quality communities. The structural quality of communities can be evaluated using different quality functions such as modularity, coverage, and conductance, and the logical quality can be evaluated based on the homogeneity of the edges inside each community. We have revealed that although different community detection algorithms use different approaches to define and extract the communities of a network, algorithms that create communities with similar granularity and size distribution also achieve similar structural and logical qualities. We have also shown that the most suitable approach for achieving high logical quality (i.e., clustering ham and spam emails into distinct communities) partitions the edges of the network rather than performing node-based clustering.

1.4 Conclusions and Future Work

In this thesis, we present how the social network analysis of email traffic captured on an Internet backbone link can reveal the differences and similarities in the network-level characteristics of legitimate and unsolicited email communications. We show that the different behavior of spam senders causes anomalies in the structural properties of email networks, and these anomalies can be detected using an anomaly detection approach. We also show that spam and ham, which are mixed in the observed traffic, can be separated into distinct communities by deploying a link community detection algorithm.

The proposed approaches in this thesis are promising and can potentially be used to complement existing anti-spam strategies. The advantage of deploying our approaches is that it provides us with the possibility of stopping spam closer to its source by merely using the communication patterns of the email traffic. However, there is more work to be done before our findings can be deployed practically as part of an anti-spam tool. Therefore, one desirable future direction is to investigate how our methods can be combined with each other to be used as a stand-alone anti-spam system or in corporation with existing tools. One possibility is to deploy a network device that monitors the traffic on a link and that is able to stop or tag suspicious traffic. The creation of email networks and computation of the most of the graph properties can be done quite fast by using graphics processing units (GPU) instead of CPUs [50]. Another possibility is to use the output of the traffic analysis to populate dynamic blacklists or whitelists which are to be consulted by the receiving mail servers as part of their pre-filtering process. It is known that spammers use fake email addresses, therefore, other identifiers of the detected spaming nodes such as their IP address should be used to blacklist these source of spam.

Moreover, a study of the robustness of our findings in order to see how easy it is for the spammers to change their sending behavior and how easy it is to invade detection, is another desirable future direction. It can also be of interest to use machine learning approaches in the process of anomaly detection to allow automatic adjustment to the dynamic behavior of the network traffic and improve the detection mechanism. In addition, there are many other anomaly detection methods that could be explored for the identification of the spamming nodes in the networks.

Graph-based analysis of backbone traffic by generating networks of IP addresses has been used previously in order to classify traffic [22, 25] and to identify P2P botnets [39]. An interesting future direction is to generate *IP networks* from the same datasets and study their structural properties and dynamics, as well as investigating whether the same anomaly detection or community detection approaches are relevant and can be applied on these type of networks.

Bibliography

[1] SUNET (Swedish University Network). *http://www.sunet.se/*.

[2] M. Abadi, M. Burrows, M. Manasse, and T. Wobber. Moderately hard, memory-bound functions. *ACM Transactions on Internet Technology*, 5(2):299–327, May 2005.

[3] Y.-Y. Ahn, J. P. Bagrow, and S. Lehmann. Link communities reveal multiscale complexity in networks. *Nature*, 466(7307):761–4, Aug. 2010.

[4] L. Akoglu and M. McGlohon. Oddball: Spotting Anomalies in Weighted Graphs. In *Proceedings of the 14th Pacific-Asia conference on Advances in Knowledge Discovery and Data Mining*, pages 410–421, 2010.

[5] R. Z. Albert. *Statistical Mechanics of Complex Networks*. PhD thesis, University of Notre Dame, 2001.

[6] H. Almeida, D. Guedes, W. Meira Jr., and M. J. Zaki. Is There a Best Quality Metric for Graph Clusters? In D. Gunopulos, T. Hofmann, D. Malerba, and M. Vazirgiannis, editors, *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD*, pages 44–59. Springer-Verlag, 2011.

[7] A. Barabási and R. Albert. Emergence of Scaling in Random Networks. *Science*, 286(5439):509, 1999.

[8] R. Beverly. Exploiting Transport-Level Characteristics of Spam. *5th Conference on Email and Anti-Spam (CEAS)*, 2008.

[9] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast Unfolding of Communities in Large Networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, Oct. 2008.

[10] P. Boykin and V. Roychowdhury. Leveraging social networks to fight spam. *Computer*, 38(4):61–68, Apr. 2005.

[11] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener. Graph Structure in the Web. *Computer networks*, 33(1):309–320, 2000.

[12] Z. Duan, K. Gopalan, and X. Yuan. Behavioral Characteristics of Spammers and Their Network Reachability Properties. In *2007 IEEE International Conference on Communications*, pages 164–171. IEEE, June 2007.

[13] H. Ebel, L.-I. Mielsch, and S. Bornholdt. Scale-free topology of e-mail networks. *Physical Review E*, 66(3):1–4, Sept. 2002.

[14] P. Erdos and A. Renyi. On The Evolution of Random Graphs. *Publication of the Mathematical Institute of the Hungarian Academy of Science*, 5:17–61, 1960.

[15] T. Evans and R. Lambiotte. Line graphs, link partitions, and overlapping communities. *Physical Review E*, 80(1):1–8, July 2009.

[16] M. Faloutsos, P. Faloutsos, and C. Faloutsos. On Power-Law Relationships of the Internet Topology. In *ACM SIGCOMM Computer Communication Review*, volume 29, pages 251–262. ACM, 1999.

[17] S. Fortunato. Community detection in graphs. *Physics Reports*, 486(3-5):75–174, Feb. 2010.

[18] L. H. Gomes, R. B. Almeida, and L. M. A. Bettencourt. Comparative Graph Theoretical Characterization of Networks of Spam and Legitimate Email. In *Conference on Email and Anti-Spam (CEAS)*, 2005.

[19] L. H. Gomes, C. Cazita, J. M. Almeida, V. Almeida, and W. Meira. Workload models of spam and legitimate e-mails. *Performance Evaluation*, 64(7-8):690–714, Aug. 2007.

[20] G. Gu, R. Perdisci, J. Zhang, and W. Lee. BotMiner: Clustering Analysis of Network Traffic for Protocol- and Structure-Independent Botnet Detection. In *the 17th conference on Security symposium*, pages 139–154. USENIX Association, 2008.

[21] E. Harris. The Next Step in the Spam Control War: Greylisting. *http://projects.puremagic.com/greylisting/whitepaper.html*, 2003.

[22] M. Iliofotou, H.-c. Kim, M. Faloutsos, M. Mitzenmacher, P. Pappu, and G. Varghese. Graption: A Graph-Based P2P Traffic Classification Framework for the Internet Backbone. *Computer Networks*, 55(8):1909–1920, June 2011.

[23] C. Kanich, C. Kreibich, K. Levchenko, B. Enright, G. M. Voelker, V. Paxson, and S. Savage. Spamalytics: An Empirical Analysis of Spam Marketing Conversion. In *Proceedings of the 15th ACM conference on Computer and communications security - CCS '08*, page 3, New York, New York, USA, 2008. ACM Press.

[24] T. Karagiannis, C. Gkantsidis, D. Narayanan, and A. Rowstron. Hermes: Clustering Users in Large-Scale E-mail Services. In *Proceedings of the 1st ACM symposium on Cloud computing - SoCC '10*, page 89, New York, New York, USA, 2010. ACM Press.

[25] T. Karagiannis, K. Papagiannaki, and M. Faloutsos. BLINC: Multilevel Traffic Classification in the Dark. In *Proceedings of the 2005 conference on Applications, technologies, architectures, and protocols for computer communications - SIGCOMM '05*, page 229, New York, New York, USA, 2005. ACM Press.

[26] T. Karagiannis and M. Vojnovic. Behavioral Profiles for Advanced Email Features. In *Proceedings of the 18th international conference on World wide web - WWW '09*, page 711, New York, New York, USA, 2009. ACM Press.

[27] G. Kossinets and D. J. Watts. Empirical Analysis of an Evolving Social Network. *Science (New York, N.Y.)*, 311(5757):88–90, Jan. 2006.

[28] C. Kreibich, C. Kanich, and K. Levchenko. Spamcraft: An inside look at spam campaign orchestration. In *the 2nd USENIX conference on Large-scale exploits and emergent threats: botnets, spyware, worms, and more*, number September. USENIX Association, 2009.

[29] C. Kreibich, C. Kanich, K. Levchenko, B. Enright, G. M. Voelker, V. Paxson, and S. Savage. On the Spam Campaign Trail. In *Proceedings of the 1st Usenix Workshop on Large-Scale Exploits and Emergent Threats*, volume 453, pages 697–8, June 2008.

[30] H.-y. Lam and D.-y. Yeung. A Learning Approach to Spam Detection based on Social Networks. In *Conference on Email and Anti-Spam (CEAS)*, 2007.

[31] A. Lancichinetti, M. Kivelä, J. Saramäki, and S. Fortunato. Characterizing the Community Structure of Complex Networks. *PloS one*, 5(8):e11976, Jan. 2010.

[32] J. Leskovec, J. Kleinberg, and C. Faloutsos. Graphs Over Time: Densification Laws, Shrinking Diameters and Possible Explanations. In *Proceeding of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining - KDD '05*, page 177, New York, New York, USA, 2005. ACM Press.

[33] J. Leskovec, J. Kleinberg, and C. Faloutsos. Graph Evolution: Densification and Shrinking Diameters. *ACM Transactions on Knowledge Discovery from Data*, 1(1):2–es, Mar. 2007.

[34] J. Leskovec, K. J. Lang, and M. Mahoney. Empirical Comparison of Algorithms for Network Community Detection. In *Proceedings of the 19th international conference on World wide web*, page 631, New York, New York, USA, 2010. ACM Press.

[35] M. McGlohon, S. Bay, M. G. Anderle, D. M. Steier, and C. Faloutsos. SNARE: A Link Analytic System for Graph Labeling and Risk Detection. In *Proceedings*

of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD ’09, page 1265, New York, New York, USA, 2009. ACM Press.

[36] S. Milgram. The Small World Problem. *Psychology today*, 2:60–67, 1967.

[37] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee. Measurement and analysis of online social networks. In *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement - IMC ’07*, page 29, New York, New York, USA, 2007. ACM Press.

[38] A. Mislove, A. Post, and P. Druschel. Ostra: Leveraging trust to thwart unwanted communication. In *Proceedings of the 5th USENIX Symposium on Networked Systems Design and Implementation*, number i, pages 15–30, 2008.

[39] S. Nagaraja, P. Mittal, C.-y. Hong, M. Caesar, and N. Borisov. BotGrep : Finding P2P Bots with Structured Graph Analysis. In *Proceedings of the 19th USENIX conference on Security*. USENIX Association, 2010.

[40] A. Nanavati, R. Singh, D. Chakraborty, K. Dasgupta, S. Mukherjea, G. Das, S. Gurumurthy, and A. Joshi. Analyzing the Structure and Evolution of Massive Telecom Graphs. *Knowledge and Data Engineering, IEEE Transactions on*, 20(5):703–718, 2008.

[41] M. Newman and J. Park. Why social networks are different from other types of networks. *Physical Review E*, 68(3), Sept. 2003.

[42] M. E. J. Newman. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences of the United States of America*, 103(23):8577–82, June 2006.

[43] A. Pathak, Y. C. Hu, and Z. M. Mao. Peeking into Spammer Behavior from a Unique Vantage Point. In *Proceedings of the 1st Usenix Workshop on Large-Scale Exploits and Emergent Threats*, pages 3:1—-3:9. USENIX Association, 2008.

[44] A. Ramachandran and N. Feamster. Understanding the Network-Level Behavior of Spammers. In *Proceedings of the 2006 conference on Applications, technologies, architectures, and protocols for computer communications - SIGCOMM ’06*, page 291, New York, New York, USA, 2006. ACM Press.

[45] A. Ramachandran, N. Feamster, and S. Vempala. Filtering Spam with Behavioral Blacklisting. In *Proceedings of the 14th ACM conference on Computer and communications security - CCS ’07*, page 342, New York, New York, USA, 2007. ACM Press.

[46] M. Rosvall and C. T. Bergstrom. Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences of the United States of America*, 105(4):1118–23, Jan. 2008.

[47] M. Rosvall and C. T. Bergstrom. Multilevel compression of random walks on networks reveals hierarchical organization in large integrated systems. *PloS one*, 6(4):e18209, Jan. 2011.

[48] S. E. Schaeffer. Graph Clustering. *Computer Science Review*, 1(1):27–64, Aug. 2007.

[49] D. Schatzmann, M. Burkhart, and T. Spyropoulos. Inferring Spammers in the Network Core. In *Proceedings of the 10th International Conference on Passive and Active Network Measurement*, pages 229–238. Springer-Verlag, 2009.

[50] J. Soman, K. Kishore, and P. J. Narayanan. A Fast GPU Algorithm for Graph Connectivity. In *2010 IEEE International Symposium on Parallel & Distributed Processing, Workshops and Phd Forum (IPDPSW)*, pages 1–8. IEEE, Apr. 2010.

[51] C.-Y. Tseng and M.-S. Chen. Incremental SVM Model for Spam Detection on Dynamic Email Social Networks. *2009 International Conference on Computational Science and Engineering*, pages 128–135, 2009.

[52] S. VAN Dongen. *Graph Clustering by Flow Simulation*. PhD thesis, University of Utrecht, The Netherlands, 2000.

[53] B. Viswanath, A. Post, K. P. Gummadi, and A. Mislove. An Analysis of Social Network-Based Sybil Defenses. In *Proceedings of the ACM SIGCOMM 2010 conference*, page 363, New York, New York, USA, 2010. ACM Press.

[54] M. Walfish, J. D. Zamfirescu, H. Balakrishnan, D. Karger, and S. Shenker. Distributed Quota Enforcement for Spam Control. In *Proceedings of the 3rd conference on Networked Systems Design & Implementation*. USENIX Association, 2006.

[55] D. J. Watts and S. H. Strogatz. Collective Dynamics of 'Small-World' Networks. *Nature*, 393(6684):440–2, June 1998.

[56] J. Xie, S. Kelley, and B. Szymanski. Overlapping community detection in networks: the state of the art and comparative study. In *Arxiv preprint arXiv:1110.5813*, number November, 2011.

[57] Y. Xie, F. Yu, K. Achan, E. Gillum, M. Goldszmidt, T. Wobber, C. C. Communication, and N. Network. How Dynamic are IP Addresses? In *Proceedings of the 2007 conference on Applications, technologies, architectures, and protocols for computer communications (SIGCOMM'07)*, pages 301–312. ACM, 2007.

[58] S. Yoo, Y. Yang, F. Lin, and I.-c. Moon. Mining social networks for personalized email prioritization. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '09*, page 967, New York, New York, USA, 2009. ACM Press.

[59] H. Yu, M. Kaminsky, P. B. Gibbons, and A. Flaxman. SybilGuard: Defending Against Sybil Attacks via Social Networks. In *Proceedings of the 2006 conference on Applications, technologies, architectures, and protocols for computer communications - SIGCOMM '06*, number September, pages 267–278, New York, New York, USA, 2006. ACM Press.

[60] W. Zachary. An information flow model for conflict and fission in small groups. *Journal of Anthropological Research*, 33:452–473, 1977.