A New Community Division based on Coring Graph Clustering

Peng Ling, Xu Ting-rong, Lan Meng
College of computer science, Soochow university, Suzhou, 215006,China
Email:epengling@163.com

***Abstract*—A new community finding algorithm, based on the greedy algorithm with graph clustering by computing the density variation sequence and identifying core nodes, number of communities, partition the certain nodes to some belonged community with the similarity of characteristics of communication behavior by continuous readjusting the centrality of the communities. The use of community density and effective diameter to measure the quality of the community partition on the real datasets of email corpus shows the feasibility and effectiveness of the proposed algorithm.**

***Index Terms*—graph clustering; mail community partition; dynamic centering**

I. Introduction

With the development of the internet, the network has become a more and more important tool in connecting with each other in our work and life. Meanwhile, it also appears the network community[1] which is based on the virtual social relationship. In this kind of network, there are more connections between nodes of the same type while less between nodes of the different ones—see Fig 1. Network community, to some extent is similar to the real community, and also satisfies Six Apart theory and 150 law[2]. So finding network communities in a large network is very helpful for us to understand the real social relationships.

As a network community, mail community is also isomorphic to the real social relationships and conforms to the small-world network model[3]. Besides, because some of the advantages of e-mail itself[4], such as: 1) with a relative standard format. 2) Email not only provides the relationship between people connected, but also records communication frequency and time. So we can use the information to build a weighted social network. 3) with the timestamp in the email, it is more convenient to find the dynamic social network. There has been an increased amount of study on identifying online communities now. The most representative algorithms are G-N algorithm, introduced by Girvan and Newman[5], is based on the edge betweenness that measures the fraction of all shortest paths passing on a given link. Layered clustering algorithm[6], introduced by Aaron and Newman and Radicchi algorithm[6] ,which is based on the number of triangles and so on. However, some of the time complexity of these algorithms is too high and difficult to handle large-scale networks. For example, in the worst case the time complexity of G-N can be achieved to $o(n^3)$.

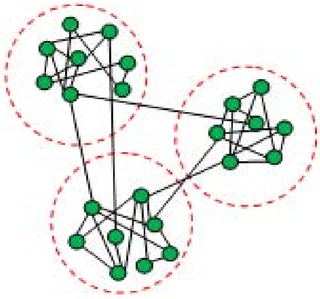

Fig 1 A small network with community structure of the type considered in this paper. In this case there are three communities, denoted by the dashed circles, which have dense internal links between which there is only a lower density of external links.

In this paper, we propose a method that finds mail community by first calculating the density variation sequence based on the greedy clustering algorithm, then identifying the number of communities and the cores of each community. lastly, based on these core nodes, we assign all the other nodes to the nearest community based on the similar communication behavior by readjusting the dynamic centering of each community. Related work is discussed in section 2.Section 3 gives some definitions that required in the paper. In section 4, we'll describe in detail of the new approach in mail community detecting. Section 5 presents some results of our experiments on the real datasets. The summary and future work will be discussed in section 6.

II. Related work

There has been an increased amount of study on identifying online communities now. It is closely related to the ideas of divisive methods in graph theory and computer science, and hierarchical clustering in sociology. Before presenting our own findings, it is worth reviewing some of this preceding work to understand its achievements and shortcomings.

(1) Divisive methods [2,4].

A simple way to identify communities in a graph is to detect the edges that connect vertices of different communities and remove them, so that the clusters get disconnected from each other. This is the philosophy of divisive algorithms. and G-N is the most representative

© 2010 ACADEMY PUBLISHER
doi:10.4304/jsw.5.10.1121-1127

method of divisive methods. It is based on the edge betweenness[5,7] that measures the fraction of all shortest paths passing on a given link. By removing links with high betweenness, we can progressively splits the whole network into disconnected components, until the network is decomposed in communities consisting of one single node. Fig 2 shows us what is the edge betweenness.

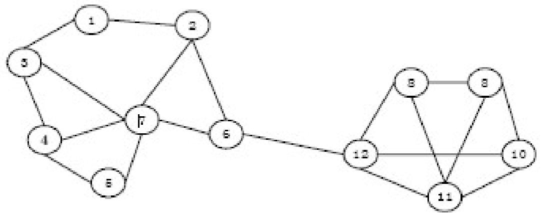

Fig 2 Shortest path centrality (betweenness) is the number of shortest paths that go through a link or node. In this simple case, the link with the largest link centrality is that, joining nodes 6 and 12

The steps of the algorithm are:

1) Calculate betweenness scores for all edges in the network.
2) Find the edge with the highest score and remove it from the network.(If two or more edges tie for highest score, choose one of them at random and remove that.)
3) Recalculate betweenness for all remaining edges.
4) Repeat from step 2)

Algorithm of Tyler et al..Tyler, Wilkinson and Huberman[3,8] proposed a modification of the Girvan-Newman algorithm[3,9], to improve the speed of the calculation and use it in the email community division with a good performance.

(2) Spectral Algorithms

Spectral properties of graph matrices are frequently used to find partitions. Traditional methods are in general unable to predict the number and size of the clusters, which instead must be fed into the procedure. Recent algorithms, reviewed below, are more powerful.

Algorithm of Donetti and Mu˜noz. An elegant method based on the eigenvectors of the Laplacian matrix has been devised by Donetti and Munoz[4].The idea is simple: the values of the eigenvector components are close for vertices in the same community, so one can use them as coordinates to represent vertices as points in a metric space. So, if one uses M eigenvectors, one can embed the vertices in an M-dimensional space. Communities appear as groups of points well separated from each other, as illustrated in Fig1.

Algorithm of Capocci et al.. Similarly to Donetti and Munoz, Capocci et al. used eigenvector components to identify communities[5].

(3) Clique Percolation.

In most of the approaches examined so far, communities have been characterized and discovered, directly or indirectly, by some global property of the graph, like betweenness, modularity, etc., or by some process that involves the graph as a whole, like random walks, synchronization, etc. But communities can be also interpreted as a form of local organization of the graph, so they could be defined from some property of the groups of vertices themselves, regardless of the rest of the graph. Moreover, very few of the algorithms presented so far are able to deal with the problem of overlapping communities[6]. A method that accounts both for the locality of the community definition and for the possibility of having overlapping communities is the Clique Percolation Method (CPM) by Palla et al[6,11].It is based on the concept that the internal edges of community are likely to form cliques due to their high density. On the other hand, it is unlikely that intercommunity edges form cliques. Palla et al. define a k-clique as a complete graph with k vertices. If it were possible for a clique to move on a graph, in some way, it would probably get trapped inside its original community, as it could not cross the bottleneck formed by the intercommunity edges. Palla et al. introduced a number of concepts to implement this idea. Two k-cliques are adjacent if they share k-1 vertices. The union of adjacent k-cliques is called k-clique chain. Two k-cliques are connected if they are part of a kclique chain. Finally, a k-clique community is the largest connected subgraph obtained by the union of a k-clique and of all k-cliques which are connected to it.

The more details of the related work about community division can be get from the reference [10, 12, 16, 20].

III. Definitions

As a kind of social network, we can import the method of community discovery in social network into mail networks [17,18]. In order to simple describe the algorithm, the mathematical description and explanation of the mail network graph and some definitions are given below.

(1) E-mail network graph. In order to describe the linkage information including communication frequency and directions of senders and receivers, we choose directed and weighted graph to show the email network graph. Set G=(V,E,W), where V is the set of all nodes that represent email senders or receivers. E is the set of all the edges connected between senders and receivers. $\boldsymbol{A_i}$ is defined as the nodes that directed connected to the node $\boldsymbol{v_i}$, and can be described as: $\boldsymbol{A_i} = \{\boldsymbol{v_j} \mid \boldsymbol{e_{ij}} \in \boldsymbol{E}\}$.W is the set of weights for each edge. Any two nodes v_i, v_j, if e=(v_i, v_j) or e=(v_j, v_i), then there exists communication linkage between v_i and v_j. w(e) ∈ W, describes the communication frequency of the node v_i and v_j. Fig 3 gives a description of a simple email network graph.

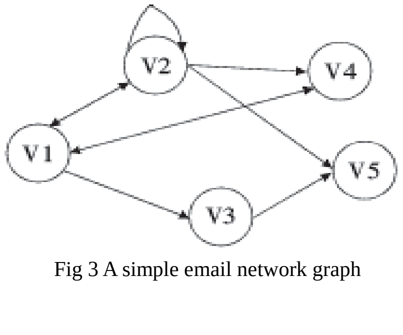

Fig 3 A simple email network graph

© 2010 ACADEMY PUBLISHER

(2) Node degree. The degree of node v_i represented by deg(v_i),is defined as the number of nodes which directed connected with it. that is, deg(v_i)= e_{ij} . while out-degree of node v_i is the number of emails it sent and in-degree is the number of emails it received, and represented by outdeg(v_i), indeg(v_i) respectively.

(3) Node density. Set node $i \in H \subseteq V$, then we define the local density at i with respect to H as.

$$d(i,H) = \frac{1}{H}(\sum_{j \in H} w_{ij} + \sum_{j \in H} w_{ji}) \qquad (1)$$

$\boldsymbol{w_{ij}}$ represents the number of emails that sent from i to j, while $\boldsymbol{w_{ji}}$ is the number of emails sent from j to i.

Function D(H) measures the local density of the weakest node of H defined by:

$$D(H) = \min_{i \in H} d(i,H) \qquad (2)$$

(4) Virtual community[1]. A community is a sub-graph of the network, which must be satisfied with the following conditions: many connections between the nodes in each subset itself while few links between nodes which are belonged to different subsets. that is, nodes in the same community have dense internal links but between which there is only a lower density of external links.

(5) The center of virtual community. A community consists of m nodes $v_1, v_2, ..., v_m$,the communication information between $\boldsymbol{v_i}$ and the other nodes of the email graph is recorded in the set $\boldsymbol{X_i}$. So we define the center of the community as (3) , which is the average connection of the nodes in the community with the other nodes, it is the representative of the community.

$$\bar{\boldsymbol{v}} = \frac{1}{\boldsymbol{m}} \sum_{i=1}^{m} \boldsymbol{X_i} \qquad (3)$$

(6) Density and effective diameter of the community[14]. Set G_k , which is a community, is the sub-graph of G.. Let D(G_k) as the density of G_k . We defined it as the ratio of the in-degree and out-degree of all the node and the number of nodes. It can be described as (4).

$$D(G_k) = \sum_{i=1}^{n} (in\deg(v_i) + out\deg(v_i)) / n \qquad (4)$$

Let R(G_k) as the effective diameter of the community, which defined as more than 90%of nodes in the community G_k , their distance is less than or equal to R(G_k).

(7) The similarity between node v and the center of community $\bar{v}$. To facilitate the description of the formula, let X records the linkage information between v and the other nodes, Y records the average linkage information between the center of the community $\bar{v}$ and the other nodes. then the similarity can be defined as (5)

$$Sim(v, \bar{v}) = \frac{XY^T}{|X||Y|} \qquad (5)$$

(8)Modularity[13]. A measure of the quality of a particular division of a network. Consider a particular division of a network into k communities, and define a $k \times k$ symmetric matrix e whose element e_{ij} is the fraction of all edges in the network that link vertices in community i to vertices in community j. (Here consider all edges in the original network—even after edges have been removed by the community structure algorithm, the modularity measure is calculated using the full network.) The trace of this matrix $Tre = \sum_i e_{ij}$ gives the fraction of edges in the network that connect vertices in the same community, and clearly a good division into communities should have a high value of this trace. The trace on its own, however, is not a good indicator of the quality of the division since, for example, placing all vertices in a single community would give the maximal value of Tre=1 while giving no information about community structure at all.

Define the row (or column) sums $a_i = \sum_j e_{ij}$ which represent the fraction of edges that connect to vertices in community i. In a network in which edges fall between vertices without regard for the communities they belong to, we would have $e_{ij} = a_i a_j$ .thus the modularity can be computed as the following.

$$Q = \sum_i (e_{ij} - a_i^2) = Tre - \| e^2 \| \qquad (6)$$

IV. Mining Social Networks

The analysis of social network based on the emails mainly consists of three modules—Email access, Data preprocessing and Network analysis. See Fig 4.

The first module—Email access can be extracted directly from the mail server message and then store into the database, or you can also extract from the individual e-mail client. Sometimes we need to take some conversion to the email address that ordinary people couldn't identify in order to protect the private information. In this paper, the dataset Enron we used is a public one while the email log information of Soochow university is accessed from the mail server by the corresponding authority. We use MD5 conversion to each email address and each address can be identified by a unique mailboxID for considering the privacy of the users of Soochow University.

In the data preprocessing, we need to preprocess the email information accessed from the first step. Since the initial acquisition of the e-mail was too diverse, so it is necessary to clear and analysis the email information[15], and compute the linkage frequency of each sender and receiver (if there exists communication between them) that needed in the following steps. Finally, we use the Mysql database to store the processed email information.

The module of network analysis can be mainly divided into the following 2 steps: firstly, construct the directed and weighted social network with the processed email log information, the nodes in the graph represent the senders or receivers of emails, and edges are the linkage information of the nodes. The second step is to use the

© 2010 ACADEMY PUBLISHER

improved algorithm to mining the social relations implicit in the network graph. While the analysis of community topic and identification of core people in the third module of the Fig 3 is our future study.

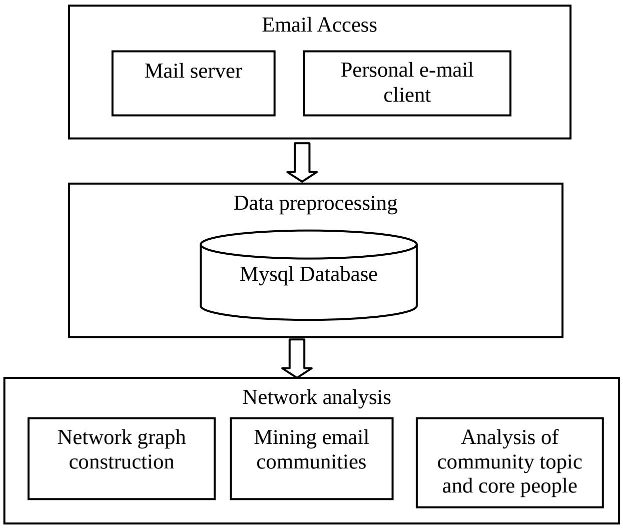

Fig 4. steps of mining email networks

A. Building social networks

Construct a directed and weighted graph based on the email addresses. The node in the graph represents the sender or receiver, while edges between nodes are the linkage frequency. To reduce the influence of the noise on the network graph, we will first set a threshold, and then choose the nodes whose linkage frequency larger than the threshold to build the network graph. In order to save the memory space and speed up when computed in-degree and out-degree of each node, we use adjacency list and inverse adjacency list to store the constructed graph. Here, we select the in-degree and out-degree of each node is larger than 6 respectively, that is, the threshold is 6 according to the experience.

B. E-mail community partition

From the point of graph partition and clustering, by analyzing the sequence of density variation and the similarity between nodes and readjusting the centering of each community, E-mail community partition can be divided into the following 2 steps:

(1) By analyzing the variation of the minimum density value D, we can identify core nodes and further identify the number of communities and the representative nodes of each community

(2) The allocation of non-representative nodes and readjustment of centering of each community

a) Algorithm of computing the number of clustering and coring nodes of each cluster.

We assume that every cluster of the input E-mail graph has a region of high density called a ‘cluster core’, surrounded by sparser regions (non-core) just like the Fig 5. The nodes in cluster cores are denoted as ‘core nodes’, the set of core nodes as the ‘core set’, and the sub-graph consisting of core nodes as the ‘core graph’ and also the original community. In this step, the work to be done is to find such a set of core nodes.

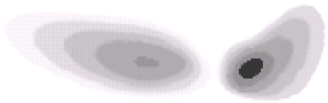

Fig 5 The graph G has a region of high density and surrounded by sparser regions

The local density of each node and the density of the weakest collection of nodes can be computed by the formula (1) (3). So by analyzing the variation of the minimum density value D, we can identify core nodes located in the dense cores of clusters. Specifically, if the weakest node is in a sparse region, the D value will increase when this node is removed, in other words, the next weakest node to be removed will be in a region with higher density. On the other hand, if the removal of the weakest node causes a significant drop in D value, then this node is highly connected with a set of stronger nodes in a high density region. It is potentially a core node because its removal greatly reduces the density of nodes around it. The step of computing the sequence of density variation is described in algorithm 1.

Algorithm1: algorithm of computing the sequence of density variation

Input: E-mail graph G=(V,E,W);

Output: the variation of node density D and the corresponding set of nodes M

```
initialization, t ← 1,H ←V
repeat
    $d(i,H)=\frac{1}{H}(\sum_{j\in H} w_{ij}+\sum_{j\in H} w_{ji})$ ,
    $D(H)=\min_{i\in H} d(i,H)$ ,  $E_c$ ;
If $M_t$ consists of more than one connected component
    then $M_t$ ← the smallest connected component
H=H-$M_t$ ,  t=t+1;
until H is empty
```

Elements of M_t are core nodes if D_t satisfies:

$$R_t=(D_t-D_{t+1})/D_t>\partial \qquad (6)$$

∂ is an adjustable parameter which between 0 and 1,and the parameter selection of ∂ must ensure that the community division meet the following two rules[10]: 1) the smallest components rules: the number of nodes in the community must be greater than or equal to 6 ; 2) community stability rules: it is most stable when nodes in a community are around 120.

After the qualified core nodes identified, there are some methods to partition the core nodes into core graph and finally identify the final number of communities and the representative nodes of each community. E-mail network graph, as a sparse graph, the core graph can be

© 2010 ACADEMY PUBLISHER

found from the connected components, and each component is considered as a cluster core or the representative nodes of the core graph—see Fig 6.

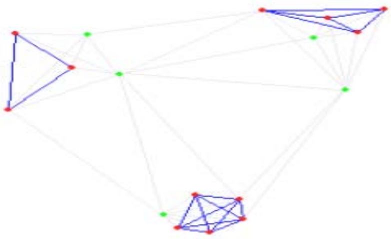

Fig 6. A sparse netwok including 3 core graphs.

b) community partition

After computing the number of clustering and coring nodes of each cluster, now we will discuss the community partition. Algorithm 2 described the steps of E-mail network community partition.

Algorithm 2: ENCD（Email Network Community Detecting）E-mail network community partition

```
Input: G=(V,E,W)
Output: Community ID and the nodes of each community
1: input the number of core graph K and core nodes of each core graph computed by 2.2.1;
2: repeat
3: for t=(T,T-1,…,2,1)
4: for (centering of each of community) //find the centering of community $c_j$ which is the greatest similar to $M_t$
5: if $M_t$ includes non-core nodes labeled $x_i$ , then compute the similarity between $x_i$ and $c_j$ ,if sim( $x_i$ , $c_j$ )>= $\beta$ ,then the community ID of $x_i$ is j and add the node into community $c_j$ //threshold $\beta$ helps consider the situation that one node belongs to multi-communities
6: for each community community[j]
7: readjust the centering of the community[j]
   //computed by formula (3)
8: util the centering of each community not change again.
```

V. Experiment of Community Partition on the Real Datasets

We demonstrate the performance of our algorithm on the Enron email corpus and the email log information of Soochow university between February 2009 and May. The Enron email corpus is a set of emails belonging to 151 users, and consists of 252,759 email messages, it is now about a public network analysis corpus. which can be obtained from http://www-2.cs.cmu.edu/~enron/. Email log information of Soochow university(ELIS) included 183,925 nodes and 391,347 edges after processing, which including the communication information of intramural mailboxes and extramural mailboxes. Considering the influence of the noise, after selecting the in-degree and out-degree of each node larger than 6 respectively, Email log information of Soochow university(ELIS) included only 5948 nodes and 23,479.

In the experiment, we took MD5 [19]conversion to each email address(mailbox) and each mailbox can be identified by a unique mailboxID for considering the privacy of the users of Soochow university.

Experiment environment: 2.80GHz Pentium CPU, 1G RAM, 80 GB hard drive; OS: Microsoft Windows XP; development platform :Myeclipse. The results of community partition are composed by mailboxID, which represents each mailbox, and communityID, which stands for the community labeling.

Fig 7 is the Visualization of the whole *Enron* Email graph. It constructs a social network. Fig 8 is the visualization of the community 6 computed by G-N on Enron, and the detail information of the community is depicted in Table II. Fig 9 is the visualization of the community 3 computed by ENCD on Enron, and the detail information of the community is depicted in Table III. We can see that either partition method, nodes in the same community are connected densely while between are much looser.

Fig 10 shows the results of community partition of Enron with different values of ∂. We can see that the number of communities is quite similar although with different ∂, So the influence of ∂ on the final results of community partition is not great.

Table I shows comparison of the results on Enron email corpus and email log information of Soochow University computed by our algorithm ENCD and G-N algorithm. Here ∂=0.26 for the Enron and 0.125 for the email log information of Soochow university, modularity is one of the indicators for the evaluation of algorithms, usually it is a decimal between 0 and 1, and the greater the modularity is, the higher the quality of that community partition is. Its definition can be seen from the formula (6).

Table I
Comparison of the results computed by ENCD and G-N on the same datasets

algorithm	dataset	modularity	Number of communities
G-N	Enron	0.372	8
	ELIS	0.296	47
ENCD	Enron	0.369	9
	ELIS	0.301	50

Table II and III show the details about the results of community partition of Enron computed by G-N and ENCD algorithm respectively.

From the table II and III, we can see that the results computed by G-N and ENCD are similar. But the distribution of the nodes in each community is of some difference. For example, there is only one node in the three communities computed by G-N algorithm and only two nodes in another on Enron while the distribution of nodes computed by ENCD is mode even. So community partition of our algorithm on Enron is more natural and stable.

Table II
The detailed results computed by G-N algorithm on Enron

Community ID	Nodes in the community
Community 1	146,37,80,71,96,90,41,127,112,122,61,44,8,101,145,117,128,56,26,1,139,148,40,17,24,59,125,77,104,100,62,107,140,38,10,126,91,118,103,108,105,106,120,81,73,25,150,111,58,69,129,49,92,54,83,78,84,47,34,110,114,151,51,95,39,113,124,22,88,45,46,64,109,89,63,36,123,119,121,82,42,60,5
Community 2	2,3,4,18,19,20,28,29,30,32,55,66,68,72,74,137,141
Community 3	6,7,9,11,12,13,14,16,23,27,48,50,52,57,65,67,75,76,98,136,142,147
Community 4	33
Community 5	79
Community 6	15,85,86,87,93,97,99,115,130,131,132,133,134,135,138,143,149
Community 7	31,35
Community 8	43

Table III
The detailed results computed by ENCD algorithm on Enron

Community ID	Nodes in the community
Community 1	2,3,4,6,9,13,16,18,19,20,23,27,28, 29,30,32,44,48,49,50,52,55,57,65 66,67,68,69,70,72,74,91,102,111 136,137,139,140,141
Community 2	10,17,21,25,26,36,37,58,75,77, 80 90,101,112,118,125,127,142
Community 3	85,86,87,97,99,115,130,131,133,134 135,149
Community 4	24,79,83,88,103,105,107,109,114,117 119,123,126,151
Community 5	7,11,12,33,38,76,98,147
Community 6	5,14,15,22,51,73,81,89,108,121,138, 143
Community 7	54,78,84,92,100,122,129
Community 8	31,34,35,39,43,45,82,94, 113,124
Community 9	1,8,40,41,42,46,47,56,59,60,61,62,63 64,71,93,95,96,104,106,110,120,128 132,145,146,148,150

Fig 11 describes the community density of Enron computed by ENCD and G-N algorithm. Fig 12 is a description of the community effective diameter comparison on Enron by the two algorithms.

From the table I, II, III, we can see that the number of communities computed by ENCD on Enron email corpus is close to that of G-N, while nodes belonged to communities computed by G-N are not distributed average and even exists only one node in 2 communities and 2 nodes in another community which is conflicted with rule1, but results of ENCD are relatively average. Fig 5 shows that the lowest community density computed by G-N is 5 and the highest is 20, while the lowest and highest of ENCD is 10 and 25 respectively. so community partition of our algorithm ENCD on Enron are more dense than that of G-N. Fig 6 describes that community effective diameter computed by the two algorithms are quite nearly. So the algorithm proposed in this paper is feasible and effective in community partition.

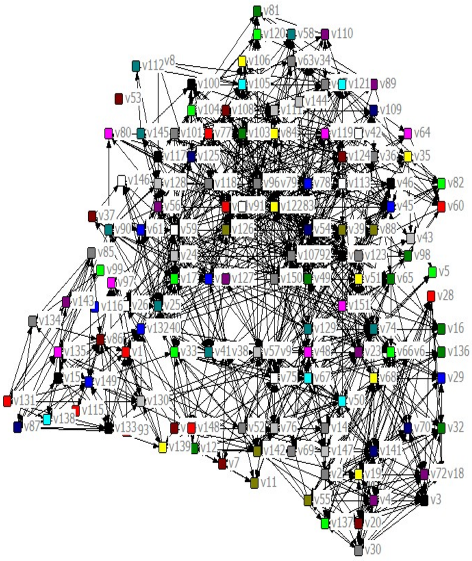

Fig 7 Visualization of the whole *Enron* Email graph

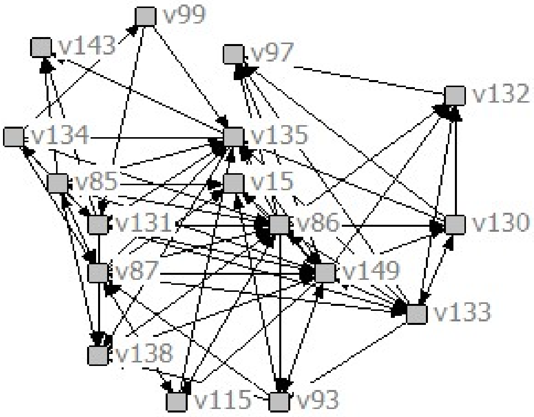

Fig 8 Visualization of the community 6 computed by G-N on Enron

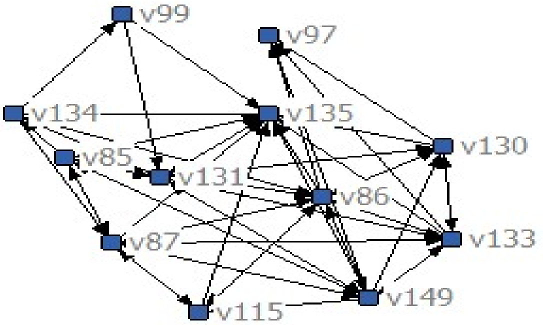

Fig 9 Visualization of the community 3 computed by ENCD on Enron

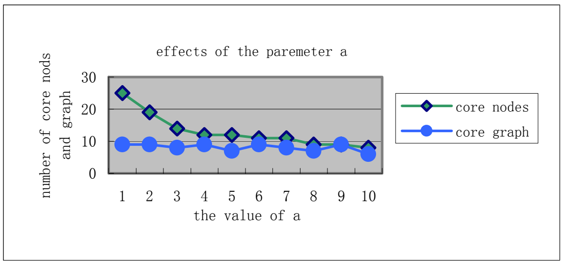

Fig 10 Effects of the parameter ∂

© 2010 ACADEMY PUBLISHER

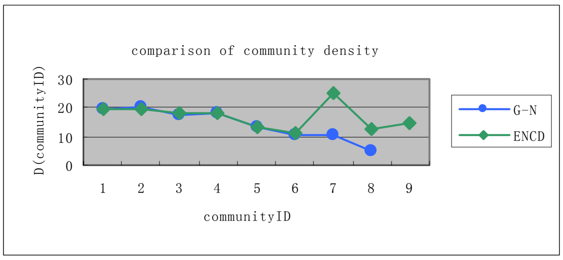

Fig 11 Comparison of community density on Enron

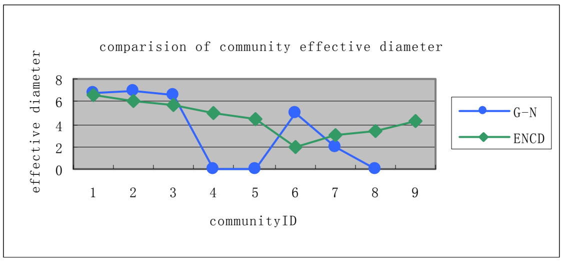

Fig 12 Comparison of community effective diameter on Enron

VI. Analysis and Evaluation

The essence of E-mail network community partition is the clustering of a sparse graph, while this kind of division is an NP-complete problem[12]. The G-N algorithm introduced by Girvan and Newman has achieved very good results on community partition, but the high time complexity of $o(E^2V)$ is hard to apply to the large scale network community finding. While the time complexity of our algorithm in the fist step is only O(|E|+|V|log|V|)+O(|V|)+O(| E_c |)because the use of adjacency list and inverse adjacency list. E_c is the number of edges in the core graphs. Time complexity of the second step is O(E). The total time is dominated by O(|E| + |V| log|V|) of step 1 which is executed only once for all settings of parameters ∂.

VII. Conclusion

In this paper, we have described a new class of algorithms for partition the E-mail network community based on clustering a directed and weighted graph. First identify the satisfactory core nodes by calculating the density variation sequence, then partition the core nodes into core graph, finally put the undivided non-core nodes into the corresponding sub-graph by computing the similarity of the communication behavior. Experiment on Enron corpus and e-mail log information of Soochow university shows that our algorithm ENCD is quite equivalent to the G-N algorithm in the quality of community partition while the execution efficiency is higher than G-N. In addition, ENCD also support the situation that a node belonging to multiple communities and this is extremely common in our real life. The future work is ready to study and discuss the topic and coring people of each specific community.

References

[1] Zhang Yan-Chun.Yu Xj,Hou.Ling-Yu.Web communities: Analysis and construction[M]. Berlin: Springer,2005:56-92.

[2] Steven H. Strogatz. Exploring complex networks[J]. Nature,410:268-276,2001.

[3] Tyler J R,Wilkinson D M, Huberman B A. Email as spectroscopy: Automated discovery of community structure within organizations[c]. In HuysmanM, Wenger E,Wulf V.(eds.)Proceedings of the first international conference on communities and technologies, Kluwer,Dordrecht(2003)

[4] Ding, C. H. Q.; He, X.; Zha, H.; Gu, M.; and Simon, H. D. 2001. A min-max cut algorithm for graph partitioning and data clustering.

[5] Donetti L, Mu˜noz MA (2004) Detecting network communities: a new systematic and efficient algorithm. Journal of Statistical Mechanics: Theory and Experiment, P10012

[6] Capocci A, Servedio VDP, Caldarelli G, Colaiori F (2004) Detecting communities in large networks. Physica A, Vol 352, No 2-4, pp 669-676

[7] Palla G, Der'enyi I, Farkas I, Vicsek T (2005) Uncovering the overlapping community structure of complex networks in nature and society. Nature, Vol 435, pp 814-818[23]

[8] M. E. J. Newman, “Analysis of weighted networks.” Phys. Rev. E 70, 056131 (2004)..

[9] T. Ito, T. Chiba, R. Ozawa, M. Yoshida, M. Hattori,and Y. Sakaki, A comprehensive two-hybrid analysis toexplore the yeast proteininteractome. Proc. Natl. Acad.Sci. USA 98, 4569–4574 (2001).

[10] F. Wu and B. A. Huberman, Finding communities in linear time: A physics approach. Eur. Phys. J. B 38, 331–338 (2004).

[11] S.Wasserman, K. Faust. Social Network Analysis[M]. Cambridge University Press,Cambridge,1994

[12] Marshall van Alstyne and Jun Zhang. EmailNet: A System for Automatically Mining Social Networks from Organizational Email Communication. In NAACSOS2003, 2003.

[13] Mark E.J.Newman. Finding and evaluating community structure in networks[J]. Physical Review E,69. 026113,2004.

[14] Scott J. Social Network Analysis: A handbook[M]. Sage,London,2nd edition,2000

[15] Donetti L, Munoz M A. Detecting Network Communities: a new systematic and efficient algorithm. cond -mat/0404652(2004)

[16] Girvan, M., & Newman, M. (2002) “Community structure in social and biological networks”, Proc. Natl. Acad. Sci. USA 99, 8271-8276.

[17] van Alstyne, M., and Zhang, J. 2003. Emailnet: A system for automatically mining social networks from organizational email communication. In NAACSOS2003.

[18] Lodhi, H.; Shawe-Taylor, J.; Cristianini, N.; and Watkins, C. Neural Information Processing Systems (NIPS), 563–569

[19] B. W. Kernighan and S. Lin, An efficient heuristic proce-dure for partitioning graphs. Bell System Technical Journal 49, 291–307 (1970).

[20] A. Clauset, M. E. J. Newman and C. Moore, “Finding community structure in very large networks.” Phys. Rev.E 70, 066111 (2004)