A Strategy to Build Connection with a Stranger Based on Social Network Topology

Shaojun Zhang

December 6, 2015

Motivation

As we know, people are connected in various social networks. An interesting question is whether we can improve our chance of making friends with some particular stranger in the network? Here the stranger can be our potential spouse, colleague, advisor or boss.

One straightforward solution is to link one by one along the shortest path to the stranger. However, the problem becomes complicated when the shortest path is still long or when there exists more than one shortest path, especially in a dense network.

Background

Since there is no ground truth or related social experiments, it is difficult to evaluate different approaches. Therefore, there is few research on this problem. I choose it as my topic for the network class project because of my personal interest.

Background

Yen's algorithm

Yen's algorithm computes single-source K-shortest loopless paths for a graph with non-negative edge cost. It employs any shortest path algorithm to find the best path, then proceeds to find $K-1$ deviations of the best path.

Yen's algorithm is known to be the efficient and widely used algorithm for determining K-shortest loopless paths. Its time complexity at worst case is $O(KN_v(N_e+N_v\log N_v))$.

Background

Louvain method

The Louvain method for community detection is a method to extract communities from large networks. It is today one of the most widely used method for detecting communities in large networks.

The advantage of Louvain method lies in its time complexity, which is $O(N_v^2)$ for a general graph and $O(N_v)$ for a sparse graph. However, it is a greedy algorithm and the solution depends on the order of which node or community is selected.

Background

label propagation algorithm

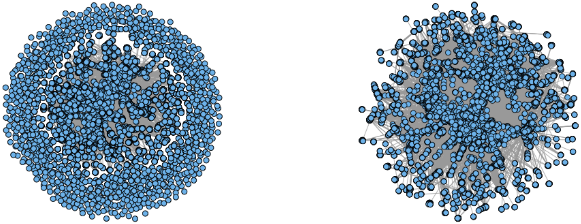
The key idea of label propagation algorithm is that each node finds a neighboring community of largest size to join until the system reaches equilibrium. Its time complexity is $O(N_v)$.

In comparison with other algorithms, label propagation has advantage in its running time, amount of a priori information needed about the network structure (no parameter is required to be known beforehand) while its disadvantage is that it produces no unique solution, but an aggregate of many solutions.

Data

The Enron email dataset was collected and prepared by the CALO project. It contains data from about 150 users, mostly senior management of Enron organized into folders. This data was originally made public, and posted to the web, by the Federal Energy Regulatory Commission during its investigation. This data is valuable because it is the only substantial collection of “real” email that is public.

Data



(a) whole graph

(b) giant component

Figure 1:The network of the Enron email communication.

Data

Table 1:Statistics for the giant component

Nodes	33696
Edges	180811
Diameter	13
Average degree	10.73
Average path length	4.02
Average clustering coefficient	0.71

Data

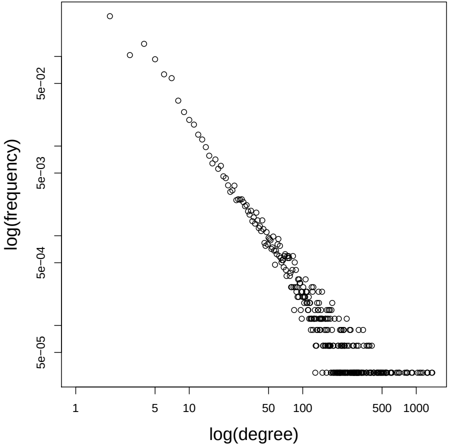


Figure 2:Degree distribution.

Data

Table 2: Top 10 vertices for three centralities

Rank	Closeness	Betweenness	Eigenvector
1	137	5025	137
2	77	141	196
3	47	567	77
4	141	589	371
5	371	1140	1029
6	293	274	274
7	196	459	735
8	735	47	417
9	176	1029	176
10	417	293	293

Assumptions

- Friendships are based on edges in the graph and each user can only make new friends by being introduced.
- Users in the same community have a closer relationship than those from different communities.
- The more importance a user has, the more likely his or her friends will be introduced successfully.
- Each user can only make at most 3 new friends excluding his or her direct friends.
- The network is fixed.

Method

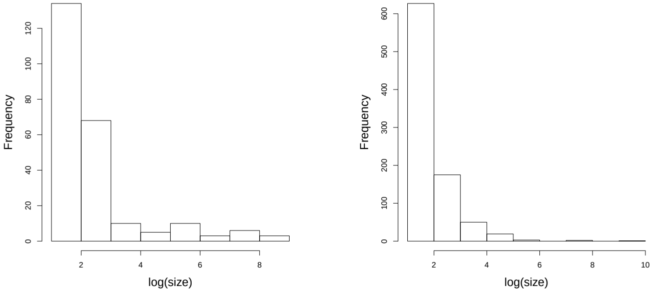
- Find all shortest paths between the user and the stranger using Yen's algorithm. All shortest paths are truncated if their length $d > 4$.
- Detect communities using Louvain method or label propagation algorithm. Select the paths with highest community score.

$$\begin{aligned} C_1 &= \text{number of pairs of neighboring users on the path} \\ &\quad \text{that belong to the same community} \\ C_2 &= \text{number of users on the path that belong to the} \\ &\quad \text{community containing the stranger} \end{aligned}$$

- Select the path with highest importance score.

$$I = \text{ total eigenvector centrality scores on the path}$$

Results



(a) Louvain method

(b) label propagation

Figure 3:Distribution of community sizes.

Results

Table 3:Results when Louvain method is used for community detection

From	To	SP#	C_1	C_2	Success
17085	9609	7	17085 → 1029 → 2287 → 2295 → 9609	17085 → 274 → 77 → 343 → 9609	✓
12885	27890	5	12885 → 274 → 10 →8543 → 27890	12885 → 274 → 10 → 8543 → 27890	✓
32349	20592	4	32349 → 14051 → 1140 → 77 → 1589	32349 → 14051 → 1140 → 274 → 1589	✗
1764	12172	7	1764 → 96 → 12172	1764 → 96 → 12172	✓
11412	2063	5	11412 → 233 → 482 → 2063	11412 → 233 → 482 → 2063	✗
8270	25406	3	8270 → 75 → 57 → 10 → 5007	8270 → 75 → 57 → 10 → 5007	✗
7156	11071	4	7156 → 196 → 4026 →11071	7156 → 196 → 442 →11071	✓
19878	19567	4	19878 → 1769 → 152 → 4397 → 19567	19878 → 1769 → 152 → 4397 → 19567	✓
24041	16112	4	24041 → 287 → 424 → 226 → 16112	24041 → 287 → 79 → 226 → 16112	✓
12405	7156	6	12405 → 459 → 196 → 7156	12405 → 459 → 196 → 7156	✓

Results

Table 4:Results when label propagation is used for community detection

From	To	SP#	C_1	C_2	Success
17085	20387	5	17085 → 1029 → 28 → 5020 → 20387	17085 → 1029 → 28 → 5020 → 20387	✓
12885	31330	7	12885 → 274 → 10 → 5574 → 31330	12885 → 274 → 10 → 5574 → 31330	✓
32349	20591	4	32349 → 14051 → 1140 → 77 → 1589	32349 → 14051 → 1140 → 77 → 1589	✗
1764	22177	5	1764 → 214 → 823 → 22177	1764 → 293 → 3126 → 22177	✓
11412	4662	5	11412 → 233 → 79 → 102 → 4662	11412 → 233 → 79 → 102 → 4662	✓
8270	29942	4	8270 → 75 → 567 → 5025 → 29942	8270 → 75 → 567 → 5025 → 29942	✓
7156	20152	5	7156 → 1140 → 7978 → 20152	7156 → 1140 → 7978 → 20152	✓
19878	27221	4	19878 → 1769 → 1341 → 4151 → 27221	19878 → 1769 → 1341 → 4151 → 27221	✓
24041	23720	4	24041 → 287 → 648 → 3310 → 23720	24041 → 287 → 517 → 3310 → 23720	✓
12405	17544	5	12405 → 459 → 77 → 274 → 17544	12405 459 77 274 17544	✓

Conclusion

The strategy in this report combines community structure and eigenvector centrality to deal with multiple shortest paths. Two versions of community score and one importance score are defined as criteria for path selection. The assumptions made are reasonable and the method works well for the Enron email dataset.

The computer program for this project is written in R with igraph package. All computations for the report, including plotting, cost about 22.5 minutes on my laptop (2.5 GHz Intel Core i5, 4 GB 1600 MHz DDR3).

Discussion

- The upper bound for the number of users allowed to link is chosen subjectively. I choose 3 here based on my personal experience.
- Friendship in real life is much more complicated. Sometimes people are not willing to share their friends for various reasons. Moreover, friendship may change over time. A friend today may not be a friend tomorrow. Therefore, an interesting problem is how to find the best path in a dynamic network. Another interesting problem is how to integrate multiple social networks, for example, Facebook and Twitter.