Towards Modeling Legitimate and Unsolicited Email Traffic Using Social Network Properties

Farnaz Moradi Tomas Olovsson Philippas Tsigas
Computer Science and Engineering
Chalmers University of Technology, Göteborg, Sweden
{moradi,tomasol,tsigas}@chalmers.se

Abstract

Identifying unsolicited email based on their network-level behavior rather than their content have received huge interest. In this study, we investigate the social network properties of large-scale *email networks* generated from real email traffic to reveal the properties that are indicative of spam as opposed to the expected legitimate behavior.

By analyzing the structural and temporal properties of the email networks we confirm that legitimate email traffic generates a small-world, scale-free network similar to other social networks. However, email traffic as a whole contains unsolicited email, thus the structure of email networks deviates from that of social networks. Our study points out the distinctive characteristics of spam traffic and reveals that the anomalies in the structural properties of email networks are due to the unsocial behavior of spam.

Categories and Subject Descriptors C.2.3 [*Network Operations*]: Network Monitoring; C.2.2 [*Network Protocols*]: Applications (SMTP, FTP, etc.)

General Terms Measurement

Keywords Email networks, social network properties, spam

1. Introduction

Eliminating the excessive amount of unsolicited *spam* which is consuming network and mail server resources is quite challenging. These email communications are mostly originated from botnets of compromised machines [8, 15] that are also likely the source of other malicious activities on the Internet. Although current anti-spam tools are efficient in hiding spam from users' mailboxes, there is a clear need for moving the defense against spam as close to its source as possible. Therefore, it is necessary to understand the network-level behavior of spam and how it differs from legitimate traffic in order to design anti-spam mechanisms that can identify spamming bots on the network. In this paper, we study the network-level behavior of email by examining real email traffic captured on an Internet backbone link. From the collected traffic, we have generated *email networks* in which the nodes represent email addresses and the edges represent email communications. To the best of our knowledge, this is the largest email traffic dataset used to study the structure of email networks which contain both legitimate (*ham*) and unsolicited email traffic.

In this study, we show that the legitimate email traffic exhibit the same structural properties that other social and interaction networks (e.g., on-line social networks, the Internet topology, the Web, and phone call graphs) typically exhibit, therefore, it can be modeled as a *scale-free*, *small-world* network. We also show that the email traffic containing spam cannot be modeled similarly, and because the unsocial behavior of spam is not hidden behind the social behavior of legitimate traffic, the structure of email networks containing both ham and spam differ from other social networks. Moreover, we show that the temporal variations in the social network properties of email traffic can reveal more distinct properties of the behavior of spam.

In this study our goal is to identify the differences in the social network properties of spam and ham traffic, and leverage these differences to spot the abusive nodes in the network.

The remainder of this paper is organized as follows. Section 2 presents the related works. The collected email datasets and their properties are discussed in Section 3. Section 4 presents and discusses the observed structural and temporal properties of our email networks. Section 5 presents a method to spot spam senders in the structure of email networks. Finally, Section 6 concludes the paper.

This is a pre-proceeding version of the paper. The proceedings of the *SNS'12* is published by ACM.

Table 1. Summary of the datasets of related works in comparison to our datasets.

Reference	Nodes $\|V\|$	Edges $\|E\|$	Email types	Dataset
Ebel et al. [5] (2002)	59,812	86,130	ham	log files of the mail server at Kiel University
Gomes et al. [7] (2005)	265,144	615,102	ham & spam	log files of a university mail server in Brazil
Boykin et al. [2] (2005)	-	-	ham & spam	headers of emails in one user's inbox
Lam et al. [10] (2007)	9,150	-	ham & simulated spam	Enron dataset and simulated spam
Tseng et al. [17] (2009)	637,064	2,865,633	ham & spam	a mail server in National Taiwan University
Leskovec et al. [11] (2007)	35,756	123,254	ham	emails of a EU research institution
Kossinets et al. [9] (2006)	43,553	*14,584,423	ham	emails at a large university
This paper, *dataset A*	10,544,647	21,562,306	ham & spam	Internet backbone SMTP traffic
This paper, *dataset B*	4,525,687	8,709,216	ham & spam	Internet backbone SMTP traffic

* Total number of emails exchanged during 355 days (separate graphs within time windows of 60 days)

Table 2. Statistics of the collected data for *dataset A*.

	Packets	Flows	Email	Ham	Spam	Rejected	Senders[1]	Receivers[1]	Domains[2]
Incoming	$626.9M$	$34.9M$	19,302,206	1,319,273	1,663,698	16,319,235	7,780,897	3,169,712	446,694
Outgoing	$170.1M$	$11.9M$	729,553	213,306	202,879	313,368	324,657	408,429	167,907

[1] Distinct email addresses. [2] Distinct domain names in email addresses.

2. Related Work

Social network analysis has been widely used in order to study the structural properties of real-world networks such as the Web graph [3], the Internet topology [6], phone call and SMS networks [14], and online social networks [12]. The structure of email networks was first studied by Ebel. et al. [5] showing that an email network generated from mail server log files of a university is a scale-free, small-world network. Leskovec et al. [11] studied the evolution of a variety of real networks, including an email network of a large institution, and observed that these social networks densify over time and their diameter shrinks, while their power law degree distribution exponent remains constant.

Deployment of social network analysis for discriminating spammers and legitimate users was first proposed in Boykin et al. [2]. They generated an email network from email headers in one user's mailbox and found distinguishing structural properties of spam and ham messages. Gomes et al. [7] generated distinct graphs from ham and spam email collected from mail server log files of their university department, and found graph theoretical metrics that structurally and dynamically differ for spam and ham. Lam et al. [10] and Tseng et al. [17] extracted different structural features from email social networks and deployed them in building learning-based spam detection systems.

Table 1 summarizes the properties of the email networks studied in the related works. All of the above studies have taken place on relatively limited email datasets. In addition to previous studies, we perform an analysis of the structural and temporal characteristics of email networks, reveal properties that distinguish ham from spam, compare our observations with previous studies, and show how our findings could reveal the spam sending nodes in the email networks.

3. Data Collection and Pre-processing

In this study we have used two distinct email datasets to generate email networks. The datasets were created from passively captured SMTP packets on a 10 Gbps link of the core-backbone of the SUNET[1]. Each dataset was collected during 14 consecutive days with a year time span between the collections. Throughout the paper, we refer to the larger dataset as *dataset A*, and the smaller dataset as *dataset B*.

The unusable email flows, including those with no payload or missing packets and encrypted communications were pruned from the datasets. The remaining emails were first classified as being either *accepted* (delivered by the receiving mail server) or *rejected* (unfinished SMTP command exchange phase and consequently not containing any email headers and body). Rejection is generally the result of spam pre-filtering strategies deployed by mail servers (e.g., blacklisting, greylisting, DNS lookups). Then, all accepted email communications were classified to be either *spam* or *ham* to establish a ground truth for our study. Similar to [7, 17], the classification was done by a well-trained filtering tool[2]. Finally, all email addresses were anonymized and email contents were discarded in order to preserve privacy.

After data collection and pre-processing, a number of email networks have been generated from the datasets. In an email network the email addresses, which are extracted from the SMTP commands ("MAIL FROM" and "RCPT TO"), represent the nodes, and the exchanged emails represent the edges. In order to study and compare the characteristics of different categories of email, from each dataset we have generated a *ham network*, a *spam network*, and a *rejected network*, in addition to the complete *email network*.

Table 2 summarizes the properties of the dataset *A* as an example. More details on the measurement location, data collection, and pre-processing can be found in [13].

[1] Swedish University Network (`http://www.sunet.se/`) serves as a backbone for university traffic, student dormitories, research institutes, etc. exchanging large amount of traffic with commercial companies.

[2] The SpamAssassin (`http://spamassassin.apache.org`) was in use for a long time in our University mail server and it incurs a false positive rate of less than 0.1%, and the detection rate of 91.4% after 94% of the spam being rejected by blacklists.

4. Structural and Temporal Properties of Email Networks

In this section we briefly introduce the most significant structural and temporal properties of social networks.

Degree distribution. The degree distribution of a network is the probability that a randomly selected node has k edges. In a *power law distribution*, the fraction of nodes with degree k is $n(k) \propto k^{-\gamma}$, where γ is a constant exponent. Networks characterized by such degree distribution are called *scale-free* networks. Many real networks such as the Internet topology [6], the Web [3], phone call graphs [14], and on-line social networks [12] are scale free.

Average path length. In *small-world* networks any two nodes in the network are likely to be connected through a short sequence of intermediate nodes, and the network diameter shrinks as the network grows [11].

Clustering coefficient. In addition to a short average path length, *small-world* networks have high clustering coefficient values [18]. The clustering coefficient of a node v is defined as $C_v = 2E_v/(k_v(k_v-1))$, where, k_v denotes the number of neighbors of v, $k_v(k_v-1)/2$ the maximum number of edges that can exist between the neighbors, and E_v the number of the edges that actually exist. The average C_v of a social network shows to what extent friends of a person are also friends with each other and its value is independent of the network size [16].

Connected components. A connected component (*CC*) is a subset of nodes of the network where a path exists between any pair of them. As social networks grow a giant CC (GCC), which contains the vast majority of the nodes in the network, emerges in the graph and its size increases over time [16]. Moreover, the distribution of CC size for some social networks follows a power law pattern [3, 14].

4.1 Measurement Results

In the following the observed structural and temporal properties of our email networks are presented. These properties can be used in order to model the behavior of legitimate traffic and to find the distinguishing properties of the unsocial behavior of spam. Although the duration of our data collections is not long enough to study the evolution of email networks, it is still possible to track the changes in the structure of email networks in order to better understand the distinct characteristics of ham and spam traffic.

Degree distribution. Figures 1(a) and 1(e) show that none of the email networks generated from datasets *A* and *B* exhibit a power law degree distribution in all points. However, the ham networks generated from each of the datasets are scale free as their degree distribution closely follow the distribution $n(k) \propto k^{-\gamma}$ with $\gamma_A = 2.7$ and $\gamma_A = 2.3$, respectively [3]. The in-degree (out-degree) distribution for ham networks, which are shown in Figures 1(b) and 1(f), also follows a power-low distribution with $\gamma_{A_{in}} = 3.2$ ($\gamma_{A_{out}} = 2.3$) and $\gamma_{B_{in}} = 3.2$ ($\gamma_{B_{out}} = 2.1$), respectively. Moreover, in contrast to previous studies [2, 7], neither the spam, nor the rejected networks are completely scale free (Figures 1(c), 1(g), 1(d), and 1(h)).

Figure 2(a) and 2(e) show that the shape of the degree distributions of the complete email networks may change over time as the networks grow. The shape of the degree distribution of spam and rejected networks can also change over time (Figures 2(c), 2(g), 2(d), and 2(h)). However, the ham networks always follow a power law distribution with an almost constant exponent (Figures 2(b) and 2(f)).

Clustering coefficient. The observed average clustering coefficients for our ham (spam) networks generated from both dataset are quite similar: $C_{A_{ham}} = 9.92 \times 10^{-3}$ ($C_{A_{spam}} = 1.59 \times 10^{-3}$) and $C_{B_{ham}} = 9.80 \times 10^{-3}$ ($C_{B_{spam}} = 1.54 \times 10^{-3}$). These values, similar to small-world networks, are significantly greater than that of random networks with the same number of nodes and average number of edges per node, and as Figures 3(b) and 3(f) show they remain relatively constant as the networks grow.

Average path length. The ham and spam networks generated from both datasets have short average path lengths, $\langle l \rangle$, as expected in small-world networks: $\langle l_{ham_A} \rangle = 7.0$, $\langle l_{spam_A} \rangle = 8.5$, $\langle l_{ham_B} \rangle = 6.7$, and $\langle l_{spam_B} \rangle = 7.8$. Figures 3(a) and 3(e) show that $\langle l \rangle$ decreases for all networks as they grow, confirming the shrinking diameter phenomenon observed in [11] for other social networks.

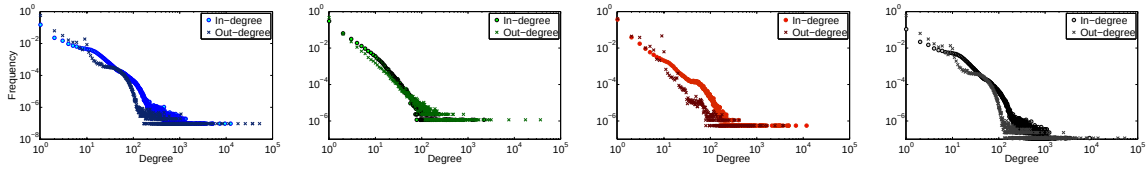
Connected components. Figure 4.2 shows the distribution of the size of the CCs for ham and spam networks. It can be seen that the GCCs of the networks are orders of magnitude larger than other CCs. The distribution of the CC size for the ham network, similar to Web [3] and phone call graphs [14], follows a power law pattern, but the spam network can have outliers in their distribution. Figures 3(d) and 3(h) show that the number of CCs in all of the ham and the spam networks increases over time, but this increase is much faster for spam. Moreover, as shown in Figure 3(c), the respective size of the GCC of the networks generated from dataset *A* increases for the ham but does not change much for the spam network. However, although the ham network generated from dataset *B* shows exactly the same behavior (Figure 3(g)), the spam network shows an increase in the percentage of nodes in its GCC over time.

4.2 Discussion

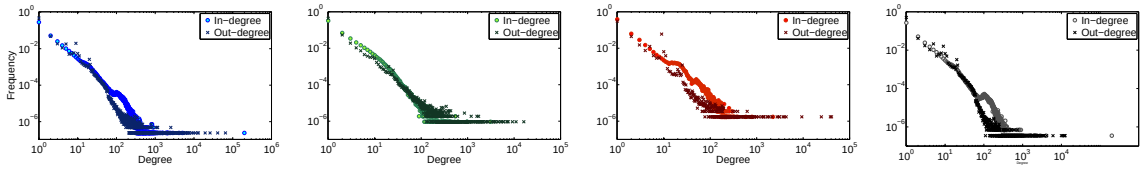
In the following paragraphs we briefly discuss our observations regarding the structure of email networks and discuss to what extent our dataset is representative for the structural and temporal analysis of email networks.

Table 3 summarizes the observed similarities and differences in the structure of the ham and spam networks. Although the studied datasets differ in size and collection time,

[3] The power law fits were calculated using the Maximum Likelihood estimator for power law and Kolmogorov-Smirnov (KS) goodness-of-fit as presented in [4].

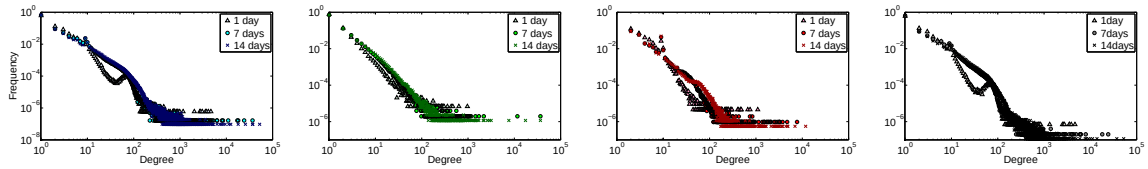


(a) Email network (*A*) (b) Ham network (*A*) (c) Spam network (*A*) (d) Rejected network (*A*)

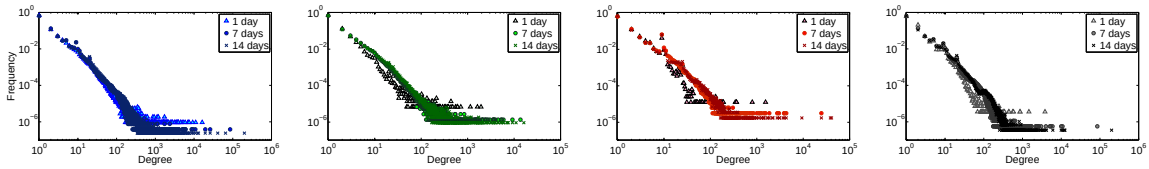


(e) Email network (*B*) (f) Ham network (*B*) (g) Spam network (*B*) (h) Rejected network (*B*)

Figure 1. Only the ham network is scale free as the other networks have outliers in their degree distribution.



(a) Email network (*A*) (b) Ham network (*A*) (c) Spam network (*A*) (d) Rejected network (*A*)



(e) Email network (*B*) (f) Ham network (*B*) (g) Spam network (*B*) (h) Rejected network (*B*)

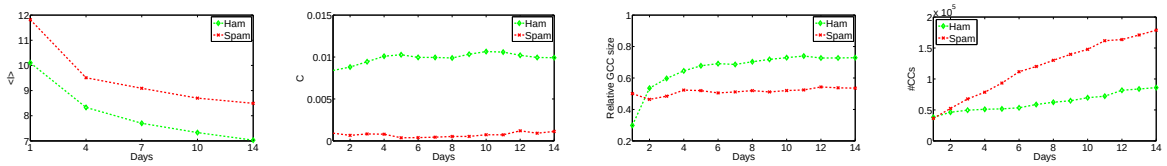
Figure 2. Temporal variation of in the degree distribution of the email networks.

our observations reveal that legitimate email always exhibit the structural properties that are similar to other social and interaction networks. Previous studies on the structure of legitimate email networks have also shown that these networks can be modeled as scale free, small-world networks [2, 5, 7, 9, 11]. In contrast, a vast majority of spam are automatically sent, typically from botnets, and it is expected that they show unsocial behavior. We have shown that the structural and temporal properties of spam networks can reveal their anomalous nature. Although spam networks show some properties that are similar to ham (i.e., small-world network properties), they can still be distinguished from ham networks as they have significantly smaller average clustering coefficient and larger average path length, regardless of the size of the networks. Overall, we have shown that although the behavior of spam might change over time, its unsocial behavior is not hidden in the mixture of email traffic, even when the amount of spam is less than ham (dataset *B*).

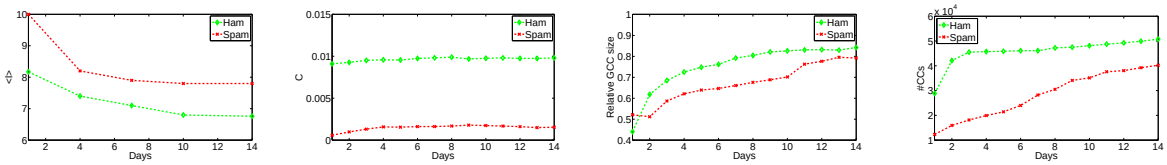
The datasets used in this study to analyze the characteristics of spam do not contain the email communications that do not pass the measurement location. Due to asymmetric routing and load-balancing policies deployed by the network routers, not all traffic travels the link, and less traffic is seen in the outgoing than the incoming direction of the link (Table 2). However, our goal is to perform a comparative analysis of the distinguishing behavior of spam and ham traffic that are observed over the link. Therefore, it is not required to generate a complete email network of all exchanged emails to be able to study the differences in the social network properties of legitimate and spam traffic.

In addition, the "missing past" problem, which is not limited to our dataset, exists since it is not possible to gather data reaching all the way back to a network's birth. Leskovec et al. [11] showed that the effect of missing past is minor as we move away from the beginning of the data observation. We investigated the effect of missing past by constructing an email network which lacked the first week of data from dataset *A* and comparing it with the network containing both weeks. We have observed that the structural properties of the email networks was relatively similar for both of the networks particularly for the legitimate email.

Earlier studies [2, 5, 7, 9, 10, 17] have also used incomplete email networks to study the structure of email networks or to deploy a social network-based approach to mitigate



(a) Average path length (*A*) (b) Average clustering coefficient (*A*) (c) Relative GCC size (*A*) (d) Number of CCs (*A*)

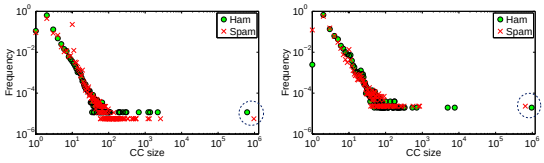


(e) Average path length (*B*) (f) Average clustering coefficient (*B*) (g) Relative GCC size (*B*) (h) Number of CCs (*B*)

Figure 3. Both networks are small-world networks (a,b,e,f), however, ham has a higher average clustering coefficient. The ham networks become more connected over time (c,g), and the number of CCs increases faster for the spam networks (d,h).

Table 3. Structural properties of the ham and the spam networks.

Dataset	Network	Nodes	Edges	C	$\langle l \rangle$	relative GCC size	No. CCs	γ degree distribution
A	Ham	859,623	1,060,380	9.92×10^{-3}	7.0	72.90%	85,992	2.7
	Spam	1,795,197	2,506,298	1.59×10^{-3}	8.5	53.53%	178,754	-
B	Ham	1,077,042	1,593,042	9.80×10^{-3}	6.7	84.24%	50,742	2.3
	Spam	578,158	1,044,714	1.54×10^{-3}	7.8	79.21%	40,236	-



(a) CC distribution (*A*) (b) CC distribution (*B*)

Figure 4. The distribution of size of CCs. The GCCs of the networks are orders of magnitude larger than other CCs.

spam. Even though our measurement duration was shorter than previous studies [5, 7, 9, 11], we have generated the largest and most general datasets used for this type of analysis. The 14 days of data collection might not be large enough to study the evolution of email networks, but our analysis of the temporal variation in the structure of email networks provides us with evidence on how their structure might change with longer periods of measurements.

Overall, this work has provided us with very large datasets of real traffic traversing a high speed Internet backbone link. These datasets allow us to model the behavior of email traffic as observed from the vantage point of a network device on the link and reveal the differences in the network-level behavior of ham and spam traffic.

5. Anomalies in Email Network Structure

The structural properties of real networks that deviate from the expected properties for social networks, suggest anomalous behavior in the network [1]. In this section, we show that the anomalies caused by the unsocial behavior of spam can be detected in the email networks by using an outlier detection mechanism.

We have shown in Section 4 that the ham networks exhibit power law out-degree distributions with $\gamma_{A_{out}=2.3}$ and $\gamma_{B_{out}=2.1}$ for dataset *A* and *B* respectively. The outliers in the out-degree distribution of the email networks are of particular importance, as we are interested in finding the nodes that send spam.

Procedure 1 presents the process of detecting outliers from the out-degree distribution. First the ratio of the out-degree distribution of the email network, containing both ham and spam, and our model is calculated. Then the Median Absolute Deviation (MAD) method is deployed to calculate the median of the absolute differences of the obtained ratios from their median. The nodes in the network that have an out-degree that deviates a lot (based on a threshold value) from the median are marked as outliers.

Table 4 shows the percentage of spam that were sent in different networks and the percentage of spam sent by the identified outlier nodes. The nodes in the email networks generated from dataset *A* (*B*) have sent in average around 70% (40%) spam and the identified outlier nodes have sent just slightly more spam than the average node. The reason is that the outlier detection method tends to mark both nodes that have sent only one email and those that have sent a large number of email as outliers. However, we have observed that the nodes which have sent only one email had sent ham and spam with the same probability, and the nodes with high out-degree have mostly sent legitimate email (e.g., mailing lists). By excluding the nodes that have a high out-degree (100 in our experiments) from the outliers as well as the nodes that have only sent one email during the collection period, we can see that more than 95% (81%) of the email sent by the identified outliers in dataset *A* (*B*) have actually been spam.

Procedure 1 Finding out-degree distribution outliers

OUTLIERS_DETECTION(G)

$G\_odd \leftarrow$ out-degree distribution for graph G
$M\_odd \leftarrow Ck^{-\gamma}$ (the power law distribution model)
$r \leftarrow$ the ratio between G_odd and M_odd
$m \leftarrow MAD(r)$
for all nodes $v \in G$ **do**
  if $r(k_v) > m \times threshold$ **then**
    add v to the list of $outliers$
  end if
end for

Table 4. Percentage of total spam, spam sent by all the identified outlier nodes, and those with degree between one and 100, in email networks containing both ham and spam.

Dataset	Network	Total spam	Spam sent by outliers	Spam sent by outliers with $1 < k < 100$
A	1 day	68%	69.9%	95.5%
	7 days	70%	74.0%	96.8%
	14 days	70%	74.8%	96.9%
B	1 day	40%	43.6%	82.7%
	7 days	35%	42.8 %	81.3%
	14 days	39 %	46.7%	87.3%

Moreover, these nodes have actually sent around 25% (35%) of the total spam in the network.

The outliers in the out-degree distribution of the complete email network which in addition to ham and spam contains rejected email can be identified similarly. As an example, the nodes in the complete email network generated from one day of email traffic in dataset *A* have sent in average 94.8% spam and rejected email. The emails sent by the outlier nodes detected by our method have been 99.3% spam or rejected.

6. Conclusions

In this study we have investigated the social network properties of email networks to study the characteristics of legitimate and unsolicited emails. The email networks were generated from real email traffic which was captured on an Internet backbone link. We have analyzed the structural and temporal properties of the email networks and have shown that legitimate email traffic generates a small-world, scale-free network that can be modeled similar to many other social networks. Moreover, the unsocial behavior of spam, which might change over time, is not hidden in the mixture of email traffic. Therefore, email networks that contain spam do not exhibit all properties commonly present in social networks.

Moreover, we have shown that by identifying the anomalies in the structural properties of email networks, it is possible to reveal a number of abusive nodes in the network. More specifically, we have shown that the outliers in the out-degree distribution of email networks to a large extent represent the spamming nodes in the network. Therefore, the social network properties of email networks can potentially be used to detect malicious hosts on the network.

Acknowledgments. This work was supported by .SE – The Internet Infrastructure Foundation and SUNET. The research leading to these results has also received funding from the European Union Seventh Framework Programme (FP7/ 2007-2013) under grant agreement no. 257007.

References

[1] L. Akoglu, M. McGlohon, and C. Faloutsos. OddBall: Spotting Anomalies in Weighted Graphs. In *PAKDD*, 2010.

[2] P. O. Boykin and V. P. Roychowdhury. Leveraging social networks to fight spam. *Computer*, 38(4), 2005.

[3] 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(1-6), 2000.

[4] A. Clauset, C. R. Shalizi, and M. E. J. Newman. Power-law distributions in empirical data. *SIAM Reviews*, June 2007.

[5] H. Ebel, L. Mielsch, and S. Bornholdt. Scale-free topology of e-mail networks. *Physical Review E*, 66, 2002.

[6] M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationships of the internet topology. *SIGCOMM Comput. Commun. Rev.*, 29, 1999.

[7] L. H. Gomes, R. B. Almeida, L. M. A. Bettencourt, V. Almeida, and J. M. Almeida. Comparative graph theoretical characterization of networks of spam and legitimate email. In *Proc. of the Conference on Email and Anti-Spam*, 2005.

[8] C. Kanich, C. Kreibich, K. Levchenko, B. Enright, G. M. Voelker, V. Paxson, and S. Savage. Spamalytics: an empirical analysis of spam marketing conversion. *Proc. of the ACM conf. on computer and communications security*, 52(9), 2009.

[9] G. Kossinets and D. J. Watts. Empirical analysis of an evolving social network. *Science*, 311(5757), 2006.

[10] H. Lam and D. Yeung. A learning approach to spam detection based on social networks. In *Proceedings of the Conference on Email and Anti-Spam*, 2007.

[11] J. Leskovec, J. Kleinberg, and C. Faloutsos. Graph evolution: Densification and shrinking diameters. *ACM Transactions on Knowledge Discovery Data*, 1(1), 2007.

[12] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee. Measurement and analysis of online social networks. In *IMC*, 2007.

[13] F. Moradi, M. Almgren, W. John, T. Olovsson, and P. Tsigas. On collection of large-scale multi-purpose datasets on internet backbone links. In *Proc. of Building Analysis Datasets and Gathering Experience Returns for Security Workshop*, 2011.

[14] A. 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. *IEEE Trans. on Knowledge & Data Engineering*, 20(5), 2008.

[15] A. Ramachandran and N. Feamster. Understanding the network-level behavior of spammers. *SIGCOMM Comput. Commun. Rev.*, 36, 2006.

[16] A. Reka and Barabási. Statistical mechanics of complex networks. *Rev. Mod. Phys.*, 74:47–97, June 2002.

[17] C. Tseng and M. Chen. Incremental SVM model for spam detection on dynamic email social networks. In *Conf. on Computational Science and Engineering*, 2009.

[18] D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684), 1998.