ST5225 Statistical Analysis of Network Project

Chong Woon Kiat A0209349X

In this report, we study the network of blogs about the 2004 US presidential election with data set from Adamic and Glace (2005). The edges were extracted from a crawl of the front page of the blog. The nodes are labelled as 0 for liberal and 1 for conservative. Our aim is to study the degree of the graph, the ranking of pages, the different sampling type on the graph, as well as the detection of community in the graph.

1. Introduction

The network is described in Figure 1. It is a directed graph and not a simple graph. The network has 1,490 nodes and 19,090 edges.

It is not a connected graph with 268 components, one with 1222 node, one with 2 nodes and the rest of 266 components are isolated nodes. It is unweighted multi-graph, consisting of 3 self-loop edges, 65 multiple edges. Each node has information on the website URL, the political stance and the source.
We are interested to study the simplified graph with self-loop and multi-edges removed. The term graph refers to the simplified graph from now on. The simplified graph has 1,490 nodes and 19,022 nodes.

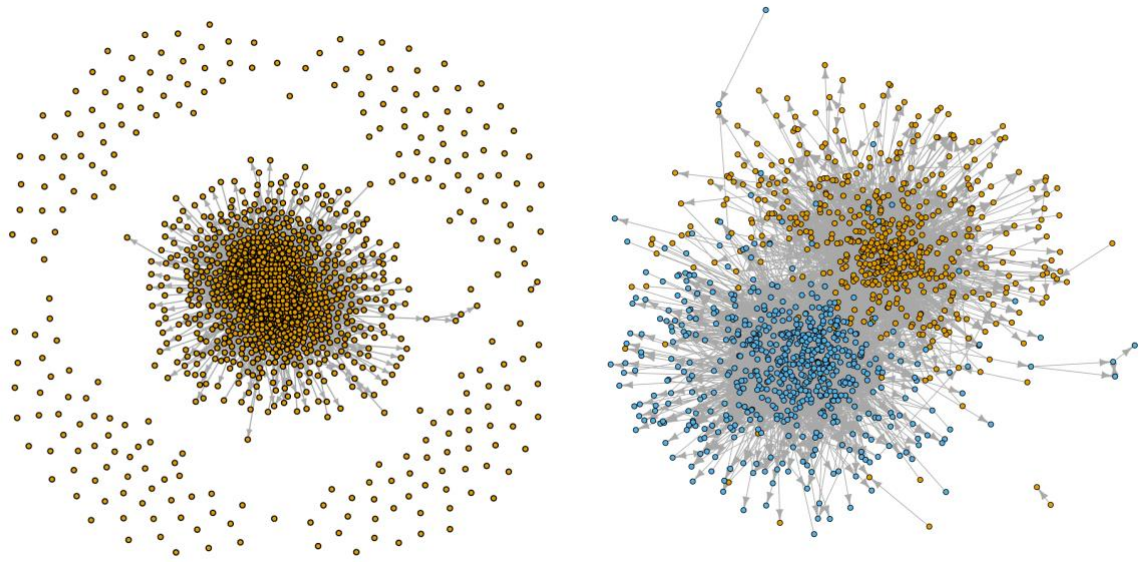

Figure 1: The left graph represents the simplified graph, the right graph represents the simplified graph with at least one in or out degree. Each node represents a blog and each edge represents the fact that one blog refers the other. Blue nodes on the right graph represents the conservative, whereas the orange represents the liberal.

The graph has a diameter of 9 nodes, a mean distance of 3.390, and a density of 0.00857. Since the graph has 19,022 edges, the total degree is 38044. From Figure 2, the degree distribution

of the graph has the shape of power law distribution, hence, we find the corresponding parameter alpha α.

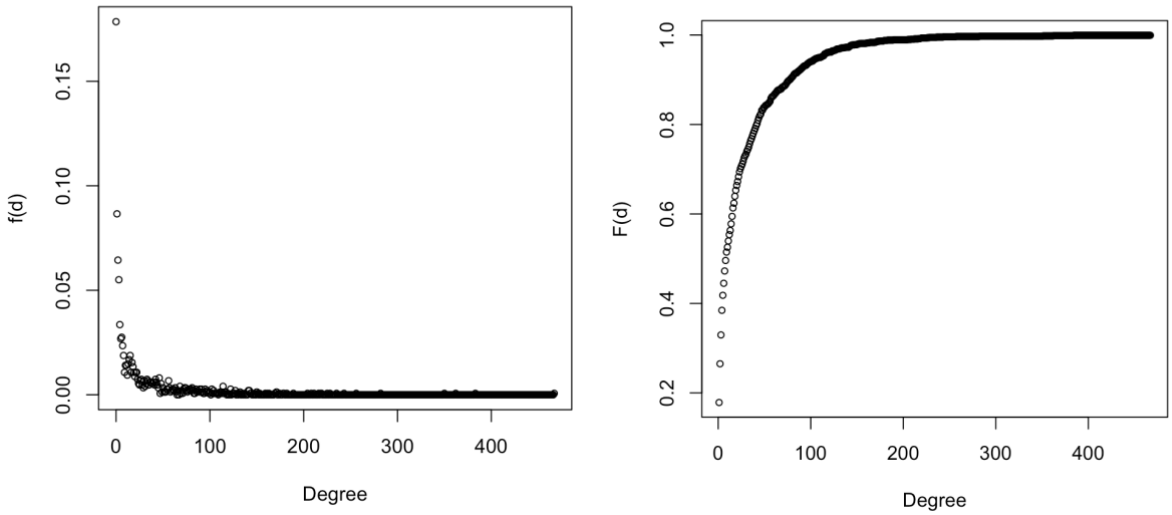

Figure 2: probability density function f(d) of degree (left) and cumulative density function F(d) od degree (right)

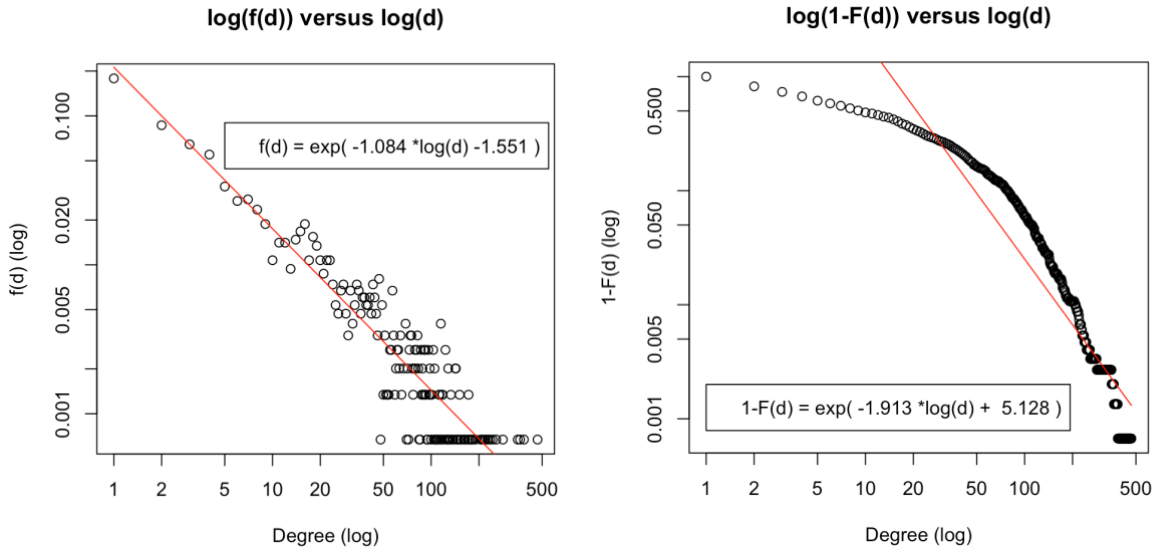

Figure 3: Log scale plot of f(d) versus d (left) and log scale plot of 1-F(d) and d (right)

From figure 2 above, the probability density function f(d) has an inverse correlation with degree d:

$$f(d) \propto \frac{1}{d^{\alpha}}$$

$$\log\big(f(d)\big) \approx -\alpha log(d) + C$$

We can estimate α by fitting a linear regression model for log(f(d)) and log(d) and obtain α = 1.084. However, from Figure 3 below, it is observed that the noise is high at higher degree and the estimation of α might not be accurate.

Another approach to improve α estimation and to reduce the noise is to fit α with cumulative density function F(d)

$$1 - F(d) \propto \frac{1}{d^{\alpha-1}}$$

$$\log(1 - F(d)) \approx -(\alpha - 1)log(d) + C$$

With this method, we obtain α = 2.913. From Figure 3 above, it is observed that the α estimation can be further improved by fitting only the tail probability. We use the maximum likelihood principle to determine α for a given xmin (minimum degree to be included in estimation). Xmin is determined by its optimal value for which the p-value of a Kolmogorov-Smirnov test between the fitted distribution and the original sample is the largest.

It is found that α = 3.933 with xmin = 112. This means that at degree beyond xmin, the probability density decreases at a high rate of $d^{-3.933}$.

We wish to study the top nodes with the highest centrality. From table 1 below, it is observed that in-degree centrality, out-degree centrality, closeness centrality, and betweenness centrality give us different top nodes.
However, nodes with both high in degree and high out degrees tend to have high betweenness centrality as it serves as a good path for other nodes.

Table 1: The top 6 most important nodes with in-degree centrality, out-degree centrality, closeness centrality, and betweenness centrality.

Methods	In-degree	Out-degree	Closeness	Betweenness
1	155	855	293	855
2	1051	454	117	55
3	641	387	217	1051
4	55	512	418	155
5	963	880	357	454
6	1245	363	1057	387

Closeness of node i is defined as $C(i) = \frac{(|V|-1)}{\sum_{j\ connected\ to\ i} dist(i,j)+m\alpha}$ to take care of m nodes unconnected to i, where α is the total number of nodes.

Betweenness of node i is defined as $B(i) = \sum_{(u,v)\in E, u\neq i, v\neq i, u\neq v} \frac{\sigma(u,v|i)}{\sigma(u,v)}$ where $\sigma(u,v)$ is the number of shortest paths between u and v, and $\sigma(u,v|i)$ is the number of shortest paths between u and v that go through i.

Table 1(a): The top 6 most important nodes with in-degree centrality

rank	node	In-degree	value	label
1	155	337	0	dailykos.com
2	1051	276	1	instapundit.com
3	641	268	0	talkingpointsmemo.com
4	55	263	0	atrios.blogspot.com
5	963	238	1	drudgereport.com
6	1245	220	1	powerlineblog.com

Table 1(b): The top 6 most important nodes with out-degree centrality

rank	node	Out-degree	value	label
1	855	256	1	blogsforbush.com
2	454	140	0	newleftblogs.blogspot.com
3	387	131	0	madkane.com/notable.html
4	512	131	0	politicalstrategy.org
5	880	123	1	cayankee.blogs.com
6	363	115	0	liberaloasis.com

Table 1(c): The top 6 most important nodes with closeness centrality

rank	node	Closeness	value	label
1	293	0.0018843	0	itlookslikethis.blogeasy.com
2	117	0.0018823	0	bushmisunderestimated.blogspot.com
3	217	0.0018817	0	etherealgirl.blogspot.com
4	418	0.0018815	0	michaelphillips.blogspot.com
5	357	0.0018812	0	lennonreport.blogspot.com
6	1057	0.0018809	1	isdl.blogspot.com

Table 1(d): The top 6 most important nodes with betweenness centrality

rank	node	Betweenness	value	label
1	855	218464	1	blogsforbush.com
2	55	90986	0	atrios.blogspot.com
3	1051	76270	1	instapundit.com
4	155	54982	0	dailykos.com
5	454	45896	0	newleftblogs.blogspot.com
6	387	45022	0	madkane.com/notable.html

2. Page ranking

Table 2: The top 10 most highest ranking nodes with hub score, authority score, page rank algorithm, and page rank algorithm with damping parameter 0.85.

Methods	Hub score	Authority Score	PageRank	PageRank with damping = 0.85
1	512	155	1159	155
2	387	641	1293	55
3	363	55	155	1051
4	618	729	55	855
5	99	642	1051	641
6	144	323	641	1153
7	56	1051	729	963
8	454	756	1153	729
9	644	493	855	1245
10	55	180	323	798

From table 2, it is observed that most of the nodes with high hub score are the tops nodes with high out-degree in table 1, as hub score measure the reliability of webpage; high out-degree to quality webpages tends to increase hub score. Similarly, most of the nodes with high authority score are the top nodes with high in-degree, as authority score measures how many quality webpages recognized it as important; high in degree from quality webpages tends to increase authority score.

PageRank algorithm suggests that node 1159 and node 1293 are the only most important nodes while the rest of nodes have almost negligible PageRank score. As shown in figure 4 below, this is because these 2 nodes are out-components which absorbs all the scores. Hence the algorithm is modified with scaling to redistributed 15% of the scores to each nodes to avoid scores being trapped in some out-components. With scaled version of PageRank algorithm, we observe that node 1159 and 1293 are removed, and the nodes with high ranking tends to have high in degree (higher probability of users randomly visiting the node).

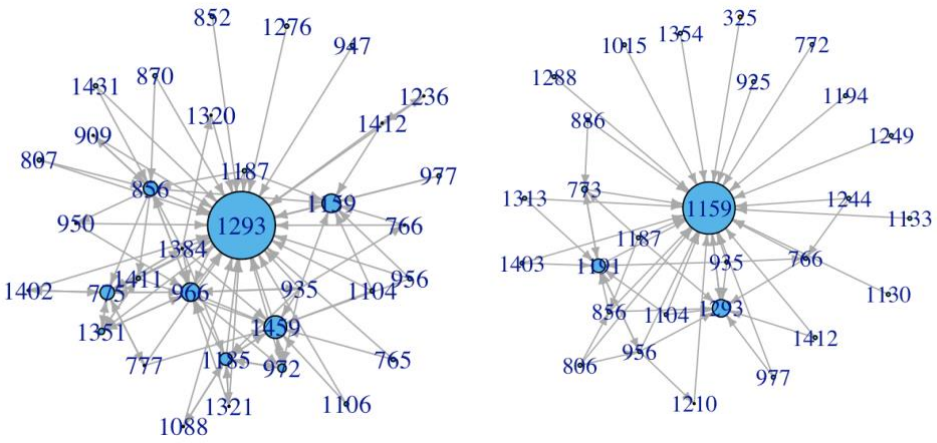

Figure 4: induced subgraphs of node 1293 and its neighbours (left) and induced subgraphs of node 1159 and its neighbour (right). Both nodes absorbs all the scores in PageRank algorithm

3. Undirected graph

In this section, we study the undirected graph of this network by removing the direction of the edges. The undirected graph has 1490 nodes, 16,715 edges (2,307 edges removed due to nodes having bidirectional edges).
The undirected graph has a diameter of 8, mean distance of 2.737, edge density = 0.0151; lower diameter and mean distance and higher edge density as compared to the directed graph.

Since the degree distribution has the shape of power-law distribution. Again, we use the maximum likelihood principle to determine α for a given xmin (minimum degree to be included in estimation), where xmin is determined by its optimal value for which the p-value of a Kolmogorov-Smirnov test between the fitted distribution and the original sample is the largest.

We get α = 3.692 with xmin = 80. This means that at degree beyond 80, the probability density decreases at a rate of $d^{-3.692}$, a slower rate as compared to the directed graph with α = 3.933

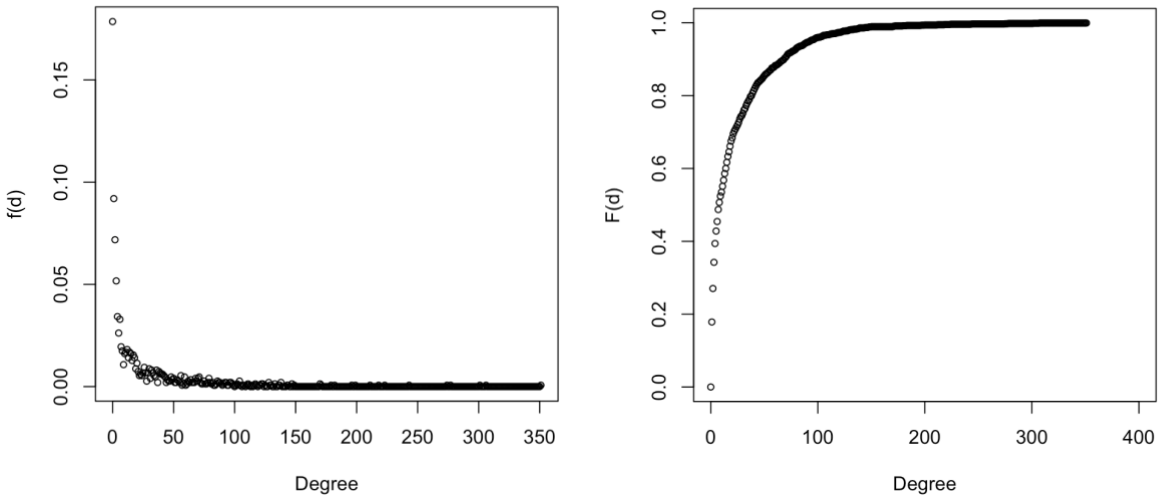

Figure 5: probability density function f(d) of degree (left) and cumulative density function F(d) od degree (right)

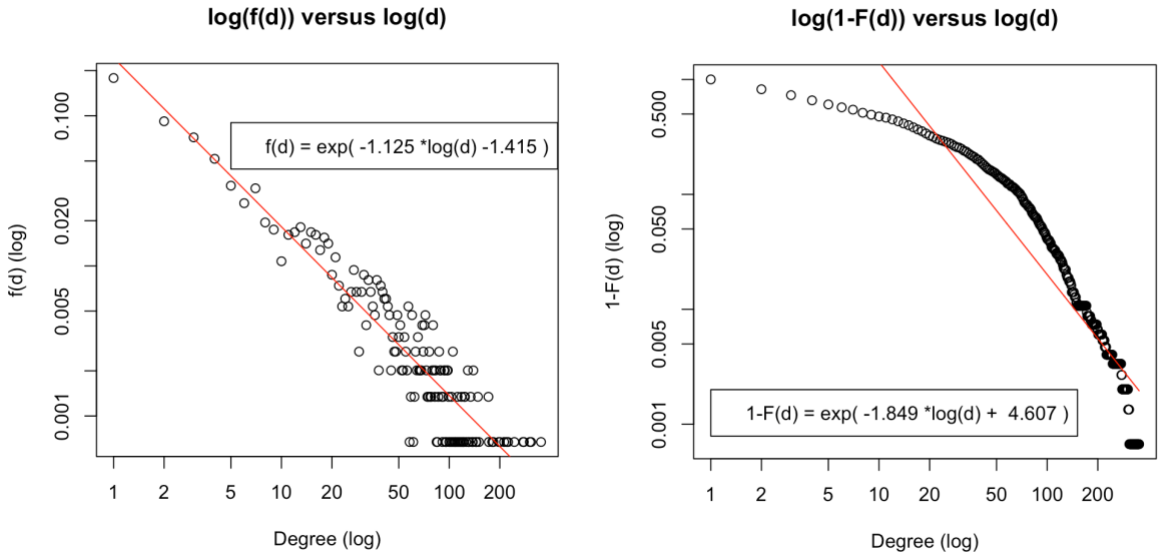

Figure 6: Log scale plot of f(d) versus d (left) and log scale plot of 1-F(d) and d (right)

In this section, we wish to study the components of the undirected graph. The graph is not connected and with 268 components, one component with 1,222 node, one component with 2 nodes and the rest of 266 components are single node.

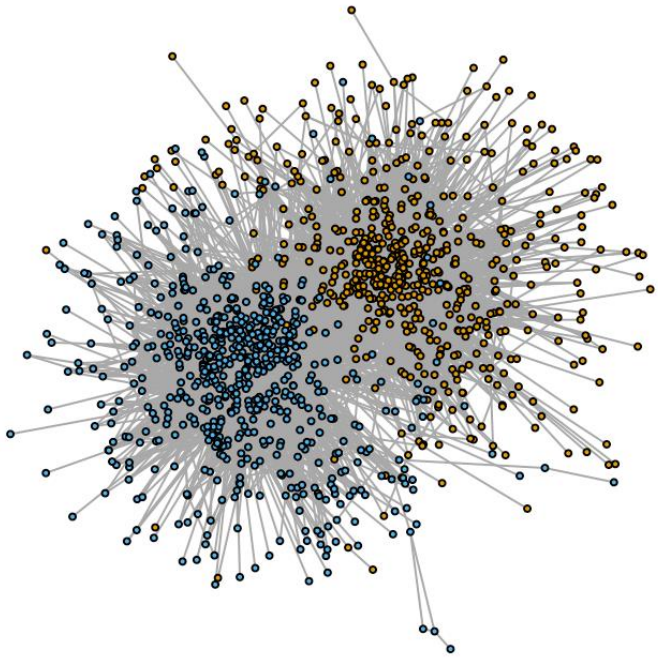

Figure 7: Giant component $\tilde{G}$ with 1,222 nodes and 16,714 edges.

In Figure 8, we show the coreness of each nodes on the left plot and the subgraph of largest k-core, where k = 36 on the right plot. There are 314 nodes in the largest k-core.

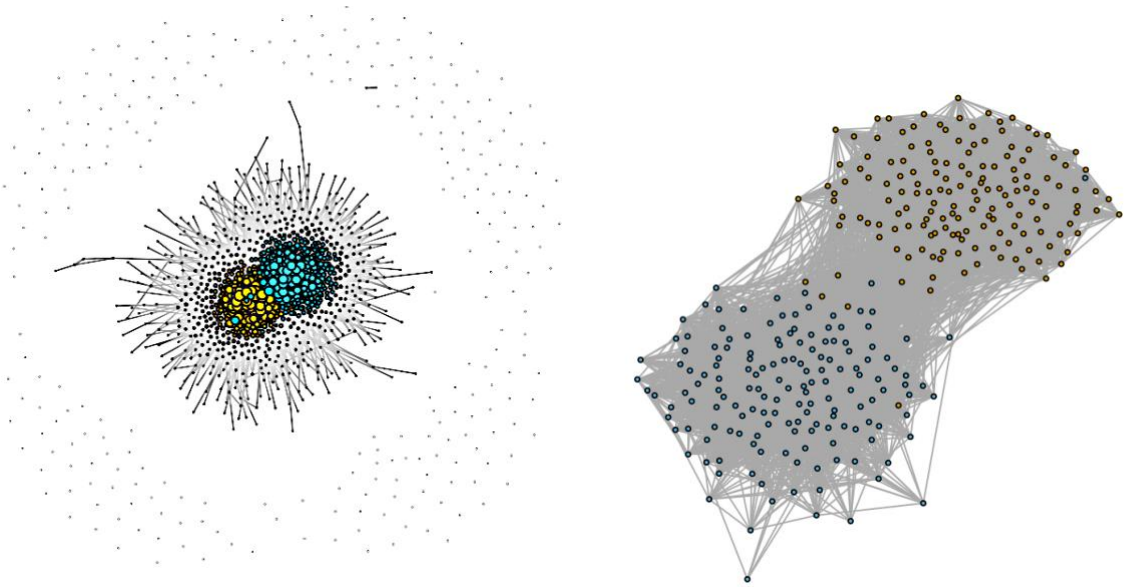

Figure 8: Undirected graph with size representing the coreness of each node (left) and subgraph of maximal k-core where k =36 with 314 nodes (right)

4. Sampling

In this section, we compare the properties of four subgraph of size around 500 nodes sampled using four sampling methods: induced-subgraph sampling, incident-subgraph sampling, snowball sampling, respondent-driven sampling. Five initial seeds are chosen for snowball sampling and respondent-driven sampling.

The results are shown in Figure 9 to 12 respectively. The plots on the left are the entire undirected graph with red nodes representing the sampled nodes while grey representing the unsampled nodes. The induced subgraphs are plotted on the right with orange nodes representing the liberal and blue representing the conservative.

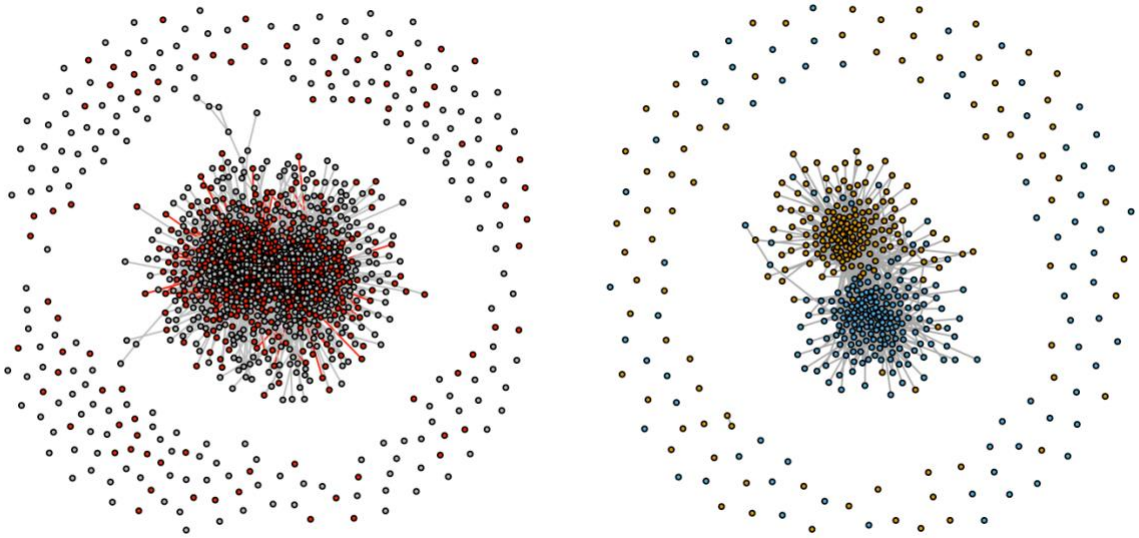

Figure 9: Induced-subgraph sampling

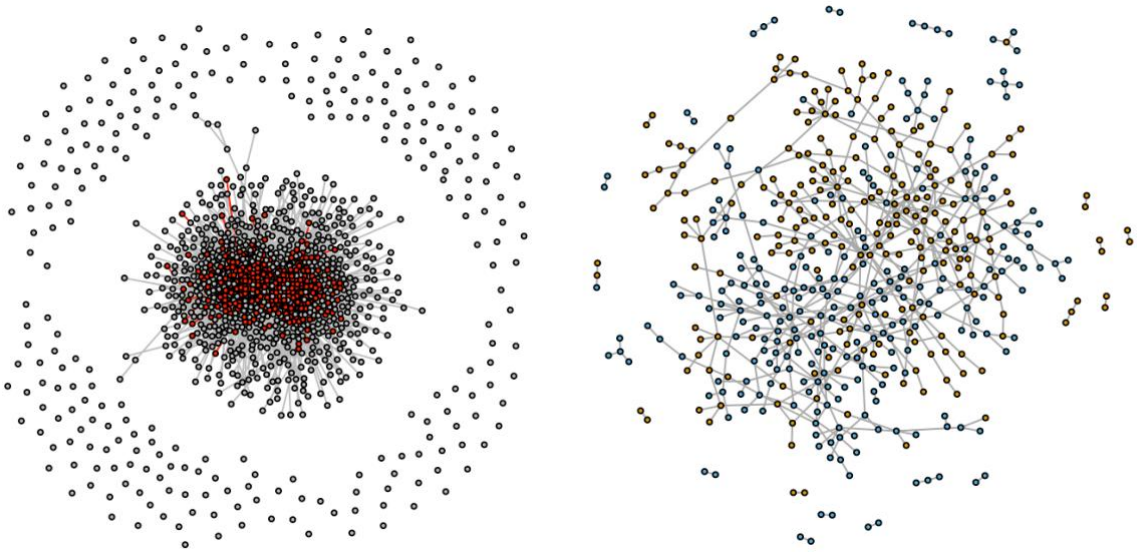

Figure 10: Incident-subgraph sampling

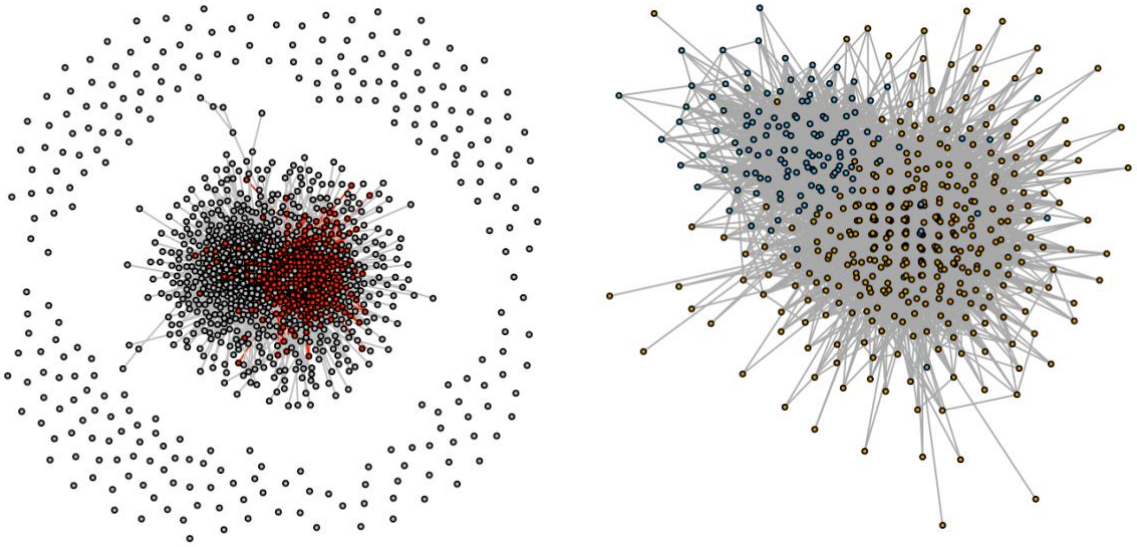

Figure 11: Snowball sampling

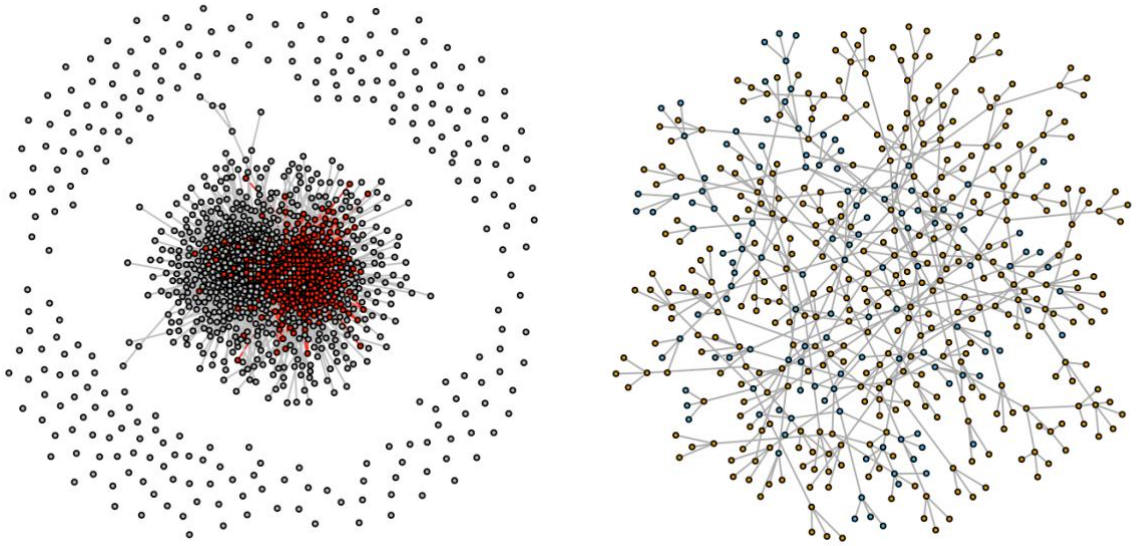

Figure 12: Respondent-driven sampling

Table 3: density, clustering coefficient and connectivity of sampled subgraph

Sampling	Initial seeds	density	clustering coefficient	connected
Original graph	-	0.0151	0.2260	FALSE
Induced-subgraph	-	0.0134	0.2153	FALSE
Incident-subgraph	-	0.0040	0.0023	FALSE
Snowball	5	0.0739	0.3011	TRUE
Respondent-driven	5	0.0040	0	FALSE**

**Respondent-driven subgraph would be a connected graph if there is only one initial seed used for generating the subgraph.

Comparing the properties of the subgraph, it is observed that incident-subgraph gives a much lower density subgraphs. Incident-subgraph method selects a set of random edges and observes the nodes incident to the edges, hence the nodes tend to have lower degree than the nodes originally has, and therefore a lower density. Since the original graph has 16,715 edges while the subgraph has only 521 edges, the probability of selecting an edge that can form a triangle at a connected triple is extremely low. Hence, the clustering coefficient is very low.

Similarly, for respondent-driven method, each node have a maximum of four edges (each respondent passes the token to three other connected nodes), much lower than average degree of 22 from the original graph, and therefore the density is lower. Since respondent does not receive a second token, triangle formation is not possible, and hence the clustering coefficient is zero.

Snowball method generate a subgraph where the nodes are very close to each other and the local density is much higher than the original graph. Therefore, the probability of triangle formation is much higher and hence clustering coefficient is higher than original graph.

Induced-subgraph samples 500 nodes randomly and has edges incident to the selected nodes, therefore, its density and clustering coefficient are closer to the original graph.

Both induced-subgraph and incident-subgraph methods yield an unconnected graph as nodes and edges are sampled randomly. Snowball method tends to give a connected graph when the initial number of seed used is low. Respondent-driven method gives an unconnected graph when the initial number of seed is greater than 1, otherwise it would be a connected graph.

5. Graph Partition

In this section, we try to partition the giant component $\tilde{G}$ into two, in hope to find two clusters that are close to the truth (liberal and conservative). We compare the divisive method (remove edges with highest betweenness), the agglomerative method (hierarchical clustering with single, complete and average linkage) and the fast greedy modularity method and the k-means method based on stochastic block model. We compare their error rate defined as

$$Error\ rate\ (\hat{l}_i) = min\left\{\frac{1}{|V|}\sum_{i\in V} 1\left\{l_i \neq \hat{l}_i\right\}, 1 - \frac{1}{|V|}\sum_{i\in V} 1\left\{l_i \neq \hat{l}_i\right\}\right\}$$

We use blue and orange to denote the two clusters obtained in following figures.

The two components obtained from removing the edges with highest betweenness give a very high error rate of 48.3%. One of the component has only 4 nodes (plotted as blue nodes in Figure 13).

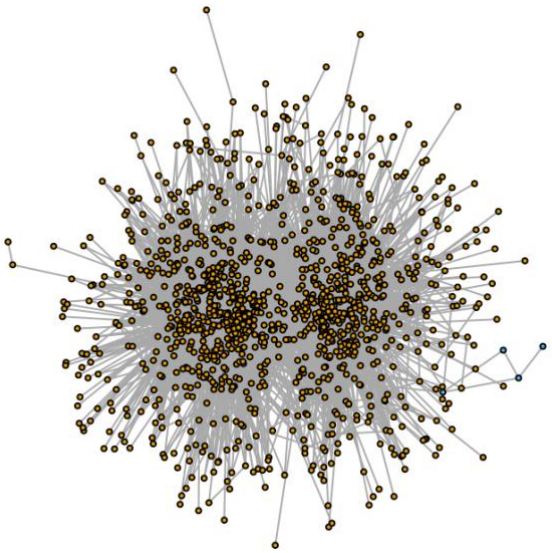

Figure 13: Graph partition obtained from removing the edges with highest betweenness

Table 4: Truth table obtained from removing the edges with highest betweenness

edge betweenness	0	1
0	586	632
1	0	4

The three hierarchical clustering method with single, complete and average linkage give high error rate too at 48.0%, 41.8%, 48.0% respectively. Single linkage suffers from chaining and hence when we cut the tree into two components, one of the components has only one node. This is because in order to merge two groups, only need one pair of nodes to be close, irrespective of all others. Complete linkage avoids chaining and has better performance but still fails to have a good split. Average linkage tries to strike a balance between single and complete linkage, however it still suffer from chaining on this dataset.

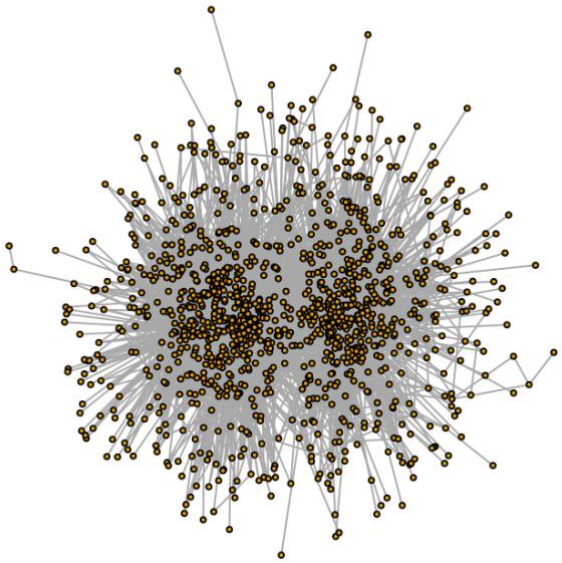

Figure 14: Graph partition obtained from hierarchical clustering with single linkage

Table 5: Truth table obtained from hierarchical clustering with single linkage

single	0	1
0	586	635
1	0	1

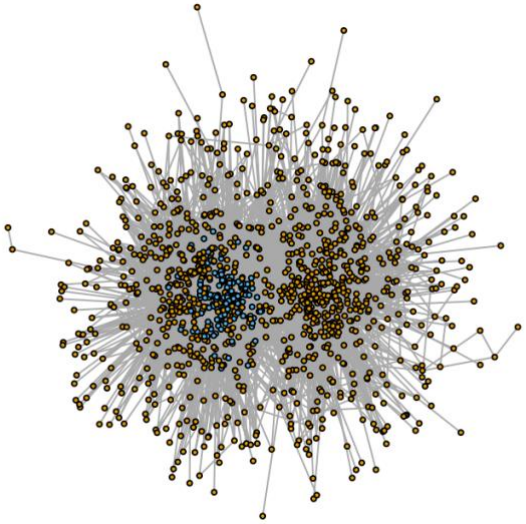

Figure 15: Graph partition obtained from hierarchical clustering with complete linkage

Table 6: Truth table obtained from hierarchical clustering with complete linkage

complete	0	1
0	583	508
1	3	128

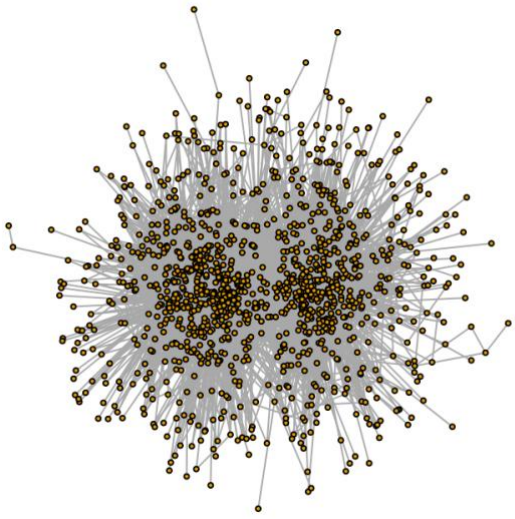

Figure 16: Graph partition obtained from hierarchical clustering with average linkage

Table 7: Truth table obtained from hierarchical clustering with average linkage

average	0	1
0	586	635
1	0	1

With the fast greedy modularity method, we obtain a partition with error rate of 5.65%, suggesting that it is close to the truth. The modularity of the partition is 0.425, higher than the modularity of the truth of 0.405.

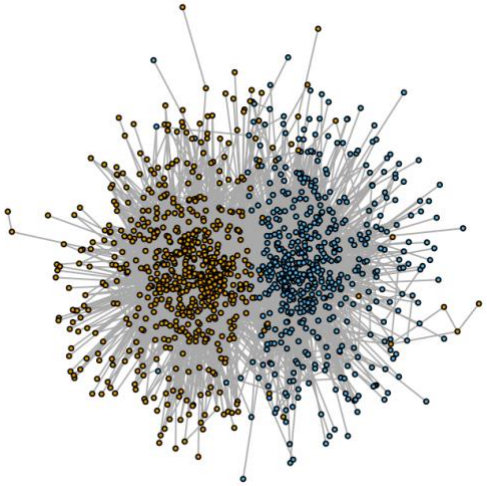

Figure 17: Graph partition obtained from fast greedy modularity method

Table 8: Truth table obtained from fast greedy modularity method

Modularity method	0	1
0	536	19
1	60	617

With partition method based on stochastic block model, we first find the first and second eigenvector of the adjacency matrix of $\tilde{G}$, denoted as v_1 and v_2. The entry-wise ratio $v_1(i)/v_2(i)$ for $i = 1, \cdots, n$, resultant vector is calculated and denoted as r. K-means method is then applied to r in order to find the labels of the nodes.

The error rate obtained is 4.99%, performing better than all the methods above.

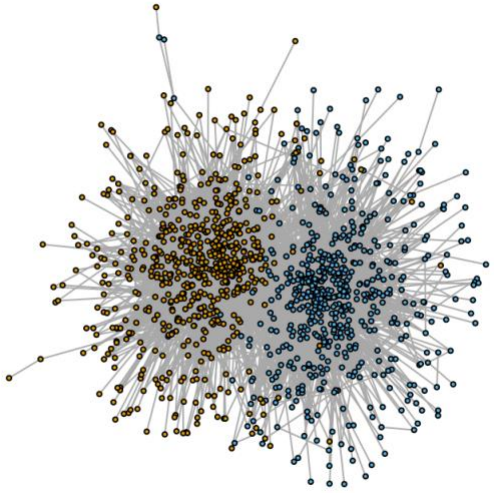

Figure 18: Graph partition obtained from k-means method based on stochastic block model

Table 9: Truth table obtained from k-means method based on stochastic block model

k-means based on SBM	0	1
0	34	609
1	552	27

6. Fitting Stochastic Block Model

From the result in section 5, stochastic block model is a good fit for the graph. Hence, in this section, we try to estimate the probability matrix of connections between/within the community, B under the assumption of stochastic block model. We have

Number of nodes in community 0, $n_0 = 586$
Number of nodes in community 1, $n_1 = 636$
Number of edges in community 0, $e_{00} = 7300$
Number of edges in community 1, $e_{11} = 7839$
Number of edges between community 0 and 1, $e_{01} = e_{10} = 1575$

Under the stochastic block assumption, the MLE of b is given by

$$\hat{b}_{00} = \frac{e_{00}}{\binom{n_0}{2}} = \frac{7300}{586*585/2} = 0.04259$$

$$\hat{b}_{11} = \frac{e_{11}}{\binom{n_1}{2}} = \frac{7839}{636*635/2} = 0.03882$$

$$\hat{b}_{01} = \hat{b}_{10} = \frac{e_{01}}{n_0 n_1} = \frac{1575}{586*636} = 0.00423$$

Hence, we have $B = \begin{pmatrix} \hat{b}_{00} & \hat{b}_{01} \\ \hat{b}_{10} & \hat{b}_{11} \end{pmatrix} = \begin{pmatrix} 0.04259 & 0.00423 \\ 0.00423 & 0.03882 \end{pmatrix}$

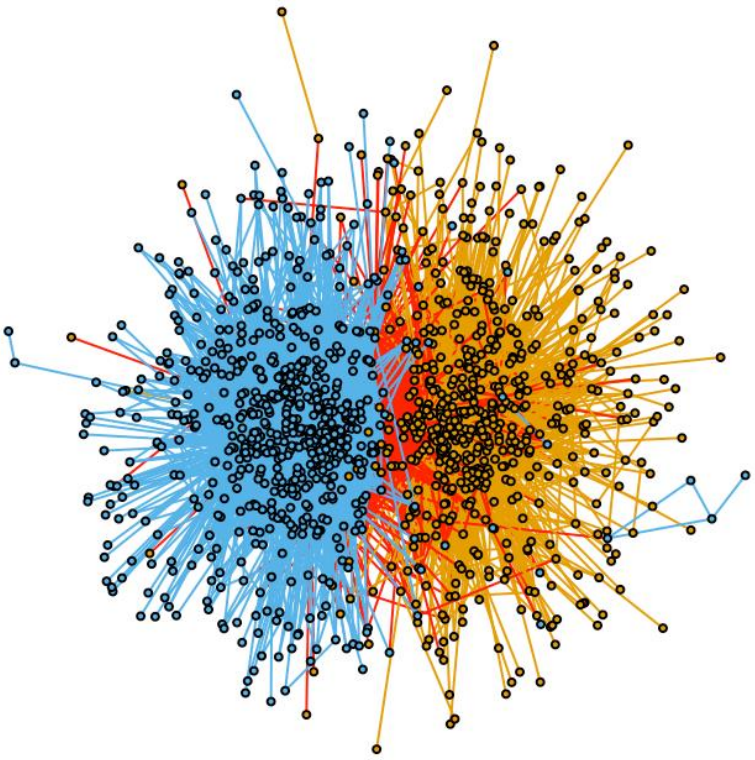

Figure 19: Giant component represented in stochastic block model