Modeling the Information Propagation in an Email Communication Network Using an Agent-Based Approach

Bin Jiang
College of Information Science and Engineering,
Hunan University,
Changsha, China.
jiangbin@hnu.edu.cn

Lei Wang
College of Information Science and Engineering,
Hunan University,
Changsha, China.
stone4827321@163.com

Chao Yang
Business School,
Hunan University,
Changsha, China.
yangchaoedu@hnu.edu.cn

Shuming Peng
College of Information Science and Engineering,
Hunan University,
Changsha, China.
pengshuming@hnu.edu.cn

Renfa Li
College of Information Science and Engineering,
Hunan University,
Changsha, China.
lirenfa@vip.sina.com

ABSTRACT
Development of Internet technology has made the use of email to be one of the predominant means of communication in the information society. Information exchange among people via email service has produced lots of communication data, which have been widely used in research about information propagation on virtual social networks. The focus of this paper is on the "Enron Email Dataset". The ideas discussed gave thorough consideration to the diversity of organizational positions' attributes, the dynamic behaviors of users to select information contents and communication partners via email service. We then established a quantitative analysis on the multiple interactive relationships of the email communication network. Further, an agent-based model for modeling the information diffusion in an organization via email communication network was proposed, by relating the microscopic individual behaviors and the macroscopic system evolution. Based on the simulation experiments, we analyzed and compared the topological characteristics and evaluative patterns of our model with the Enron Email Dataset. The experimental results proved that our model was beneficial to uncover the implicit communication mechanisms of a real organization.

Categories and Subject Descriptors
J.4 [**Computer Applications**]: Social and Behavioral Sciences

General Terms
Human Factors, Measurement, Experimentation

∗This work was supported by grant 13JJ3049 of the Natural Science Foundation of Hunan Province, China and grant 2012FJ4131 of the Science and Technology Research Plan of Hunan Province, China.

GECCO '14, July 12 - 16 2014, Vancouver, BC, Canada

ACM 978-1-4503-2881-4/14/07…$15.00.
http://dx.doi.org/10.1145/2598394.2610013

Keywords
Email analysis, Information Propagation,
Agent-based Model, Communication Mechanisms

1. INTRODUCTION
Due to its robustness, economic efficiency, speed and fairly unrestricted access, social network platform generated by Internet service has lowered the cost of the interpersonal communication. As a result, it has made it easier for humans to realize their connections and interactions with each other. Email communication was chosen as the most useful and preferred tool for computer-mediated communication for being particularly expressive in a greater detail [1]. Enterprise organizations widely use email platform for various communication purposes such as scheduling tasks, issuing notices, submitting reports, data transfer, and other works, with significantly improved work efficiency. The underlying phenomena of information flows are mostly the consequence of complex organizational networks of interactions (sending or receiving information) among numerous users.

There have existed obvious hierarchical differences among users, such as Top-level (C.E.O., President etc.), Middle-level (Director, Manager etc.) and Bottom-level (general Employees).This lead to the variance in communication mechanisms among the users, as they belong to either equal or unequal levels. Daily interactions among users of the equal levels cause frequent communication and form similar communication groups that are composed of users from the same level positions. Generally, such users in each communication group are socially familiar with each other. Besides, email users of each group usually prefer to propagate information to authority users in other groups. Therefore, familiarity decides the user's influence of communication within the group, and authority determines the influence outside of the group. Meanwhile, due to work needs or the need to save time, quite a number of emails are transferred through P2MP (peer-to-multipeer), instead of P2P (peer-to-peer). At the same time, users can also forward the received emails to other users, who also forward them in turn. Consequently, these interactions of emails form an information chain, in the directional graph of propagation. The graph has just only one node with its in-degree equal to zero. These can be observed in the Figure 1.

As the above mentioned, communication networks in specified groups (such as a company) are special to some extent. Their

interactions in the real world make them to know each other, but in different levels of closeness. Besides, organizational position and some additional properties also cause diversity in the content, frequency and mode of communication. All of these reasons increase the difficulty of modeling and simulating the propagation of information in the communication network. If we take the involved staff /user as an entity, the group relationship as a link and emails interaction among users as a series of rules based on the relationships, then the information propagation is the dynamic evolution process of interpersonal relationships via social interactions, which can be observed by linking the edges between entities. In the paper, we proposed an agent-based model which considered the mentioned specialty for modeling and simulating the information propagation in an organizational communication network, through which to compare and analyze the interaction mechanisms among users, information and network, and their influence on system evolution.

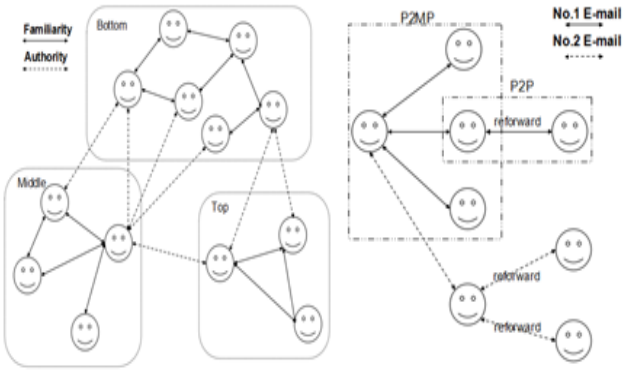


Figure 1. Email communication network among organizational users

Our research consists of four steps: first, extract facts from actual communication data, in order to define entities of individual users and email information and their relevant state variables; next, establish the virtual communication mechanisms on agent-based simulation (ABS) and carry out the experiments; third, analyze and compare the results of the model with the dataset using social network analysis techniques; and finally conduct the experiments on the agent model with different parameters set by changing features of the interaction, through which to observe and compare the different impacts of users interaction mechanisms on information propagation.

The data set we adopted was the Enron Email Dataset, which included a large set of email messages and was made public by a legal investigation concerning the Enron organization. This data set was available from the site http://www.cs.cmu.edu/~enron/, which contained email data from January, 2000 to June, 2002 [2]. We utilized the emails of 151 people with their sender/receiver links. We also got the files which described the organizational positions of these 151 staffs in the Enron organization. This dataset was found to be very detailed and complete, hence, was a perfect test bed for studying the communication mechanism and diffusion result with emails.

The rest of this paper is organized as follows: Section 2 briefly reviews the related research works on email; Section 3 describes our agent-based model through the ODD protocol; Section 4 gives the simulation experiments and then discusses the results; and finally, we conclude our work and present the future works in Section 5.

2. RELATED WORKS

There have been lots of research works on the email communication network. In this paper, we roughly divided these contributions into two categories:

The first category mainly employed the social network analysis techniques to analyze the static topological structure. For example, Ebel et al. studied the topology of email networks from server log files and found that the resulting network exhibited a scale-free link distribution [3] and pronounced small-world behavior [4]. They further concluded that the spreading of email viruses was greatly facilitated in scale-free network structure compared with random architectures [5]. Uddin et al. used the social network analysis measures of degree centrality, betweenness centrality, closeness centrality and reciprocity for exploring a longitudinal email communication network among students [1][6]. Diesner et al. extracted communication networks from Enron corpus by refining the relations, and then applied various quantitative indicators to explore structural properties of the networks in Enron and to identify key players across time [7]. Karagiannis et al. studied the behavioral patterns of email usage in a large-scale enterprise by focusing on pair-wise interactions; they had examined various factors that could potentially affect email replies [8]. There were also some other techniques used to analyze the email network, such as Genetic Algorithm (GA) [9] and Visual Analytics [10], in order to explore interesting patterns of email interactions and key actors.

The other category aimed to explore how the network evolved dynamically over times. To deal with this issue, the Agent-Based Modeling (ABM) [11] had been employed with great acceptance in the simulation of the social network. For example, Matsuyama et al. analyzed the implicit mechanisms of peer-to-peer communications among people of a large organization by using an agent-based simulation model based on the Enron email dataset. Additionally, they validated the influence of the changes of the members in the group on an artificial society [12][13]. Wang et al. utilized a simple stochastic branching model to capture the structural properties of email spreading trees, i.e., to how many people a user forwarded the email and the total coverage the email reached. The result indicated that the spreading process followed a random yet reproducible pattern, largely independent of context [14]. Menges et al. used an agent-based approach to model growth of email-based social networks, in which individuals established, maintained and allowed atrophy of links through contact-lists and emails. Their approach enhanced both common neighbors and preferential attachment in order to model the connection between the nodes at a deeper level [15]. Wu et al. introduced a model with decay in the transmission probability of information as a function of the distance between the source and the target. They found the decay of similarity among members had strong implications for the information propagation, so that the number of individuals that a given email message reached was very small [16].

Following their works, in this paper, we study how information propagates in an organization through multiple interaction relationships on email communication network through agent-based approach, the model description is given below.

3. MODEL DESCRIPTION WITH ODD PROTOCOL

The model description follows the ODD (Overview, Design concepts, Details) protocol for describing agent-based models [17].

3.1 Purpose

In this paper, we focus on the problem of understanding and modeling a social network media, information propagation system based on the complex interaction mechanisms among the users, the information and the network via email service. Users are classified according to the positions and emails are classified according to their participants in the propagation process, and multiple communication groups are formed. We build the quantitative indicators to describe the influence of varied relations of users, and propose an agent-based model for modeling to diffuse information in the communication network through. We then analyze and compare the resultant patterns of the model with the Enron dataset using social network analysis techniques. Further, we study how those impact on the evolution by changing feature parameters of the interaction. This model is put forward in trying to understand the social network information propagation, diffusion patterns and guide the information dissemination activities effectively.

3.2 Entities, State Variables, and Scales

As shown in Table 1, our model includes two types of entities: user and email. User is regarded as the individuals who disseminate information (sending or receiving emails), and email is regarded as the carrier of information. Table 2 defined and described the relevant state variables of user and email.

Table 1. Entity definition and description

Entity	Description	Definition
User	The individuals who disseminate the email	Agent
Email	The carriers of the information	Information

Table 2. Definition and description of state variables

State variables	Description	Variable name in the model
The agent position	The position of the user	*aPosition*
The partners position	The positions set of the user's communication partners	*cpList*
The agent state	The distance between the user and the Internet	*aState*
The communicated information set	The information set of the user has been disseminated	*infoList*
The position set of agents	The positions set of users who could disseminate the information	*cPostions*
The disseminated number	The disseminated number of the information	*cNum*

3.3 Process Overview and Scheduling

During the simulation process, after the initialization of system and parameters, each agent would carry out the following actions according to the rules in turns at each simulation step. For the current executor, namely agent A_i, it would (1) change the state, if the state value of A_i is zero, follow-up process will be performed; (2) select the mode of communication and the information; (3) set the number of receivers as n, and then select group and the n receivers from group. Finally, after executing the information propagation, the model will update the links of communication network and the related parameters. Figure 2 is the pseudo code description of the simulation process, the simulation of information propagation will be terminated when the simulation step comes to a predefined value.

```
Start
    for i = 1 to 151
        Initialize the Agents in network
    end for
    for i = 1 to 115000
        Initialize the Information in network
    end for
    for t = 1 to 10000
        for i = 1 to 151
            if (the aState of Agent Ai == 0)
                if (Ai sends new information)
                    Ai selects a new Information Ik
                end if
                if (Ai sends old information)
                    Ai selects an old Information Ik from its infoList
                end if
                The number of receivers n is set by a probability distribution function
                for j = 1 to n
                    Agent Ai selects Agent Aj for receiver through Familiarity or Authority
                    Update the links of communication network and the parameters of Ai, Aj, Ik
                end for
            end if
        end for
    end for
End
```

Figure 2. Pseudo code description of information propagation process

3.4 Design Concepts

Basic principles. The main principle of this model came from the theory of "Philos Relationships", proposed by Krackhardt [18]. He had developed the concept of influence strength in [19], and then defined the relationship as one that meets three conditions: Interaction (frequency of interaction), Affection (one feels affection for another) and Time (history of interaction). Further, the different types of influence played different roles in the diffusion process. But they also actively contributed to make a theoretical prediction. In fact, interaction created the opportunity for the exchange of information. According to Krackhardt's theory, we study the heterogeneous relationships of users from multi-dimensional influences, and then take the Interaction as Familiarity and the Affection as Authority in the history of interactions, in order to determine the probability of selection impacts on users' interactions, and build our information propagation model. Meanwhile, we construct the model based on the systemic generative mechanism within the POM (Pattern-Oriented Modeling) framework [20], which uses multiple patterns to guide the model design, test and evaluation in agent-based complex systems.

Adaptation. Each agent is the subject who conducts the behaviors of information propagation. Its adaptation is represented as dynamic behaviors for selecting receivers according to their influence strength in different communication groups.

Emergence. In the model, agents select information and receivers based on a series of rules, their dynamic behaviors and interactions would lead to different evolution of the information propagation network, and emerge different distribution patterns.

Stochasticity. The selection of the mode to propagate information (sending new information or forwarding old information), the information type, and the number of senders are all random processes. More details about stochasticity are described in the Section 3.6.

Collectives. The agent behavior of participating in the information propagation would form agent groups and communication

network structures, implying a direct or indirect relationship or influence among agents, which will affect the individual selection behaviors as a feedback.

Observation. The experimental data collected from the agent-based simulation model includes the information items, the agents and the dissemination path among the participant agents during the entire simulation process. Therefore, a communication topology network and some distribution patterns are observed, such as the quantitative distribution of users to send or receive e-mail, the size distribution of e-mail chain, the frequency distribution of communication between users in different positions and so on.

3.5 Initialization

In the initial stage, we assumed three types of agent positions: *Top*, *Middle* and *Bottom*. According to the records of Enron email communication, agent at each position may send information to one or several other agents at all positions, thus we define eight types of agents with their distinguished abilities to send information to others, including the type of agent who never sends information (the null). Here, the agent information spread ability refers to the positions collection of his potential receivers in communication event. As outlined in Table 3, we could finally get twenty-four types of agents, represented by integers from *1* to *24*.

Table 3. Agent types and distributions in Enron email dataset

N \ R / S	Top	Mid	Bot	Top Mid	Top Bot	Mid Bot	Top Mid Bot	Null
Top	0	1	0	0	0	5	10	0
Mid	0	0	0	0	1	13	39	2
Bot	0	0	8	2	0	30	37	3

In Table 3, *S* and *R* represent a pair of sender and receiver with their specific positions; and the numbers in Table 3 correspond to the actual distribution of such pairs in Enron email dataset. For example, the number *2* in the third row, fourth column denotes that the sender's position is Bottom, it can send information to the agents whose positions are Top or Middle, and this type of agents has *2* in Enron email dataset. Furthermore, we define seven types of information based on the composition of agents who participate in propagating this information; and the actual distribution of email information is given in Table 4. Where information type *T* means that only those agents with *Top* positions can disseminate this information (sending or receiving), but information type *TM* could be spread among agents with *Top* and *Middle* positions (both *S* and *R* could come from *Top* or *Mid* positions), and the rest cases can be deduced similarly.

Table 4. Information types and number distribution

Type	T	M	B	TM	TB	MB	TMB
Number	1500	15000	30000	7000	6000	50000	5000

3.6 Sub Models

In this model, the subprocesses of information propagation mainly include: (1) Change the state of an agent; (2) Select communication mode and information; and (3) Set the number of receivers and select them from agent groups.

3.6.1 Change the State of agent

In the model, we simulate the change of agent's state by extending the "Random-Walk" model proposed by Michaela [21]. We employ this model because it can explain well how human groups make their decisions in uncertain environments. Meanwhile, some users' activities for sending emails are bursty while other users only send email occasionally. Therefore, we use the random walk model instead of other simple probability models to simulate this feature. Further, we only consider a one-dimensional random walk in the model so as to avoid the situation that fewer and fewer users send emails when time progresses. For more details, in the "Random-Walk" model, at each simulation step, the agent has a state represented by an integer, namely the distance between the agent and the Internet. The agent cannot access the Internet to propagate information via social media service unless his current state value is zero. As the users may use email for a period of time, that is, the agent may maintain his current state for a few simulation steps. Therefore, we modified the original model to make the change of agent state not so fast, and the operations are presented as in the formula (1).

$$aState_t = \begin{cases} aState_{t-1} + 1 & , \quad if\ 0 \le r < 1/3 \\ aState_t & , \quad if\ 1/3 \le r < 2/3 \\ aState_{t-1} - 1 & , \quad otherwise \end{cases} \tag{1}$$

Where a$State_{t-1}$ and $aState_t$ represent the user's state at simulation step *t-1* and *t,* respectively; *r* is a random number in (0, 1]. Below, Figure 3 gives an example of the "Random-walk" curves in 20 steps: the agent's state value was equal to zero at the step 1, 9, 10, 12, 16, 18 and 20, hence, it could propagate information in these steps.

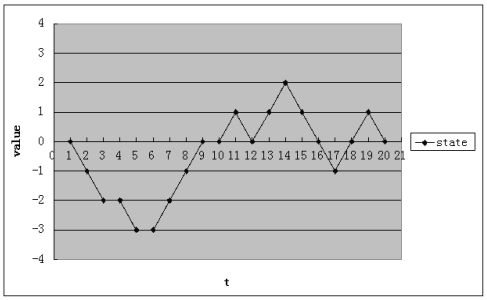


Figure 3. Random walk over the agent state

3.6.2 Select the Communication Mode and Information

When an agent A_i executes information propagation, it can select either to send new information or forward old information which was disseminated in the previous simulation step. According to the actual proportion that users choose these two communication modes to propagate information in Enron email data, we set the probability of sending new information is 0.7 and forwarding old information is 0.3 in the model. In the case that agent A_i intends to send a piece of new information, the model will generate a new information with its type satisfying the agent positions and information types described in Table 3 and 4 (For example, if the agent's position is *Top* and his potential receivers' positions are *Top* and *Middle*, then the information type could be selected from *T*, *TM* and *TMB* type), and saves it into the agent's information list (represented by the values in *infoList* in Table 1) automatically. The probabilities of the type of new information are proportional to their quantitative distribution in Enron data. Otherwise, agent A_i would forward a piece of old information. The probability of selecting a piece of old information is given in formula (2).

$$Pr[I_k] = \frac{1/cNum_{k,t}}{\sum_m^{|infoList_t|} 1/cNum_{m,t}} \tag{2}$$

Where I_k is the selected information, $infoLis_{i,t}$ is the information set sent or received by A_i before simulation step t, $cNum_{k,t}$ is the disseminated number of I_k at simulation step t.

3.6.3 Select the Communication Group and Receivers

The key objective of information propagation under a network environment is how to generate the sending and receiving behaviors among a number of users, i.e. how to select agents to communicate through the social influence. We assume if the position levels of two agents are equal, then they belong to the same communication group, or else belong to different groups. First, the agent A_i need to select the number of receivers by formula (3), in other words, send this information n ($0<n\leq 50$) times.

$$Pr[n] = \frac{1/n}{\sum_{k=1}^{50} 1/k}, 0 < n \leq 50 \tag{3}$$

Next, we define three communication groups as *Top*, *Middle* and *Bottom* groups, according to the agent's positions. Then, we explain how to select information receivers in two steps. The first step is to select the recipient's group, which could be divided into two situations: 1) if the sender's information spread ability determines that he can only select receivers from one group; then the recipient's group is this selected group. 2) Or else if the information spread ability of the sender allows him to select the receivers from different groups, then we need to decide the recipients' group based on the historic selective information list. For example, assume a sender A_i at *Top* position and its potential receivers could be those agents at *Top* and *Middle* position. If the chosen information type was T, then the receiver's group must be *Top*. Else if the chosen information type was *TM* or *TMB*, then the receiver's group could be chosen from either *Top* or *Middle* with both probabilities equal to 0.5. Other cases could be deduced similarly. After determining the receiver's group, the agent would select the recipients from this selected group. The model considers two kinds of interaction modes for selecting receivers: Familiarity mode for those participants who have equal-level positions in the same communication group; and Authority mode for those who come from different groups. These two kinds of interaction modes are quantitative calculated below:

$$\text{Familiarity}: \ Pr[A_j] \frac{FAT_t(A_i, A_j, T) + \lambda}{\sum_k^{|A|} (FAT_t(A_i, A_k, T) + \lambda)} \tag{4}$$

$$\text{Authority}: \ Pr[A_j] = \frac{AUT_t(A_i, A_j, T) + \lambda}{\sum_k^{|A|} (AUT_t(A_i, A_k, T) + \lambda)} \tag{5}$$

Where A is the universal set of all agents, A_j is the information receiver to be selected (A_j didn't disseminate the chosen information before) and T is the type of information selected by A_i in the previous steps, $FAT_t(A_i, A_j, T)$ is the total communication volume between agent A_i and A_j with information type T at simulation step t, $AUT_t(A_i, A_j, T)$ is the total volume of information that agent A_j has received from other agents in A_i's group with information type T at simulation step t; λ is the interference parameter and its role is in twofold: (1) to ensure the denominator of formulas not be zero in the simulation process, so as to prevent floating point arithmetic overflow; (2) to adjust the impact strength of each kinds of interactions by changing the value of λ, the bigger the value of λ, the smaller the strength. When λ tends to infinity, the probability of each agent to be selected tends to be the same. Namely, the agent A_i selects the sender randomly.

4. EXPERIMENTAL RESULT DISCUSSION

In this paper, we assumed that the email delivery activities of an organization had their explicit purposes and strong relevance. Therefore, the parameter λ of the model was first set to a very small value 0.001 to eliminate the influence of random strengths on the result. Further, we examined the effects of different values of parameter λ on experimental results. The simulation experiments on information propagation model generated a communication network, where the nodes represented agents and the directed edges recorded the information spread path from one agent to another. We compared and analyzed the topologic characteristics and emergent patterns of the communication network based on the simulation result. For a robust result, each simulation was executed 50 times and the average value was achieved as the final result. Further, we extracted the emergent patterns of the communication network from multiple scales, including the degree distribution of agents to send or receive information (in-degree and out-degree), the size distribution of information chain, and the frequency distribution of communication between agents with different positions.

4.1 Comparison of structural characteristics of network

First, we compared the structural characteristics of the communication network generated by Enron e-mail data and our simulation model. We also simulated the information propagation with random network and small world network model. The probabilities of rewiring in random network model and small-world network model are set to 1 and 0.2, respectively. It should be noted that, the network generated from our model has multiple edges. Therefore, we set up the random network and small-world network by starting with a ring of 151 nodes, each connected to its four nearest neighbors by undirected duplicate edges, in order to compare with each other better. However, that is not completely in conformity with these two common models, with duplicate edges forbidden.

Table 5 listed the results of three indexes which described the structural characteristics under different networks: the average degree of nodes, the average shortest path length and the average clustering coefficient.

Table 5. Structural characteristics of different network

	Number of nodes	Average degree of nodes (SD)	Shortest path length (SD)	Clustering coefficient (SD)
Enron e-mail Network	151	544	2.08	0.51
Our model Network	151	560 (38)	2.28 (0.22)	0.60 (0.05)
Random Network	151	295 (17)	2.0 (0.03)	0.13 (0.01)
WS model Network	151	570 (17)	2.44 (0.01)	0.48 (0.02)

Based on the result in Table 5, we observed that the communication network formed by the Enron e-mail data had a "small-world" property——short path lengths and high clustering coefficient. Meanwhile, the network generated by our model, with the same number of nodes, had its average degree of nodes, average shortest path length and average clustering coefficient approximate to the real Enron network. It suggested that our model could generate a social communication network of small world as in the real world.

4.2 Comparison of degree distribution of agents

Next, we compared the degree distribution of nodes in the Enron email network and the resultant network of our model.

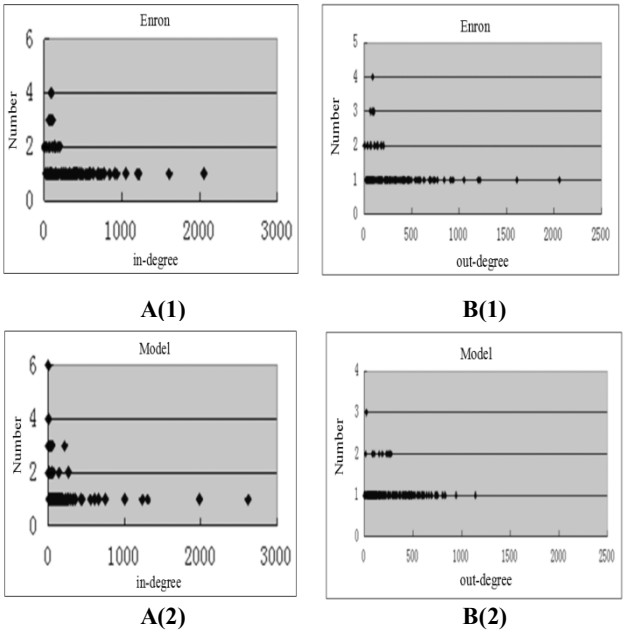


Figure 4. Degree distribution of Enron E-mail and our model

As described in Figure 4, the results proved that our agent model successfully reproduced the degree distribution of agents approximate to the result in actual Enron e-mail data. As shown in A (1) and A (2), there were some nodes with large in-degrees. By comparing with the real communication records in Enron email data, we found that these nodes corresponded to a small fraction of users who not only communicated frequently with other agents in the same group, but also often received information from agents in other different groups, thus formed large in-degrees. On the other hand, according to the authority definition described in Section 3.6.3, because we set the greater the in-degree, the easier to be selected as a recipient by users at other groups, therefore those agents who received more information from agents with different positions at the early stage would form large "in-degree" fast, but the number of these agents was small.

Further, there were a few agents with large out-degree in the Enron e-mail network (see B (1)), which meant they had sent much more information than other agents. In fact, all these agents executed several large-scale information spreading behaviors in Enron email communication. However, in our model's network (see B (2)), the largest value of out-degree was less than 1500. It was because that we set the number of receivers in once spread limited to 50, and it was almost impossible for the same agent to send information to a mass number of receivers every time, thus would not generate agents with very large out-degree.

4.3 The size distribution of the information chain

As the size distribution of the information chain also reflected the users' behaviors and preferences in their information propagation process, we therefore compared the size distribution of the information chain between the Enron email data and the simulated result of our model in Figure.5.

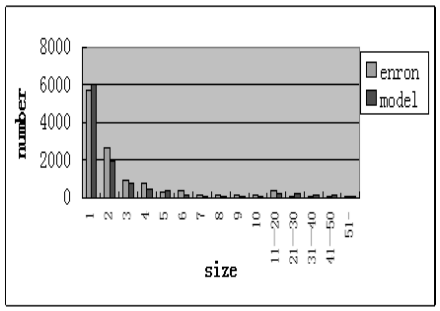


Figure 5. Size distribution of the information chain

As shown in Figure 5, the size of 1 denoted that the information had not been forwarded, but only be disseminated from one agent to another, while other information ever had been either forwarded once, or disseminated at least once with P2MP mode or several times with P2P mode. On one hand, Figure 5 showed the power laws in the distribution of the size of information chain of both Enron email ($y=5.5e3x^{-1.41}$, R^2:0.975) and our model($y=5.2e3x^{-1.46}$, R^2:0.987). The power laws of the size distributions in our model can be explained by the process of selecting the old information and the number of receivers. Different from a rich-get-richer phenomenon, the agent in our model preferred to select the old information with smaller disseminated number consistent with the real data records, and so was the receiver number. Such mechanism could also give rise to power law distribution. On the other hand, we found most information chain had small sizes. The average size of Enron email chain was 3.5, and 4.0 in the simulation model. This suggested that email was typical "Narrowcasting Media".

4.4 The Communication frequency distribution with different positions

Next, we analyzed the impact of agent's attribute of positions on information propagation in an organization. Figure 6 compared the results of communication frequency distribution of agents with different positions between the Enron email data and the simulated data of our model.

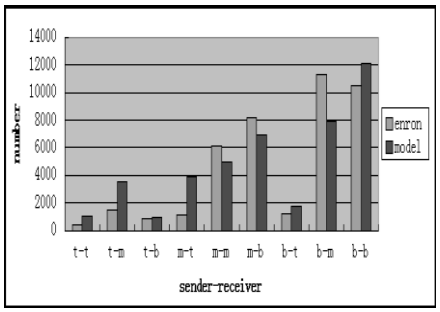


Figure 6. Frequency distribution of communication between different positions

In Figure 6, the *x* axis represented the pair of sender-receiver positions for the information propagation event, and the *y* axis denoted the specific numbers. For example, *t-t* represented the information spreading among users with *Top* positions (it corresponded to the *Top–Top* type information), and the rest can be inferred by analogy. Figure 6 showed that the *b-b, b-m, m-b* and *m-m* type of information occupied a major proportion in the real communication of Enron email data, and our model also reproduced the similar distribution. This was because Enron email data had more users with *Bottom* and *Middle* positions than those

with *Top* positions, and users with *Bottom* and *Middle* positions participated in the communication more frequently than users with other positions in Enron organization. Specifically, most of the *b-m* information was disseminated in P2MP mode, which actually expanded the spreading scale of this information. Meanwhile, the *b-b* information held the largest communication quantity of the Enron email data, mostly disseminated by P2P mode. It suggested that communication among Bottom users was very frequent. To conclude, there existed diverse communications among users across the organizational positions, and our model was good at simulating these special features.

4.5 The different influence strength of interaction

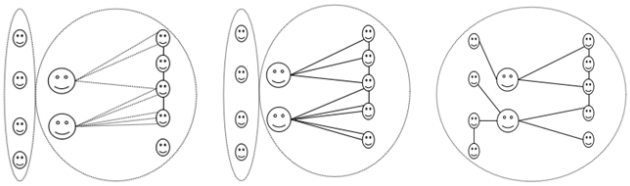
In Section 3.6.3, we mentioned that a change in the value of λ in formula (2) and (3) could adjust the influence strength of each interaction. The greater the value of λ is, the smaller the impact strength is. When the value is infinity, the probability of each agent to be selected as receivers tends to be the same, namely the selection is random. Here we carried out a set of experiments for comparing and analyzing the influence with different intensities impact on the topological characteristic of evolution network by changing the value of λ.

Table 6. Topologic characteristics of the network with different influence strength of interaction

λ	Average degree of nodes (SD)	Shortest path length (SD)	Clustering coefficient (SD)
0.001	560 (38)	2.28 (0.22)	0.60 (0.05)
0.01	576 (50)	2.31 (0.20)	0.55 (0.05)
0.1	580 (45)	1.68 (0.13)	0.45 (0.05)
0.5	587 (31)	1.44 (0.15)	0.70 (0.07)
1	591 (33)	1.35 (0.12)	0.77 (0.06)
10	609 (52)	1.24 (0.09)	0.88 (0.08)
100	618 (47)	1.24 (0.09)	0.88 (0.07)

From the result in Table 6, we found that the average degree of nodes in the communication network increased with the rising λ, but the increased rate was not obvious. While the average shortest path length decreased with the rising λ. It was because the selection of information receivers tended to be more random when the λ became bigger, which improved the probability that agents interacted with each other. Consequently formed the direct path between two agents and thus shortened the average shortest path length of the network.

Further, we observed that when λ was less than 0.1, clustering coefficient decreased with the rising λ, while when λ was greater than 0.5, the clustering coefficient increased with the rising λ. We discussed these two cases below: (1) when λ was less than 0.1, *Familiarity* and *Authority* mode played dominant roles in the selection of communication partners. Under this case, the smaller the λ was, the agents who ever communicated with each other by *Familiarity* mode would also had higher probability to choose a common agent with high Authority for spreading information, thus the average clustering coefficient of the communication network was increased. (2) When λ was greater than 0.5, the random factor was the dominant influence. Under this case, with a rising λ, the agents could select communication partner from a larger range and the success rate of communications between two agents could be improved. Thus, it was relatively easy for two agents (See Figure 7(a)) to generate a communication link and finally form a communication network with a higher clustering coefficient, and there always existed multiple links between the two agents. Interestingly, the clustering feature in this random mode was quite different from the general random network, just as the third network in Table 5. It was supposed that considering more details before random linking edges might lead to a different result. For example, the selection behaviors of information type and communication group in our random mode would limit the alternative range of attachable nodes (See Figure 7(b), only two nodes in the same circle can be linked together and almost no multiple edges), while any two nodes can be connected with a stochastic connection probability in the random graph model (See Figure 7(c)). Accordingly, we could generate a network which is consisted of several circles with relatively higher clustering coefficient under the random mode.



(a) λ<0.1 (b) λ>0.5 (c) random network

Figure 7. The network of stochastic linking two nodes under each mechanism

5. CONCLUSIONS AND FUTURE WORKS

In this paper, we proposed an agent-based model to simulate the information propagation in Enron's email communication network. We first gave the definition of user and information by extracting real data. And then, we built a series of natural subprocesses in order to model the dynamic behaviors of users who selected information and recipients during the information propagation event. By labeling the different communication groups as Top, Middle and Bottom which consisted of a group of users with equal position levels, we established a quantitative analysis of the multiple interactions: Familiarity for interaction within the same group, and Authority for interaction spans different groups. On this basis, we compared the network topological characteristics and diffusion patterns of our model with the Enron dataset using social network analysis techniques. Although the origins of the two networks were quite different, the characteristics were quite similar to each other.

The investigation suggested that the agent-based model was beneficial to uncover the characteristics of implicit communication mechanisms of the organization. It is a new attempt to study the model of information propagation in communication networks and build the measurements of influence quantization from heterogeneous group relations.

In the future work, we plan to validate and improve our information propagation model on other types of social media, such as Weibo, through which to find the internal organization mechanism for enabling and promoting information propagation on social communication networks. Weibo, however, as a more complex information dissemination system, the diversity of individual users leads to the different information subject preferences and ever-changing selection strategies, for example, in view of the multiple dimensions of social relations, users will utilize the combined influences of *Familiarity*, *Authority*, and even *Similarity* for their choices of interacting partners and information. If still using the method proposed in this paper, for example, using predefined parameters and rules in the interaction, the simulation result may not fit the actual data well. Therefore, focus on the next step of work, we'd like to use genetic algorithm to search the feature variable combination of interactive modes,

and then estimate the relative importance between them in the model, so that we can calibrate the weight proportion of each influence in the mixed interactive mechanism and other key correlation parameters for each special user group over different time periods. At the same time, we might try to use other reasonable models, such as Levy Flight model (a special random motion with an occasional larger step jump), to simulate the more complex behavior of users. We hope all of these strategies will contribute to make agent-based model more efficient and natural.

6. REFERENCES

[1] S. Uddin and M. J. Jacobson. Dynamics of email communications among university students throughout a semester. *Computers and Education*, 63:95-103, May 2013.

[2] J. Shetty and J. Adibi. The Enron email dataset database schema and brief statistical report. *Information Sciences Institute Technical Report*. University of Southern California, 2004.

[3] A. L. Barabási, R. Albert, and H. Jeong. Mean-field theory for scale-free random networks. *Physica A: Statistical Mechanics and its Applications*, 272(1):173-187, July 1999.

[4] D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world'networks. *Natur*e, 393(6684):440-442, April 1998.

[5] H. Ebel, L. I. Mielsch, and S. Bornholdt. Scale-free topology of e-mail networks. *Physical Review E*, 66(3): 035103, September 2002.

[6] S. Uddin, K. Thompson, B. Schwendimann, and M. Piraveenan. The impact of study load on the dynamics of longitudinal email communications among students. *Computers and Education*, 72:209-219, March 2014.

[7] J. Diesner and K. M. Carley. Exploration of communication networks from the Enron email corpus. In *Proceedings of the 2005 SIAM Workshop on Link Analysis, Counterterrorism and Security*, pages 3–14, April 2005.

[8] T. Karagiannis and M. Vojnovic. Behavioral profiles for advanced email features. In *Proceedings of the 18th International Conference on World Wide Web*, pages 711-720, April 2009.

[9] G. Wilson and W. Banzhaf. Discovery of email communication networks from the Enron corpus with a genetic algorithm using social network analysis. In *Proceedings of the IEEE Congress on Evolutionary Computation*, pages 3256-3263, May.

[10] M. E. Joorabchi, J. D. Yim, and C. D. Shaw. EmailTime: visual analytics and statistics for temporal email. In *Proceedings of Visualization and Data Analysis 2011*, pages 78680Q, January 2011.

[11] E. Bonabeau. Agent-based modeling: Methods and techniques for simulating human systems. *Proceedings of the National Academy of Sciences of the United States of America*, 99(Suppl 3):7280-7287, March 2002.

[12] S. Matsuyama and T. Terano. An agent simulator meets the enron for communication network analysis. In *Proceedings of 2007 IEEE International Conference on Systems, Man and Cybernetics*, pages 1999-2004, October 2007.

[13] S. Matsuyama and T. Terano. Analyzing the ENRON Communication Network Using Agent -Based Simulation. *Journal of Networks*, 3(7):26-33, July 2008.

[14] D. Wang, Z. Wen, H. H. Tong, C. Y. Lin, C. Song, and A. L. Barabási. Information spreading in context. In *Proceedings of the 20th International Conference on World Wide Web*, pages 735-744, March 2011.

[15] F. Menges, B. Mishra, and G. Narzisi. Modeling and simulation of e-mail social networks: a new stochastic agent-based approach. In *Proceedings of the 40th Conference on Winter Simulation*, pages 2792-2800, December 2008.

[16] F. Wu, B. A. Huberman, L. A. Adamic, and J. R. Tyler. Information flow in social groups. *Physica A: Statistical Mechanics and its Applications*, 337(1):327 - 335, June 2004.

[17] V. Grimm, U. Berger, D. L. DeAngelis, J. G. Polhill, J. Giske, and S. F. Railsback. The ODD protocol: a review and first update. *Ecological Modelling*, 221(23):2760 - 2768, November 2010.

[18] D. Krackhardt. The strength of strong ties: The importance of philos in organizations. *Networks and organizations: Structure, form, and action*, Harvard Business School Press, Boston, MA, 216-239, August 1992.

[19] M. Granovetter. The strength of weak ties. *American journal of sociology*, 78(6):1360 - 1380, March 1973.

[20] V. Grimm, E. Revilla, U. Berger, F Jeltsch, W. M. Mooij, S. F. Railsback, and D. L. DeAngelis. Pattern-oriented modeling of agent-based complex systems: lessons from ecology. *Science*, 310(5750):987-991, November 2005.

[21] M. Goetz, J. Leskovec, M. McGlohon, and C. Faloutsos. Modeling Blog Dynamics. In *Proceedings of Third International AAAI Conference on Weblogs and Social Media*, pages 26–33, July 2009.