MODELING AND SIMULATION OF E-MAIL SOCIAL NETWORKS: A NEW STOCHASTIC AGENT-BASED APPROACH

Fabian Menges
Bud Mishra
Giuseppe Narzisi

Computer Science Department, Courant Institute of Mathematical Sciences
New York University
New York, NY 10012, U.S.A.

ABSTRACT

Understanding how the structure of a network evolves over time is one of the most interesting and complex topics in the field of social networks. In our attempt to model the dynamics of such systems, we explore an agent-based approach to model growth of email-based social networks, in which individuals establish, maintain and allow atrophy of links through contact-lists and emails. The model is based on the idea of common neighbors, but also on a detailed specialization of the classical preferential attachment theory, thus capturing a deeper understanding of the topology of inter-node connections. In our event-based simulation that schedules the agents' actions over time, the proposed model is amenable to significant efficiency improvements through an application of the Gillespie stochastic simulation schemes. Computer simulation results are used to validate the model by showing that its unique features endow it with ability to simulate real-world email networks with convincing realism.

1 INTRODUCTION

Understanding and utilizing the topological structure of social networks has been at the forefront of computer science research in recent years. This interest has been primarily motivated by our desire to identify sub-structures and "important" individuals in social network graphs, and to then use them for a variety of different purposes and applications (e.g., commercial, security, criminal investigation, etc.). In parallel, we also try to identify and extract data that implicitly represent social networks.

Currently, the term "social network" is usually associated with Internet communities such as Facebook, MySpace, LinkedIn, Friendster, Xing and others. Artificially created communities such as these have the advantage of conveying a deep knowledge about their members and the information that connects them, but they suffer from other disadvantages, such as limitations imposed by the comparatively small number of members, as well as being designed for

small target audiences. While most of the research focus today is directed to the analysis of the World-Wide-Web networks and web-based friendship communities, our approach points to other neglected aspects of social networks: namely, modeling natural social networks (e.g., email networks) created via ubiquitous digital communication that connects people around the world, and understanding how its evolution may allow external manipulations for a desired purpose.

Investigation of social graphs generated from email datasets is a topic receiving much interest recently. For example, in (Rowe, Creamer, Hershkop, and Stolfo 2007) the authors develop a novel algorithm for automatically extracting social hierarchy data from electronic communication behavior. Newman and co-authors in (Newman, Forrest, and Balthrop 2002) discuss the implications of an email network structure for the understanding and prevention of computer virus epidemics by empirically investigating the structure of a network derived from a large computer installation. However, the research is hampered by the paucity of publicly available email datasets that can be used for research purposes. To address this problem, in this paper we present a new agent-based model, which generates simulated networks similar to real-world email networks, and thus provides a basis for further designs of applications aimed at such social networks. Furthermore, we validate it by comparing the model output to the well-known Enron email dataset (Shetty and Adibi 2004, Diesner, Frantz, and Carley 2005).

These models have many real-world applications: in particular, they can be used to create inter-net-based expert/maven-networks that can then be applied to statistically estimate various properties of the social networks in an unbiased and robust manner. These properties can then be fruitfully exploited by search engines, recommenders, collective price discovery and targeted social ad tools.

The paper is structured as follows: First, a quick background on social network modeling is given in section 2. Next, the Email Network Model (ENM) is described in de-

tail in section 3 in terms of agents behaviors and dynamics. Simulation results are discussed in section 4, comparing the model output to the structure of the Enron dataset. Finally, a conclusion section appears as section 5.

2 Background

Erd´os-Renyi style random graphs have been found to lack the generality needed to model real-world networks. By using models with more elaborate static structures while ignoring the interactions among neighbors and consequent changes to the local topologies, one can only partially address the limitations of the random networks. In certain situations, naturally occurring networks, such as metabolic networks or blood vessels, are fundamentally quasi-static and do not change their topology perceptibly over time; however, most real-world networks are open, and they form by the continuous accretion and attrition of vertices in the system. Thus, the number of vertices fluctuates throughout the lifetime of the network, with many small and large changes to the network topology.

For example, scientific collaboration networks grow by the addition of new scientists to the system; the World-Wide-Web grows exponentially over time by the addition of new pages; and the research literature constantly grows by the publication of new papers. The common feature shared by all of these systems is that the network expands through the addition of new nodes that are connected to the nodes that were already present in the system.

There are various models of network growth (Newman 2003, Albert and Barabasi 2002, Liben-Nowell and Kleinberg 2007) that propose different heuristic solutions for adding nodes and/or edges to a network over time. Two well-known and popular ones are the small world model (Watts 2003), and the model of Barabasi and Albert (Barabasi and Albert 1999).

The small world model incorporates the idea that social networks illustrate clustering (Watts 2003). This is based on the notion that, if two neighboring "nodes" or individuals within a social network have a mutual connection to a third node, then the first two individuals are far more likely to have a direct connection to each other as well (as compared to the lesser likelihood that two random individuals would form a connection).

Similarly, the Barabasi and Albert model (Barabasi and Albert 1999) is based on the idea that constantly growing networks use *preferential attachment* of new nodes, meaning that individuals on the outskirts of a social network are more likely to connect to nodes with a high degree of connectivity (i.e., large number of "friends") than to ones with fewer connections. Models founded upon this theory have been shown to be representative of the growth of many different kinds of networks: for example, that of Wikipedia (Capocci 2006), the network of

scientific collaboration (Roth 2005) and proteins networks (Eisenberg and Levanon 2003).

Finally, in (Jin, Girvan, and Newman 2001), Jin and co-authors developed two models (one simpler than the other) of growth for social networks based on classical stochastic processes that are able to reproduce many features of real social networks, including high levels of clustering, or network transitivity and strong community structure. Even though both ours as well as their models simulate the growth of social networks, they differ in many subtle ways. Jin's model uses a probability per unit time p_{ij} that two people, say i and j, will meet, which depends both on the total number of friends and the number of mutual friends of the two selected people. This way, the evolution of the structure is based on top-level rules that operate on the whole set of nodes in the network. Our model instead is based on an agent-based bottom-up approach in which the creation of new connections between nodes is due only to the specific decisions made over time by each single person in the network. This way, the final network generated is the result of the emergent behavior of the system. In order to pick pairs of individuals with the correct probability per unit time, the first of Jin's two models uses a continuous-time Monte Carlo method, while the second one, albeit more efficient, is based on few artificial rules that can be argued to not represent real-world scenarios.

As we will describe later, as opposed to Jin's framework, the scheduler of our model-simulator operates on the same working principle as that of the Gillespie algorithm (Gillespie 1977), thus allowing more efficient simulation than continuous Monte Carlo methods while still using rules of behaviors that reflect the way real people manage their emails and interact.

3 THE EMAIL NETWORK MODEL

The objective of this paper is to develop a model that is capable of simulating the behavior of an email social network. In conjunction with the development of the model, this paper derives the model parameters from real-world data within the Enron email dataset, qualitatively comparing the model outputs to the real data. For this purpose, we created extensive statistics on the Enron dataset using the MySQL database (available for download at (<www.isi.edu/~adibi/Enron/Enron.htm>) created by the University of Southern California. The database consists of 252,759 emails that were received 2,064,442 times. Extensive research has been performed to analyze this dataset, resulting in the widespread consideration of a subset of 151 employee and approximately 517,431 email receptions (Shetty and Adibi 2004). Taking only emails that were sent and received by these 151 employees into account, we reduced the number of unique emails to 14,876 over the course of roughly three years.

3.1 Model features

The Email Network Model (ENM) is an Agent-Based model implemented in Repast 3.1 (North M.J. and J.R. 2006), a popular and versatile Java-Based software toolkit that has been used to model such diverse concepts as intracellular processes, social norms, planning against catastrophes (Narzisi, Mysore, Byeon, and Mishra 2006, Mysore 2006) and business strategies. We have also integrated Repast with JUNG, the Java Universal Network/Graph Framework (<jung.sourceforge.net/>), which supports modeling, analysis and visualization for network analysis.

Our email model is built upon the idea of common neighbors and a modification of preferential attachment. We model the email traffic in a closed community. This community can represent, for example, a research network, a company, a sports club, etc.; in short, we are interested only in the connections (emails sent) to individuals within this community, as opposed to emails sent to external individuals.

The main idea of common neighbors still remains valid: namely, the likelihood of two neighbors of a third node becoming neighbors themselves is higher than that of a total random connection. However, we enhance this theory with the hypothesis that this behavior is even more likely if multiple neighbors of an individual are the recipients of the same unique email, thus implying a deeper level of connection. As a basic illustration of this hypothesis, imagine an email that is written by the CEO of a company and sent as a carbon copy to every employee within the organization. The likelihood that the recipients of this email– the employees–have connections to each other independent of their mutual connection to the CEO is far greater than if they were simply random individuals existing in an open-ended network.

We incorporate the idea of growth in the model by first identifying "a seed set," or the original cluster of individuals from which the network evolves. Because the model is based on the concept of a closed community, the network is not able to grow infinitely; thus, the growth of the network is restricted by establishing a maximum number of individuals.

We also incorporate the idea of preferential attachment, but we modify it by adding a new dimension to the Barabasi-Albert model. That original model focused exclusively on the premise of quantity: individuals with a large number of connections attract the outliers of the network solely because of this numeric "popularity." We take this basic theory and expand it to include an element of quality, in which we define an individual's attractiveness to social subordinates by specific qualitative attributes that are uniquely relevant. Specifically, an individual's degree of connectivity within a social network does correlate to the number of his connections, but also to social characteristics that make him "attractive" in a specific community. These two factors work in tandem to shape and define a social network's structure.

The last main concept that we incorporate is inspired by the Exact Stochastic Simulation (Gillespie 1977) by Daniel Gillespie (1977). The Gillespie algorithm has been developed in the context of chemical interactions in order to model the time behavior of spatially homogeneous chemical systems efficiently and accurately using limited computational power. However, its basic working principle can be applied to simulate increasingly complex systems.

Following Gillespie's ideas, we could improve the time-complexity of the simulation, even when the underlying time implicitly behaves continuously; Gillespie's techniques rely on the fact that, if the time distributions of actions are known in advance, then it is possible to schedule the actions based on the answers to the following two questions:

1. Which action occurs next?
2. When does it occur?

We apply this method in the present context on the basis that, for every node, sending an email and removing a contact from the contact list are both recurring events. These two events follow specific distributions (explained in section 4) and are organized by a schedule that invokes and organizes them at a given time based on the following simple steps:

- INITIALIZATION: Initialize the number of agents, parameters and random number generators.
- MONTE CARLO STEP: Generate random numbers (according to some distribution) to determine the next action to occur, as well as the time interval.
- UPDATE: Increase the time step by the randomly generated time in the Monte Carlo Step.
- ITERATE: Go back to the Monte Carlo Step, unless the simulation time has been exceeded.

The specific distributions used in this sequence of steps are described in section 4.1.

3.2 Agents and rules of behaviors

The model consists of the interaction between two different kinds of agents: *individuals* (also referred to as nodes) and *emails*. In the following two paragraphs, we describe the characteristics and rules of behavior of the agents both in static and dynamic settings.

Emails Nodes only interact with each other through emails. If we consider the network created by the model as a graph, the emails do not directly represent the edges. Rather, the edges only connect nodes that had a bidirectional email transfer, i.e. node A sent an email to node B, and vice versa. This definition of edge has a number of advantages, and it represents reality better than using graph definition derived

directly from emails. For example, spam and mass emails do not affect the density of the network, which influences the degree of distribution, and therefore the network growth, by preferential attachment.

Properly defined emails must have one sender and at least one receiver. They must also get a time stamp assigned to them at the time of sending. Once an email is sent, it performs only one action: invoking a receiving action (open the email, delete the email, etc.) among all recipients.

Individuals/Nodes Every node is characterized by two parameters representing the most simplified notion of personality. The following two personality traits model the "quality" of each node:

1. ATTRACTIVENESS: is a random variable attribute that follows a uniform distribution. The attractiveness is defined in the range $[0,1]$ and affects the interest of other nodes in sending emails to this node. The larger the attractiveness, the higher the probability of receiving emails from other contacts becomes. In addition, it is also the only parameter that affects the removal of contacts from the contact list.

2. INVOLVEMENT: represents the different types of behaviors of individuals in social networks based on participation and motivation. Involvement is defined as having a value in the range $[0,1]$ and follows a distribution with probability density function:

$$f(z) = \frac{1}{log(P+1)} \times \frac{1}{z+1} \times \mathbb{1}_{[0,P]}(z) \quad (1)$$

where P is the involvement parameter set to 200. Higher the involvement, higher the rate of sending emails.

Besides the personal traits, each node maintains a contact list. New nodes are added to the contact list when the list's owner sends an email to another node. Nodes are removed from the contact list as a reoccurring event at a rate determined by the probability p_{del} (defined in paragraph 3.3). Nodes with low attractiveness are more likely to be removed from the contact list than other nodes.

3.3 Description of the dynamic behavior

To explain the mechanisms for the growth and evolution of our social network, let us follow the time response of the model. For the sake of simplicity, we assume that the simulation begins with one individual as starting seed, from which the rest of network evolves.
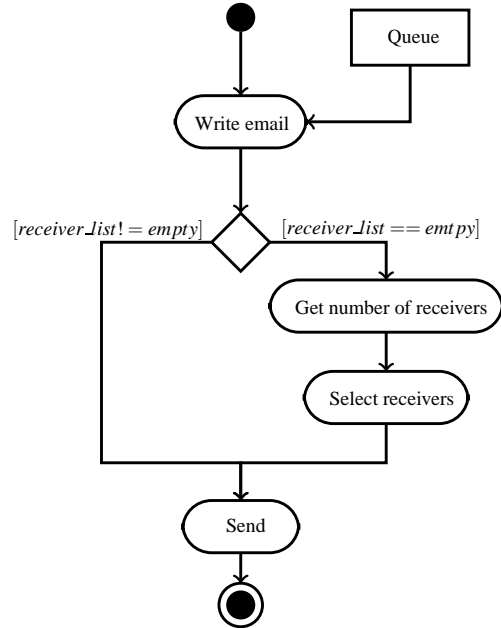

Figure 1: Activity diagram for "Send Email".

Send Email The standard action for the first node is to send an email, and this actions is scheduled at a random time following the distribution S_n (probability distribution for node *n*). Its activity diagram is shown in Figure 1. The schedule jumps to the first scheduled action, which is "Send Email." The first step this action takes is to reschedule itself at some time in the future, following the same distribution S_n. Specifically the time to next email is calculated as:

$$t_{send}^{i+1} = \frac{1}{inv_i} \cdot N(S_n), \quad (2)$$

where inv_i is the involvement-value of node *i* and $N(S_n)$ is a random number according to distribution S_n.

Next, the actual process of sending an email is invoked, which involves making the following decisions: (1) deciding the number of receivers and (2) selecting the receivers. Each node has a FIFO queue containing a list of scheduled email information. If the queue contains information about an email, this email is sent. If not, as in this case, the following actions are invoked:

- GET NUMBER OF RECEIVERS: This is a uniformly-distributed random number in the range of 1 and the size of the contact list.
- SELECT RECEIVERS: This is a node of one of following three kinds: (*i*) One from the contact list of the sender (L_{sender}), or (*ii*) one from previously selected nodes from contact list ($L_{selected}$), or (*iii*) a node which was previously unknown to the sender.

Whether (*i*), (*ii*), or (*iii*) determines a recipient at a particular selection-point is a stochastic event for which we introduce the three global parameters:

- p_{cl} for case (*i*),
- p_{sel} for case (*ii*), and
- p_{un} for case (*iii*).

Where $p_{cl}, p_{sel}, p_{un} \in [0,1]$ and $p_{cl} = 1 - p_{un}$. These parameters can be used to model different kinds of communities. See Figure 2 for the activity diagram of "Select receivers". The function $F(L_n)$ is the main function for the
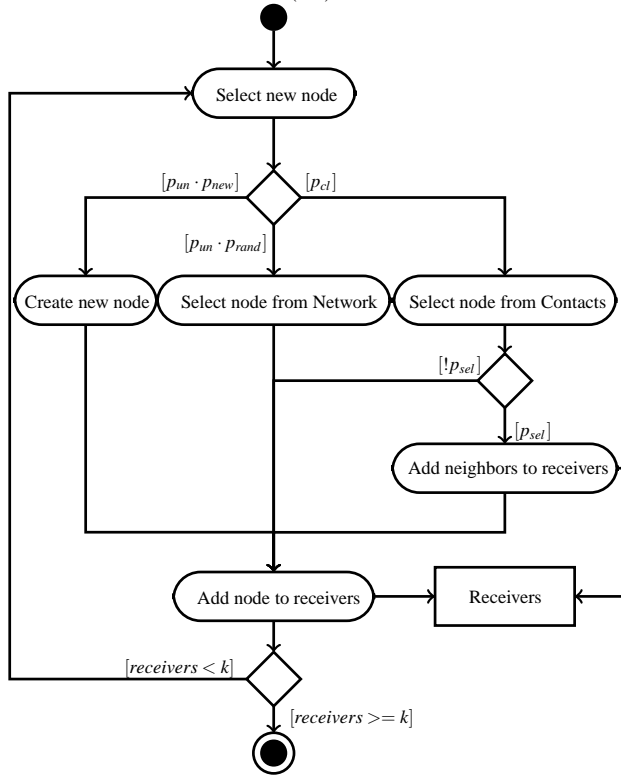

Figure 2: Activity diagram for "Select receivers".

link prediction. It samples a uniform random number of nodes from the contact list L_n according to a combined probability of the degree and attractiveness of each node. The mapping is determined by assigning to each node i a probability of being selected p_i in the following way. First the average degree weight (p_{d_i}) and the average attractiveness weight (p_{a_i}) are computed as follow:

$$p_{d_i} = \frac{d_i}{\sum_{j=1}^{N} d_j}, \ p_{a_i} = \frac{a_i}{\sum_{j=1}^{N} a_j}, \tag{3}$$

where d_i and a_i are respectively the degree and attractiveness of node i, and N is the total number of contacts in i's contact

list. Hence, the selection probability p_i is defined as follows:

$$p_i = \frac{p_{a_i} + p_{d_i}}{\sum_{j}^{N}(p_{a_j} + p_{d_j})} = \frac{p_{a_i} + p_{d_i}}{2}. \tag{4}$$

Thus, we have that:

$$p_i = \frac{\sum_{j=1}^{N}(a_i d_j + d_i a_j)}{2 \cdot \left(\sum_{j=1}^{N} a_j\right) \cdot \left(\sum_{j=1}^{N} d_j\right)}. \tag{5}$$

If one node is selected as receiver, there is a small probability p_{sel} that some nodes belonging to its contact list will be added to the receiver list as well. In other words, once the function $F(L_{sender})$ returns a node n, a uniform random number of this node's neighbors will be added to the receiver list with probability p_{sel}. This step reflects the core idea of the small world model, as it represents the chance of getting introduced to a new contact by a friend using a means of communication other than email. Those neighbors are selected according to the same mapping function $F(L_n)$.

When a new unknown node is selected, this node can be either a random node of the network or a newly created node. The probability of creating a new node p_{new} is simply the normalized difference between the estimated network size and the the actual network size.

$$p_{new} = 1 - \frac{\text{actual network size}}{\text{estimated network size}}. \tag{6}$$

Hence, the probability of selecting a random node from the network p_{rnd} is:

$$p_{rnd} = 1 - p_{new} = \frac{\text{actual network size}}{\text{estimated network size}}. \tag{7}$$

When a new node is created the network grows. The initial attributes of the node are set and its reoccurring events are scheduled. If, on the other hand, a random node from the network is selected, such selections follow a distribution that is based only on the attractiveness values of all nodes in the network and use the same technique for selection as in the case of the contact list. Since the probability for a node i to be selected is only based on its attractiveness, it follows equation (3) with N representing all nodes in the network instead of a single contact list.

At the end of this process, k receivers are selected and the email is sent. More specifically, an email agent is created with the attributes described above and scheduled for reception. The list of k receivers is then added to the list of contacts of the sender. Since our network does not contain any node other than the seed, k new nodes are created with a new set of initial random attributes. For these new nodes, the standard action of "Send Email" and
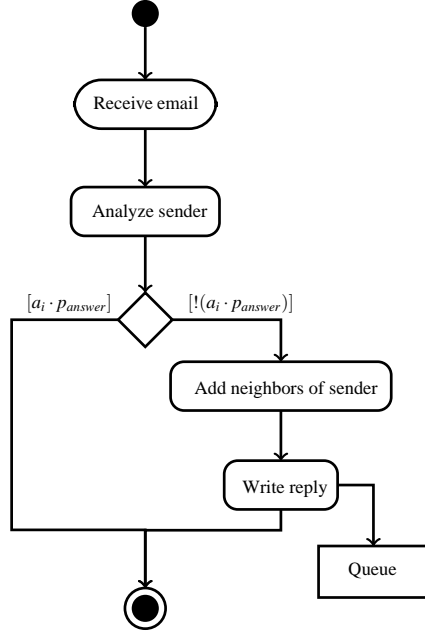
Figure 3: Activity diagram for "Receive Email".

"Remove Contact" are scheduled.

Receive Email   As the simulation proceeds, the next scheduled action of the model is invoked, which will likely be the reception of the newly created email. Figure 3 shows the activity diagram for the "Receive Email" action.

When an email is received, with some probability, the recipient responds to that email by sending a reply. This step involves a stochastic decision which is a function of the attractiveness of the sender a_{sender} and, a global parameter, the probability to answer $p_{answer} \in [0,1]$. We define the total probability of replying to an email p_{reply} as:

$$p_{reply} = p_{answer} \cdot a_{sender}. \qquad (8)$$

If the recipient replies to an email, it will reply to all attractive nodes that were also recipients of that same email. Replay emails are stored in the FIFO queue of the node and are scheduled to be sent according to the same distribution S_n described later in section 4.1.

While sending a new email shows the standard behavior of a small world model, neighbors are preferred over strangers, the emails sent as a reply enhances this effect and gives it a time dependent notion.

Remove Contact   As mentioned earlier, nodes are removed from the contact list as a reoccurring event. When invoked, this action removes one node i from the contact

list. The probability to be selected for deletion is:

$$p_{del} = \frac{\frac{1}{a_i}}{\sum_j^N \frac{1}{a_j}}. \qquad (9)$$

where a_i is the attractiveness of node i and N is the size of the node i's contact list. p_{del}, as defined, causes the less attractive nodes to be removed from the contact list with a higher probability. When this action is invoked, the time to next node-deletion event is calculated as follows:

$$t_{del}^{i+1} = \left(\frac{1}{inv_i} \cdot N(S_n) \right) \cdot \frac{1}{d_i \cdot p_{un}}, \qquad (10)$$

where inv_i is the involvement for node i, $N(S_n)$ is a random number according to the distribution S_n, and d_i and p_{un} are defined as before.

The set of rules as defined above describes the entire model.

4   SIMULATION RESULTS AND DISCUSSION

4.1   Emails time distribution

The distribution most central to this model is the distribution of the time intervals between two consecutive emails. We chose to analyze and infer this data directly form the Enron dataset. Inspecting Figure 4, it can be observed that the distribution of time intervals between two consecutive emails (S_n) sent by the same node n follows a log-normal distribution. The blue line represents the empirical data from the Enron email dataset. The fitted distribution is a log-normal distribution with the probability density function:

$$PDF_S(x, \sigma_t, \mu_t) = \frac{1}{\sqrt{2\pi}\sigma_t} \cdot e^{\frac{-(\ln x - \mu_t)^2}{2\sigma_t^2}}. \qquad (11)$$

The fitting is obtained using the following values for the log-normal distribution: $\mu_t = 3.35$ and $\sigma_t = 3.6$.

We emphasize that this is the only information that is extracted from the Enron dataset and used in the model. Furthermore, this distribution is applied locally at the level of each single agent.

Table 1: Model parameters.

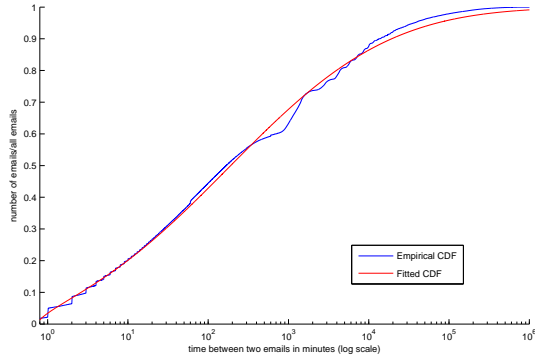| Name	Value
μ_t	3.35
σ_t	3.6
involvement	200
attractiveness	uniform in $[0,1]$
p_{cl}	0.98
p_{un}	0.02
p_{sel}	0.04
p_{answer}	0.25

Figure 4: Comparison between the CDF of the fitted distribution S and the empirical data from the Enron email dataset.

Figure 5: Cumulative distribution function (CDF) of the number of emails sent for both the ENM model and Enron.

Table 2: Comparison of statistical properties between the ENM and Enron networks.

Parameter	ENM	Enron
Mean distance	6.85	7.48
Reciprocity	0.32	0.41
clustering coefficient	0.16	0.28
Num. emails	15113	14876
mean degree	100.09	98.52

4.2 Statistical properties

After an extensive sensitivity analysis we have observed that in order to approximate the behavior of the Enron dataset the parameters must be selected as described in table 1. In order to compare the output of the ENM model with the Enron dataset we simulate the growth of the network for a maximum of 3 "virtual" years. Table 2 provides a summary of the statistical properties of the two networks. Inspecting the results in the table it is clear that the two graphs have similar properties. The values of ENM graph approximate quite well the Enron dataset except for the clustering coefficient where Enron is qualitatively higher than the ENM model. However, the value of the ENM network is definitely higher than the expected clustering of a random graph of the same size, which is typical in social network graphs. Finally, in Figure 5 it is shown that the cumulative distribution of the number of emails sent by each node closely follows that of the Enron dataset. It is clear that the two distribution match very well.

4.3 Model dynamics

Email networks, like many other social networks, are characterized by their extremely dynamic nature. New nodes and links are always introduced over time and the process that regu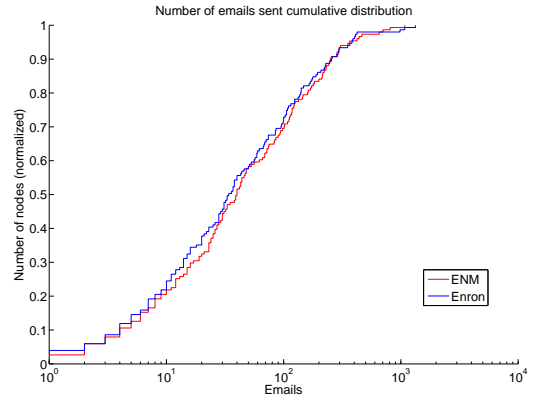lates the dynamics of such systems is of great interest and not completely understood yet. Figure 6 and 7 shows the evolution of both the in-degree and out-degree distributions as a function of time. It is clear from the plots that, as time goes on, the degree distributions of the ENM generated graph get closer to the Enron one. Similar behavior can be observed in Figures 8 where the distribution for received emails is plotted over time.

It is important to observe that the these distributions have shapes very similar to that of the Enron's email network. We may, thus, conclude that even if the total degree size are sometimes slightly different for the two networks the internal distribution is the result of a similar growth process.
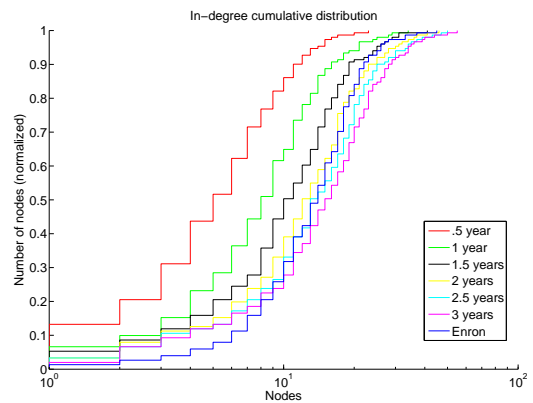

Figure 6: Cumulative distribution function (CDF) of the in-degree of the nodes as a function of time.

5 CONCLUSION

This paper presents a stochastic agent-based approach for modeling email social networks in a closed community such as the ones represented by a company, a sport club, etc. The
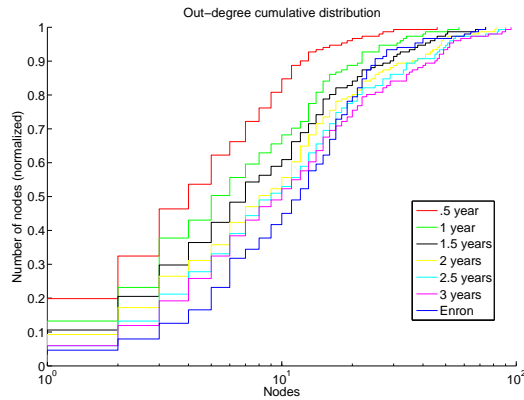
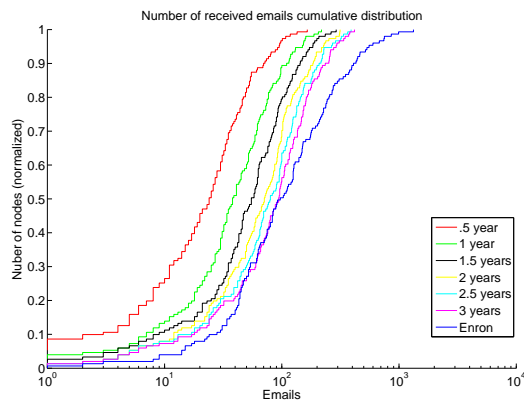Figure 7: Cumulative distribution function (CDF) of the out-degree of the nodes as a function of time.

Figure 8: Cumulative distribution function (CDF) of the number of emails received as a function of time.

model is based on the notions of common neighbors and preferential attachment, but both these concepts have been further enhanced in order to model the connection between the nodes at a deeper level. The scheduler of the model is based on the same working principle as the Gillespie algorithm, which lets the actions self-schedule themselves according to a probability distribution empirically extracted from the dataset (e.g., Enron email). Simulation results in terms of different statistical properties (degree distribution, number of email sent / received, cluster coefficient, reciprocity, etc.) have shown that the model convincingly approximates the behavior of real world email networks. Also, in contrast to previous models, in which the network is created in a top-down fashion, the graph generated by this model is the result of the interaction of many single agents, thus, the social network is created using a bottom-up approach.

This kind of model can have many different applications. For example, varying the model parameters, it can be used to understand the important factors that affect the evolution of email networks and social networks in general. After the model is validated against a specific social network, it can be run to generate different graphs that can be used as test-bed for graph algorithms that try to identify sub-structure or important nodes in the network. In this way the algorithms can be tested on reasonably smaller graphs that resemble the structure of real world social networks, which are typically difficult to obtain or particularly large to be used as test-bed.

Finally we are planning to expand and completely automate the problem of fitting/tuning the model to real network data. This problem can be modeled as a multi-objective optimization problem. The different statistical information curves (degree distribution, number of email sent/received, etc.) of the network generated by the model need to approximate the real curves of the candidate social networks. Each of this curve fitting problems can be seen as one of the objectives to minimize while the model parameters represent the variable space to explore. Many interesting computational problems arise in order to make such a task feasible and applicable in real world scenarios.

ACKNOWLEDGMENTS

REFERENCES

Albert, R., and A.-L. Barabasi. 2002. Statistical mechanics of complex networks. *Reviews of Modern Physics* 74:47.

Barabasi, A. L., and R. Albert. 1999, October. Emergence of scaling in random networks. *Science* 286 (5439): 509–512.

Capocci, A. e. a. 2006, Feb. Preferential attachment in the growth of social networks: the case of wikipedia.

Diesner, J., T. L. Frantz, and K. M. Carley. 2005. Communication networks from the enron email corpus "it's always about the people. enron is no different". *Comput. Math. Organ. Theory* 11 (3): 201–228.

Eisenberg, E., and E. Y. Levanon. 2003, Sep. Preferential attachment in the protein network evolution. *Phys. Rev. Lett.* 91 (13): 138701.

Gillespie, D. T. 1977. Exact stochastic simulation of coupled chemical reactions. *The Journal of Physical Chemistry* 81 (25): 2340–2361.

Jin, E. M., M. Girvan, and M. E. J. Newman. 2001, Sep. Structure of growing social networks. *Phys. Rev. E* 64 (4): 046132.

Liben-Nowell, D., and J. Kleinberg. 2007, May. The link-prediction problem for social networks. *Journal of the*

American Society for Information Science and Technology 58 (7): 1019–1031.

Mysore, V. e. a. 2006. Agent modeling of a sarin attack in manhattan. *In Proc. of the 1st Intl. Workshop on Agent Tech. for Disaster Management (ATDM)*.

Narzisi, G., V. Mysore, J. Byeon, and B. Mishra. 2006. Complexities, catastrophes and cities: Emergency dynamics in varying scenarios and urban topologies. *in Proc. of the International Conference on Complex Systems (ICCS)*.

Newman, M. E. J. 2003. The structure and function of complex networks. *SIAM Review* 45:167.

Newman, M. E. J., S. Forrest, and J. Balthrop. 2002, Sep. Email networks and the spread of computer viruses. *Phys. Rev. E* 66 (3): 035101.

North M.J., C. N., and V. J.R.. 2006. Experiences creating three implementations of the repast agent modeling toolkit. *ACM Trans. on Mod. and Comp. Simulation* 16 (1): 1–25.

Roth, C. 2005. Measuring generalized preferential attachment in dynamic social networks.

Rowe, R., G. Creamer, S. Hershkop, and S. J. Stolfo. 2007. Automated social hierarchy detection through email network analysis. In *WebKDD/SNA-KDD '07: Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, 109–117. New York, NY, USA: ACM.

Shetty, J., and J. Adibi. 2004. The enron email dataset database schema and brief statistical report.

Watts, D. J. 2003, November. *Small worlds : The dynamics of networks between order and randomness (princeton studies in complexity)*. Princeton University Press.

AUTHOR BIOGRAPHIES

FABIAN MENGES is a visiting student at the Computer Science department of Courant Institute of Mathematical Science. His alma mater is the University of Technology in Dresden, Germany, where he is expected to graduate (diploma) in 2009 in Electrical Engineering with a major in Automation and Control. In 2007 he joined the Bioinformatics Lab, led by Professor Bud Mishra, and worked as Junior Research Scientist for the Large Scale Emergency Readiness (LaSER) project of the NYU Center for Catastrophe Preparedness and Response (CCPR). His research interests include Automation and Control, Modeling and Simulation, Multi-Agent Systems (MAS) and Complex Systems. His email address is <fabian.menges@googlemail.com>.

BUD MISHRA is a professor of Computer Science and Mathematics at NYU's Courant Institute of Mathematical Sciences, professor of Human Genetics at Mount Sinai School of Medicine, and a professor of Cell Biology at NYU School of Medicine. He founded the NYU/Courant Bioinformatics Group, a multi-disciplinary group working on research at the interface of computer science, applied mathematics and biology. Prof. Mishra has a degree in Physics from Utkal University, in Electronics and Communication Engineering from Indian Institute of Technology Kharagpur, and MS and PhD degrees in Computer Science from Carnegie-Mellon University. He also holds adjunct professorship at Tata Institute of Fundamental Research in Mumbai, India. From 2001-04, he was a professor at the Watson School of Biological Sciences, Cold Spring Harbor Lab. He is a fellow of the ACM and a NYSTAR Distinguished Professor. His web site is <www.cs.nyu.edu/mishra>, and his email address is <mishra@nyu.edu>.

GIUSEPPE NARZISI is a PhD candidate both at the Courant Institute of Mathematical Sciences and at the University of Catania. He received the Laurea degree in computer science from the University of Catania, Italy, in 2004. From October 2005 to August 2007 he joined the Bioinformatics Lab led by Professor Bud Mishra and worked as a Junior Research Scientist for the Large Scale Emergency Readiness (LaSER) project of the NYU Center for Catastrophe Preparedness and Response (CCPR). His research interests include Multi-Objective Optimization, Evolutionary Computation, Multi-Agent Systems, Complex Systems, Numerical and Combinatorial Optimization, Protein Structure Prediction, Planning Algorithms for Large scale Emergency Response. Giuseppe Narzisi is a reviewer of several international journals and conferences on Evolutionary Computation field. His web site is <www.cims.nyu.edu/~gn387> and his email address is <narzisi@nyu.edu>.