

Empirical Network Analysis

[Qiao Zhang](#)

Sep 9th, 2012

Obtaining Data

Last semester, my friends and I created a questionnaire application called “Who knows me best” on Renren.com (the copycat version of Facebook in China), which allows users to generate their own questionnaire and invite their friends to answer. A user can create only one questionnaire while can answer several. (This app is available [here](#), but only in Chinese) After a few months we have accumulated some data, and I always wanted to construct a graph of the users, hoping to find something interesting.

The user interface looked like this:





The database structure underlying this application is:

```

202.112.87.20 - PuTTY

mysql> show tables;
+-----+
| Tables_in_questionnaire |
+-----+
| answer                  |
| question                |
| ranking                 |
| user                    |
+-----+
4 rows in set (0.00 sec)


mysql> show columns from user;
+-----+-----+-----+-----+-----+-----+
| Field | Type          | Null | Key | Default | Extra          |
+-----+-----+-----+-----+-----+-----+
| qid   | int(10) unsigned | NO   | PRI | NULL    | auto_increment |
| uid   | text          | YES  |     | NULL    |                |
| uname | text          | YES  |     | NULL    |                |
| qname | text          | YES  |     | NULL    |                |
| datetime | datetime      | YES  |     | NULL    |                |
+-----+-----+-----+-----+-----+-----+
5 rows in set (0.00 sec)

mysql> show columns from ranking;
+-----+-----+-----+-----+-----+-----+
| Field | Type          | Null | Key | Default | Extra          |
+-----+-----+-----+-----+-----+-----+
| uid   | int(10) unsigned | NO   | PRI | 0        |                |
| qid   | int(10) unsigned | NO   | PRI | 0        |                |
| correctrate | int(11)       | YES  |     | NULL    |                |
| datetime | datetime      | YES  |     | NULL    |                |
+-----+-----+-----+-----+-----+-----+
4 rows in set (0.00 sec)

mysql>

```

Because basically I only want information about the surveyers and respondents, I used the following SQL query to obtain the author IDs and the respondent IDs (For privacy reasons, I did not fetch users' real names), which are the nodes of the graph. Therefore the edges of the graph would be the author-answerer relationship.



The screenshot shows a PuTTY terminal window titled "202.112.87.20 - PuTTY". The terminal displays a MySQL command and its output:

```
mysql> select ranking.uid as respondent, user.uid as author from ranking, user where ranking.qid=user.qid into outfile '/tmp/questionnaire.csv' fields terminated by ',' enclosed by '"' lines terminated by '\r\n';
Query OK, 63 rows affected (0.00 sec)

mysql>
```

Data Analysis

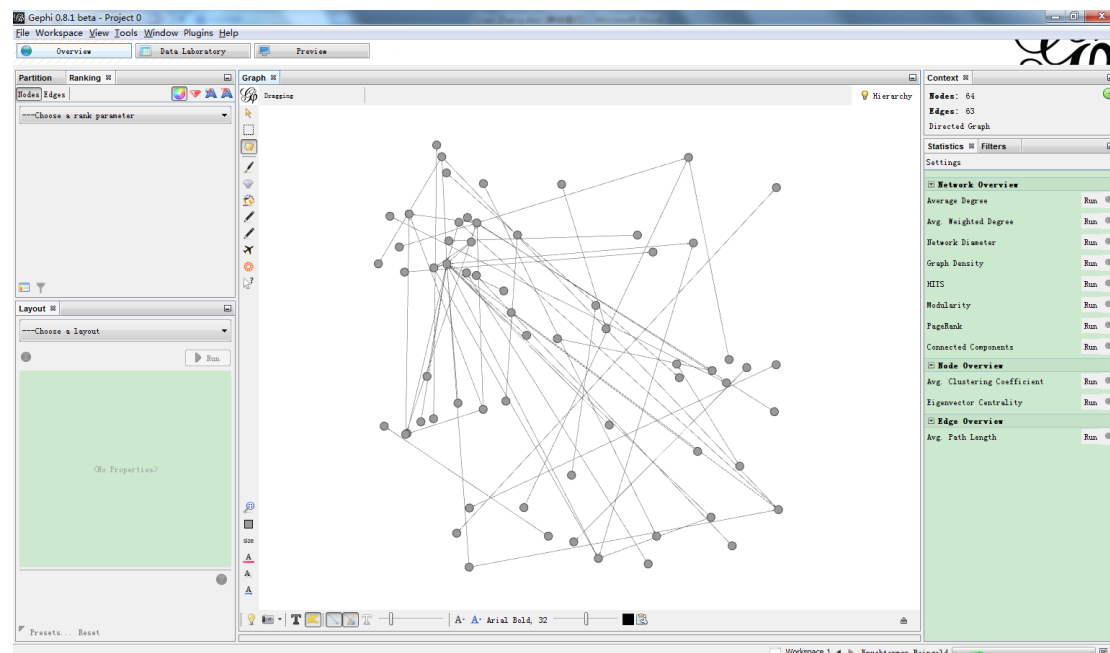
The dataset:

The screenshot shows the Microsoft Excel interface with the 'graph.csv' file open. The ribbon is set to the '开始' (Home) tab. The active cell is A1, which contains the formula '=229061005'. The worksheet displays a table with 14 rows and 11 columns (A-J). The data in the table is as follows:

	A	B	C	D	E	F	G	H	I	J
1	229061005	234450143								
2	229381125	234450143								
3	234450143	300718007								
4	234450143	285200199								
5	234450143	272302131								
6	234450143	316252582								
7	240822298	234450143								
8	240822298	272302131								
9	254146241	354062303								
10	255835870	234450143								
11	265962806	234450143								
12	272302131	234450143								
13	272302131	316252582								
14	275100166	234450143								

The status bar at the bottom indicates the file name 'graph', the zoom level '100%', and the status '就绪' (Ready).

When the file is loaded in Gephi, the graph looked like this:



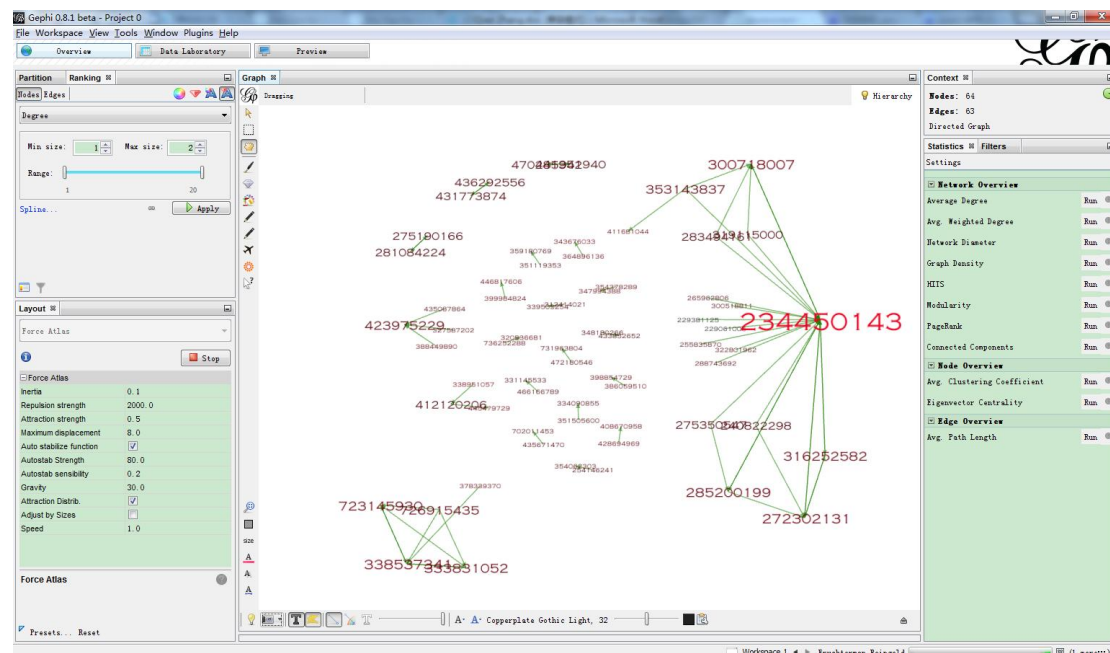
Although a bit chaotic, there is a node that seemed to have more degrees than the others, which is, me. This is because I actively called for my friends to try this app out, since I have labored over the development process.

To give the graph more clarity, I applied Fruchterman Reingold layout algorithm, a coloring algorithm (according to degree) and several other adjustments:



It turns out that 234450143 (me) is indeed the most important node.

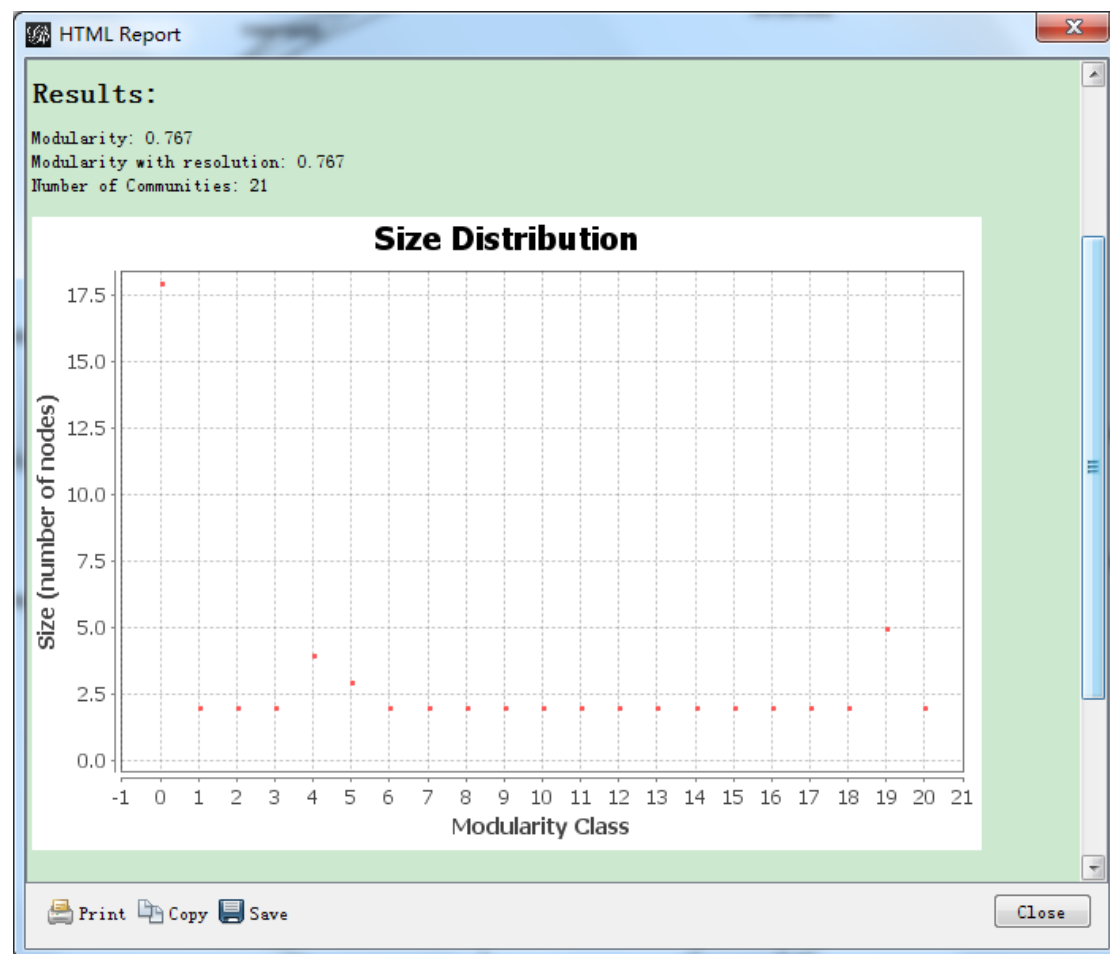
In order to see the aggregations of the network, I applied Force Atlas layout algorithm:



Then I see two clusters containing relatively more nodes, of which one is me, the other cluster might be another group of friends who had used our app. To my puzzlement, in the center shows up a number of nodes in pairs. After observing their profiles, I suddenly realized that these pairs are actually close friends and couples. Interestingly, if a pair of nodes are a couple of lovers, the answerer is always a boy. It is probably because our app's name is "Who knows me best", users are more likely to invite their best friends or girl(boy)friends rather than just random friends. After calculating average path length (treat as undirected), the average path length is about 1.8, and number of shortest paths is 378, which might not be so meaningful because the graph is not completely connected. However, it indicates that this app is indeed used by intimate group of friends, rather than average friends, because the average path length is so short that is less than 2 hops. (the following graph ranked nodes by betweenness centrality)



To take a further step on finding clusters, after running the modularity algorithm, the result shows:

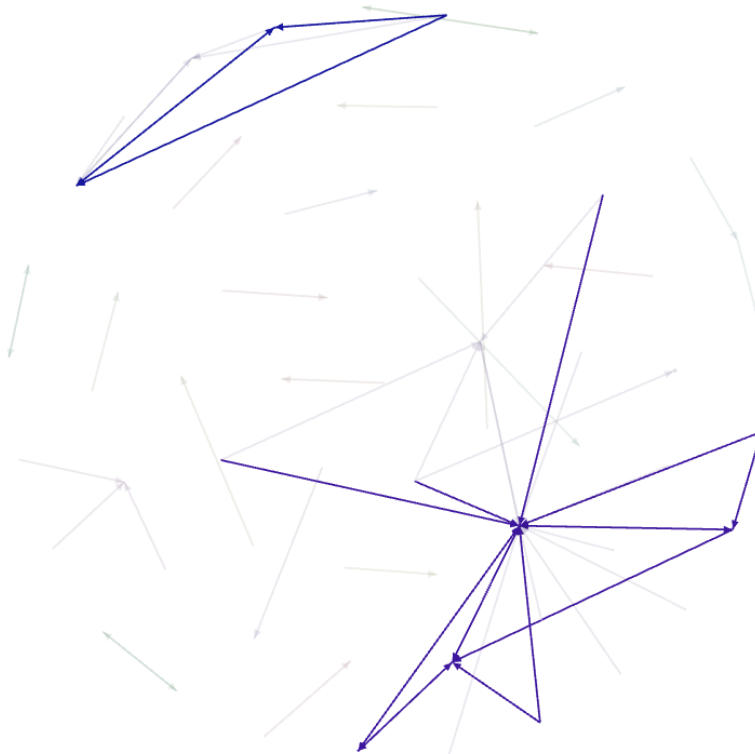


Of all modularity classes, seventeen classes consist of only 2 nodes, and one of 18 nodes. Which means the dataset might not be so good because the size distribution looks flat.

After grouping the communities:



Because I want to find out which users are more active, I filtered the nodes by its out degree, selected those who has more than 2 out degrees, and the result shows that out of 64 users, only 12 had answered more than one questionnaire.



The graph is so sparse that I want to know how sparse it is. So I calculated the graph density, which is a measurement of how close the network is to complete. A complete graph has all possible edges and density equals to 1, and mine is 0.026, which is a pityfully sparse graph. It is because one will not invite average friends to answer his/her questionnaire, nor to answer questionnaires of them. So the edges are sparse, compared to the size of nodes.

The limitation of the analysis is that the dataset is too small, because we have not yet collected enough data. Later on when there are more users and questionnaires, the analysis itself and conclusions drawn may be more persuasive.

Dataset

229061005	234450143
229381125	234450143
234450143	300718007
234450143	285200199
234450143	272302131
234450143	316252582
240822298	234450143
240822298	272302131
254146241	354062303
255835870	234450143
265962806	234450143
272302131	234450143
272302131	316252582
275190166	281084224
275350547	234450143
275350547	285200199
281084224	275190166
283494161	234450143
283494161	300718007
285200199	234450143
285200199	272302131
288743692	234450143
300518811	234450143
300718007	234450143
312414021	339503254
316252582	234450143
316252582	272302131
319115000	234450143
319115000	300718007
322801962	234450143
327587202	423975229
333831052	338537341
333831052	723145930
338537341	723145930
338537341	333831052
338951057	412120206
348180266	433852652
351119353	359180769
351505600	334090855
353143837	411681044
353143837	234450143
353143837	300718007

354378289	347994388
364896136	343676033
378339370	338537341
386059510	398854729
388449890	423975229
399984824	446817606
412120206	445479729
428694969	408670958
431773874	436292556
435087864	423975229
435671470	702011453
436292556	431773874
441351940	470285942
466166789	331145533
470285942	441351940
472180546	731963804
723145930	338537341
726915435	338537341
726915435	723145930
726915435	333831052
736252288	320836681