Comparative Graph Theoretical Characterization of Networks of Spam and Legitimate Email

Luiz H. Gomes†, Rodrigo B. Almeida†, Luis M. A. Bettencourt‡, Virgílio Almeida†, Jussara M. Almeida†

† Computer Science Department
Universidade Federal de Minas Gerais
Belo Horizonte - Brazil
{lhg, barra, virgilio, jussara}@dcc.ufmg.br

‡ Computer and Computational Sciences
Los Alamos National Laboratory
Los Alamos - USA
lmbett@lanl.gov

Abstract

Email is an increasingly important and ubiquitous means of communication, both facilitating contact between individuals and enabling rises in the productivity of organizations. However, the relentless rise of automatic unauthorized emails, a.k.a. spam is eroding away much of the attractiveness of email communication. Most of the attention dedicated to spam detection has focused on the content of the emails or on the addresses or domains associated with spam senders. Although, methods based on these - easily changeable - identifiers work reasonably well, they miss on the fundamental nature of spam as an opportunistic relationship, very different from the normal mutual relations between senders and recipients of legitimate email. Here we present a comprehensive graph theoretical analysis of email traffic that captures these properties quantitatively. We identify several simple metrics that serve both to distinguish between spam and legitimate email and to provide a statistical basis for models of spam traffic.

1 Introduction

Spam is quickly becoming the leading threat to the viability of email as a means of communication and a leading source of fraud and other criminal activity worldwide. Much is known about spam traffic. According to the Spamhaus project [16] the vast majority of spam emails presently originate in the USA and China, hosted by well known ISPs and generated by identified individuals. Nevertheless an increased effort in criminal investigation and waves of high profile legislation have not yet succeeded at reducing the relentless increase in spam traffic [10], which now accounts for about 83% of all incoming emails, up from 24% in January 2003 [12].

It is often said that the problem of spam email is that it is an extremely asymmetric threat. While it is technically easy and very cheap to send a spam email it requires sophisticated organization and much higher costs at the receiving end to sort out legitimate emails from junk.

This asymmetry is of course not directly manifest in the sender's email address, on the domain he/she uses, nor certainly on the simplest characteristics of the message (e.g. its size). It is rather a property of structural relationships - spammers tend to be senders to a socially unrelated set of receivers - while legitimate email tends, instead, to reflect the variety of mutual personal, professional, institutional ties among people. Thus by identifying the comparative structural and dynamical nature of email traffic, we expect to find good discriminators between legitimate email and spam traffic. The goal of this work is to present the modeling of email - legitimate and spam - traffic as networks, in order to identify graph theoretical metrics that can be used to differentiate between the two. We are also interested in providing a unified view of several metrics characterizing the relationships between senders/recipients and their evolution for legitimate and spam traffics in order to formulate, in the future, a predictive model of spam dissemination.

Our study goes beyond several recent analyses [4, 7] on the graphical nature of spam traffic. We deal with a different database, involving a much larger number of users and messages, and analyze a wider set of metrics, both static and dynamic. We will show that there is no single graphical metric that unequivocally distinguishes between legitimate and spam email. There are, however, several graph theoretical measures that can be combined into a probabilistic spam detection framework. These are then identified as candidates for the construction of a future spam filtering algorithm.

The remaining of this paper is organized as follows. In section 2 we introduce the modeling of email traffic in terms of two graph classes and present the types of metrics to be studied. Section 3 presents the email workload used in this work. We present several graph theoretical metrics and evaluate this workload according to them in Section 4. In Section 5 we present related work. Finally, we present our conclusions in section 6 and discuss open questions left for future work.

2 Graph-Based Modeling of E-mail Workloads

In order to characterize spam email traffic versus non-spam we define two types of graphs: a *user graph* and a *domain graph*. The vertices of the *user graph* are email senders and recipients of our log. An email sent by A to receiver B results in a link between A and B. The *domain graph* has as vertices the domains of the external senders being analyzed, and users inside the local domain. In this case, if an user B of the local domain receives an email from any user in an external domain D, we define a link between D and B. Note that, sets of users external to the local domain who share an domain are aggregated together into a single node. Note, also, that the domain graph is a simpler bipartite graph and not all characteristics studied will be valid in it.

The edges of both graphs can take one of four forms: directed or undirected; binary (or unweighted) or weighted (e.g. by the number of emails exchanged or by the total size of the emails exchanged in bytes). These options cover most of the possibilities for direct graphical construction out of the email logs at our disposal (described in Section 3).

The user graph is, in principle, the most useful in identifying the individual nature of users as spam or non-spam senders. In some cases these characteristics extend to the whole external domain (particularly if the spammer changes his name[1] more often than its domain) and the domain graph produces a useful aggregation of the user data. We believe that user graphs will be more effective in identifying senders of non-spam since spam senders tend to change their full email address very frequently.

The user or domain graphs can be constructed exclusively out of spam traffic, non-spam traffic, or the aggregate set of all emails. Some of the graph theoretical properties studied below will be analyzed in terms of the graphs constructed when considering the different traffics separately while others will be evaluated on selected nodes from the aggregated traffic. The selected nodes represent senders in the aggregated graph and can be divided in two classes - spam and non-spam - based on the type of emails they send. These classes do not form disjoint sets, see Table 2.

Given these two graph types we will analyze two types of properties: (i) structural and (ii) dynamical. The former capture the structure of social relationships between users exchanging emails, while the latter relate to how graphical properties evolve over time. As we shall show below there are distinct independent signatures of spam traffic in both structural and dynamical properties. As a consequence they should be taken together to generate a better detection procedure.

3 E-mail Workloads

Measure	Non-Spam	Spam	Aggregate
# e-mails	336,580	278,522	615,102
Size of e-mails	11.00 GB	1.70 GB	12.71 GB
# sender users	94,985	170,664	263,144
# sender domains	20,414	48,087	59,971
# recipients	26,450	12,867	35,471

Table 1: Workload summary

The construction of the graphs introduced in Section 2 is subject to several practical constraints. Our knowledge of email traffic comes from Postfix logs of the central SMTP incoming/outgoing servers of an academic department from a large University in Brazil. Incoming emails only contain the recipients internal to the department's domain. Outgoing emails contain the full list of recipients. Moreover our data set does not contain information about emails exchanged between users external to the domain.

The logs were collected between 11/18/2004 and 12/31/2004 and contain the following data for each email: (i) received time and date; (ii) a reject flag, indicating whether connection was rejected during e-mail acceptance (iii) Size of email[2]; (iv) sender address; (v) list of recipients and (vi) a spam flag, indicating if it was classified as spam or not by Spam-Assassin [15].

The logs were sanitized and anonymized to protect the users' privacy. Statistical characteristics of the workload are in agreement with previous email traffic analyses [9, 6, 17]. Table 1 summarizes the data set.

Spam-Assassin [15] is a popular spam filtering software

[1] The first part of the address, located before the @.

[2] Only for the accepted emails.

that detects spam emails based on a changing set of user-defined rules. These rules assign scores to each received e-mail based on the presence in the subject or in the e-mail body of one or more pre-categorized keywords. Spam-Assassin also uses other rules based on email size and encoding. Highly ranked emails, according to these criteria, are flagged as spam.

4 Spam Networks vs. Legitimate Email Networks

Type	External		Internal	
Spam	169931	(277535)	733	(987)
Non-Spam	93666	(186607)	1319	(151973)
Spam & Non-Spam	2366	(-)	139	(-)
Total	263231	(462142)	1913	(152960)

Table 2: Number of unique email addresses by origin (internal or external to the domain) and classified as spam, non-spam or both. Numbers in parentheses indicate the total number of emails sent by each class.

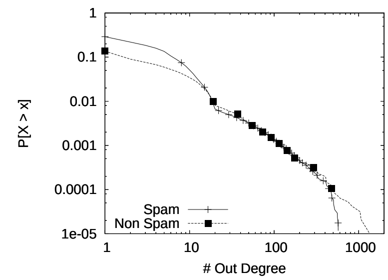
Although spam emails originate mainly from users outside the local domain spam senders use several techniques to falsify or steal local addresses (e.g. crawling the web for email addresses available at web pages, network sniffing, name dictionaries). As a result spam email does originate from the local domain both from real users and from forged ones. This mixing between regular email users and spam senders can lead to more complex email networks than might have been naively expected and poses a challenging problem for detection.

Table 2 summarizes the number of addresses and emails by node classes and by internal or external origin. Node classes are as defined in Section 2 plus a third category - Spam & Non-Spam - which is the intersection of the former two. The size of this overlap shows the impact of email address spoofing.

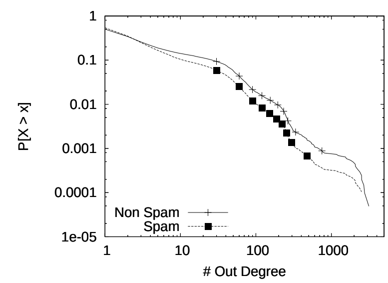
Most emails originate outside the domain. In our log most outside users are spam senders and account for the majority of the emails. Because it is very easy for a spammer to forge an address spam senders use many addresses simultaneously and/or frequently switch between them. This strategy is visible in our database as non-spam internal users send many more emails per user than spam internal users. We expect that this is a general feature of spam versus non-spam traffic.

The number of spam senders that are internal is very small. The fraction of these that send exclusively spam is 81%. These addresses correspond presumably to internal emails that have been forged and do not actually exist[3]. The remaining addresses send both spam and non-spam and are probably genuine users whose addresses have been spoofed.

4.1 Structural analysis of spam vs. non spam email graphs



(a) User Graph



(b) Domain Graph

Figure 1: Distribution of the node degrees for sender classes in the aggregated graphs.

One of the most common structural measures analyzed in complex networks is the distribution of the number of the incoming and outgoing node connections, or degree [14, 13, 2]. Figure 1 shows the distribution of the out-degrees of the different sender classes for the user and domain graphs.

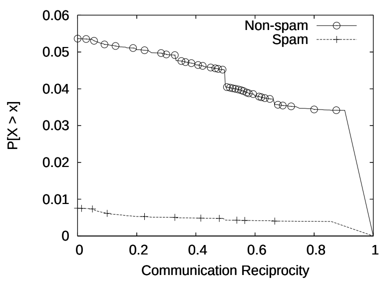
The out degree distributions approximately follow a power law (C/x^α). By using a simple statistical linear regression we estimated the exponent α that best models the data. For the user graph we obtained α = 1.497 (with R^2 = 0.965.) for spam senders and α = 1.359 (R^2 = 0.981) for non-spam senders. We conclude that the spam sender's out degree distribution is slightly more skewed. We conjecture that this is because spammers have a limited knowledge of the set of users in each specific domain. Since in our analysis we only observe a fraction of the spammers' lists (the one composed by

[3]This suggests that a simple effective way to filter out spam originating from internal domain addresses is to verify that they correspond to an existing user.

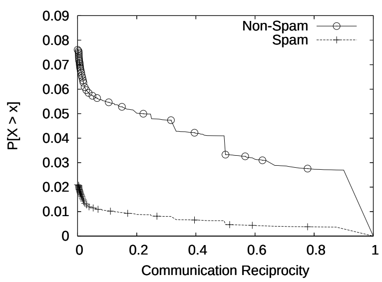
the messages sent to the domain studied) there are no spammers with recipients' lists as large as those found for non-spam senders.

Degrees from 1 to near 20 are much more probable for spam senders than for non spammers, while very large degrees are more likely in non-spam. There is no difference between the two sender classes in the body of the distribution, for degrees from about 20 to 400. The mean out-degrees, are 3.56 and 1.63 for non-spam and spam, respectively (see Table 2).

In the domain graph the out-degree distribution shows a much higher probability for nodes with low out-degree in spam traffic than in non-spam.



(a) User Graph



(b) Domain Graph

Figure 2: Distribution of Communication Reciprocity

In order to evaluate discrepancies between in and out sets of addresses for a given node we create a simple metric called Communication Reciprocity (CR) of x as:

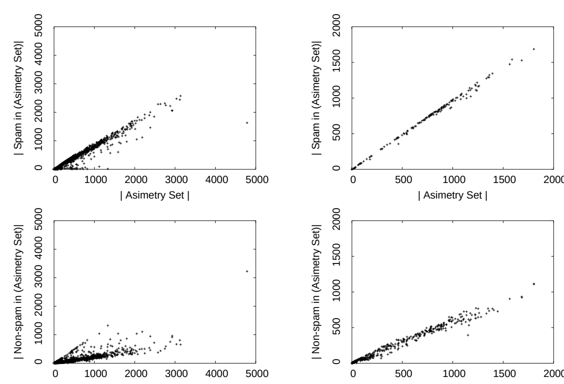
$$CR(x) = \frac{|OS(x) \cap IS(x)|}{|OS(x)|}, \tag{1}$$

where $OS(x)$ is the set of nodes that receive a message from node x and $IS(x)$ is the set of addresses that send messages to x. With our choice of normalization this metric measures the probability of a node receiving a response from each one of his addresses.

Figure 2 shows the distribution of the Communication Reciprocity. This metric is able to effectively differentiate users associated with spam from non-spam. The grouping of users in the domain graph makes this differentiation more difficult. However, even in the domain graph the difference is very clear.

The analysis of the communication reciprocity suggests that a strong signature of spam is its structural imbalance between the set of senders and receivers associated with a spam sender. However whenever there is an imbalance, how many of the unmatched addresses correspond to spam senders?

To address this question, let the asymmetry set for a node be the difference of its in and out sets. Figure 3 shows the number of spam addresses in the asymmetry set versus the size of the asymmetry set itself. The resulting relation is very well fit by a straight line at 45^o, showing a strong correlation between the two numbers. The statistical correlation is $\rho = 0.979$ for user graph and $\rho = 0.998$ for the domain graph. So, almost all senders in the asymmetry sets are spammers indifferently of the graph analyzed. The non spam data is not very well modeled by a 45^o straight line. These correspond to the non spam senders that were not answered (or to whom we could not see an answer in our log). The correlation is $\rho = 0.8723$ and $\rho = 0.9932$ for the user and domain graphs respectively. As expected from the result of the spam data the non spam data has a higher correlation for the domain graph.



(a) User Graph (b) Domain Graph

Figure 3: Number of spammers/non spam senders in the asymmetry set vs. the number of nodes in it

This result can be made sharper if we analyze the correlation between the number of spammers in the incoming set of a node and spammers in its asymmetry set. We find $\rho = 0.999$ and $\rho = 0.994$ for the user and domain graphs, respectively. There is a slightly worse correlation in the domain graph. We conjecture this is due to the external reliable domains used by spammers (e.g. through spoofing and forging techniques). These may not be counted in the asymmetry set since they are replied through their legitimate emails but are part of the incoming set as spammers.

These results show that spam messages are almost never replied to, except in cases of spoofed or forged domains or users' ids and rarely, we assume, intentionally.

Asymmetry sets can in principle be used as a component in a probabilistic spam detection mechanism. The arrival of an email from a sender that has already been contacted by an internal recipient is an indication that it has high probability of being a non spam.

Another common characteristic of social networks is a high average clustering coefficient (CC) [8]. The CC of a node n, denoted C_n, is defined as the probability of any two of its neighbors being neighbors themselves. This metric is associated to the number of triangles that contain a node n. For an undirected graph, the maximum number of triangles connecting the N_n neighbors of n is $N_n \times (N_n - 1)/2$. Thus, the CC measures the ratio between actual triangles and their maximal value. During clustering coefficient analysis we only consider the nodes with $N_n > 1$, since this is a necessary condition for the CC to be nonzero.

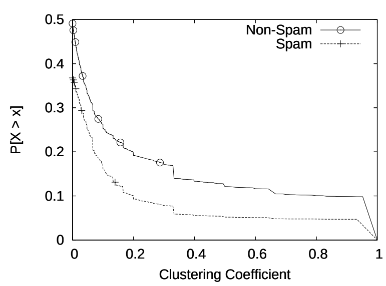


Figure 4: Distribution of the clustering coefficient for the different classes in the aggregated user graph.

Figure 4 shows the distribution for the CC of nodes in the aggregated graphs. The clustering coefficient measures cohesion of communication, not only between two users but among *friends of friends*. This is a pervasive characteristic of social relations that is absent from spam sender receiver connections. As a result regular email users have higher CC than spam senders. In terms of the average value, regular email also has a higher value (0.16 against 0.08).

Some recent studies [4] have studied graphical metrics of the strongly connected components (SCC) of email graphs. A SCC is a subset of the nodes of a graph, such that one node can be reached from any other node in the set following edges between them. A complementary measure to the CC and SCC is the average path length between two nodes. The CC and average path length properties are generally related to the so-called small world networks, which display high CC (higher than a random graph with the same connectivity) and short path length, usually comparable to $\log N$, where N is the number of nodes in the graph.

In our experiments both the SCC and the average path length have not been able to convincingly differentiate spam from legitimate traffic. All of the graphs studied are small world networks to some extent. Also all of the graphs have giant connected components. Other studies have used the clustering coefficient of SCCs to identify spam in networks constructed from the correspondence of a single user [4]. However for data from servers that aggregate the communication between different senders and recipients we find that these metrics do not suffice to perform a clear identification of spam.

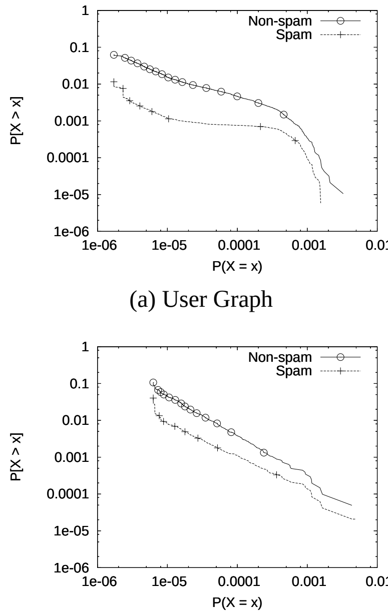
Another interesting structural characteristic of graphs is the probability of visiting a node during a random walk through the graphs [5]. At each step of the random walk we need to select the next node to be visited. This can be done in two ways. The next node can be randomly selected from the out set of the current node or we can perform a jump. For a jump, one of the nodes of the graph is selected randomly as the next node. Note that, this measure is related to node betweenness[4] since higher node betweenness tends to generate a higher probability of visitation. Nevertheless this probability is much easier to compute than node betweenness for large graphs. The probability $P(x)$ of finding a node x in a random walk is computed iteratively as follows:

$$P(x) = \frac{d}{N} + (1-d) * \sum_{z \,\in\, IS(x)} \frac{P(z)}{|OS(z)|}, \quad (2)$$

where d is the probability of performing a jump during a random walk, N is the number of nodes in the graph. The parameter d is a dumping factor that can be varied. A value usually used in the literature is 0.15 [5], that is also the value we use in our measurements.

The results are shown in Figure 5. The difference between spam and non-spam behavior is less noticeable in the domain graph than in the user graph. Spam nodes show generally lower probabilities of being visited, as might have been expected because of the asymmetry of their communication. Visiting probabilities for spam nodes in the user graph are localized to the initial and final parts of the distribution and are less pronounced in the middle range.

[4]The number of shortest paths that pass through a node.



(a) User Graph

(b) Domain Graph

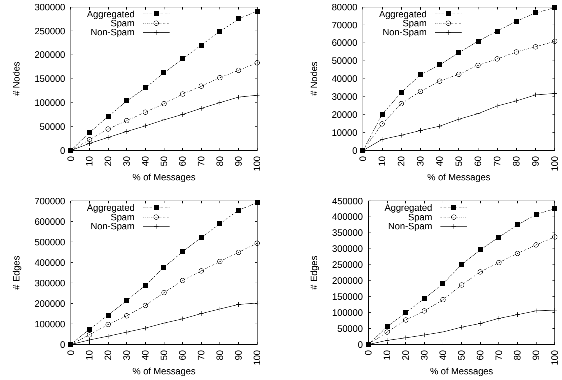
Figure 5: Distribution of the probability of finding a node during a random walk.

The node visitation probability distributions can be modeled by a power law. We estimate the corresponding exponent at $\alpha = 0.694,\ 1.097$ and 0.975 for the non-spam component of the user graph, and for the non-spam and spam components in the domain graph, respectively. The R^2 associated with the fits varies between 0.959 and 0.998. The R^2 for the spam curve of the user graph is 0.853, showing that it is not well modeled by a power law, as visual inspection suggests.

4.2 Dynamical analysis

Beyond the structural characteristics of the graphs of spam and non-spam email other metrics related to the dynamics of communication and graph evolution may help model spam traffic.

A large amount of effort has been devoted recently to creating realistic growth models for complex networks. One of the key characteristics of such models is the evolution of the number of nodes and edges, as well as the probabilistic connection rules for the new nodes to those already in the graph. Figure 6 shows the evolution of the graph in terms of number of nodes and edges. We plot these quantities against percentage of messages evaluated for each graph, to avoid the influence of the rate of message arrival, which varies with time depending on the type of the traffic being considered (e.g. the bell

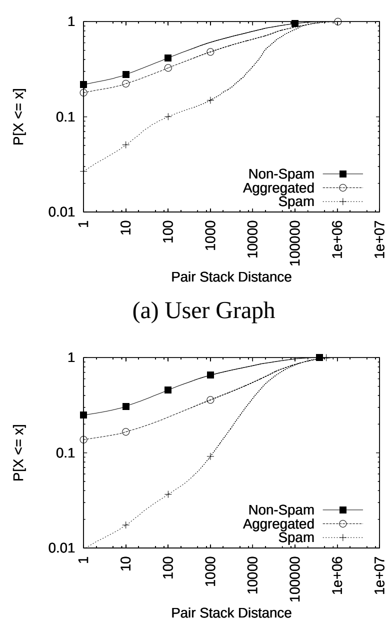


(a) User graph (b) Domain graph

Figure 6: Graph evolution by percentage of messages.

shaped behavior for the non spam traffic against the almost constant rate for spam traffic [9, 3, 6]).

The growth of the aggregated graph (a composition of the spam and the non-spam graphs) results from the growth in both the spam and non-spam components. The spam subgraph is a much more rapidly growing structure.Over the time of the log we find no saturation effect in these numbers. Instead the number of addresses and edges grows almost linearly with the number of emails. An eventual saturation in the non-spam component might be expected for longer times.



(a) User Graph

(b) Domain Graph

Figure 7: Distribution of stack distances for the pairs in the different traffics.

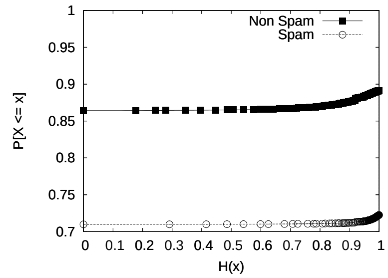
Another important dynamical graph characteristic is how the weights of edges evolve. An interesting metric that can be used to measure this is the *stack distance* [1] of connected pairs in terms of the emails they exchange over time. The stack distance measures the number of distinct references between two consecutive instances of the same object in a stream. We take the total email log as the stream and each pair (sender,receiver) as the object, disregarding the order. Figure 7 shows the pairs' stack distance distributions. We see that temporal locality is much stronger in non-spam traffic. This means in practice that legitimate users exchange emails over small concentrations of time.

We were also interested in studying how do the nodes communicate with their peers in terms of the number of messages. Because of the impersonal nature of spam we expect that spam senders communicate in a more structured way with their recipients. Not only will legitimate senders show more variation in the number of messages they send to each person in their out sets, they will also show variability of the messages themselves in terms of their sizes. In order to quantify these effects we evaluated the normalized entropy of the in and out flows for each node, defined as

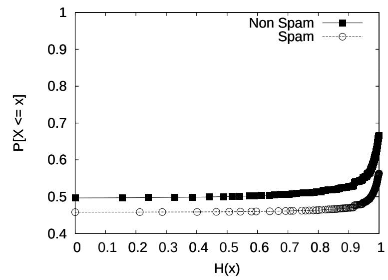
$$H(x) = \frac{\sum_{y \in OS(x)} -p(y) * log(p(y))}{log(|S(x)|)}, \qquad (3)$$

where $p(y)$ is the probability of y receiving a message from x and and $|S(x)|$ is the number of distinct elements in the set being considered.

Figure 8 shows the normalized entropy for the out flow of the nodes in the different sender classes for the aggregated graphs. As expected, spammers communicate with their recipients with much less variability (higher entropy). A similar analysis was conducted considering the bytes that each node sends with similar results.



(a) User Graph



(b) Domain Graph

Figure 8: Distribution of entropy of the number of messages in the flow of e-mails for the aggregated graph.

5 Related Work

Several studies have recently analyzed the statistical properties of email workloads [6, 11, 9, 3, 17]. These studies consider the messages as a flow and study metrics such as inter-arrival times, e-mail sizes, and number of recipients per e-mail. Although spam and legitimate email show differences in terms of these metrics little has been done about using them to filter out spam. The work of the present manuscript takes a different tack by creating a graph theoretical higher level representation of email traffic and attempting to differentiate spam from legitimate email in this abstraction. We believe that this approach, based on graph theoretical metrics, proves to be much better suited to the filtering problem.

Other recent papers have focused on models of email traffic as graphs [4, 7]. For example in Ref. [4] a graph is created representing the email traffic captured by the mailbox of an individual user. The subsequent structural analysis is based on the fact that such a network possesses several disconnected components. The clustering coefficient of each of these components is then used to characterize messages as spam or non-spam. Their results show that 53% of the messages were classified using the proposed approach and they obtained 100% of accuracy in this subset. Our graphs are based on a different type of dataset, i.e. the logs of SMTP servers, and as such do not take the perspective of the individual user. As a result for our data set the approach proposed in [4] can not be used successfully since there is a giant SCC in all of the graphs shown. In [7] the authors used the approach of detecting machines that behave as spam senders by analyzing a border flow graph of sender and recipient machines. Moreover, they analyzed the evolving graph structures over a period of time, based on a single metric using the HITS algorithm. Our workload differs from theirs since we do not have access to the underlying overlay network formed by email relays.

6 Conclusions

In this paper, we have shown that legitimate and spam email graphs differ in two fundamental classes of characteristics: structural, which capture the graphs' architecture, and dynamical, concerning node communication and graph evolution.

We have shown that the spam and non spam subgraphs are structurally characterized by different distributions of the clustering coefficient of their nodes. Legitimate email users display on average higher clustering coefficients than spam senders. Node visitation probability is a measure of the centrality of a node relative to other nodes in the graph. Legitimate email nodes have higher visitation probability than spam nodes. We also defined a new metric called communication reciprocity. It measures the probability that a node receives a response from any of its addressees. There is a strong difference in the probability distributions of the communication reciprocity in the legitimate and spam graphs; legitimate nodes have a much higher probability of being responded to. Another metric introduced in this paper is the email asymmetry set, which represents the difference between the sets of in and out edges of a node. We showed that there is a strong correlation between the size of asymmetry sets and the number of spammers in the set. Dynamically the spam graph grows much faster than the legitimate email graph. The legitimate email graph grows more slowly both in the number of nodes and edges, manifesting the higher stability of relations in a social group. Two other dynamical metrics, entropy and stack distance, are used to reveal the temporal characteristics of communication among nodes. Spam nodes display a much higher entropy than legitimate email users, and a much longer stack distance.

We have shown that differences in both classes of graph characteristics can be explained by the same hypothesis, namely that legitimate email graphs reflect real social networks, while spam graphs are technological networks, devoid of a sense of community. Although no single metric can unequivocally differentiate legitimate emails from spam, the combination of several graphical measures paint a clear picture of the processes whereby legitimate and spam email are created. For this reason they can be used to augment the effectiveness of mechanisms to detect illegitimate emails.

Acknowledgements

Luiz H. Gomes is supported by Banco Central do Brasil.

References

[1] V. Almeida, A. Bestavros, M. Crovella, and A. de Oliveira. Characterizing Reference Locality in the WWW. In *Proc. IEEE Conference on Parallel and Distributed Information Systems (PDIS)*, pages 92–103, Miami Beach, FL, December 1996.

[2] J. Balthrop, S. Forrest, M. E. J. Newman, and M. M. Williamson. Technological Networks and the Spread of Computer Viruses. *Computer Science*, 304, April 2004.

[3] L. Bertolotti and M. C. Calzarossa. Workload Characterization of Mail Servers. In *Proc. SPECTS 2000*, pages 301–307, Vancouver, Canada, July 2000. Elsevier Science Publishers B. V.

[4] P. O. Boykin and V. Roychowdhury. Leveraging Social Networks to Fight Spam. *IEEE Computer*, 38(4):61–68, April 2005.

[5] S. Brin and L. Page. The Anatomy of a Large-Scale Hypertextual Web Search Engine. In *Proc. 9th International World Wide Web Conference*, 1998.

[6] L. F. Cranor and B. A. LaMacchia. Spam! *Communications of the ACM*, 41-8:74–83, August 1998.

[7] P. Desikan and J. Srivastava. Analyzing Network Traffic to Detect E-Mail Spamming Machines. In *Proc. ICDM Workshop on Privacy and Security Aspects of Data Mining*, pages 67–76, Brighton UK, November 2004.

[8] S. N. Dorogvtsev and J. F. F. Mendes. Evolution of Networks. 51(66):1079–1187, 2002.

[9] L. H. Gomes, C. Cazita, J. Almeida, V. A. F. Almeida, and W. M. Jr. Characterizing a Spam Traffic. In *Proc. 4th ACM SIGCOMM Conference on Internet Measurement*, pages 356–369, Taormina, Italy, 2004. ACM Press.

[10] S. Hansell. Dozens Charged in Push Against Spam and Scams. http://www.nytimes.com/2004, August 2004.

[11] J. Jung and E. Sit. An Empirical Study of Spam Traffic and the Use of DNS Black Lists. In *Proc. 4th ACM SIGCOMM Conference on Internet Measurement*, pages 370–375, Taormina, Italy, 2004. ACM Press.

[12] Message Labs Home Page. http://www.messagelabs.co.uk/.

[13] M. E. Newman, S. F. S., and J. Balthrop. Email Networks and the Spread of Computer Viruses. 66-035101:1–4, September 2002.

[14] M. E. J. Newman and J. Park. Why Social Networks are Different from other Types of Networks. *Physical Review E*, 68-036122, September 2003.

[15] SpamAssassin Home Page. http://www.spamassassin.org.

[16] Spam Haus Project. http://www.spamhaus.org, 2005.

[17] R. D. Twining, M. M. Willianson, M. Mowbray, and M. Rahmouni. Email Prioritization: Reducing Delays on Legitimate Mail Caused by Junk Mail. In *Proc. Usenix Annual Technical Conference*, pages 45–58, Boston, MA, June 2004.