Technical Report no. 2011-18

Structural and Temporal Properties of E-mail and Spam Networks

Farnaz Moradi *Tomas Olovsson* *Philippas Tsigas*

CHALMERS | GÖTEBORG UNIVERSITY

Department of Computer Science and Engineering
Chalmers University of Technology and Göteborg University
SE-412 96 Göteborg, Sweden

Göteborg, 2011

Technical Report in Computer Science and Engineering at
Chalmers University of Technology and Göteborg University

Technical Report no. 2011-18
ISSN: 1652-926X

Department of Computer Science and Engineering
Chalmers University of Technology and Göteborg University
SE-412 96 Göteborg, Sweden

Göteborg, Sweden, 2011

Abstract

In this paper we present a large-scale measurement study and analysis of e-mail traffic collected on an Internet backbone link. To the best of our knowledge this is one of the largest studies of network-wide behavior of e-mail traffic. We consider e-mail networks connecting senders and receivers that have communicated via e-mail, capturing their social interactions. Our study focuses on temporal and structural properties of these e-mail networks.

By analyzing the structural properties of e-mail networks, first we confirm that legitimate e-mail traffic generates a small-world, scale-free network that can be modeled similarly to many other social networks, but we also show that, contrary to previous work, e-mail traffic as a whole does not exhibit a scale-free behavior. We show that this deviation is caused by the unsocial behavior of unsolicited e-mail traffic. We also analyze how various structural properties of e-mail networks change over time and reveal the structural properties that are indicative of the unsocial behavior of spam traffic. Finally, we show that our findings can be used to identify spam traffic in regular e-mail traffic without inspecting the e-mail contents.

Keywords: E-mail characterization, social networks, spam detection, measurement, statistical properties of e-mail networks

1 Introduction

In this paper, the behavior of e-mail traffic has been studied and analyzed by examining network traffic captured on an Internet backbone link of a national university network. From the collected packet level data, we have generated an *e-mail network*, i.e. a social network with e-mail sender and receiver addresses as vertices and edges that connect vertices that have communicated via e-mail. We have studied both the structural and the temporal properties of the generated e-mail network and compared them with properties previously observed for other types of interaction networks such as on-line social networks, the Internet topology, the web, and phone call graphs. We show that the existing models for the structure of e-mail and social networks cannot accurately model the behavior of e-mail traffic containing unsolicited e-mail (*spam*). To the best of our knowledge, this is the largest e-mail traffic dataset ever used to study the structure of e-mail networks.

We show in this paper that the difference between the structure of e-mail networks and the structure of other social and interaction networks is caused by the spam traffic. The legitimate e-mail (*ham*) traffic exhibits the same structural statistical properties that other social networks typically exhibit, therefore, a ham network can be modeled as a scale-free, small-world network. We also show that the unsocial behavior of spam cannot be hidden behind the social behavior of legitimate traffic and that spam networks cannot be modeled similar to other social networks.

We have also studied how the structure of e-mail networks change over time. Although a study of a single snapshot provides useful information about its structural properties, metrics describing the temporal variations in the structure of the networks provide us with a better understanding of the behavior of e-mail traffic. We show that a variety of structural metrics exists that can distinguish between ham and spam traffic.

Finally, we show that these distinctive structural properties can be used to detect misbehaving nodes on the network as opposed to the methods that inspect content of e-mails.

The purpose of this study is not to generate a real social network that captures the friendship relation between e-mail senders and receivers. Rather, the aim is to use existing social network-based analysis methods to understand the characteristics of e-mail communications and how they can be deployed to combat spam. Our contributions in this paper are as follows.

- We have generated e-mail networks from SMTP packets captured on a high-speed Internet backbone link.
- We have performed a structural and temporal analysis of the generated e-mail networks in order to understand their topological characteristics.
- We have (surprisingly) found that although e-mail networks are considered scale free in the literature, this is not the case, implying the need for development of new models for this type of traffic.
- We have analyzed and compared the structural and temporal properties of ham and spam networks, and have identified network-wide properties that are indicative of the unsocial behavior of spam traffic.
- We have presented and tested a novel approach to spam detection on the network level by deploying the distinguishing structural properties of ham and spam without a need to inspect and classify e-mail contents.

Table 1: Summary of datasets of related works

Reference	Year	Nodes $\|V\|$	Edges $\|E\|$	Dataset
Ebel et al. [10]	2002	59,812	86,130	log files of the mail server at Kiel University
Gomes et al. [12]	2005	265,144	615,102	log files of the mail server in a university in Brazil
Gomes et al. [13]	2009	279,504	562,664	emails arriving at a departmental mail server
Boykin et al. [6]	2005	-	-	headers of e-mail messages in one user's inbox
Lam et al. [19]	2007	9,150	-	Enron dataset* and simulated spam accounts and e-mails
Kong et al. [16]	2006	56,969	84,190	the GSCC of the dataset used in [10]
Tseng et al. [29]	2009	637,064	2,865,633	mail server of the computer center in National Taiwan University
Leskovec et al. [21]	2007	35,756	123,254	e-mails from a large European research institution
Kossinets et al. [17]	2006	43,553	**14,584,423	e-mails exchanged at a large university
This paper	2010	10,544,647	21,537,314	SMTP traffic captured on an Internet backbone link

* `http://www.cs.cmu.edu/~enron/`
** Total number of e-mails exchanged during 355 days, used to generate separate graphs within time windows of 60 days

The remainder of this paper is organized as follows. Section 2 describes the related work. The methodology for data collection and generation of e-mail networks is described in Section 3. Section 4 presents the structural properties of the e-mail networks and in Section 5 temporal variations in the structure of e-mail networks are studied. In Section 6 the possibility of deploying our findings for spam detection is evaluated and Section 7 provides a discussion of the dataset used for this study. Finally, Section 8 concludes the paper.

2 Related Work

The problem of understanding the structure of social and interaction networks such as the Internet topology [11, 22], the web [1, 3, 7], phone call graphs [25], e-mail networks [10, 17, 21], and on-line social networks [23], has received huge interest. Studies of structural properties such as the *small-world* phenomenon and the *scale-free* behavior of networks, have revealed that social and interaction networks are fundamentally different from random networks [26].

The structure of e-mail networks was first studied by Ebel. et al. [10]. They studied a network generated from mail server log files of a university and showed that this network is scale-free and exhibits properties of small-world networks. Leskovec et al. [21] studied the evolution of a variety of real networks including an e-mail network of a large research institution. They observed that e-mail networks similar to other social networks densify over

time and their diameter shrink, while their power law degree distribution exponent remain constant.

Deployment of social networks for discriminating spammers and legitimate e-mail senders was first proposed in Boykin et al. [6] and gained a lot of interest afterwards. They generated an e-mail network from e-mail headers in one user's mailbox and found distinguishing structural properties of spam and legitimate e-mail messages. Gomes et al. [12, 13] generated two distinct graphs from ham and spam e-mails collected from mail server log files of their university department, and found graph theoretical metrics that structurally and dynamically differ for spam and ham. Kong et al. [16] suggested using the topological properties of e-mail social networks to generate efficient distributed collaborative spam filters. Lam et al. [19] extracted different structural features from e-mail social networks generated from the Enron dataset and deployed them in building a learning-based spam detection method. A similar set of features was studied in Tseng et al. [29] to construct a spammer detection system based on an incremental SVM model.

Table 1 summarizes the properties of the e-mail networks studied in the previous works. All of the above studies have taken place on relatively limited e-mail datasets. To the best of our knowledge, our study is the first study of the social structure of e-mail networks on the Internet on a large scale. In this study we also perform an analysis of the structural and temporal characteristics of our e-mail networks, reveal properties that distinguish ham from spam, compare our observations with previous studies, and discuss how our findings could be deployed as a complement to existing anti-spam strategies.

3 Data Collection Methodology

In this section we describe the methodology used to collect data and generate e-mail networks.

3.1 Collection and Pre-processing

The dataset used in this paper was generated from SMTP packets captured passively on an 10 Gbps link of the core-backbone of the Swedish University Network (SUNET)[1]. During a period of 14 consecutive days in March 2010, we captured more than 797 million SMTP packets (filtered on port 25) in both directions. Then, we aggregated these packets into more than 46.8 million flows.[2] Finally, we extracted around 24.4 million e-mails and 12 million mail server replies from these flows.

The unusable e-mail flows including those with no payload (mainly due to port scanning), encrypted SMTP communications, delivery notifications, and flows missing proper SMTP commands were pruned from the dataset. Moreover, e-mail transactions missing intermediate packets were considered as incomplete since further classification of them was not possible.

The remaining e-mail transactions were first classified as being either *accepted* or *rejected*. An accepted e-mail is an e-mail delivered by the receiving mail server and contains the basic SMTP commands, e-mail headers and an e-mail body. A rejected e-mail is one that does not

[1] SUNET `http://www.sunet.se/` serves as a backbone for university traffic, a substantial number of student dormitories, research institutes, as well as some museums and government agencies and it contains a large amount of exchange traffic with commercial companies.

[2] We used the tcpflow program (`http://www.circlemud.org/~jelson/software/tcpflow/`) which understands sequence numbers and correctly handles out-of-order and retransmitted packets to reconstruct data streams.

Table 2: Dataset statistics.

	Incoming	Outgoing
Packets	626.9×10^6	170.1×10^6
Flows	34.9×10^6	11.9×10^6
Emails (Emails per recipient)	19,302,206 (25,279,555)	729,553 (1,434,159)
Ham (Ham per recipient)	1,319,273 (1,328,545)	213,306 (288,363)
Spam (Spam per recipient)	1,663,698 (2,291,813)	202,879 (574,608)
Rejected (Rejected per recipient)	16,319,235 (21,659,197)	313,368 (571,188)
Distinct sender (receiver) e-mail addresses	7,780,897 (3,169,712)	324,657 (408,429)
Distinct domains in email addresses	446,694	167,907

finish the SMTP command exchange phase and consequently does not contain any content. Rejection is generally the result of spam pre-filtering strategies deployed by mail servers including blacklisting, greylisting, DNS lookups, and user database checks.

Finally, we have classified all accepted e-mails to be either *spam* or *ham*. This allows us to establish a ground truth for our study. Similar to [12,13,29], the classification was done by a well-trained SpamAssassin filtering tool.[3] After classification, all IP and e-mail addresses were anonymized and e-mail contents were discarded in order to preserve privacy.

The SMTP replies from rejected e-mails could be used to further classify rejected traffic based on the rejection reasons. However, due to asymmetric routing[4] only approximately 10% of the collected flows contain both the e-mail and the corresponding mail server reply. Therefore, we decided to not further classify rejected e-mail transactions.

Table 2 summarizes the collected dataset. More details on the measurement location, data collection procedure, and pre-processing steps can be found in [24].

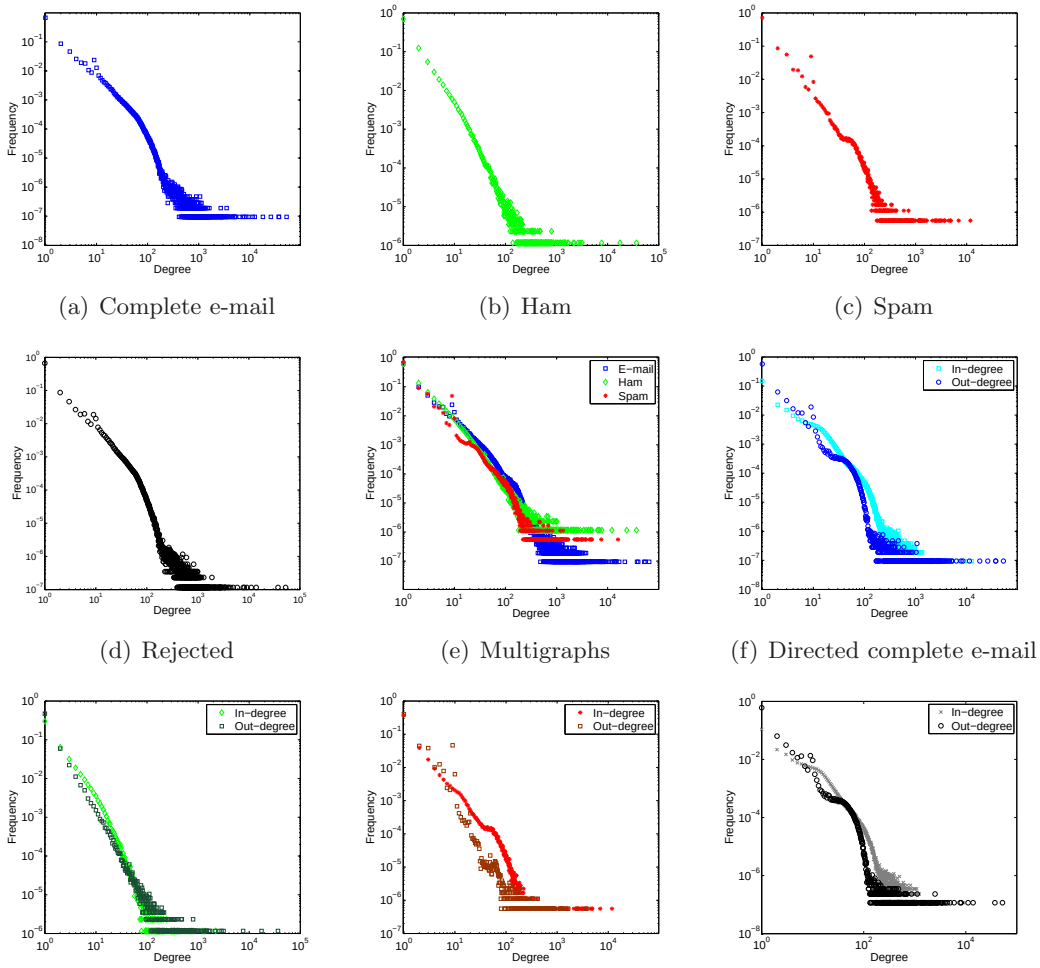
3.2 E-mail Networks

After data collection and pre-processing, an e-mail network was generated from e-mail addresses extracted from the SMTP commands (i.e. "MAIL FROM" and "RCPT TO") as vertices and exchanged e-mails between them as edges. In order to study the structural and temporal characteristics of the e-mail networks and to compare different categories of e-mail, we generated different types of e-mail networks, namely a *complete e-mail network*, a *ham network*, a *spam network*, and a *rejected network*. Only vertices participating in e-mail exchange and the edges between them were included in each network. The complete e-mail network contains all the communications including ham, spam, and rejected.

To the best of our knowledge, the complete e-mail network with 10,544,647 vertices and 21,537,314 edges, is the largest dataset used so far for structural analysis of e-mail networks.

[3]The SpamAssassin `http://spamassassin.apache.org` was in use for a long time in our University mail server and it incurs approximately 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, leading to 99.5% detection rate in total. This high detection rate allows us to use it as a ground truth for our analysis.

[4]The traffic is load balanced between alternative links which introduces routing asymmetry, as we can sometimes only see the traffic going in one direction of a TCP connection.



(a) Complete e-mail (b) Ham (c) Spam

(d) Rejected (e) Multigraphs (f) Directed complete e-mail

(g) Directed ham (h) Directed spam (i) Directed rejected e-mail

Figure 1: Degree distribution of e-mail networks. Only the ham network is scale-free.

4 Structural Characteristics of e-mail Networks

In this section we study the most significant structural properties of social networks, namely the small-world effect and the scale-free degree distributions. Our main focus is on the ham and spam networks and the characteristics that can differentiate them.

4.1 Small-World Networks

A network exhibits the small-world phenomenon, often referred to as "six degree of separation", if any two vertices in the network are likely to be connected through a short sequence of intermediate vertices [15]. In real networks, the small-world property is often associated with the presence of clustering [5]. Thus, Watts and Strogatz [30] proposed that in addition to a short average path length, small-world networks have high clustering coefficient values.

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

	Ham	Spam
Number of vertices ($\|V\|$)	859,623	1,795,197
Number of edges ($\|E\|$)	1,060,380	2,506,298
Average clustering coefficient (C)	9.92×10^{-3}	1.59×10^{-3}
Percentage of vertices in GSCC	72.90%	53.53%
Number of SCCs	85,992	178,754
Average path length ($\langle l \rangle$)	7.03	8.49
Power law exponent (γ)	2.7	-

4.1.1 Average Path Length

We have calculated the average shortest path length over the ham and the spam networks.[5] The calculations were done on the giant strongly connected component of each of the networks (see Section 4.3) where a path exists between any pair of vertices of the component [30]. Although the number of vertices in the connected components of the studied ham and spam networks are different, it is still clear that they both have short path lengths similar to small-world networks: $\langle l_{ham} \rangle = 7.03$ and $\langle l_{spam} \rangle = 8.49$.

4.1.2 Clustering Coefficient

The clustering coefficient C_v of a vertex v is given by $C_v = 2E_v/(k_v(k_v - 1))$, where, k_v denotes the number of neighbors of v, $k_v(k_v - 1)/2$ denotes the maximum number of edges that can exist between the neighbors of v, and E_v the number of the edges that actually exist.

The clustering coefficient in a social network shows to what extent friends of a person are also friends with each other. Therefore, the legitimate vertices are expected to have a larger clustering coefficient than the spamming vertices [6, 12, 29]. The observed average clustering coefficients for our ham and the spam networks which are $C_{ham} = 9.92 \times 10^{-3}$ and $C_{spam} = 1.59 \times 10^{-3}$, respectively, confirm this assumption. These observed values are also significantly greater than that of random networks with the same number of vertices and average number of edges per vertex.

4.2 Degree Distribution

The most fundamental structural property of a network is its degree distribution. The degree distribution of a network is the probability that a randomly selected vertex has k edges. In a *power law distribution*, the fraction of vertices with degree k, $n(k) \propto k^{-\gamma}$, where γ is a constant. Networks characterized by a power law degree distribution are also known as *scale-free* networks [2,11], since power laws have the property of having the same functional form at all scales [5]. Many real social and interaction networks such as the Internet topology [11,22], the web [1, 7], phone call graphs [25], and on-line social networks [23] exhibit a power law degree distribution. Ebel et al. showed for the first time that e-mail networks are scale-free

[5]Since the calculation of average shortest path length over all pairs of vertices is computationally prohibitive, similar to [23], 10,000 vertices were selected randomly in the networks and the average shortest path from these vertices to all other vertices in the networks were calculated.

with the power law degree distribution $n(k) \propto k^{-1.81}$ [10]. Others have also observed a power law degree distribution in their e-mail networks [6, 8, 12, 17, 21].

Surprisingly, in contrast with previous studies the degree distribution of the e-mail network studied here does not exhibit a power law distribution. Figure 1(a) reveals that e-mail networks are not scale-free since the connectivity of the vertices do not follow a power law distribution in all points. The figure is plotted in logarithmic scale and the x-axis denotes the degree of a vertex and the y-axis the fraction of vertices in the network with that degree.

We can also show that this deviation from power law distribution is due to the excessive amount of unsolicited e-mail communications. Figure 1(b) shows that the degree distribution for the ham network closely follows the distribution $n(k) \propto k^{-2.72}$, making the ham network scale-free.[6] However, Fitting a power law distribution for example to the node degrees of the spam network is only possible for nodes with degrees higher than 55 (with $\gamma = 3.4$ and Kolmogorov-Smirnov goodness-of-fit of 0.0274 compared to 0.0075 for the ham network) which are less than 0.4% of the total nodes in that network. Figures 1(c) and 1(d) also show that neither the spam, nor the rejected e-mail network is a scale-free network.

The observation that legitimate e-mails obey the statistical structural properties similar to other social networks is not unexpected. However, a vast majority of spam e-mails are automatically sent, typically from botnets of compromised machines [14, 27], and therefore spam traffic do not exhibit a normal social behavior. Consequently, in mixed e-mail traffic, this unsocial behavior of unsolicited e-mail is clearly visible.

The degree distribution can also be analyzed on directed e-mail networks where the direction of the e-mail communication is also considered. For the directed network we have to consider the distribution of both the in-degree and the out-degree of the vertices. For the ham network, both in- and out-degree distributions are a power law distribution (Figure 1(g)) and for the spam (Figure 1(h)), the rejected (Figure 1(i)), and the complete e-mail network (Figure 1(f)) the deviation from power law distribution is even more clear than what can be observed for the corresponding undirected networks.

Finally, it is also possible to consider e-mail networks as multigraphs where multiple edges can exist between any two vertices (capturing multiple e-mail communications between two users). Figure 1(e) shows that the degree distribution of the multigraph networks exhibits the same properties as we have observed for undirected and directed networks; the ham network is scale-free, but not the spam and the complete e-mail networks.

4.3 Strongly Connected Components

Another important structural property of complex networks, which quantifies the connectivity of a network, is the size of the giant strongly connected component of the network [10, 25]. A strongly connected component (SCC) is a subset of vertices of the network where a path exists between any pair of them in the same set. A giant SCC ($GSCC$) contains a significant fraction of the vertices in the network.

The ham network studied in this paper has 85,992 separate components with a GSCC containing 72.9% of the total vertices in the network. The second largest SCC of this network contains only 0.26% of the total vertices, so is orders of magnitudes smaller than the GSCC. The spam network has 178,754 disconnected components with a GSCC containing 53.5% and a second largest SCC with 0.15% of the total number of the vertices in this network. This

[6] We have used the maximum likelihood method and Kolmogorov-Smirnov goodness-of-fit [9] to calculate the best power law fit.

property can also be analyzed on a directed e-mail network, where an SCC is a subset of vertices that can both reach and be reached from any other vertex in the same set. The GSCC of the directed ham network is much smaller than that of the corresponding undirected network, but still is dramatically larger than the GSCC of the directed spam network.

The distribution of the size of the SCCs for the ham and spam networks is plotted in Figure 2 in logarithmic scale. The x-axis denotes the size of each SCC (in number of vertices), and the y-axis the fraction of SCCs in the network with that number of vertices. It can be seen that the GSCCs of the networks are orders of magnitude larger than other SCCs of the corresponding networks. The distribution of the SCC size for the ham network is similar to web [7] and phone call graphs [25] and follows a power law pattern. However, for the spam network many outliers are clearly visible in the distribution.

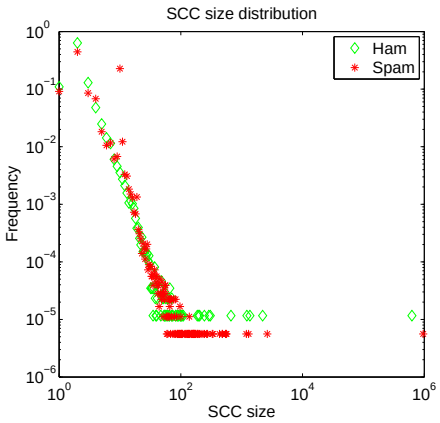
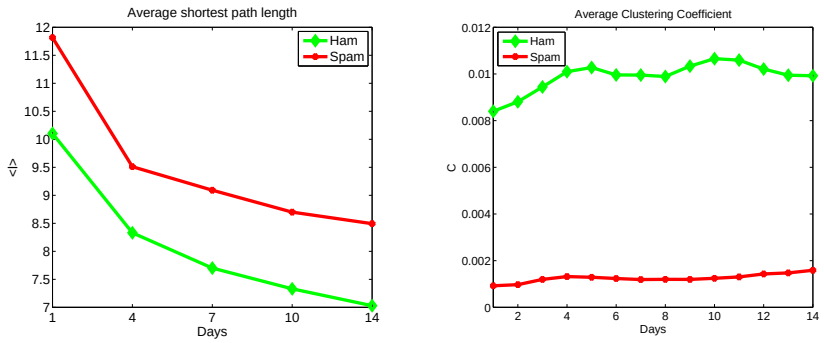


Figure 2: Distribution of size of the strongly connected components. The GSCCs of the networks are orders of magnitude larger than other SCCs. The distribution for the spam network compared to the ham network has many outliers.

4.4 Summary

The high clustering together with the short path length between vertices of the networks confirms that both the ham and the spam networks are small-world networks. However, these properties on their own are not enough to conclude that e-mail networks in general are structurally similar to social networks.

The non-power law degree distribution of e-mail networks has not previously been observed. In [21] spam e-mails were removed from the dataset before the analysis of the e-mail network. The e-mail network used in [10] was limited to log files of a university mail server, and probably only contained delivered e-mails with no or little spam. Although the network studied in [12] contained rejected and accepted spam, they also did not observe a non-power law behavior in the degree distributions. We conjecture that it was due to their limited dataset and that the method they used to generate their spam networks was based on spam vertices (in contrast to our method that uses spam e-mails as edges). It was also observed in [16] that their spam network had a power law degree distribution with an exponent larger than that of legitimate e-mails. However, their e-mail network was limited to e-mail communications of a single user. Our e-mail network which is based on real e-mail traffic seen on an



(a) Average shortest path length (b) Average clustering coefficient

Figure 3: Temporal variation in the small world properties of the networks. Both ham and spam are small world networks, but ham has a larger average clustering coefficient.

Internet backbone link, in contrast, is not limited to a single domain and is significantly larger than all previous studies and contains a large amount of unsolicited traffic. Therefore, it has allowed us to clearly observe the unsocial behavior of spam and its effect on the structure of the complete e-mail network.

Overall, the longer average path length value, the lower average clustering coefficient value, the non-power law degree distribution, and the smaller GSCC in the spam network compared to the ham network, has the potential to be used for detection of unsolicited e-mail traffic (see Table 3).

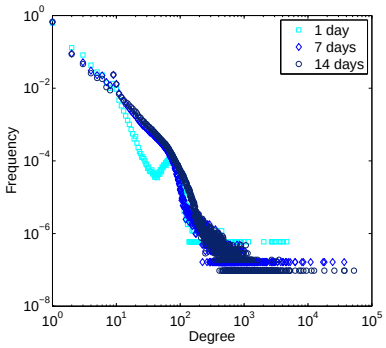
5 Temporal Characteristics of E-mail Networks

E-mail networks have an inherently evolving dynamic structure. In this section we show how the structural properties of our e-mail networks changes over time. Tracking these changes can lead to some insights into how the e-mail networks evolve. A similar study was done in [25] where they studied the evolution of phone call graphs using one week of data. Our observations are based on two weeks of data collection.

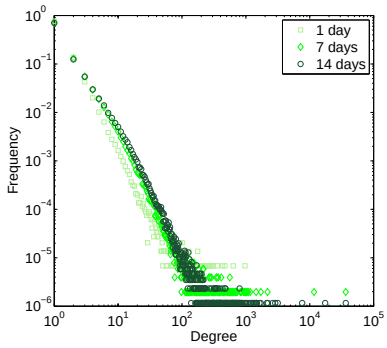
5.1 Densification Power Laws

As time progress, new vertices and edges are added to the e-mail network. The number of vertices in our undirected complete e-mail network increased from 1,688,020 vertices in the first day, to 10,544,647 vertices after 14 days, and the number of edges from 2,239,560 to 21,537,314. Both the ham and the spam networks have also grown in terms of number of vertices and edges, but the growth was faster for the spam network. The number of edges was growing faster than the number of vertices; consequently the average degree of the networks also increased over time. Leskovec et al. [21] also observed that in many real networks such as the Internet AS graph and citation graphs, the networks become denser over time and this densification follows a power law pattern.

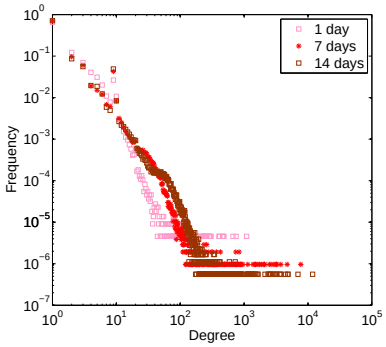
The number of edges $E(t)$ versus the number of vertices $N(t)$ in our ham and spam networks follow the densification law $E(t) \propto N(t)^{\alpha}$, where the exponent α corresponds to the slope of the line. The slopes for the ham and spam networks are $\alpha_{ham} = 1.33$ and $\alpha_{spam} = 1.42$ respectively, indicating a large deviation from linear growth.



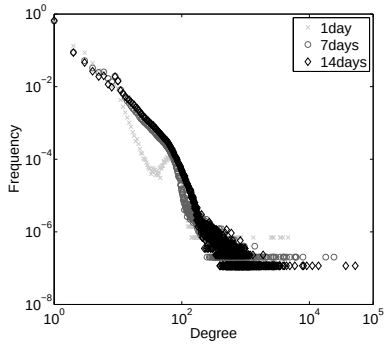
(a) Degree distribution of complete e-mail network



(b) Degree distribution of ham network



(c) Degree distribution of spam network



(d) Degree distribution of rejected network

Figure 4: Temporal variation in the degree distribution of the networks. The ham network is always scale-free and its power law exponent remains almost constant.

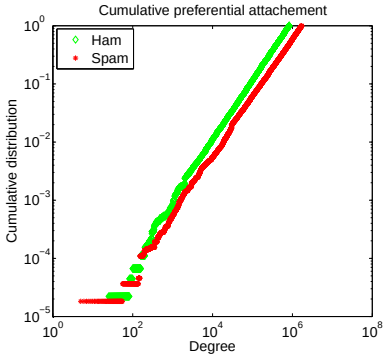
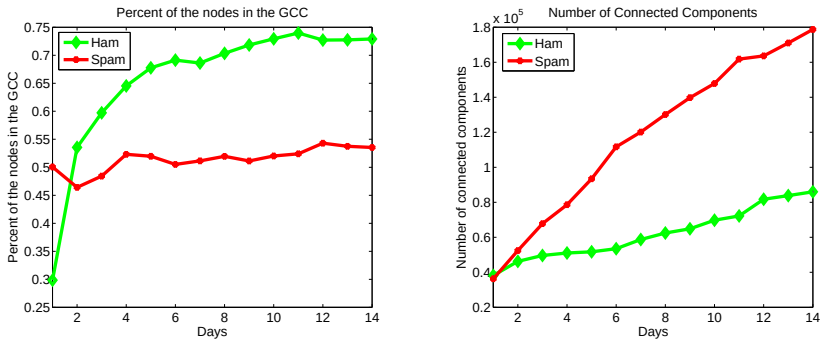


Figure 5: Evidence of preferential attachment in the ham and the spam networks.



(a) Percentage of nodes in giant strongly connected component

(b) Number of strongly connected components

Figure 6: Temporal variation in the strongly connected components of the networks. The ham network becomes more connected over time.

5.2 Shrinking Diameter

As networks grow and become denser, their diameter decrease.[7] This property was first observed by Leskovec et al. [20, 21] who calculated the effective diameter of different real networks over time. Figure 3(a) provides the evidence that the average shortest path length decreases in both the ham and the spam networks.

5.3 Clustering Coefficient

In small-world networks the value of the average clustering coefficient C is independent of the network size [28]. Figure 3(b) shows that C remains relatively constant as the networks grow, and that it is always significantly larger for the ham network than for the spam network.

5.4 Degree Distribution

Leskovec et al. [21] observed that the power law degree distribution of their e-mail network has a constant exponent $\gamma = 1.76$ which doesn't change much as the network evolves. Kossinets et al. [17] also observed that the shape of the degree distribution of their e-mail networks is relatively constant across the duration of their measurements.

The shape of the degree distribution of our e-mail network changes from the first day to the 14^{th} day as the network grows (Figure 4(a)). The spam network (Figure 4(c)), and the rejected e-mail network (Figure 4(d)) also have changing shapes of their degree distributions over time. The ham network, however, always follows a power law distribution with an almost constant exponent $\gamma = 2.7$ (Figure 4(b)). Similar properties also hold for the in- and out-degree distributions of the directed networks and the degree distribution of the multigraph network.

[7] This decrease is bounded from below

5.5 Preferential Attachment

Many evolving network models are based on preferential attachment [2]. This model states that the probability of getting new edges by a vertex increases with its degree [28]. This property which is also known as "the rich get richer" rule is associated with the power law degree distribution of social networks.

To find out if preferential attachment is coming into play in the evolution of e-mail networks, we used the same method as described in [28] to calculate the evidence of preferential attachment. First, we considered the network generated after 13 days of measurement and recorded the vertices existing in that network and their degrees. Then, we measured the increase in the degree of these vertices after one day. Finally, we calculated the relative increase in the degree of each vertex k_i to get $\Pi(k_i)$. Plotting the cumulative distribution $\kappa(k) = \sum_{k_i=0}^{k} \Pi(k_i)$ as in Figure 5 shows that evidence of preferential attachment is present in both the ham and the spam networks.

5.6 Strongly Connected Components

As social networks grow, a giant cluster emerges in the graph and its size increases [28]. This property is not specific to social networks; random networks have the same property, as well.

As shown in Figure 6(a), the percentage of the vertices in the GSCC of the ham network increases as the network grows, but the size of the GSCC of the spam network does not change significantly. The same property holds for the directed ham and spam networks. Figure 6(b) shows that the number of disconnected SCCs in both the ham and the spam networks increase over time. However, this increase is much faster for spam meaning that there are more disconnected components in the spam network.

5.7 Summary

By analyzing the static structure of ham and spam networks we have revealed that both these networks can be modeled as small-world networks. Analyzing the temporal variation in the structure of these networks also confirms this observation. The average shortest path length of both networks decrease with similar pace, and their average clustering coefficient does not change with the increase in the number of vertices in the networks. However, the values of these metrics dramatically differ for ham and spam networks, and thus can potentially be used for discriminating them.

The analysis of the temporal variation of the degree distribution of the ham and the spam network proves that a spam network is not a scale-free network. In a scale-free network, the power law exponent of the degree distribution does not change with the growth of the network. This is not the case for the spam network, the rejected network, and the complete e-mail network containing massive amount of unsolicited e-mail traffic. Although it seems that preferential attachment is associated with the growth of both ham and spam networks, the structure of these networks differ significantly. Finally, the size and the distribution of the SCCs of the ham network are similar to other social networks, but this is not true for the spam network.

In conclusion, temporal analysis of the e-mail networks gives us a better insight into the structure of these types of networks and reveals characteristics that can potentially be deployed for discriminating ham and spam traffic.

Table 4: Detection rate (DR) and the false positives (FP) incurred by the proposed detection methods

	Spamming nodes		Spam e-mails	
Method	DR	FP	DR	FP
Out-degree distribution outliers	7.70%	4.04%	25.3%	4.17%
SCC size distribution outliers	21.33%	9.19%	21.5%	7.38%
Combined	25.5%	8.78%	34.4%	6.58%

6 Social Network-Based Spam Mitigation

Our main goal with analyzing the social structure of e-mail networks is to produce realistic models for legitimate e-mail traffic, models that can be deployed to distinguish unsolicited traffic from a mixture of real e-mail traffic on the network level.

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 to reduce the amount of unwanted traffic and the waste of mail server resources. In addition, it has been observed [18, 27] that spam mainly originates from botnets that are also likely active in other malicious activities on the Internet. Therefore, detecting spamming nodes may also help to detect other malicious traffic from the same origin. In this section we investigate the possibility of early spam detection using the observed differences in the structural properties of ham and spam networks. The goal is to solely use the e-mail-sending patterns that are characteristic of spamming behavior rather than having to inspect and classify the e-mail contents.

6.1 Outlier detection

As we have previously shown in Section 4, an e-mail network is not scale-free since the degree distribution of the network contains many outliers. To verify that these outliers are caused by the unsocial behavior of spamming nodes, we have inspected the outliers in the out-degree distribution of an e-mail network containing both ham and spam.

To find the outliers, we first calculate the ratio between the out-degree distribution of the network and the power law distribution of the ham network ($n(k) \propto k^{-2.33}$), which is our model for the expected social behavior. Then we deploy the Median Absolute Deviation (MAD) method, which calculates the distance from the median of the obtained ratios to mark the outliers. Other methods to find outliers in the distribution can also be used.

Using this method we have detected more than 97,320 outlier nodes in the e-mail network generated from 14 days of collected ham and spam data. The detected outliers had sent 906,976 e-mails of which 96% were actually spam, confirming that the outliers were mostly spamming nodes.

We have also shown in Section 4 that the distribution of the size of the SCCs of the ham network follows a power law distribution with $\gamma = 2.38$, but not the spam network. By using the same method we have detected 73,324 outlier SCCs in the network. These outlier components contained 469,803 edges of which 97% were actually spam e-mails.

6.2 Spam detection

Knowing that nodes deviating from social behavior are sending spam, we have conducted a number of tests to investigate the possibility of using our social network-based analysis to detect spam e-mails. Our tests were designed to address the following three questions: *"Is it possible to detect spamming nodes solely based on the outliers in the structural properties of e-mail networks?"*, *"Is it necessary to construct a complete e-mail network capturing all the communications between users during a long period of time to detect spam, or is it enough to construct a partial e-mail network based on shorter traffic snapshots?"*, and *"Is it possible to perform the spam detection in real-time?"*

To be able to answer the above questions, we have split the captured e-mail traffic into snapshots of one hour which allows us to generate the e-mail networks and test the effect of a shorter time window. The generated e-mail networks contained only accepted spam and ham e-mails since the rejected e-mails are already detected by existing pre-filtering mechanisms (e.g. IP blacklisting). We have then evaluated our outlier detection method using the first 32 hours of our dataset by finding degree distribution and SCC size distribution outliers for each one-hour e-mail network.

The one-hour time window was arbitrarily chosen to verify the applicability of our approach on smaller networks. Smaller time windows might also be possible as we discuss later. The window size together with the model parameters (i.e. the out-degree distribution exponent and the SCC size distribution exponent) should be fine-tuned based on the vantage point of the tool deploying it and the volume of traffic passing through it. This is outside the scope of this paper and left for future work.

In order to evaluate the results, we have calculated the *detection rate* and the number of *false positives* incurred by the outlier detection mechanisms. Similar to Intrusion Detection Systems (IDS) the detection rate is calculated as the percentage of correctly identified true spam by the system and the false positive rate as the number of legitimate e-mails incorrectly classified as spam. We have calculated the number of false positives and the detection rates based on our classification using SpamAssassin.

Table 4 shows the average detection rate (DR) and the number of false positives (FP) using two of the proposed methods with a window size of one hour. As can be seen, the degree distribution outliers detected as spamming nodes are just 7.7% of the total number of spamming nodes in the network (a node is called a spamming node if it has sent at least one spam e-mail during the collection time window). However, by combining the output of both the degree distribution and the SCC size distribution outliers, we achieve a detection rate of 25.5% of all spamming nodes, however with the cost of more false positives.

Furthermore, by inspecting the e-mails transmitted by the detected outlier nodes (using a combination of both methods), we have observed that more than 93.4% of the e-mails sent by these nodes were actually spam e-mails. As Table 4 shows, these nodes in average have sent more than 34.4% of the total spam in the network.

To be able to use our spam detection method incrementally in real-time (i.e. without the use of any time window), each newly arriving e-mail should be added to an incrementally generated e-mail network and then be verified whether the sender is part of the outliers or not. To test the effectiveness, first, prior the start of the detection, we have generated an e-mail network from a one-hour traffic snapshot as the base. Then we have classified the e-mails collected during the following hour. The arrival of a new email could only affect the out-degree of the sending node, so there is no need to re-calculate the out-degree of other

nodes in the network to re-generate the degree distribution. The results show that using only outliers of the degree distribution, on average the detection rate for newly arriving e-mails has slightly decreased (from 25.3% to 23.0%) and the false positive rate has increased (from 4.17% to 6.39%) compared to the non-incremental approach that delays the detection until the end of the one-hour time window. Therefore, the proposed detection method is also effective if deployed in real-time.

6.3 Incorporating with existing systems

We have shown that it is possible to detect spamming nodes by using the distinguishing properties obtained from social-network based analysis of e-mail traffic. In the following we discuss two possibilities for deploying our spam mitigation strategy: (1) on the wire deployment as a small network device and (2) integration with existing spam detection techniques.

We can deploy a network device that can examine traffic on the wire by inspecting the SMTP headers and generating e-mail networks on the fly. Such a device can be placed anywhere on the Internet that has a vantage point similar to the one we had in our measurements. So it is not necessary to place it on the receiving mail servers. The proposed collection and analysis strategies introduced in this paper could be used by such a device. The advantage is that such a system implementation will allow us to detect and stop spam solely by inspecting e-mail senders and receivers.

We can also incorporate our approach into existing spam filtering tools (e.g. SpamAssassin) on the receiving mail servers to contribute to the spam scores generated by these tools. The advantage is the easy deployment in existing infrastructure.

Regardless of where the system is located, our method is a pre-acceptance strategy that can be implemented either as a reputation based system, or as a spam detection tool.

A reputation-based system assigns a spamming score to the detected spamming nodes. These scores can be used for example to populate existing IP blacklists with the IP addresses of the detected spammers. They can also be used in conjunction with existing greylisting strategies, i.e. e-mails sent by nodes that are detected as outliers are greylisted (temporarily rejected) and are accepted only after being re-transmitted by the sender in appropriate time.

A spam detection tool marks each arriving e-mail as spam or ham on the fly. Our incremental detection approach performs the same task and marks e-mails sent by detected spamming nodes as spam.

Overall, our system is a pre-acceptance strategy that can complement existing systems including IP blacklists. Assuming that all the rejected e-mail traffic observed in our dataset were stopped by existing pre-acceptance strategies, more than 84% of the spam traffic has already been detected. By deploying our approach, approximately 34% of the remaining spam e-mails would also be detected, increasing the detection rate to more than 89% before any content is being transmitted.

6.4 Summary

In this section we have proposed a novel method to exploit structural properties of ham and spam networks to detect spamming nodes from a mixture of traffic. We show that by using only two basic structural properties, we can still detect a significant number of spamming nodes that were not detected by existing pre-acceptance strategies like IP blacklisting, improving the overall detection rate. The advantage of our method is that it can be deployed

as a stand-alone system on the wire or incorporated into existing tools and detect spam e-mails in real-time as they pass through the link. Our study shows that it is not necessary to capture all the e-mails communications between users over long periods of time to be able to take advantage of the distinguishing properties of legitimate and unsolicited traffic for spam detection. By introducing the rest of the studied structural and temporal metrics into the system, and fine-tuning the parameters (e.g. time window size), it is most likely that the number of false positives and the detection rates improve.

7 Discussion

In this section we discuss to what degree the dataset we have collected and used in our study is representative for the purpose of structural and temporal analysis of e-mail networks.

The e-mail dataset used in this paper is based on real SMTP traffic captured on a high speed Internet backbone link. This dataset has provided us with a better understanding of the characteristics and behavior of e-mail and spam traffic. The captured traffic does not, for obvious reasons, include e-mail communications not traversing the link, i.e. e-mail communications between "inside" users and communication between external users. Due to asymmetric routing and the use of parallel load-balanced links, we were not always able to see all traffic. In the rest of this section, we discuss the possible effects the data collection procedure and network structure may have on the studied metrics.

7.1 Asymmetric Routing and Missing Data

Due to asymmetric routing and load-balancing policies deployed by the network routers, not all traffic was captured and less traffic was also traveling in the outgoing than the incoming direction of the link (Table 2). This has limited the possibility to analyze graph metrics that are specific to the direction of the edges. However, this limitation is not that important since the goal of this work is to find methods to analyze network traces using social network-based techniques in order to detect misbehaving (spamming) nodes. We want to find distinguishing parameters of spam and ham traffic, preferably with observing as little traffic as possible. Earlier studies [6, 10, 12, 17, 19, 29] have also shown that it is not required to generate a complete e-mail network of all e-mails exchanged between all users during long periods of time to be able to study the structure of e-mail networks or to deploy a social network-based approach to mitigate spam. It is also reasonable to assume that similar limitations will be present in real-life scenarios when traffic is analyzed in the search for spammers.

Boas et al. [4] studied the effect of perturbations on the structure of complex networks, and their results can help us to understand what possible effects missing e-mails in our datasets may have had and how they might have affected the structure of the studied e-mail networks. They showed that scale-free networks are highly robust to edge additions or deletions. Therefore, working with snapshots of communications should have not affected our findings regarding the degree distribution of the e-mail networks. They also showed that the average shortest path length in complex networks is sensitive to addition of edges to the network, since a new edge can connect vertices that are far away, resulting in a decrease in this value. However, the values for the average shortest path length of our ham and spam networks are already very small despite missing the internal/external traffic. Boas et al. also showed that the average clustering coefficient is a robust metric that does not change by adding new

edges to a complex network. In our study, this value similarly remained unchanged for both of the ham and the spam network.

7.2 Missing Past

A source of missing data, which is not limited to our dataset, is the "missing past" problem. This problem appears since there is no possibility to gather data reaching all the way back to a network's birth. Leskovec et al. [21] showed that the effect of missing past is minor as we move away from the beginning of the data observation. We have also investigated the effect of missing past in our analysis by constructing an e-mail network which lacked the first week of data to see if it would affect our results. The value of the average clustering coefficient of the ham and spam networks with missing past was similar to that of the corresponding complete networks. The average shortest path value became slightly higher, however it was still small enough to support our observation that both ham and spam network are small-world networks. The size of the GSCC of the ham and the spam networks with missing past were smaller but still contained a large fraction of nodes of the network, and still the GSCC size of the ham network was substantially larger than that of the spam network.

Finally, the degree distribution of the ham network with missing past was identical to the complete one indicating that missing past had no major effect on this property. However the degree distribution of spam had a different shape when the past was missing, indicating that structure of the spam network is not robust against perturbations, hence it further strengthens our findings that a spam network is not scale-free.

7.3 Measurement Duration

Another limitation that could have affected our analysis is the data collection duration. Although our measurement duration was shorter than previous studies [10, 12, 17, 20], we have generated the largest and most general dataset used for this type of analysis. Our analysis of the temporal variation in structural properties of e-mail networks have provided us with evidence for how the structure of different e-mail networks changes with longer periods of measurements.

As shown in Figures 4, 3, and 6, for most of the structural properties of ham and spam networks, i.e. the average path length, the average clustering coefficient, and the degree distribution, even one day of data could have discriminated ham from spam traffic. After one week of data collection, all the observed properties were applicable to differentiate spam from ham. With more data, the difference between ham and spam gets even more obvious. However, there is a trade-off between analyzing less data for faster detection and the number of false positives. Our tests in Section 6 have shown that even by using a snapshot of one hour of the data, it is still possible to detect spam senders based on the structural differences of spam and ham networks.

Overall, this work has provided us with a large dataset containing real traffic between regular users, spammers and receiving mail servers. Although a full and complete e-mail network cannot be generated from this data, we can model the behavior of e-mail traffic as observed from the vantage point of a network device that can be placed on a link to stop spam as close to its source as possible.

8 Conclusions

In this paper we have performed a large scale data collection and analysis of e-mail traffic captured on an Internet backbone link. To the best of our knowledge this is the largest study on the social behavior of e-mail traffic that has been performed so far. We have analyzed the structural and temporal properties of e-mail networks. Our study shows that:

- legitimate e-mail traffic generates a small-world, scale-free network that can be modeled similar to many other social networks.
- contrary to previous studies, e-mail traffic in general does not exhibit all of the structural and temporal properties common in social networks due to spam traffic which is unsolicited.
- structural and temporal properties of e-mail networks can distinguish between ham and spam, and spamming nodes can be detected based on their unsocial behavior.

Moreover, we have proposed and tested a method for detection of spamming nodes by using two of these distinguishing structural properties. Our results suggest that it is possible to implement this method on a network device to classify e-mail traffic based on simple operations in real-time without having to inspect the e-mail contents. By incorporating our method with existing pre-acceptance strategies used in mail systems today, an extra 34% of spam can be stopped on the network level, bringing it up to 89% stopped unsolicited traffic.

We are currently incorporating all distinguishing structural characteristics into the proposed spam detection approach, and also examining how and in what way buffering and timing parameters (e.g. time window size) can improve the effectiveness of spam mitigation on the network level.

9 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] R. Albert, H. Jeong, and A. L. Barabasi. The diameter of the world wide web. *Nature*, 401, 1999.

[2] A. L. Barabasi and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439), 1999.

[3] A. L. Barabasi, R. Albert, and H. Jeong. Scale-free characteristics of random networks: the topology of the world-wide web. *Physica A: Statistical Mechanics and its Applications*, 281(1-4), 2000.

[4] P. R. V. Boas, F. A. Rodrigues, G. Travieso, and L. da F Costa. Sensitivity of complex networks measurements. *Statistical Mechanics*, 2010.

[5] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D. Hwang. Complex networks: Structure and dynamics. *Physics Reports*, 424(4-5):175–308, February 2006.

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

[7] 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.

[8] G. Caldarelli, F. Coccetti, and P. De Los Rios. Preferential exchange: Strengthening connections in complex networks. *Physical Review E*, 70(2), 2004.

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

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

[11] M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationships of the internet topology. In *Proceedings of the conference on Applications, technologies, architectures, and protocols for computer communication*, SIGCOMM, 1999.

[12] 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 *Proceedings of the Conference on Email and Anti-Spam*, 2005.

[13] L. H. Gomes, V. A. F. Almeida, J. M. Almeida, F. D. O. Castro, and L. a. M. A. Bettencourt. Quantifying social and opportunistic behavior in email networks. *Advances in Complex Systems (ACS)*, 12(01):99–112, 2009.

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

[15] J. Kleinberg. The small-world phenomenon: An algorithmic perspective. In *in Proceedings of the ACM Symposium on Theory of Computing*, 2000.

[16] J. S. Kong, B. A. Rezaei, N. Sarshar, V. P. Roychowdhury, and P. O. Boykin. Collaborative spam filtering using e-mail networks. *Computer*, 39(8), 2006.

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

[18] 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*, 2008.

[19] 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.

[20] J. Leskovec, J. Kleinberg, and C. Faloutsos. Graphs over time: Densification laws, shrinking diameters and possible explanations. In *Knowledge Discovery and Data Mining*, 2005.

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

[22] P. Mahadevan, D. Krioukov, M. Fomenkov, X. Dimitropoulos, k. c. claffy, and A. Vahdat. The internet as-level topology: three data sources and one definitive metric. *SIGCOMM Comput. Commun. Rev.*, 36, January 2006.

[23] 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*, 2007.

[24] 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 the 1st Workshop on Building Analysis Datasets and Gathering Experience Returns for Security (BADGERS)*, 2011.

[25] 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 Transactions on Knowledge and Data Engineering*, 20(5), 2008.

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

[27] 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, pages 291–302. ACM, 2006.

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

[29] C. Tseng and M. Chen. Incremental SVM model for spam detection on dynamic email social networks. In *Proceedings of the International Conference on Computational Science and Engineering*, 2009.

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