A Novel E-Mail Network Evolution Model based on user Information

Lejun ZHANG
College of Information Engineering
Yangzhou University, Yangzhou, 225127
Yangzhou, China

Chunhui ZHAO
College of Information Engineering
Yangzhou University, Yangzhou, 225127
Yangzhou, China

Tongxin ZHOU
College of Computer Science and Technology
Harbin Engineering Universit
Harbin, China

Zilong JIN
School of Computer and Software
Nanjing University of Information Science and Technology
Nanjing, China

***Abstract*—E-mail is one of the main means of communication in society today, and it is a typical social network. Studying the evolution of the social network structure by constructing an e-mail network evolution model is of great significance to the literature. In this paper, we first analyze the e-mail network by constructing an e-mail network communication model; this mainly includes analysis of the structure of the e-mail network and analysis of the user information in the e-mail network; then, we propose an e-mail network evolution model based on the characteristics of user information and give the specific evolutionary steps; finally, the simulation experiments are carried out to analyze the characteristics of the model. Experiments show that the nodes are characterized by a power-law distribution, and compared with other models; the model is closer to the real network, so it has important practical significance.**

Keywords—Information characteristics; e-mail; network evolution; complex network

I. INTRODUCTION

With the rapid development of information technology, people's lives have been fully integrated into a complex network world, tangible, intangible, various, ubiquitous. Networks are like large systems; each node in the network is a different element in the system, and the relationship between different elements forms the edge between nodes. So scientists want to find a certain law to construct the network topology and thus to better understand the network, finally determining the value of the network. Application of a complex network analysis method can better reveal the characteristics of the network; it has important significance for network formation and expansion, information dissemination and other research. With the development of computer technology and the Internet, many scholars at home and abroad have studied the network model in many ways. In different fields, the particularity of the structure of the network model is also different.

With the rapid development of networks and computers, people have studies networks in a more profound and comprehensive way. They have found that regular networks are not applicable to the universality of real networks. Then, the first to initiate change were Watts and Strogatz [1], who proposed a WS network named after them. In the process of network generation, the edge will be randomized to reconnection, when the network reaches a certain scale; the average path of the network is small, and the clustering coefficient is high; the network has the characteristics of a "small world"; subsequently, Newman and Watts [2] found that the defects of the WS network, random network rewiring, may lead to more independent nodes appearing; in response to this phenomenon, they used adding edges instead of randomized reconnection. Later, Barabasi and Albert [3], [4] found that nodes in a network have the power law characteristic through a large number of actual networks; then they put forward a scale-free network, which is known as the BA network; then, more and more models were proposed; Flammini [5] propose a criterion of network growth that explicitly relies on the ranking of nodes according to any prestige measure; Sun et al. propose a model driven by events and interests [6]; Alireza et al. validate the significance of betweenness centrality in the evolution of research collaboration networks [7]; Barrat et al. study the complex weighted network [8], [9]; Sun et al. propose a topological evolution to simulate social activities [10]; Song et al. build up a class of edge-growing network models and provide an algorithm for finding spanning trees of edge-growing network models [11]; Huang et al. use spanning trees and other graphs to illustrate some results and phenomenon and try expressing mathematically key notions from researching scale-free networks [12]; a likelihood analysis is provided about evaluation models by Wang et al [13]; Zou et al. propose an evolving network model growing fast in units of module [14]; article [15] finds that with the increment of network interdependence, the evolution of cooperation is dramatically promoted on the network playing Prisoner's Dilemma, and the cooperation level of the network playing Snowdrift Game reduces correspondingly; Zhuang et al. study the problem of maximizing influence diffusion in a dynamic social network [16]; Llhan et al. present a framework for modeling community evolution prediction in social networks [17].

This paper proposes an e-mail communication network named "User-Information-keyword" through network analysis of an Enron dataset and analyzes the structure and user information of an e-mail network; finally, an e-mail network evolution model based on user information is proposed. Through the above methods, the research on e-mail networks for a certain range of people has a certain significance and value in reflecting the communication between people in real society.

II. E-Mail Network Analysis: A Case Study of an Enron Dataset

A. *Introduction to the Enron Dataset*

In this paper, we use the Enron data set; the Enron dataset is from the former energy giant Enron Corp in the United States, which went bankrupt because of bad management and a bribery scandal. The United States Department of Justice conducted an investigation of Enron Corp, including their email. Later, MIT purchased the dataset in order to implement the CALO project, SRI Laboratory Start to sort out the mail; each mail will be stored in the format of the SMTP protocol, and the mail attachments are removed. The email mainly includes the following: mail ID, post time, send mailbox, receive mailbox, mail theme, mail content and so on. The paper uses the Enron_20150507 version; the dataset contains 150 users, with a total of 517374 emails.

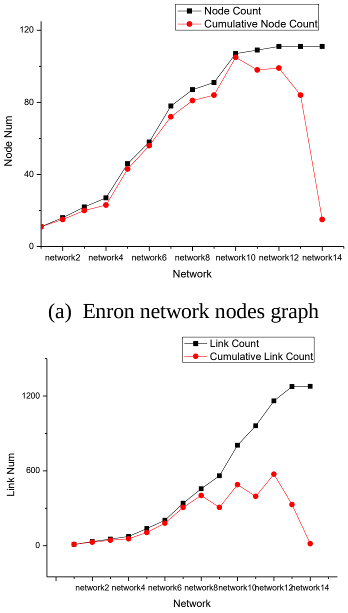
The Enron dataset is huge; after treatment by researchers, the dataset still has many mistakes or useless information. For example, the mail records email information from January 1999 to June 2002, but some emails are from 1980 and 2044; we delete these; in addition, the dataset contains 150 user mail folders; each folder corresponds to a user, but some folders do not correspond to the right email address; for example, for the folder crandell-s, the e-mail address is *.crandall@enron.com; the user name and e-mail address do not match, and we need to manually change them; it is not possible to avoid by getting the corresponding email address in later; and there are other situations like message ID repeats or sending their own e-mail. The above examples show the vulnerabilities of email, which must be addressed.

Because the main analysis contains the internal mail information records of 150 employees of Enron, we delete email addresses that do not belong to the company's internal e-mail address and keep the communication records of the 150 Enron employees. The messages contain many mass emails or copied emails; because the communication record should be transformed into a graph or matrix later, the paper separates the mass addresses and copy addresses; if a message corresponds to multiple e-mails, this will only increase the total number of communication records, but the communication between employees remains unchanged, and does not affect the subsequent analysis.

B. *E-mail Network Structure Analysis*

In this paper, we use the ORA software developed by the CASOS Laboratory of Carnegie Mellon University, which can transform data in a corresponding format to the network structure diagram, and we can perform dynamic analysis of the data.

In order to analyze the generation process of the network, this paper divides the network into different time periods according to the sending time. The Enron dataset is from January 1999 to June 2002; we take three months for a period of time; to establish a network of dynamic books, each book time corresponds to a communication network, for a total of 14 networks. For example, the time from January 1999 to April 1999 corresponds to network1, followed by analogy.

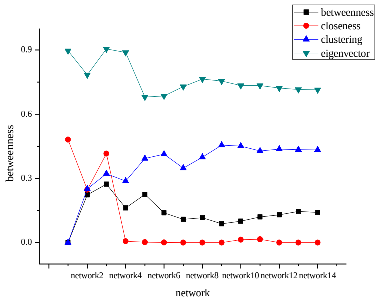


(a) Enron network nodes graph

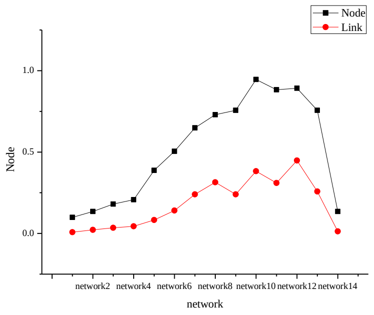
(b) Enron network edges graph

Fig. 1. Enron network diagram.

Fig. 1(a) and (b) show the distribution of the number of nodes and edges in the whole network, respectively, and in each period network as time goes on in the e-mail network; it can be seen from the graph that the number of nodes and the number of edges in the whole network is an increasing process; in the case of network1-network11 in the figure, there are new nodes joining the network and the activity of nodes in each time book is very high; then, the trend of the nodes in the network tends to be gentle; as shown in figure network11-network14, the total number of nodes and edges changes slowly, and the activity of each period decreases sharply. It is during the time of the outbreak of the Enron Corp crisis.



(a) Enron network centrality analysis



(b) Enron information occupancy

Fig. 2. Enron network parameter analysis.

As shown in Fig. 2(a), betweenness, closeness, clustering, eigenvector express the four centralities of network analysis. The process of the whole network from growth to smoothness can be seen. Fig. 2(b) shows the occupancy of nodes and edges in each period network, which represents the activity of 111 users in the network. It can be seen from the figure that user activity exceeds 50% from network5 to network13; the activity of users is very high at this time.

Based on the above analysis, this section summarized the following points: (1) the process of mail network growth; in the case of a certain network size, the node change is increased rapidly at first and then the rate of increase tends to slow until the number of nodes does not change; (2) in the case of a certain network size, traffic in the network begins to increase, reaches a threshold, and the amount of communication is maintained at a fluctuating value; (3) the structure of the mail network will stabilize after the traffic is stable and the network structure has been formed.

C. *Email Information Analysis*

In order to avoid the problem of sparse data in the analysis of mail messages, the paper selects the data from network5 to network13 to analyze. Analysis from the previous section shows that communication traffic is very stable, and user activity is higher; the network structure is stable; the messages of this period can reflect the basic features of each user.

1) *"User-Information feature-Keyword" Model*

In order to analyze the message sending behavior of email users, this paper establishes a model of “User-Information feature-Keyword”. As shown in Fig. 3, User1 sends an email to User2; then, a directed edge is generated from User1 to the Feature Network; the Feature Network generates a feature vector $(F_1, F_2, F_3 \dots)$; F_i represents the weight of this type feature in the current mail; the higher the proportion of the weight in the whole feature, the more the text tends to this kind of characteristic; then a directed edge is generated between the feature network and User2. The relationship model between the user and information feature is constructed. In the paper, the feature class is composed of the key words, and the different feature classes are composed of key words with different attributes; keywords with similar attributes are regarded as the same feature class. Thus, the mail communication model of “User-Information feature-Keyword” is constructed.

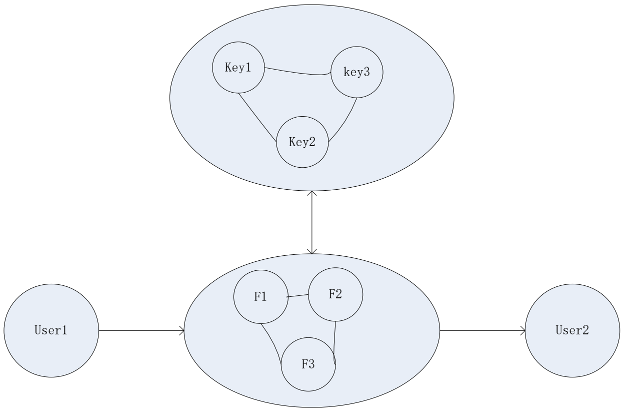


Fig. 3. Schematic diagram of mail communication network

Due to the large number of messages, how to extract the user information effectively is a problem. Because the user information characteristics are excavated from the emails, the problem of extracting the user information characteristics is changed to how to extract the text feature, that is, keyword. In the mail message, the subject and the text are the two main parts of the message content. The subject can often show the main content of the text, but the subject is generally composed of a short sentence; the number of words are usually not more than 10; it cannot express the whole message; in this paper, through the integration of subject and text information, the most representative keywords in an email are extracted, and the feature class is formed by the similarity between these keywords.

2) *Keyword Extraction Method*

The paper uses the DF (Document Frequency) method to select features. The DF algorithm is a relatively simple feature selection algorithm; it refers to the amount of text including words in the dataset. The general document frequency algorithm sets the threshold to remove the feature according to which the document frequency is particularly high or the document frequency is particularly low; these two features represent the two extremes, that is, “no representation” and “no use”. The evaluation function of the DF is a text correlation method, which makes the establishment of stoplist critical, and this will directly affect the classification feature. Because the dataset is English, the stop word is (http://jmlr.csail.mit.edu/papers/volume5/lewis04a/a11-smart-stop- list/english.stop). The DF method is also flawed; for example, some scarce words can reflect the good performance of certain characteristics, but they are removed because the amount is too small. In order to reduce the interference caused by these types of factors, in this paper, in addition to considering the text feature, we also take into account the key words of the subject to achieve a complementary and to maximize the expression of e-mail information characteristics.

Based on the above methods, we select 150 key words as text feature vectors; although the stop word is considered in the algorithm, for the problem of data, there are a lot of useless words, such as “pm”, “st” and some numbers; by deleting these useless words, we get a total of 92 key words, and the following Table I shows the top of the list of words:

TABLE I. KEYWORD LIST

KeyWord	Count	KeyWord	Count
meeting	2056	report	887
credit	1521	master	882
trader	1427	california	859
presentations	1285	power	855
gas	1238	capacity	655
enron	1225	trading	648
responses	1062	notification	500
agreement	1027	company	489
new	908	summary	446
energy	903	draft	445

Enron was the biggest energy giant in the United States; it was one of the most active stocks traded in the 2000-2002 the period; from the first few keywords, “meeting”, “credit”, “trader” and “gas” fit the company’s image; taking as an example “presentations”, from the words “enron”, “responses” and “agreement”, we can see there is a great deal of work communication in the emails. These keywords can be used to express the email information characteristics to a certain extent.

There are many keywords that will impact the analysis of sparse data, so the K-means method is used to cluster the keywords; the parameters of the algorithm are as follows: initialize type = Random cluster, cluster method = Euclidean distance, cluster num = 8; we classify all keywords into 8 feature classes by clustering, which eliminates the interference caused by sparse data. After clustering, we name the eight kinds of clusters as the feature class, and the following Table II shows the key words of the eight feature classes:

TABLE II. FEATURE CLASSES

Feature	Size	Member	Primary Key
Feature1	5	america,received,north,entity,master	america
Feature2	12	thanks,price,energy,deals,deal,day, group,need,know,sara,gas,time	energy
Feature3	18	texas,type,approved,isda,contract, distributed,transaction,executed,products date,Stephanie,copies,border,susan, exchange,financial,effective,counterparty	products
Feature4	2	enron,hou	enron
Feature5	11	work,forwarded,john,trading,power, business,subject,new,market,pm,mark	trading
Feature6	4	meeting,office,west,mike	meeting
Feature7	50	eol,carol,let,fax,going,want,tomorrow, morning,phone,following,jones,don,does Email,Monday,think,forward,change, report,information,through,mail,smith, use,question,available,contact,legal, credit,taylor,like,list,sent,look, regards,regarding,attached,week,houston street,tana,today,Friday,working, company,david,just,make	credit
Feature8	2	agreement,corp	corp

3) *Analysis of User Participation in Feature Class*

In the paper, we analyze the situation of different users participating in the feature class, from which we can find that users have different preferences for different classes. As shown in Table III, the paper compares several users with a large amount of communication.

TABLE III. COMMUNICATION RATIO OF DIFFERENT FEATURE CLASSES

	1	2	3	4	5	6	7	8
jones-t	5.16%	7.39%	13.67%	28.25%	11.48%	0.54%	28.71%	4.80%
shackleton-s	5.33%	20.76%	7.50%	16.70%	11.87%	2.22%	31.78%	3.84%
grigsby-m	0.63%	21.51%	4.96%	14.41%	17.23%	15.25%	24.69%	1.32%
stclair-c	1.19%	10.23%	8.01%	16.86%	12.48%	1.65%	46.55%	3.04%
williams-w3	0.38%	47.00%	2.67%	2.51%	9.94%	3.22%	33.99%	0.30%
phanis-s	23.35%	5.16%	40.57%	7.81%	2.13%	0.17%	8.75%	12.03%
delainey-d	1.04%	11.21%	1.33%	22.80%	24.61%	4.16%	31.66%	3.18%
taylor-m	1.84%	9.94%	9.00%	20.52%	22.53%	2.67%	29.95%	3.56%
keiser-k	0.49%	10.20%	1.11%	11.77%	4.19%	31.68%	31.59%	8.97%
symes-k	0.21%	31.32%	4.75%	4.50%	11.28%	4.10%	42.55%	1.30%
whalley-l	0.00%	18.53%	3.44%	10.12%	41.36%	0.48%	24.93%	1.15%
perlingiere-d	13.36%	6.95%	8.01%	15.84%	5.39%	0.87%	43.93%	5.65%
heard-m	23.37%	7.98%	26.62%	7.27%	5.14%	0.62%	16.54%	12.47%
skilling-j	3.18%	5.98%	2.73%	9.41%	8.71%	13.50%	52.73%	3.78%
haedicke-m	2.18%	8.07%	5.92%	22.71%	27.87%	3.08%	28.12%	2.06%

From the table, we can see that most users show big differences in feature classes in communication. We can see that stclair-c, symes-k, perlingiere-d and skilling have great interest in the seventh type of features, and the seventh feature class occupies more than 40% of the proportion; similarly, the communication traffic of williams-w3 occupies 47% of the proportion in the second feature class; the communication traffic of whalley-l occupies 41.36% of the proportion in the fifth feature class; for other users, such as jones-t, delainey-d and keiser-k, the communication traffic occupies about 30%; other users can reach more than 20%.

Fig. 4 shows the communication ratio of different feature classes by users with large communication traffic. We can see that several users in the linear graph have obvious uplift in a certain feature class while having a low proportion in other feature classes; in addition, some users’ line drawings have two distinct bumps, indicating that the users are interested in more than one category, while they also have a low interest in other feature classes. Fig. 4 is a very good response to the feature of email messages.

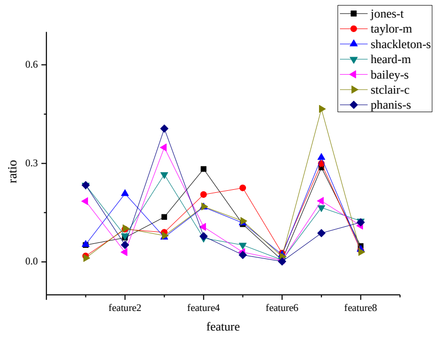


Fig. 4. Communication ratio of different users in different feature classes.

The paper selects the user with the largest communication traffic to analyze, for example, user jones-t and stclair-c; his communication table and corresponding feature relation graph is as Fig. 5 and 6, and Tables IV and V.

TABLE IV. COMMUNICATION RELATIONSHIP OF STCLAIR-C

name	shackleton-s	taylor-m	Jones-t	bailey
stclair-c	299	235	227	269

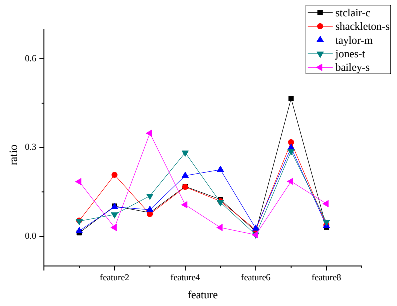


Fig. 5. Comparison between user Stclair-C and its communication characters.

TABLE V. COMMUNICATION RELATIONSHIP OF JONEST-T

name	Taylor	shackleton-s	heard-m	bailey-s	stclair-c
jones-t	402	313	302	269	261

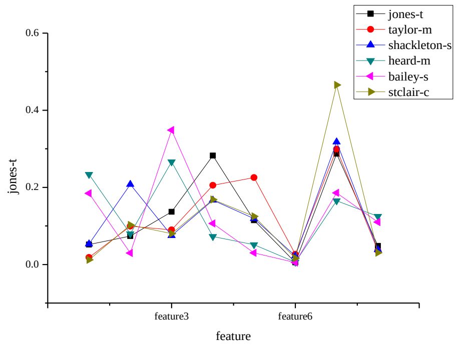


Fig. 6. Comparison between user jones-t and its communication characters.

As shown in Fig. 5 and 6, which show the communication relationship of users with larger communication traffic, the black line represents the user being compared in the figure; other colors represent the users involved in the comparison. From the figure, we can find that there is more than one feature class that takes a larger proportion between users participating in the comparison and users being compared.

Through the above analysis, we find that the "User-Information features-keyword" model can effectively analyze the e-mail dataset; we can find the user information characteristics and what they are interested in to classify the users. These findings can provide a theoretical and data basis for the identification, recommendation and evolution of the network in the future.

III. E-MAIL NETWORK EVOLUTION MODEL BASED ON USER INFORMATION CHARACTERISTICS

In the last chapter, we find that there is a correlation between the user's communication intensity and the user's information characteristics through the "User-Information Feature-keyword" model, so this paper proposes an E-mail Network Evolution Model Based on User Information, the abbreviated UIEM model. The UIEM model is proposed based on Undirected Weighted Network Model (BBV) and Local World Model; it considers the node selection idea of BBV and the group of LocalWorld.

A. Related Knowledge Theory

Define abbreviations and acronyms the first time they are used in the text, even after they have been defined in the abstract. Abbreviations such as IEEE, SI, MKS, CGS, sc, dc, and rms do not have to be defined. Do not use abbreviations in the title or heads unless they are unavoidable.

1) BBV Network Model

The BBV [18] model considers the edge weight based on the scale-free network; it can be represented by the adjacency matrix of the network; w_{ij} represents the weight of the edge between node i and node j; if the network is undirected, the matrix is symmetric, $w_{ij} = w_{ji}$; if the network is directed, w_{ij} and w_{ji} are considered separately. Here, we introduce the BBV model.

In the BBV model, the degree of the node is called the strength or tendency of the node, in which the strength of the node s_i is defined as:

$$s_i = \sum_{j\in Neighbor(i)} wij \qquad (1)$$

The *Neighbor(i)* represents the neighbor node set of node *i*;

The evolution rules of BBV model are as follows:

a) The initial network: the network is a unity coupling network with m_0 nodes; each edge is given the initial weight w_0.

b) The growth of the network: In each time step, there is a new node N with $m(m<m_0)$edges; the new node selects a node from the original network to connect according to a certain probability; the probability that the node is selected is as follows:

$$\Pi = \frac{s_i}{\sum_j s_j} \qquad (2)$$

c) The dynamic evolution of edge weights: in the BBV model, each time the new edge (*n,i*) is given the initial weights of w_0, and to facilitate the analysis, the new edges will only impact the weight between node *i* and its neighbor *j*, so the weight of the edge between node *i* and node *j* readjusts as:

$$w_{ij} \to w_{ij} + \Delta w_{ij} \tag{3}$$

$$\Delta w_{ij} \to \delta_i \frac{w_{ij}}{s_i} \tag{4}$$

It can be seen from the above equation that while node *i* adds a new edge, the edge between node *i* and its neighbor will increase. δ_i represents the rate of change; the edge weight between node *i* and node *j* increases Δw_{ij}, which shows that the neighbors of node *i* allocate additional traffic according to the edge weight. Therefore, the weight of node *i* finally adjusts to:

The δ_i said the rate of change, the weight of node i and its neighbor nodes and the edges of *j* increased Delta w_{ij}; Delta w_{ij} said *i* node and its neighbors through the weights of edges to allocate additional traffic. Therefore, the weight of the node *i* is finally adjusted to:

$$s_i \to s_i + \delta_i + w_0 \tag{5}$$

Compared with the scale-free model, the BBV model considers the edge weights, which makes the network strength distribution obey the power-law distribution; by adjusting the δ_i values, it can change the same feature measurement, which makes great progress compared with scale-free network.

2) Local World network model

In the real world, many people have a specific circle, and they only live in this circle, which is the origin of the local world network; the local world is only a part of the entire network. The local world model [19], [20] is used to describe the situation.

The evolution rules of the local world network model are as follows:

a) The initial network: a network with m_0 nodes and e_0 edges;

b) The growth of the network: each time step a new node N joins into a network, node N with $m(m<m_0)$ edges. $M(m<M)$ nodes are selected randomly from the network as the local world of the new node N, and the new node N selects the m nodes from the local world network according to the node priority probability formula.

$$\Pi_{local} = \frac{M}{m_0+t} \cdot \frac{k_i}{\sum_{j \in local} k_j} \tag{6}$$

The local world network model is suitable for some specific networks, when the network size is large enough, the cluster coefficient is close to 0.

B. Basic Concepts of the Model

Based on the research and analysis of the Enron e-mail network in the third chapter, the paper proposes an e-mail network model based on the user information characteristics. Prior to this, we first give some definitions of basic concepts.

Definition 1: Feature vector of node: The feature attributes of nodes are represented by the tendency of nodes to fall under different feature classes. The information feature vector of the node is represented mathematically as:

$$f_i = \{F_1, F_2, F_3, F_4 \dots F_k\} \tag{7}$$

f_i represent the feature vector of node *i*, F_k represents the weight of the first K feature class of node *i*.

Definition 2: The similarity between nodes: The degree of similarity between the nodes, the higher similarity expresses the nodes are more likely to belong to the same class, and they have a high possibility of connecting with each other. The similarity between nodes is expressed mathematically as:

$$Similarity(i,j) = \frac{F_{i1} \cdot F_{j1} + F_{i2} \cdot F_{j2} + \dots + F_{ik} \cdot F_{jk}}{\sqrt{{F_{i1}}^2 + {F_{i1}}^2 + \dots + {F_{ik}}^2} \cdot \sqrt{{F_{j1}}^2 + {F_{j1}}^2 + \dots + {F_{jk}}^2}} \tag{8}$$

In this paper, we take the cosine similarity of vector space, which does not take the distance between two vectors into account.

Definition 3: Feature similarity network: A collection of M nodes with higher degree of features similarity from the original network after the new node is added. Mathematical representation:

$$V_{Feature}(i) = \{v_1, v_1, v_1 \dots v_M\} \tag{9}$$

Definition 4: Node strength: In a directed weighted network, the in strength of node *i* is the sum of edge weights, while node *i* is the in node; the out strength of node *i* is the sum of edge weights, while the node *i* is the out node.

The in strength of node *i* is:

$$s(in)_i = \sum_{j \in Neighbor(i)} w_{ji} \tag{10}$$

The out strength of node *i* is:

$$s(out)_i = \sum_{j \in Neighbor(i)} w_{ij} \tag{11}$$

The strength of node *i* is:

$$s_i = s(in)_i + s(out)_i \tag{12}$$

C. Model Evolution Rule

In the paper, according to the email transmission, sending, forwarding and replying, we consider the characteristics of the user in the message communication process; when a new node is added, selecting a certain number of nodes from the original network to form a feature similarity network, the new node selects a node from the feature network to connect; at the same time, there is internal evolution in the original network.

The specific construction algorithm of the e-mail network model based on the characteristics of user information is as follows:

1) *Initial Network*

The initial network contains m_0 nodes, and each node initializes a feature vector. For the sake of simplicity, this paper defines the m_0 class in the feature vector, and the initial value of each feature class in the feature vector of m_0 nodes is:

$$F_i(\mathrm{k}) = \begin{cases} 0, & i \neq k; \\ 1, & i = k; \end{cases} \tag{13}$$

$F_i(\mathrm{k})$ represent the value of the first k class of node i; the m_0 nodes form a fully coupled network, and the initial edge weight is w_0.

2) *The Growth of the Network*

in each time step, a new node n joins the network, and the node n is randomly assigned a feature vector; at the same time, $M(M<m_0)$ nodes is selected from the original network according to the similarity of the feature vector to form the feature similarity network and with a certain probability to proceed as follows:

a) The new node with *M (m<m0)* edge joins the feature similarity network with probability *p1*; some edges are out edges with probability q, and others are edges with probability q. The probability of node i as in node is:

$$\Pi = \frac{s(in)_i}{\sum_j s(in)_j} \tag{14}$$

The probability of node i as the out node is:

$$\Pi = \frac{s(out)_i}{\sum_j s(out)_j} \tag{15}$$

Among them, j is a set node that forms the feature similarity network.

b) The evolution in a feature similar network: to add m edges into a feature similarity network with probability *p2* to achieve internal growth; in the feature similarity network, the new edge is <*i,j*>; if there is a connection between node i and node j, we increase their weight; or, we establish a new edge and assign the initial weights w_0. The probability of choosing node i is (15); the probability of choosing node j is (14).

c) Connection between the feature similarity network and the external network: to add m edges with probability *p3* between the feature similarity network and the external network. In terms of the feature similarity network, m edges is as out edges with probability q, and m edges is as in edges with probability *1-q*. The choice of node is according to the operation *2*).

3) *The dynamic Evolution of the Weight*

The generation of the new edge will trigger the readjustment of the weight between the node and the neighbor node. If the new edge is the in edge, the weight associated with the node i is changed to:

$$w_{ji} = w_{ji} + \Delta w_{ji} \tag{16}$$

$$\Delta w_{ji} = \delta_i \frac{w_{ji}}{s(in)_i} \tag{17}$$

Parameter δ_i is the additional traffic burden while new edge <*n,i*> is added; the neighbor nodes of node i share this traffic. So, the strength of node i is adjusted to:

$$s(in)_i = s(in)_i + w_0 + \delta_i \tag{18}$$

If the new edge is the out edge, the weight of the node i changes to:

$$w_{ij} = w_{ij} + \Delta w_{ij} \tag{19}$$

$$\Delta w_{ij} = \delta_i \frac{w_{ij}}{s(out)_i} \tag{20}$$

The out strength of node i is adjusted to:

$$s(out)_i = s(out)_i + w_0 + \delta_i \tag{21}$$

The growth of the network in steps *2*) and *3*) belongs to the evolution of the inner network, and the model does not consider additional traffic burden, so the corresponding weight of the edge and node strength are increased by w_0.

4) *Node Feature Vector Adjustment*

While a new edge <*i.j*> is added, the node i delivers information to the node j, and it only causes the feature vector adjustment of node j; that is, each node that changes is an in node. The paper uses vector space cosine similarity, and it only considers the direction gap but not the distance gap, so the feature class in the feature vector of node j is adjusted to:

$$F_{jk} = F_{jk} + F_{ik} \tag{22}$$

After t time steps, the network contains $m_0 + t$ nodes.

IV. Experiment and Analysis

According to the evolution model based on user information characteristics, we use Java programming to achieve the evolution of the network; then we get a network topology matrix; finally, we use MATLAB to calculate the parameters distribution. In the following, we analyze the model from two aspects, that is, average path length and cluster coefficient.

A. *Average Path Analysis*

Network path and diameter are important parameters of network transmission delay, and network transmission delay is an important factor of network performance and information dissemination. In order to represent the performance of the whole network, the concept of average path is introduced. First, the shortest path for each node to other nodes is obtained, each node is only allowed access once. After finding the shortest path between all the nodes, the average path length of the current network can be calculated.

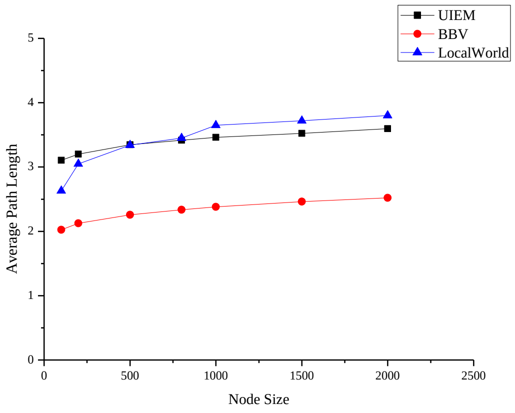


Fig. 7. Average path comparison of different models when the network size is 2000.

Fig. 7 shows the changes in network average path while the network size ranges from 1 to 2000. We can see that the average path of UIEM is similar to the local world network. Compared to the BBV network model, the average path of UIEM is larger, but the growth of the average path length is slower than that of the Local World network.

B. Network Cluster Coefficient

Cluster coefficient is the relationship between the node and its neighbors; in general, it is used to express the possibility that people's friends are also friends. Because our network is a weighted network, we need to calculate the weighted clustering coefficient of the network [21]-[23]. The cluster coefficient represents the clustering degree of nodes in the network and is an important feature of a network [24]-[26]. A large number of studies have shown that real networks have high clustering characteristics.

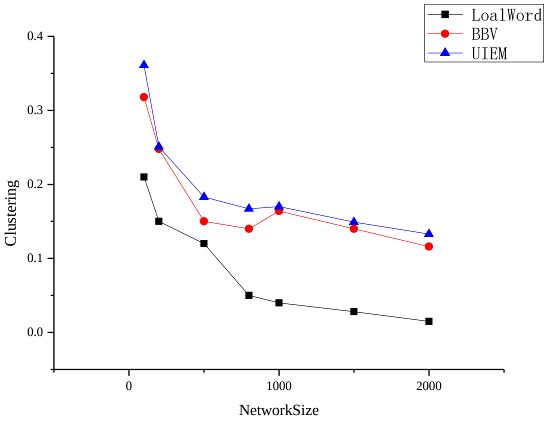


Fig. 8. Comparison of cluster coefficients of different models when the network size is 2000.

As shown in Fig. 8, the abscissa represents the size of the network, and the ordinate represents the cluster coefficient of the network. It can be seen that the cluster coefficient of UIEM is similar to the BBV model, but it is higher than the Local World network. The cluster coefficient is the degree of the network group. The cluster coefficient of the Local World network tends to 0 when the network size is large enough, and it is clearly inconsistent with the actual network. The cluster coefficient of UIEM declines slowly with the growth of the network; this is consistent with the different actual networks to better reflect the authenticity of the network

C. Contrast Experiment

In this paper, in addition to using the Enron dataset, we also downloaded a mail data uploaded by the Department of Automation, Shanghai University to analyze the two data sets and perform a comparison with UIEM. For practical reasons, the Enron dataset contains 150 nodes, and the mail dataset from the Department of Automation, Shanghai University contains 1133 nodes, so this paper uses the corresponding number of nodes to compare with UIEM to ensure fairness.

TABLE VI. COMPARISON OF ENRON NETWORK WITH THREE MODELS

	Average Path	Cluster Coefficient
Enron	6.3	0.433
BBV	2.053	0.302
LocalWorld	2.82	0.18
UIEM	3.125	0.36

As shown in Table VI, when the scale is small, the average path of UIEM is longer, which is closer to the real Enron network; and the cluster coefficient of UIEM is higher; it is close to the actual network model.

The following is a comparison between the e-mail dataset and the three network models, which are shared by the Department of automation, Shanghai University.

TABLE VII. COMPARISON OF E-MAIL NETWORK FROM DEPARTMENT OF AUTOMATION, SHANGHAI UNIVERSITY WITH THE THREE MODELS

	Average Path	Cluster Coefficient
Email From Department of Automation, Shanghai University	3.606	0.22
BBV	2.381	0.155
Localworld	3.802	0.042
UIEM	3.482	0.168

From Table VII, we can see that when the network size is 1133, the average path length of the Local World network is growing too fast, more than the average path of real email network, but our UIEM is closer to a real email network than the other two models.

V. CONCLUSION

In this paper, we take the idea of using the Local World network and the dynamic evolution of the BBV model; then, according to the relationship between user information characteristics and communication that is found in chapter three, we present an e-mail network evolution model based on the characteristics of user information and give the construction rules and related definitions. Finally, realizing the evolution of the network by programming, we find that the strength and degree of nodes are in accordance with the power law distribution. And compared with the BBV model and the Local World, UIEM is closer to the actual network and has practical significance.

ACKNOWLEDGMENT

The authors would like to thank the reviewers for their detailed reviews and constructive comments, which have helped improve the quality of this paper. This work is sponsored by the Natural Science Foundation of Heilongjiang Province of China under Grant No. LC2016024, the Natural Science Foundation of the Jiangsu Higher Education Institutions Grant No. 17KJB520044.

REFERENCES

[1] D.J. Watts, S.H. Strogatz. "Collective dynamics of 'small-world' networks". Nature, vol. 393, pp. 440-442, 1998

[2] M.E.J. Newman, D.J.Watts. "Renormalization group analysis of the small-world network model". Physics Letters A, vol. 263, pp. 341-346, 1999.

[3] Barabasi A L, Albert R. "Emergence of scaling in random netwoks" Science, vol, 286, pp. 509. 1999.

[4] R. Albert, A.L. Barabasi. "Topology of evolving networks: local events and universality". Physical Review, vol. 85, pp. 5234, 2000.

[5] S. Fortunato, A. Flammini, F. Menczer. "Scale-free network growth by ranking". Physical Review, vol. 96, 2006

[6] X.L. Sun, H.F. Lin, K Xu. "A social network model driven by events and interests". Expert Systems With Applications, vol. 42, pp. 4229-4238, 2015.

[7] A. Abbasi, L. Hossain, L. Leydesdorff. "Betweenness centrality as a driver of preferential attachment in the evolution of research collaboration networks". Journal of Informetrics, vol. 6, pp. 403-412, 2012.

[8] A. Barrat, M. Barthélemy, A. Vespignani. "Weighted evolving networks: coupling topology and weight dynamics". Physical Review, vol. 92, 2004

[9] A. Barrat, M. Barthélemy, R. Pastor-Satorras, A. Vespignani. "The architecture of complex weighted networks". Proceedings of the National Academy of Sciences of the United States of America .vol. 101, pp. 3747-3452, 2004

[10] X. Sun, J.Y. Dong, R.C. Tang, M.T. Xu, L. Qi, Y. Cai. "Topological evolution of virtual social networks by modeling social activities". Physica A: Statistical Mechanics and its Applicat, vol. 433, pp. 259-267, 2015.

[11] B. Yao, M. Yao, X.E. Chen, X. Lin, W.J. Zhang. "Applied Mechanics and Materials. Research on Edge-Growing Models Related with Scale-Free Small-World Networks". Applied Mechanics & Materials, vol. 513, pp. 2444-2448, 2014.

[12] B. Yao, C. Yang, M. Yao, et al. "Graphs as Models of Scale-Free Networks'. Applied Mechanics & Materials, vol. 380, pp. 2034-2037, 2013.

[13] W.Q. Wang, Q.M. Zhang, T. Zhou. "Evaluating network models: A likelihood analysis" EPL (Europhysics Letters)， vol. 98, pp. 28004-28009, 2012.

[14] Z.Y. Zou, P. Liu, L. Lei, J.Z. Gao. "An evolving network model with modular growth". Chinese Physics B, vol. 21, pp. 603-609, 2012.

[15] B.K. Wang, Z.H. Pei, L. Wang. "Evolutionary dynamics of cooperation on interdependent networks with the Prisoner's Dilemma and Snowdrift Game". EPL (Europhysics Letters), vol. 107, pp. 58006, 2014.

[16] H. Zhuang,Sun Y,Tang J,et al. "Influence maximization in dynamic social networks". 2013 IEEE 13th International Conference on Data Mining (ICDM) .pp. 1313-1318, 2013.

[17] N. Ilhan, I.G. Oguducu. "Community Event Prediction in Dynamic Social Networks". 2013 12th International Conference on Machine Learning and Applications (ICMLA) , pp. 191-196. 2013.

[18] A. Barrat, M. Barthelemy, A. Vespignani. "Modeling the Evolution of Weighted Networks". Physical Review E, vol. 70, pp. 1-13, 2004.

[19] X. Li, G. Chen. "A local world evolving network model". Physical A. vol. 328, pp. 274-286, 2003.

[20] Z.f. Pan,X. Li, X.F. Wang. "Generalized local-world models for weighted networks". Physical review. E, Statistical, nonlinear, and soft matter physics. Vol. 73. 2006

[21] X. Xue, S. Wang, B. Gui, et al. "A computational experiment-based evaluation method for context-aware services in complicated environment". Information Sciences, vol. 373, pp. 269-286, 2016.

[22] X. Xue, Y.M. Kou, S. Wang, et al. "Computational experiment research on the equalization-oriented service strategy in collaborative manufacturing". IEEE Transactions on Services Computing, vol. 11, pp. 369-383, 2018.

[23] X. Xue, H. Han, S. Wang, et al. "Computational Experiment-based Evaluation on Context-aware O2O Service Recommendation". IEEE Transactions on Services Computing, 2016

[24] T. Wang, .Y. Wu, X. He, et al. "A Cross Unequal Clustering Routing Algorithm for Sensor Network". Measurement Science Review, vol, 13, pp. 200-205, 2013.

[25] T. Wang, Y. Cao, Y. Zhou, et al. "A Survey on Geographic Routing Protocols in Delay/Disruption Tolerant Networks (DTNs)". International Journal of Distributed Sensor Networks, vol.6 , 2016.

[26] Y. Cao, T. Wang, O. Kaiwartya, et al. "An EV Charging Management System Concerning Drivers' Trip Duration and Mobility Uncertainty". IEEE Transactions on Systems Man & Cybernetics Systems, vol. 48, pp. 596-607, 2018.