Studying E-mail Graphs for Intelligence Monitoring and Analysis in the Absence of Semantic Information

Petros Drineas, Mukkai S. Krishnamoorthy, Michael D. Sofka
Bülent Yener

Department of Computer Science, RPI, Troy, NY 12180, USA.
{drinep, moorthy, sofkam, yener}@rpi.edu [⋆]

Abstract. This work describes a methodology that can be used to identify structure and communication patterns within an organization based on e-mail data. The first step of the method is the construction of an e-mail graph; we then experimentally show that the adjacency matrix of the graph is well approximated by a low-rank matrix. The low-rank property indicates that Principal Component Analysis techniques may be used to remove the noise and extract the structural information (e.g. user communities, communication patterns, etc.). Furthermore, it is shown that the e-mail graph degree distribution (both with respect to indegrees and outdegrees) follows power laws; we also demonstrate that there exists a giant component connecting 70% of the nodes.

1 Introduction and motivation

E-mail communications play an important role in information society as a means for collaboration and knowledge exchange. It dominates business, social and technical transactions and it is an attractive area for research on community formation and evolution in the social networks context. Since individuals in an organization create formal or adhoc groups, their e-mail communication patterns usually carry implicit information regarding their common activities and interests.

This paper describes an experiment conducted in our institution (RPI) using e-mail message logs obtained over several days. Using this data we construct an *e-mail graph* that captures the communication patterns. In our work we examine the properties of e-mail graphs and study a variety of metrics; first of all, we validate that the distributions of the indegrees and outdegrees in our graph follow power laws. Next, we present a spectral analysis of the graph based on the Singular Value Decomposition (SVD) and demonstrate that the adjacency matrix of the graph is quite low rank. The low-rank property and the existence of power laws in the e-mail graph indicate that Principal Component Analysis

[⋆] This research is supported by NSF ITR Award #0324947.

techniques can be used for the discovery of user communities and communication patterns. Additionally, we observe that the graphs have a giant connected component, which could be used to help reduce the complexity of our algorithms.

Our work essentially validates the use of models such as the ones proposed in [1, 2] to model the e-mail graph. Thus, from such models, we can easily estimate the probability that a user sends/receives k e-mails within a period of time T (i.e. $p(k)$ will be $k^{-\alpha}$ for suitable choices of the exponent α, as we shall demonstrate in the experimental results section). We can also predict how the e-mail graph evolves over time, thus predicting future e-mail communications. Finally, such models might be used to design and evaluate strategies for the spreading of infectious software, such as worms and viruses, in the Internet via e-mail.

1.1 Related Work

Understanding on-line social networks and analyzing their structures has been a focus of intense research in both Social Science and Computer Science (please see [3] and references therein). E-mail is the predominant means of communication in on-line society. Most works have focused on information flow via e-mail and, in particular, the spread of a viral epidemic and the design of effective strategies to prevent such a spread. In [4] the authors analyze e-mail logs and constructed the corresponding e-mail graph over a period of 2 months for senders sending e-mail to or from the HP lab e-mail server; they demonstrate a power law distribution for the outdegrees of this graph (the exponent of the power law was close to 2.0). Similar work was done in [5]; the e-mail graph was generated from the address books of users in a large university system. The authors demonstrate that the degree distributions follow exponential distributions; notice though that they examine a different graph, since the address book of an e-mail user does not necessarily generate links to all e-mail recipients.

In [6], the authors analyze e-mail communication between members of an HP lab over a period of 2 months; their goal is to identify efficient strategies for searching such networks for a specific individual (node). They form a “social network” from this graph by putting an edge between two users if and only if the two users exchanged at least 6 e-mail messages. The resulting graph exhibits an exponential (and not a power law) distribution on the degrees of its nodes. However, in [7], the authors analyze a similar e-mail log, again from the HP labs, over a period of 2 months; they explicitly state that the “raw” data (namely, the graph created by adding an edge between user i and user j if user i sent an e-mail to user j) exhibits power-law degree distributions. Results of [8] imply that such graphs consist of a giant connected component and many smaller isolated components. Also, results of [9] imply that the eigenvalues of the adjacency matrix of the graph follow a power law distribution; we expand on these points in section 3.1.

Finally, we note that e-mail communications can be addressed at the organizational or workgroup level as suggested in [10]. Furthermore, application of

SVD to social networks are previously considered for discovering communities [11] [1].

Our Contribution The main contribution of our work is the application of SVD on a graph created from e-mail data and the spectral analysis of this graph. By showing existence of the power law and low-rank properties on e-mail graphs, we establish a basis for using SVD based data mining approaches to discover hidden communities in e-mail communications.

2 Model and Methodology

2.1 Data Collection and Processing

Data for the e-mail samples were taken from a full SMTP (Simple Mail Transport Protocol) feed at Rensselaer Polytechnic Institutes' central mail servers from one full week.

The SMTP protocol minimally identifies a connecting SMTP relay, the envelope sender, the envelope recipients and the DATA. The DATA contains the e-mail message body and additional e-mail headers. Sendmail logs a "from" line for the envelope sender and a series of "to" lines for each message recipient. These appear in the log file in the order that multiple, parallel, sendmail daemons processed the messages. When a new SMTP connection is established, sendmail assigns a unique queue id to the message. The queue id can be used to extract the linked log entries corresponding to a single e-mail message.

Logging is further complicated by e-mail forwarding, and alias expansion, which result in "ctladdr" and "clone" entries, respectively. RPI's e-mail logs also record the results of virus and Unsolicited Commercial E-mail (spam) filtering.

It is important to make clear the distinction between envelope sender and recipients, and the e-mail "From" and "To" headers. E-mail headers are DATA as far as the SMTP protocol is concerned, and are not recorded sendmail. Sendmail logs the envelope information, which appears in the SMTP dialog, and which may differ from the message headers. For example, a message sent with a Blind Carbon Copy (bcc) header will have envelope recipients which do not appear in the message headers. Likewise, a mailing list may use a envelope sender which is a special bounce detection address with a unique identification tag, while the message header provides the true list name and reply-to address. The provided data included only envelope information.

Personally identifiable information in the logs was obscured using the HMAC message authentication protocol with a 128bit SSH1 hash. Identifying information included the envelope sender and recipients, the connecting mail relay, the message id (a unique id generated by e-mail clients), and delivery status information.

[1] The authors thank to the anonymous referees for pointing out additional references.

Information about spam scores and viruses were not included in the logs used for the study. The format of virus and spam messages has gone through numerous changes over the past year, making relatively safe obfuscation error prone. In addition, messages from spammers and virus infected machines is intentionally misleading, and attempt to exploit bugs in client e-mail readers. As a result, the log entries often contain non-standard characters (especially in envelope sender, and message id and attachment names) making consistent parsing difficult. However, such messages can often be inferred from delivery status errors, unknown users and missing recipients (when the spam is dropped by Rensselaer's spam filter).

2.2 The e-mail graph

From the e-mail data we construct a graph represented by an adjacency matrix A. In the graph each node corresponds to a concealed e-mail address (i.e., the rows and columns of A) and there is an edge (i, j) in the graph if node i sends an e-mail to node j.

2.3 SVD and Spectral Analysis

We briefly review the singular value decomposition of matrices; we will use some of its properties in our discussion in Section 3.1. Any $m \times n$ matrix A can be expressed as

$$A = \sum_{t=1}^{r} \sigma_t(A) u^{(t)} {v^{(t)}}^T,$$

where r is the rank of A, $\sigma_1(A) \geq \sigma_2(A) \geq \ldots \geq \sigma_r(A) > 0$ are its singular values and $u^{(t)} \in \mathcal{R}^m, v^{(t)} \in \mathcal{R}^n, t = 1, \ldots, r$ are its left and right singular vectors respectively. The $u^{(t)}$'s and the $v^{(t)}$'s are orthonormal sets of vectors; namely, ${u^{(i)}}^T u^{(j)}$ is one if $i = j$ and zero otherwise. We also remind the reader that

$$\|A\|_F^2 = \sum_{i,j} A_{ij}^2 = \sum_{i=1}^{r} \sigma_i^2(A),$$
$$\|A\|_2 = \max_{x \in \mathcal{R}^n : \|x\|=1} \|Ax\| = \max_{x \in \mathcal{R}^m : \|x\|=1} \|x^T A\| = \sigma_1(A).$$

In matrix notation, SVD is defined as $A = U \Sigma V^T$ where U and V are orthogonal (thus $U^T U = I$ and $V^T V = I$) matrices of dimensions $m \times r$ and $n \times r$ respectively, containing the left and right singular vectors of A. $\Sigma = \mathbf{diag}(\sigma_1(A), \ldots, \sigma_r(A))$ is an $r \times r$ diagonal matrix containing the singular values of A.

One of numerous applications of SVD is in recovering the structure of matrices and in noise removal in a variety of settings. The underlying idea is very simple: since $A = \sum_{t=1}^{r} \sigma_t(A) u^{(t)} {v^{(t)}}^T$, we can create approximations to A by keeping only the top k "principal components" (i.e. the top k $\sigma_t(A)u^{(t)}v^{(t)^T}$) for various values of k. Essentially, discarding the "smallest" principal components (the ones corresponding to the smallest singular values), results in a small loss in accuracy, and we might justifiably consider these components as "noise". This procedure is commonly referred to as "Principal Component Analysis" and the following theorem (usually attributed to Eckart and Young [12]) quantifies the loss of accuracy incurred by keeping only the top k components for various values of k.

Theorem 1. *Let $A_k = \sum_{t=1}^{k} \sigma_t u^{(t)} v^{(t)^T}$ (for any $1 \leq k \leq r$). A_k is the "best" rank k approximation to A with respect to the 2-norm and the Frobenius norm; namely, for any matrix D of rank at most k,*

$$\|A - A_k\|_2 \leq \|A - D\|_2 \qquad \textit{and} \qquad \|A - A_k\|_F \leq \|A - D\|_F.$$

Also,

$$\|A - A_k\|_F^2 = \sum_{t=k+1}^{r} \sigma_t^2(A) \qquad \textit{and} \qquad \|A - A_k\|_2 = \sigma_{k+1}(A).$$

We say that a matrix A has a "good" rank k approximation if the 2-norm and the Frobenius norm of $A - A_k$ is small; for a detailed treatment of Singular Value Decomposition see [12].

3 Experimental Results

3.1 Spectral Analysis of the e-mail graph

Not surprisingly, the e-mail graphs exhibit low-rank structure. As a result, Principal Component Analysis techniques would be successful if applied on such graphs to extract communities of users, remove noise, etc. In Figure 3 we plotted the spectral characteristics of the 3 graphs. The graphs are obtained from splitting one week mail logs into three sets for managing the complexity of SVD algorithm. More specifically, the plot demonstrates how accurately we can approximate the given graphs by keeping a small percentage of the maximal number of principal components (see Section 2.3 for background) that comprise the graphs; notice that for a graph with n vertices, we might have up to n principal components.

We now describe the findings of Figure 3. Notice that the x-axis represents the percentage of principal components (out of the maximal possible number of principal components) that are kept, while the y-axis represents the percentage of the spectrum of the graph contained within these principal components; more specifically, let A denote the adjacency matrix of a graph. Then, the y-axis represents the ratio

$$\frac{\|A - A_k\|_F^2}{\|A\|_F^2},$$

which may be viewed as the relative error of the approximation A_k to A using a certain percentage of the maximal number of principal components. Recall that the above ratio is equal to

$$\frac{\sum_{t=k+1}^{r} \sigma_t^2(A)}{\sum_{t=1}^{r} \sigma_t^2(A)}.$$

Thus, in order to compute the ratio, we computed a large number of singular values for the adjacency matrix A of each graph using MatLab. Notice that a very small percentage (say 5%) of the principal components is enough to cover almost 80% of the spectrum for all three graphs. Thus, we conclude that the graphs have a low-rank property, which directly implies that Principal Component Analysis could be used to remove the noise and extract their structure. We should also mention that our findings are consistent with the power-law distribution of the indegrees and outdegrees of the graph nodes. In [9], the authors studied the distribution of the eigenvalues of graphs whose degree distributions (on the nodes) follow power laws, and they demonstrated that the eigenvalues exhibit a power law distribution themselves; we defer a more thorough analysis of this connection to the full version of the paper.

3.2 Power Laws in E-mail Graph

In this section we show that *(i)* degree distribution of e-mail graph obeys power laws, *(ii)* there is a giant component within each graph that spans approximately 70% of the nodes, *(iii)* the diameter of the giant component is small, and *(iv)* the neighborhood connectivity of a node is quite sparse.

The **diameter** is the longest shortest path between any pair of nodes in a connected graph. It reflects how far apart two nodes are (from each other) in the e-mail graph. We computed the diameter of the giant component and the results are shown in Table 1.

The **clustering coefficient** reflects the connectivity information in the neighborhood environment of a node [13]. It provides the transitivity information [14], since it controls whether two different nodes are connected or not, assuming that they are connected to the same node. The clustering coefficient C_i is defined as the percentage of the connections between the neighbors of node i, i.e.

$$C_i = \frac{2 \cdot E_i}{k \cdot (k-1)}, \tag{1}$$

where k is the number of neighbors of node i and E_i is the number of existing connections between its neighbors. We compute C_i for all nodes i in the giant component and take the average to get a global value. Table 1 shows that the overall clustering coefficient is quite low.

In Figures 1 and 2 we show that there exists a giant component, and that both the whole graph and the giant component have power law degree distributions. For the whole graph, the in-degree exponent is around 1.9, while for the giant component it is approximately 1.7. The out degree exponents of the giant component and the whole graph are also close (3.1 and 3.2, respectively).

E-mail Graph Size	**Giant Component (GC) Size**	**GC Clustering Coefficient**	**GC Diameter**
22776	15260	0.0013	23
21614	14996	0.0025	20
21732	15400	0.0012	26

Table 1. E-mail Graph Properties.

The power-law distribution and the in-degree, out-degree exponent difference intuitively emerge from the observation that there are a few individuals inside the organization getting many external e-mails (e.g., member of mailing lists), and there are also some individuals inside the organization sending out announcements to many people, thus having very high outdegrees (e.g., program committee members for a conference, etc.).

4 Discussion and Conclusions

This work establishes a basis for applying Principal Component Analysis (PCA) to e-mail communication graphs, to discover groups and communication patterns, by showing existence of the low-rank property and power laws in these graphs. It also shows existence of the giant component property which helps identifying the groups in other connected components thus reduces the complexity of the data mining algorithms. Based on the power laws one can now determine the probability that a user sends/receives k e-mails with in a period of time T. This probability can be used to model how e-mail graphs evolve over time (i.e., prediction of future e-mail communications).

However, we also note the limited applicability of this approach to discovering malicious groups such as terrorist cells (a.k.a., the adversary in security jargon). In particular, the adversary may use freely available e-mail servers like Yahoo or Hotmail with fake e-mail IDs. Furthermore, the adversary can conceal its communication patterns by inducing noise into the system in the form of unsolicited mail or self e-mailing.

References

1. Babarasi, A., Albert, R.: Emergence of scaling in random networks. Science **286** (1999)
2. Kleinberg, J., Kumar, S., Raghavan, P., Rajagopalan, S., Tomkins, A.: The web as a graph: measurements, models and methods. In: Proceedings of the International Conference in Combinatorics and Computing. (1999)
3. Garton, L., Haythornthwaite, C., Wellman, B.: Studying online social networks. http://www.ascusc.org/jcmc/vol3/issue1/garton.htm (1997) (accessed March 16, 2004).

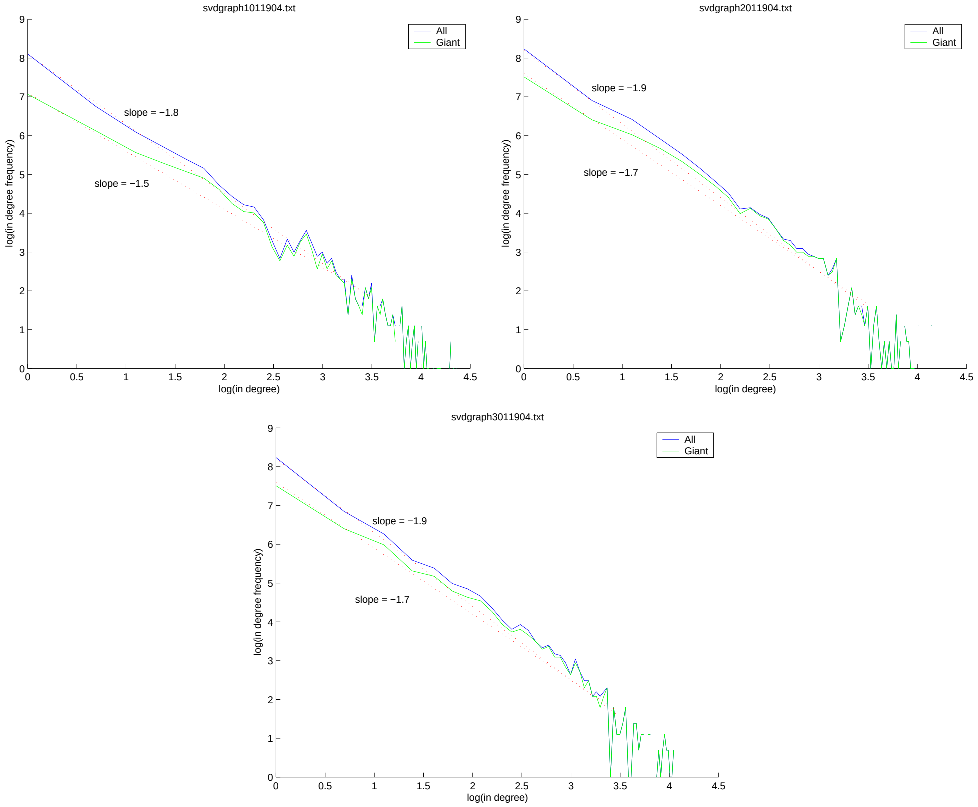

Fig. 1. The in-degree distribution of e-mail graph

4. Wu, F., Huberman, B., Adamic, L., Tyler, J.: Information flow in social groups (2003) manuscript.
5. Newman, M., Forrest, S., , Balthrop, J.: E-mail networks and the spread of computer viruses. Physical Review **(E)66** (2002)
6. Adamic, L., Adar, E.: How to search a social network (2003) manuscript.
7. Tyler, J., Wilkinson, D., Huberman, B.: E-mail as spectroscopy: Automated discovery of community structure withing organizations. In: Proceeding of the International Conference on Communities and Technologies, Netherlands, Kluwer Academic Publishers (2003)
8. Aiello, W., Chung, F., Lu, L.: A random graph model for massive graphs. In: STOC. (1999) 171–180
9. Chung, F., Lu, L., Vu, V.: Eigenvalues of random power law graphs. Annals of Combinatorics **7** (2003)
10. Nonaka, I.: A dynamic theory of organizational knowledge creation. Organization Science **5** (1994) 14–37
11. Freeman, L.C.: Visualizing social groups. In: 1999 Proceedings of the Section on Statistical Graphics, American Statistical Association (2000) pp:47–54

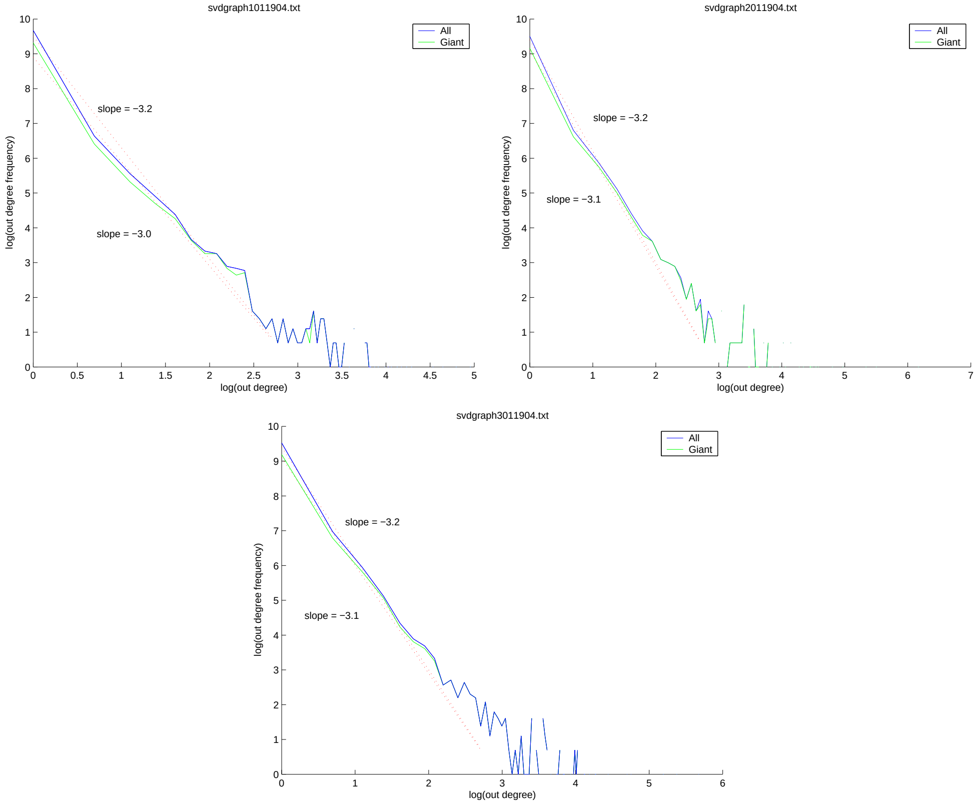

Fig. 2. The out-degree distribution of the e-mail graph

12. Golub, G., Loan, F.V.: Matrix Computations. Johns Hopkins University Press (1984)
13. Dorogovtsev, S.N., Mendes, J.F.F.: Evolution of Networks. Advances in Physics **51** (2002) 1079–1187
14. Newman, M.: Who is the best connected scientist? a study of scientific coauthorship networks. Physical Review (2001)

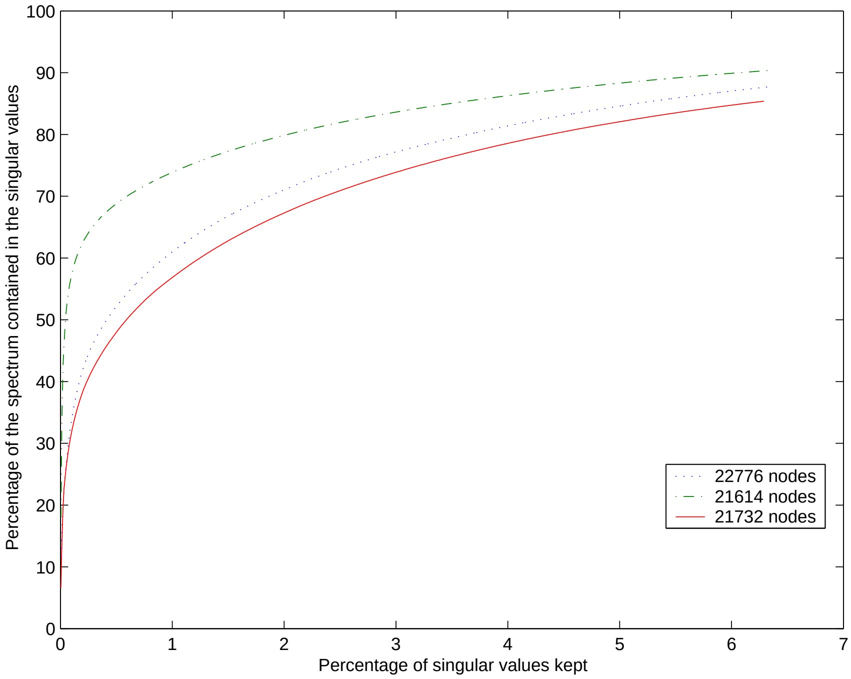

Fig. 3. Adjacency matrix spectrum vs. rank