Big Data Community Detection using Girvan-Newman Algorithm

**on**

DNC\_EMAILS

GROUP 9

LU YANZHANG

HE YAQING

LI JIANI

2016/08/05

# CATALOG

# 

[Big Data Community Detection using Girvan-Newman Algorithm 0](#_Toc458202720)

[CATALOG 1](#_Toc458202721)

[1.Introduction 2](#_Toc458202722)

[2.Problem Statement 2](#_Toc458202723)

[3.Algorithm 3](#_Toc458202724)

[4.Visualization and Analysis 3](#_Toc458202725)

[4.1 Data visualization 5](#_Toc458202726)

[4.2 Analysis 7](#_Toc458202727)

[5.Conclusion 10](#_Toc458202728)

[6.Lessons Learnt 11](#_Toc458202729)

[7.Teamwork in this project 11](#_Toc458202730)

[８.Reference 12](#_Toc458202731)

# 1.Introduction

Friday 22 July 2016 at 10:30am EDT, WikiLeaks releases 19,252 emails and 8,034 attachments from the top of the US Democratic National Committee -- part one of our new Hillary Leaks series. The leaks come from the accounts of seven key figures in the DNC: Communications Director Luis Miranda (10770 emails), National Finance Director Jordon Kaplan (3797 emails), Finance Chief of Staff Scott Comer (3095 emails), Finanace Director of Data & Strategic Initiatives Daniel Parrish (1472 emails), Finance Director Allen Zachary (1611 emails), Senior Advisor Andrew Wright (938 emails) and Northern California Finance Director Robert (Erik) Stowe (751 emails). The emails cover the period from January last year until 25 May this year.

We collect the Clinton's message about 2W from the wiki, through the 2W messages, the relationship between the number of data, we obtained with the e-mail address which are closely associated with Hillary ; through community discovery algorithm for data set analysis and the resulting data is divided into several sub-communities. By the social analysis between the mailboxes, Hillary life social circle can be analyzed, they involve all aspects of daily life; and, We can get unexpected results from the data analysis.

# 2.Problem Statement

As for data, we get information though a crawler. The information including Email address, Name, the number of send and receive. We make data visualization by the GN algorithm so that know the relationship between data obviously.

And in the forensic analysis of the mass e-mail, due to the number is too large, the proportion of meaningless data increase, information extraction and classification of e-mail data becomes particularly important. Information extraction means automatic search to find the metadata preset type information, such as sender address, recipient address, IP address, to send and receive time, e-mail address or domain name contained in the e-mail data information. For this study, we extract the relevant information of the sender address, recipient address, we will analyze from several aspects.

# 3.Algorithm

The Girvan–Newman algorithm detects communities by progressively removing edges from the original network. The connected components of the remaining network are the communities. Instead of trying to construct a measure that tells us which edges are the most central to communities, the Girvan–Newman algorithm focuses on edges that are most likely "between" communities.

The algorithm's steps for community detection are summarized below

1.Calculate  betweenness of all edges

2.Remove the edge(s) with highest betweenness

3.Repeat steps 1 and 2 until graph is partitioned into as many regions as desired

Edge betweenness of edge e:(full version) Total amount of“ﬂow” an edge e carries between all pairs of nodes where a single unit of ﬂow between two nodes divides itself evenly among all shortest paths between the nodes (1/k units ﬂow along each of k shortest paths)

# 4.Visualization and Analysis

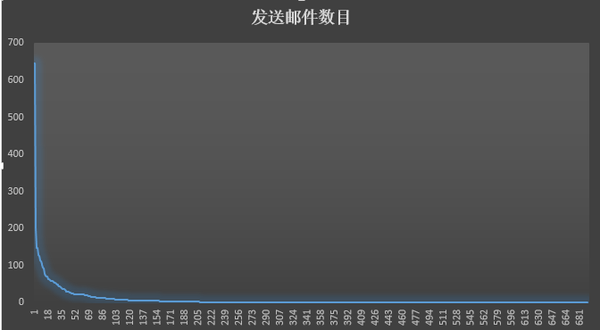
Data analysis

As you see, when we just focus on send, top 3 are

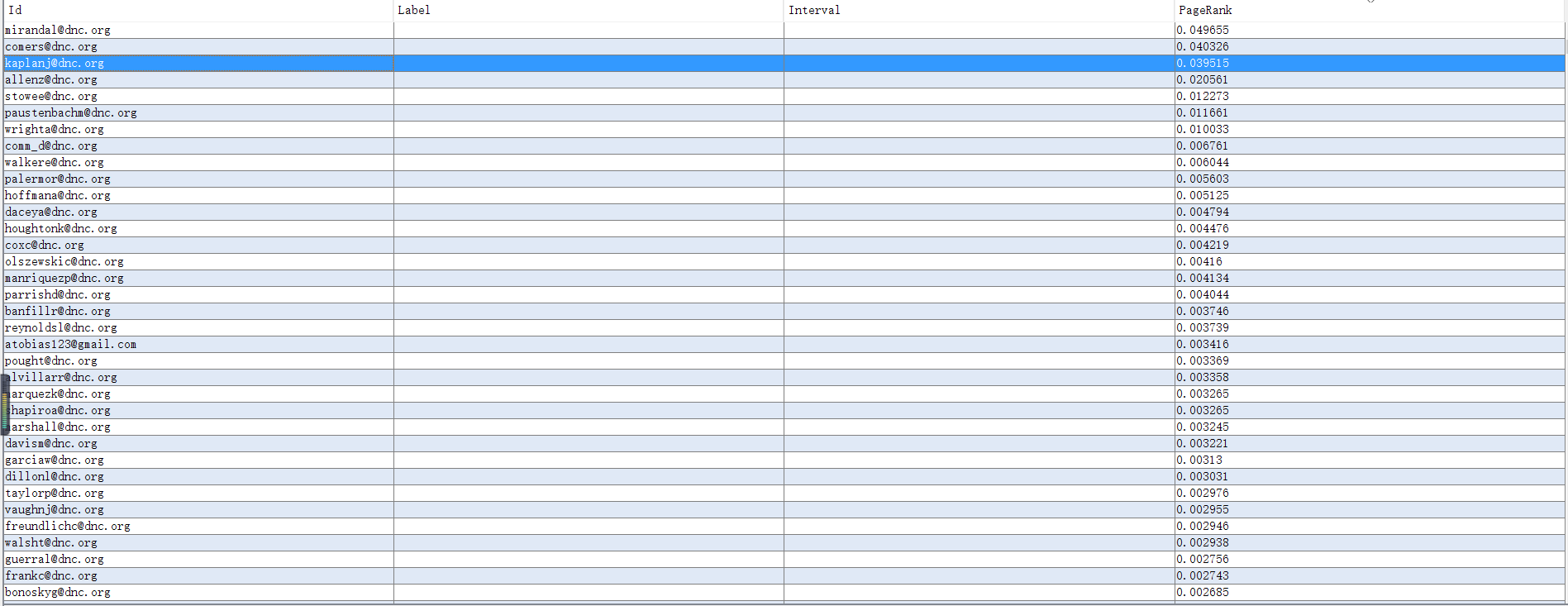
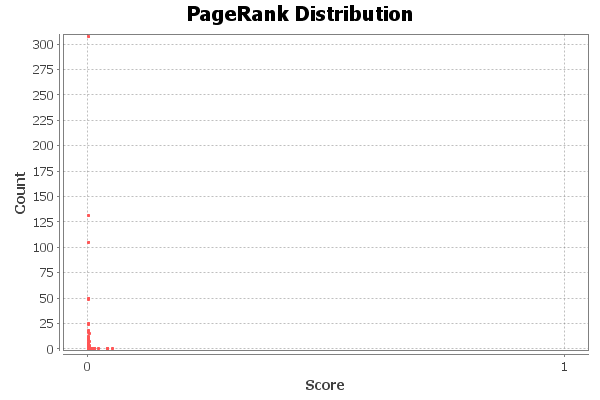
[kaplanj@dnc.org,mirandal@dnc.org](mailto:kaplanj@dnc.org,mirandal@dnc.org) and dncpress@dnc.org().



Add more spice of visualization into datasheet. As we wished, a [inverse proportional function](http://www.baidu.com/link?url=ke7-67jT6JoeoKZOBYKv0RPzqMSRyCmYZq6Nz7HeorWbVB0uQ78O0LzsKrmUSNdHkwaFFCff8y6qXyFDWk4mjIfhTPITb_AyABP0uTlNYHYDtprmgd2zRXT4Lk0F-if0osxUMAKjy1zGOQUif6H7Uq).

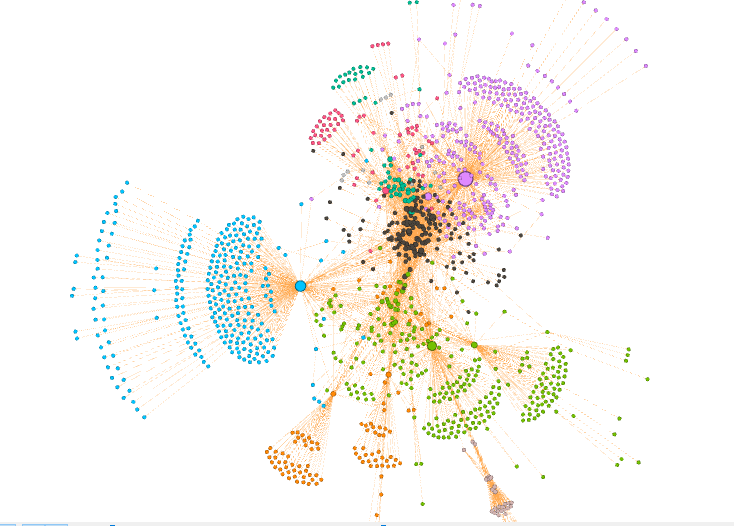


When we use PagePank algorithm,as the graph above shown, top three are mirandal@dnc.org, comers@dnc.org, [kaplanj@dnc.org.It](mailto:kaplanj@dnc.org.It) makes sense like the important one always say less.



## 4.1 Data visualization

Based on data visualization, the graph showed by Gephi:



By using the community discovery algorithm on the basis of the second step, the entire campaign can be divided into several sub-communities.

A kind of color represents a sub-community. Through this we can see, the algorithm obviously more accurate than the naked eye.

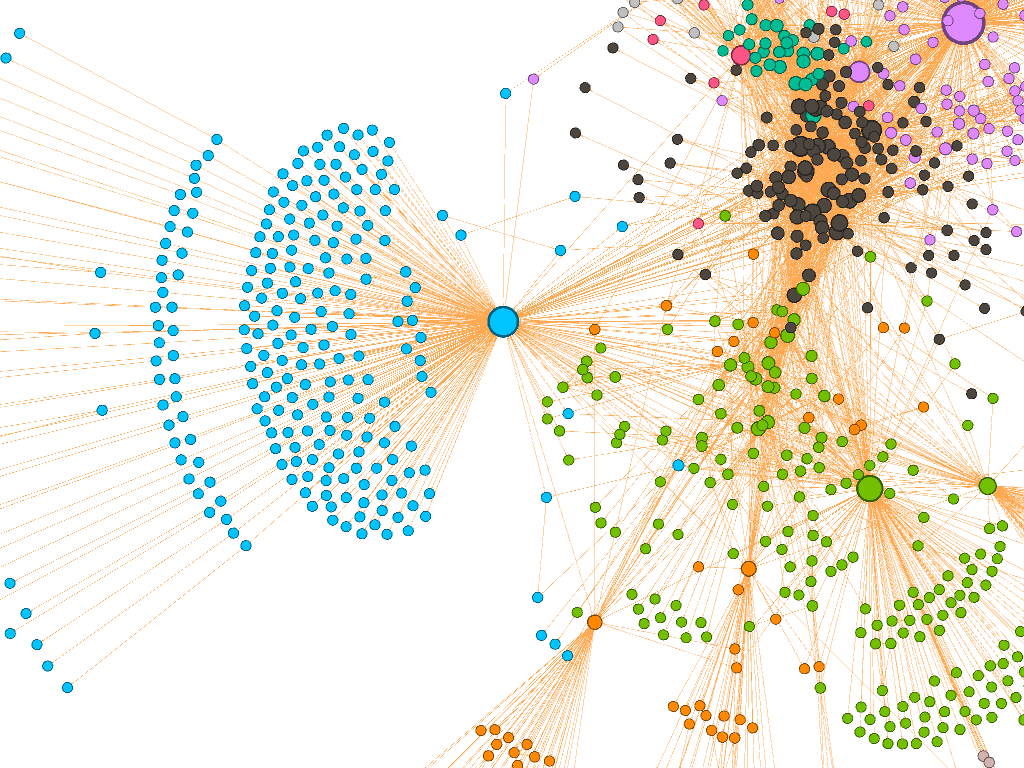
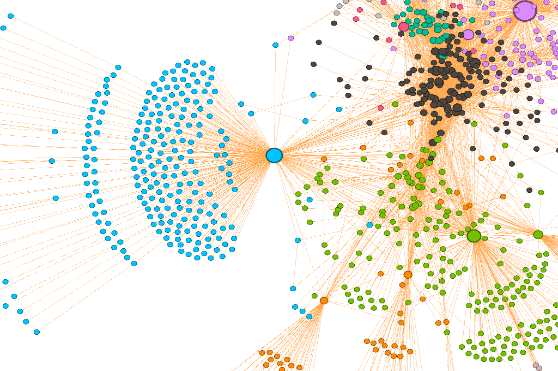
In which, the nodes represent mailbox and the edges represent messages sent between mailboxes. Color depth is related to the number of receiving and responsing e-mail, while the size of the point is ​​associated with the value PageRank. You can see there are a few deep nodes above, which are [comers@dnc.org](mailto:comers@dnc.org) (blue), [kaplanj@dnc.org](mailto:kaplanj@dnc.org) (green) and [mirandal@dnc.org](mailto:mirandal@dnc.org) (purple), and the largest node that is [mirandal@dnc.org](mailto:mirandal@dnc.org).

You can also see from the above campaign team obviously divided into blue, purple and green in three parts

Another point is there are a lot of points arranged in dense for the above figure, like the blue part, purple part and green part.

## Analysis

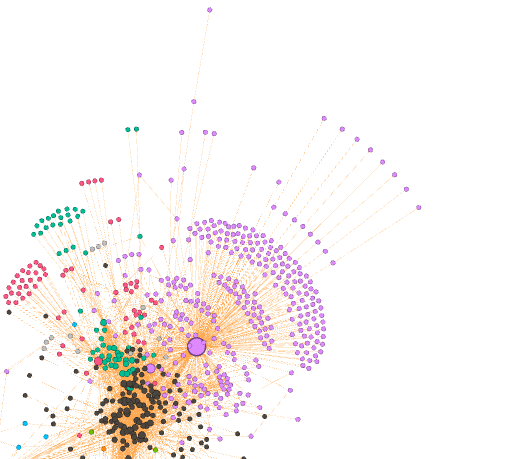
###### 4.2.1 Blue part of graph

The main point is comers@dnc.org, Finance Chief of Staff Scott Comer. The graph shows that Scott Comer is responsible for maintaining the campaign finance activity. Many of these are external mailboxs, and usually those mailboxes rowed together with only a single contact mailbox. Although most mailboxes were just a one communication, but it can also be seen that each person is responsible for what part of the work. This portion of the mailboxes are more complex, including a number of software companies services, strategy consulting companies, financial services companies, clothing companies. In general, campaign team is to maintain

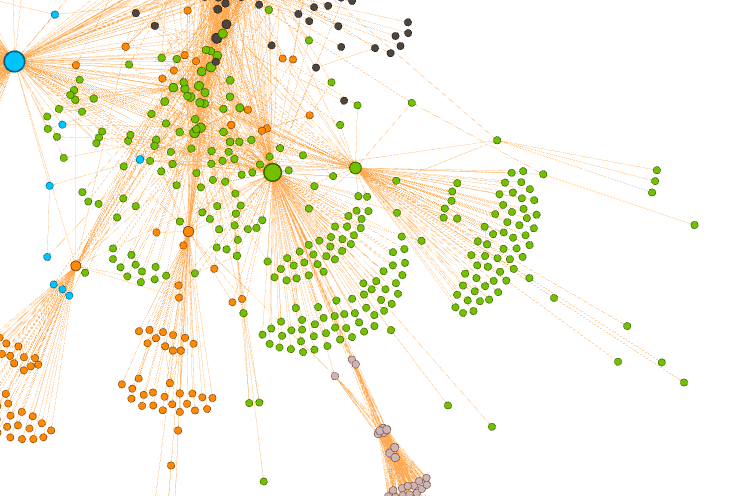
Normal operation, mainly contact by Scott Comer.

###### 4.2.2 Purple part of graph:

In this graph, the main point is [mirandal@dnc.org(Id)](mailto:mirandal@dnc.org(Id)), Luis Miranda (Name) . This part includes media, such as network media and traditional media，and LGBT website，Teacher Unions and so on. The main contact is Communications Director Luis Miranda.

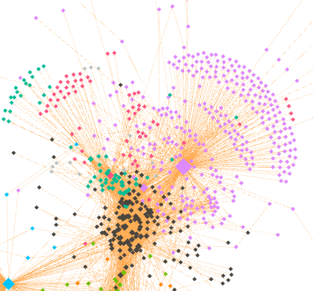
Because of this event, Luis Miranda announced his resignation after a period of time.

###### 4.2.3 Green part of graph:

 The main point is [kaplanj@dnc.org(Id)](mailto:kaplanj@dnc.org(Id)), National Finance Director Jordan Kaplan(Name).

This graph shows that Jordan Kaplan is responsible for maintaining the campaign normal activity. It follows that he contacted some company about service software, tactics counsel, financial and garment.

###### 4.2.4 Red part of graph

The main point is [manriquezp@dnc.org](mailto:manriquezp@dnc.org), in the previous diagram, there is a point we have not speaking, which is the middle part of that lump of red spots.

It can be seen very closely the internal communication frequently between red spots, but little contact with the outside world only by a few people contact with Kaplan J. Very mysterious look.

I checked some of the information, because I am not familiar with it, it is probably only some information.

It can be seen, most of these people belong to an advisory body, it appears to be devoted to the proposed idea.

The famous character of the green part are:

KaplanJ@dnc.org,CoxC@dnc.org,etc.

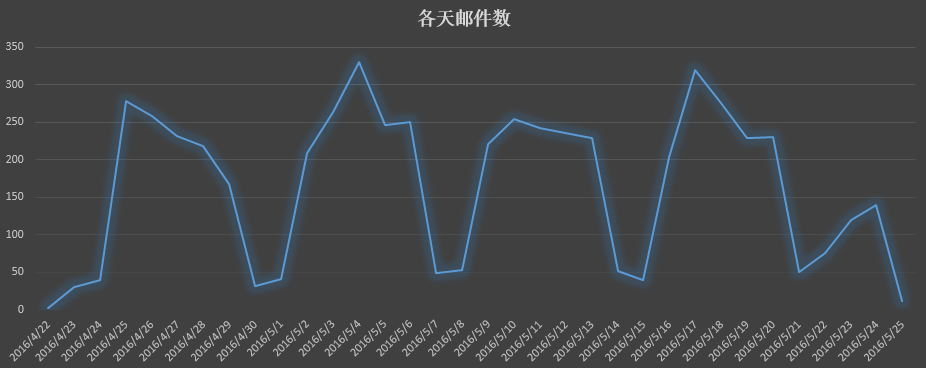
In purple section,the most obvious is [mirandal@dnc.org(Id)](mailto:mirandal@dnc.org(Id)),

Orange part, the most important is weis@dnc.org

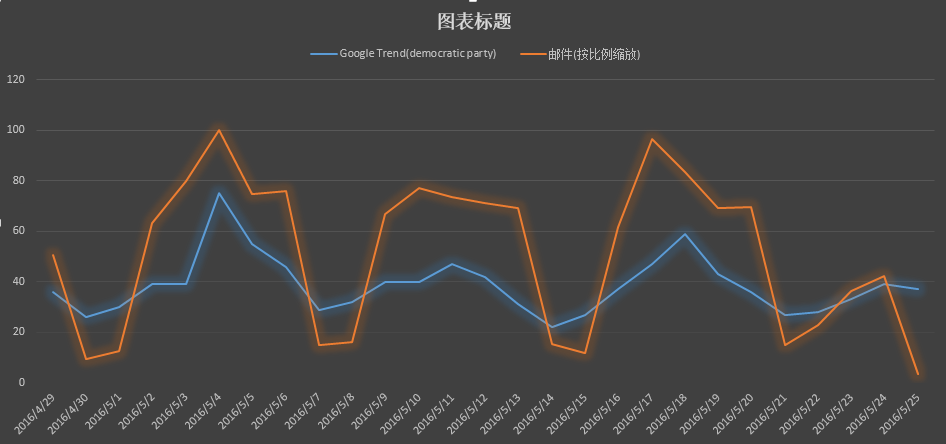
There is no particularly conspicuous figure in red part. But from the location perspective, it is estimated that the ordinary people of the campaign team, mainly in charge of daily affairs and analysis.

###### 4.2.5 The relation between the number and date of the mail

Most of the mail is start from April 20 of this year. First, drawing the number of messages daily changes

  
It can be seen from April 20 to May 25 month, the number of messages showed significant period of change. This is due to the weekend. Almost all of the trough occurred in Saturday and Sunday. It seems they perform a good weekend in their system, and overtime work is not obvious.

After reading the trough, and then look at a few dates relatively high number of messages: 5/4, 5/10, 5/17. Obviously, these days there are certainly happened big events. We went GoogleTrend found search index of the democratic party, and found that are highly consistent with the number of messages.



In which, orange is on behalf of the number of messages, blue represents the search index. To set the two sets of data on a map, I have been proportionally reduced the number of messages.

# 5.Conclusion

1. Hillary email Communications Director Luis Miranda frequently since April

of this year.  
2. Every fluctuation means that a big event happened according to charts.  
3. The analysis of the colored part is useful for finding the person who played an important role and what his main responsibilities for the communication.

4. the Scott Comer is responsible for maintaining the finance activity. And Luis Miranda is responsible for media such as network media and traditional media，and LGBT website，Teacher Unions and so on. They contact theCommunications Director Luis Miranda frequently. Because of this event, Luis Miranda announced his resignation after a period of time.

5.National Finance Director Jordan Kaplan is responsible for maintaining the campaign normal activity. It follows that he contacted some company about service software, tactics counsel, financial and garment. Hillary’s message is part of the normal exchange of information

6.There are relations between the number and date of the mail.When high number of emails represents these days there are certainly happened big events. We found search index of the democratic party, and found that are highly consistent with the number of messages.

7. Analysis of algorithms help to make predictions. Community detection helps to know the truth.

8 By analyzing the communities detected can develop or confirm relationships between the data, the entities and the groups, it can derive new insights from community relationships.The Pattern is about groups of people and the individuals in relation to others. Big data are threaded with connections.

# 6.Lessons Learnt

Though we came across some problems, but our team try our best to solve the problem, and know more about the usage of Community Detection.  
Analyzing the communities detected can develop or confirm relationships between the data, the entities and the groups, it can derive new insights from community relationships.From the project, we can learn more information from the Big Data and turn into competitive advantage. The Pattern is about groups of people and the individuals in relation to others. We learn that big data are threaded with connections.  
At the beginning, we have collected a lot of information and finally found a good point about the Hillary’s email. By doing this project,we are familiar with Python.we learn that the community algorithms help us to understand scalability and performance often draws the line between what is feasible and what is impossible. We understand that the algorithmic mathematics help us to find the truth of the events.  
Analysis of algorithms help us make predictions. Community detection  helps us know the truth

# 7.Teamwork in this project

LU YANZHANG is responsible for the code and presentation.

HE YAQING is responsible for the report of PPT

LI JIANI is responsible for part of the report and PPT.，the report of analysis

Problem:

1 We encountered difficulties when determine the subject at first.

Our team cooperate with each other and search for google to gather information.

2 In the process of data Extraction and data visualization

Through our efforts,as for data, we get information though a crawler. The information including Email address, Name, the number of send and receive. We make data visualization by the GN algorithm so that know the relationship between data obviously.

# ８.Reference

[1] Kernighan B W and Lin S.An efficient heuristic procedure for partitioning graphs[J].Bell system Technical Journal,1970,49(2):291一307

[2] Karypis G and Kumar V.A Fast and High Quality Multilevel Scheme for Partitioning Irregular Graphs[J].SIAM Journal on Scientific Computing,1998,20:359

[3] Flake G W, Lawrence S and Giles C L. Efficient identification of web communities [C]. In Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 2000:160.

[4]    Flake G W, Lawrence S and Giles C L, et al. Self-organization and identification of web communities [J]. Computer, 2002, 35(3):66-70.

[5]    Girvan M and Newman M E J. Community structure in social and biological networks [J] . Proceedings of the National Academy of Sciences of the United States of America, 2002, 99(12):782.

[6]    Newman M E J and Girvan M. Finding and evaluating community structure in networks [J]. Physical review E, 2004, 69(2): 26113.

[7]     Donath W E and Hoffman A J. Lower bounds for the partitioning of graphs [J]. IBM Journal of Research and Development, 1973, 17(5):420-425.

[8]     Pothen A, Simon H D and Liou K. Partitioning sparse matrices with eigenvectors of graphs [J]. SLAM Journal on Metrix Analysis and Applications, 1990, 11(3).

[9]     Newman M E J. Fast algorithm for detecting community structure in networks [J]. Physical review E, 2004, 69(6):066133.

[10]   Clauset A, Newman M E J and Moore C. Finding community structure in very large networks [J]. Physical Review E, 2004, 70(6):6611.

[11]   Bron C and Kerbosch J. Algorithm 457: finding all cliques of an undirected graph [J]. Communications of the ACM, 1973, 16(9):575-577.

[12]   Makino K and Uno T. New algorithms for enumerating all maximal cliques [C]. Algorithm Theory-SWAT, 2004:260-272.

[13]   Liu G and Wong L. Effective pruning techniques for mining quasi-cliques [C]. Machine Learning and Knowledge Discovery in Databases, 2008:33-49.

[14]   Batagel J V and Zaversnik M. An O(m) algorithm for cores decomposition of networks [C]. CoRR (Computing Research Repository), 2003.

[15]   Blondel V D, Guillaume J L and Lambiotte R, et al. Fast unfolding of communities in large networks [J]. Journal of Statistical Mechanics : Theory and Experiment, 2008:PI0008.

[16]   Bu D, Zhao Y, Cai L, et al. Topological structure analysis of the protein-protein interaction network in budding yeast [J]. Nucleic Acids Research, 2003, 31(9):2443.

[17]   Wang N, Parthasarathy S, Tan K L, et al. CSV: visualizing and mining cohesive subgraphs [C]. In Proceedings of the 2008 ACM SIGMOD international conference on  Management of data, ACM, 2008:445-458.

[18]   Charikar M. Greedy approximation algorithms for finding dense components in a graph [C]. Approximation Algorithms for Combinatorial Optimization, 2000:139-152.

[19]   Andersen R and Peres Y. Finding Sparse cuts locally using evolving sets [C]. In Proceedings of the 41st annual ACM symposium on Theory of computing, ACM, 2009:235-244.

[20]   Andersen R. A Local Algorithm for Finding Dense Sub-graphs [J] . ACM Transactions on Algorithms(TALG), ACM, 2010, 6:1-12.