CSP301 Assignment 1

- -

Social Network Visualisation

Abhishek Kumar (2011CS50272)
Akhil Jain (2011CS50274)
Shivanker Goel (2011CS10298)

29-Aug-2012

Contents

1 Social Network Visualisation

PROBLEM : We were given a dataset on purchase patterns of politics-related books on Amazon. In the dataset, books were nodes, labeled as whether the book is left leaning, right leaning, or neutral in its political stance. Edges existed between two books if some people have purchased both the books. This dataset was called **polbooks**.gml. Another dataset called **polblogs**.gml on blog affiliations was also given.

We were asked to build visualizations for both datasets using various layout algorithms in *Prefuse*, that will show if people like to read diverse books that touch upon several different affiliations or they rather like to read stuff that possibly resonates with their own viewpoints.

We were to state any interesting observations in our report.

2 Project

A soft copy of the source code of the our project is available at
https://github.com/CSDesign/secret-bear.git

This project consist of the following four main files:

1. **graphAlpha.java** : This file runs the visualisation of *polblogs*.
2. **graphBeta.java** : This file runs the visualisation of *polbooks*.
3. **graphGamma.java** : This file merges the *Strongly Connected Components* of the *polblogs* dataset into super-nodes and shows its visualisation. Clicking on any super-node will bring up the graph of its components.
4. **graphDelta.java** : This file merges all 3-cliques of polblogs whose nodes are of the same ideology into single super node. Right leaning cliques are shown by triangles pointing to the right while left leaning cliques point to the left. Clicking on any triangle opens up the subgraph of the clique.

For purposes of comparison, we made 20,000 random graphs for each dataset and carried out similar analysis on all of them.

3 Definitions

3.1 Edge Ratio

Edge Ratio *is defined as ratio of number of edges between nodes of the same affiliation to total number of edges.*

3.2 Clustering Coefficient

Local Clustering Coefficient

In undirected networks, the *Local Clustering Coefficient* C_n *of a node n is defined as* $C_n = 2e_n/(k_n(k_n\text{-}1))$, where k_n is the number of neighbors of n and e_n is the number of connected pairs between all neighbors of n.
In directed networks, the definition is slightly different: $C_n = e_n/(k_n(k_n\text{-}1))$.
In both cases, the clustering coefficient is a ratio N / M, where N is the number of edges between the neighbors of n, and M is the maximum number of edges that could possibly exist between the neighbors of n.

Average Network Clustering Coefficient

The *Average Network clustering coefficient is the average of the local clustering coefficients for all nodes in the network.*

Global Clustering Coefficient

The *Global Clustering Coefficient is the ratio of number of triangles present in the graph to the maximum number of triangles that could possibly exist in a graph with the same number of nodes.*

3.3 Pearson's Correlation Coefficient

The Pearson's Correlation Coefficient (ρ) is a measure of linear dependence between two variables X and Y. It is defined as follows

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$$

3.4 Strongly Connected Component

A component of the directed graph is called a ***Strongly Connected Component*** *if it is the maximal component in which there is a path from each vertex in the component to every other vertex.*

4 Results

4.1 polbooks

Looking at the visualisation we can easily infer that the graph is highly polarised with a lot of clustering in two major sections. One of these clusters mainly consists of liberal books while the other consists of conservative books.
There were 49 Conservative books, 43 Liberal books and 13 Neutral books. The maximum degree among all the nodes is 23 while the median is 6. The diameter of the graph is 7. The graph was found to be connected as a whole.

4.1.1 Edge Ratio

The **Edge Ratio** of graph *polbooks* was found to be 0.84. Such a high ratio is significant of the fact that most people like to read books which suit their own political ideologies. Only few venture to read up about the other existing ideologies.
When we analysed the same ratio for random graphs, we found that the Edge Ratio ranged from 0.30 to 0.49. This reinforces our inference above that most people tend to remain in their ideological shell rather than explore other ideologies as we expect a neutral ideal person to do.
The variation in edge ratio can be seen in the folowing Illustration.

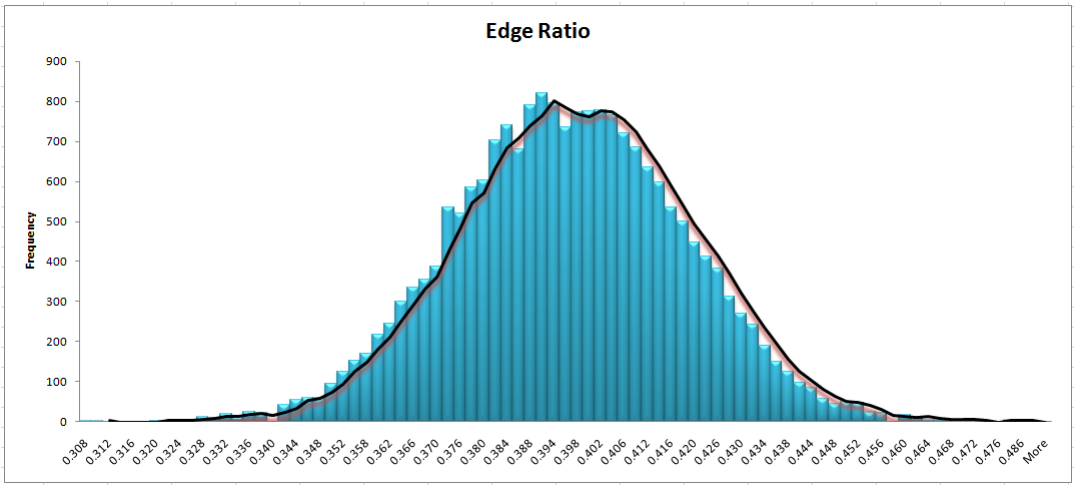

Fig 4.1.1 : The graph depicts a typical Gaussian curve for random distribution.

polbooks lies to the extreme right of the above Gaussian distribution. Thus most people tend to stay within their ideological shells rather than explore other ideologies.

4.1.2 Clustering Coefficient

Average Network Clustering Coefficient

The *Average Network Clustering Coefficient* of graph *polbooks* was found to be 0.88. When we analysed the same for random graphs. We found that the Average Network Clustering Coefficient ranged from 0.05 to 0.12. The variation in Average Network Clustering Coefficient can be seen in the folowing Illustration.

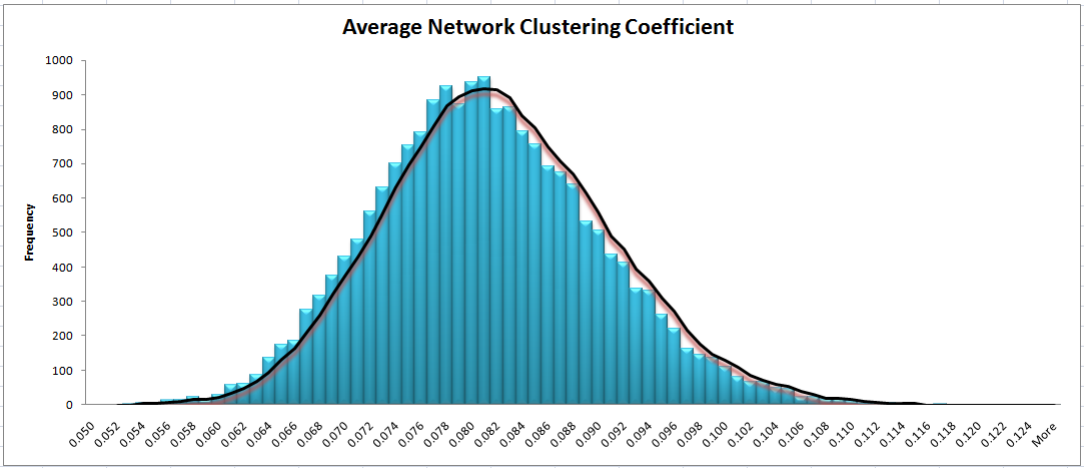

Fig 4.1.2 : The graph depicts a typical Gaussian curve for random distribution.

polbooks lies to the extreme right of the above Gaussian distribution. This shows that in polbooks there is extremely dense distribution of triangles within the clusters which cannot be seen in a random distribution of edges where no such distinct clustering exists.

Global Clustering Coefficient

The *Global Clustering Coefficient* of graph *polbooks* was found to be 0.003. When we analysed the same for random graphs. We found that the Global Clustering Coefficient ranged from 0.0003 to 0.0008. The variation in Global Clustering Coefficient can be seen in the folowing Illustration.

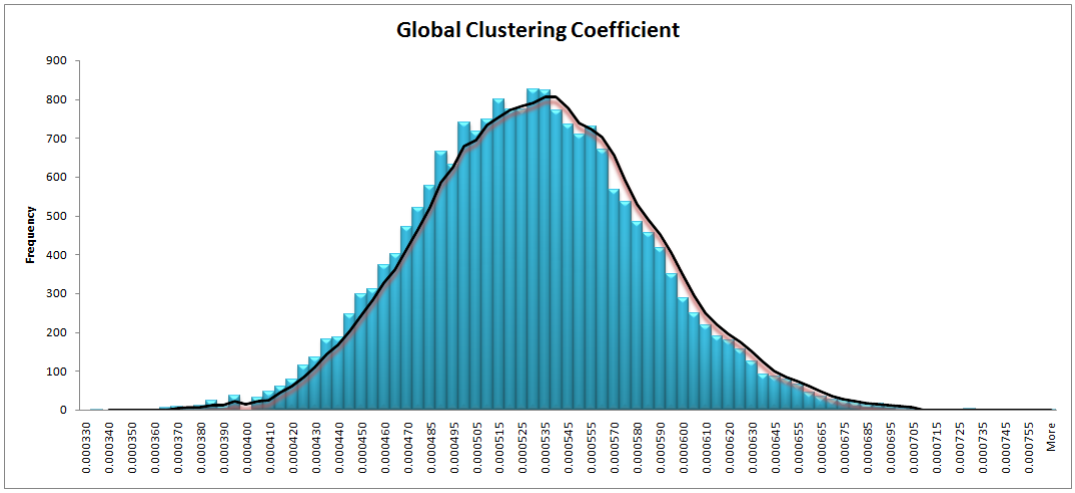

Fig 4.1.3 : The graph depicts a typical Gaussian curve for random distribution.

polbooks lies to the extreme right of the above Gaussian distribution. This shows that in polbooks there is extremely dense distribution of triangles which cannot be seen in a random distribution of edges.

4.1.3 Pearson's Correlation Coefficient

We calculated the correlation coefficient for dependence between the number of triangles a particular node is involved in and its degree.
The *Pearson's Correlation Coefficient* of graph *polbooks* was found to be 0.96. When we analysed the same for random graphs. We found that the Pearson's Correlation Coefficient ranged from 0.49 to 0.90. The variation in Pearson's Correlation Coefficient can be seen in the folowing Illustration.

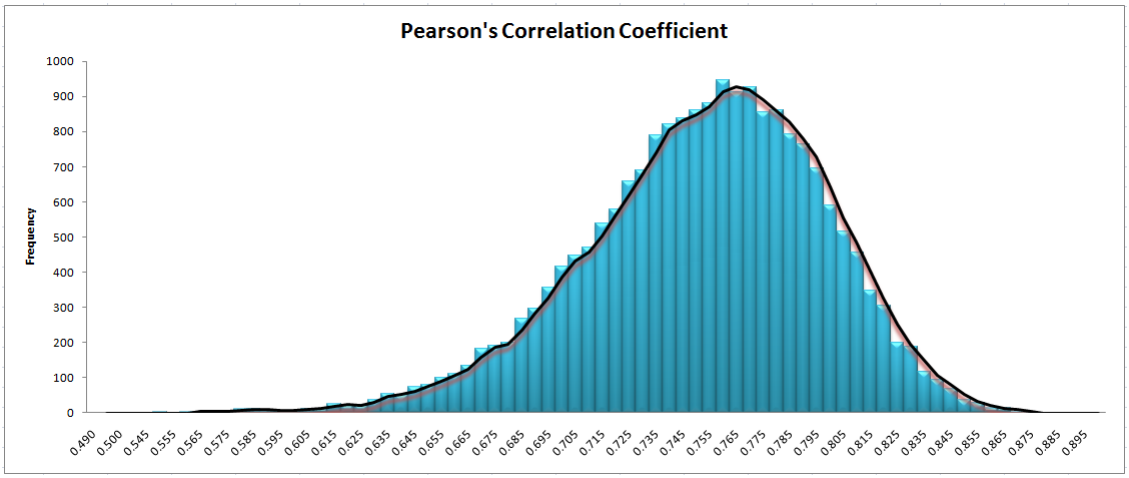

Fig 4.1.4 : The graph depicts a typical Gaussian curve for random distribution.

polbooks lies to the right of the above Gaussian distribution. This shows that in polbooks there is an almost complete linear dependence between the number of triangles formed by a node and its degree.

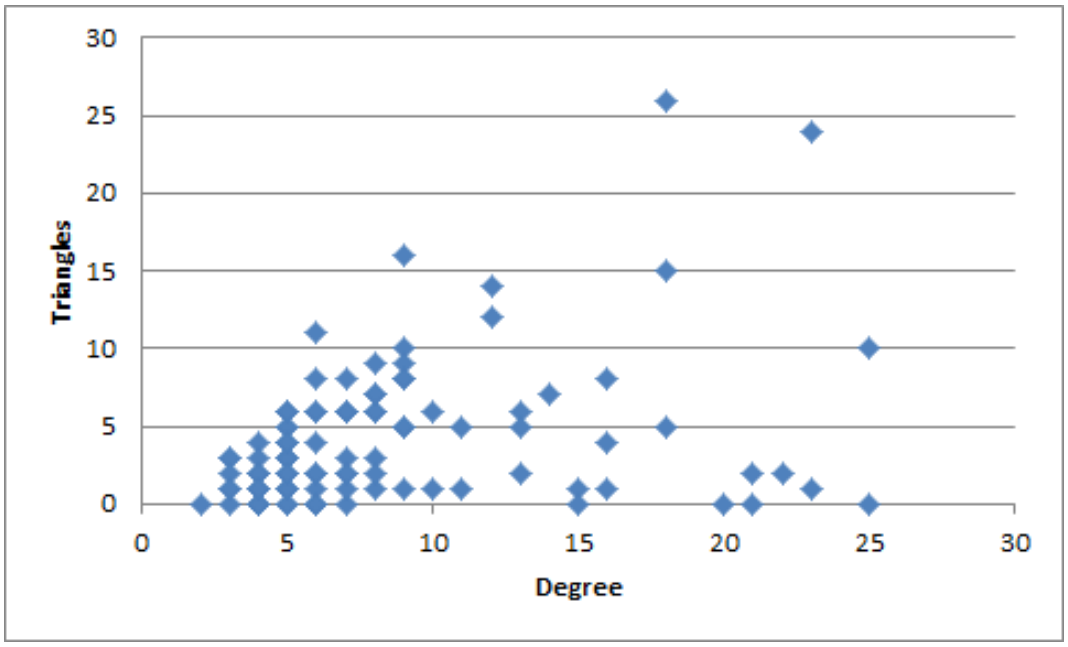

Fig 4.1.4 : Scatter plot between number of triangles and degree of a node.

A similar analysis was also done on social network graphs generated using scale-free graph generation algorithms. However the analysis generated was also quite far away from the analysis for polbooks. An analysis of the same can be seen on our github repo at: https://github.com/CSDesign/secret-bear/blob/master/csp301/randSocbooksGraphAnalysis.xlsx

4.2 polblogs

Looking at the visualisation we can easily infer that the graph is highly polarised with a lot of clustering in two major sections. One of these clusters mainly consists of liberal blogs while the other consists of conservative blogs.
There were 758 Liberal blogs and 732 conservative ones. The maximum degree among all the nodes is 468 while the median is 8. It was found that polblogs is a disconnected graph with 269 disconnected components of which one is of size 1221, one is a pair while all others are singlets.
The diameter of the largest connected component was found to be 9. This in essence points to the **Small World Network** concept wherein most nodes can be reached from every other node in a small number of hops.

4.2.1 Edge Ratio

The **Edge Ratio** of graph *polblogs* was found to be 0.92. Such a high ratio is significant of the fact that most blogs link to other blogs which suit their own political ideologies. This is expected since most of them would want to highlight their own political ideology rather than the alternate.
When we analysed the same ratio for random graphs. We found that the Edge Ratio ranged from 0.48 to 0.52. The variation in edge ratio can be seen in the folowing Illustration.

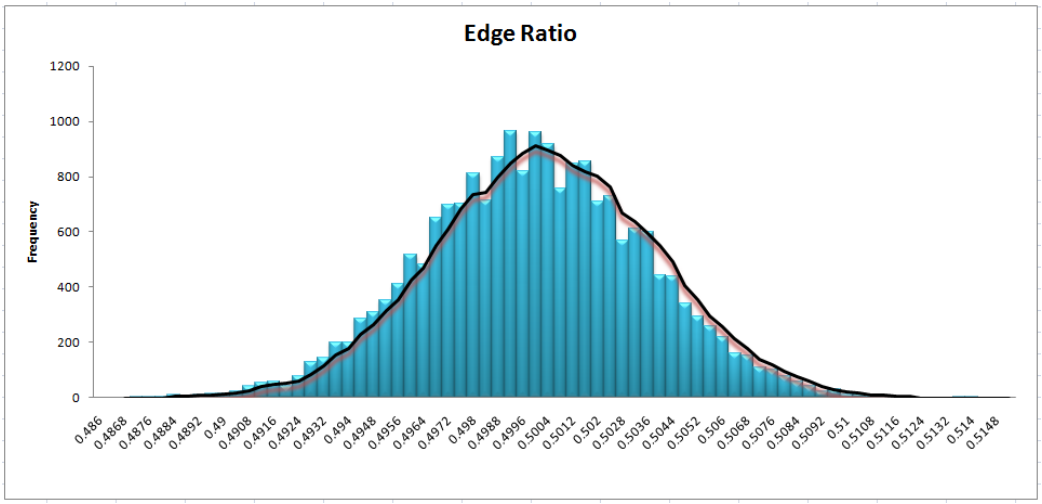

Fig 4.2.1 : The graph depicts a typical Gaussian curve for random distribution.

polblogs lies to the extreme right of the above Gaussian distribution. Thus most blogs link to other blogs within the same ideology.

4.2.2 Clustering Coefficient

Average Network Clustering Coefficient

The *Average Network Clustering Coefficient* of graph *polblogs* was found to be 0.114. When we analysed the same for random graphs. We found that the Average Network Clustering Coefficient ranged from 0.0039 to 0.0046. The variation in Average Network Clustering Coefficient can be seen in the folowing Illustration.

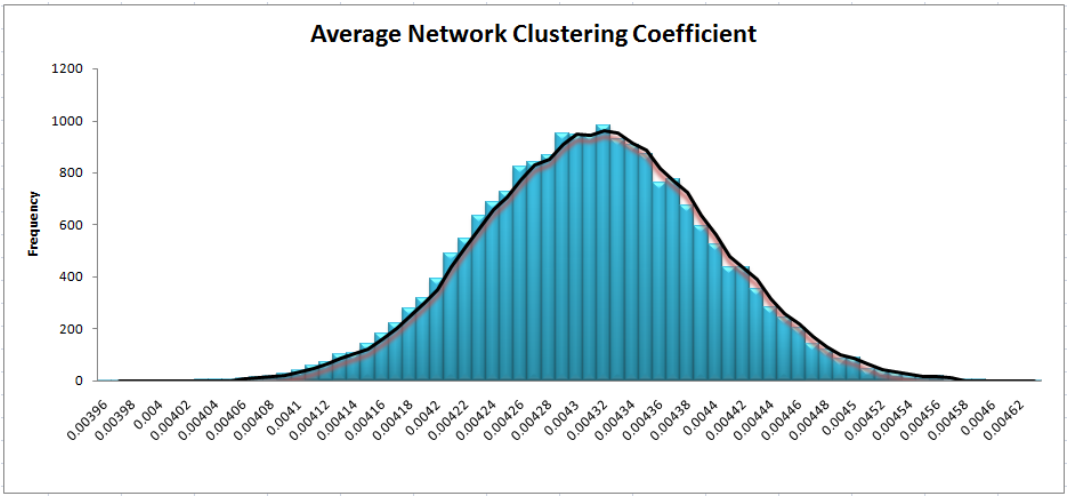

Fig 4.2.2 : The graph depicts a typical Gaussian curve for random distribution.

polblogs lies to the extreme right of the above Gaussian distribution. The general values of the average network clustering coefficient for the analysed data is low because of the very high degree of most nodes and thus the large number of triangles that they are capable of forming, which increases the denominator to a large extent and hence reduces the value of the coefficients.

4.2.3 Pearson's Correlation Coefficient

We calculated the correlation coefficient for dependence between the number of triangles a particular node is involved in and its degree.
The *Pearson's Correlation Coefficient* of graph *polblogs* was found to be 0.92. When we analysed the same for random graphs. We found that the Pearson's Correlation Coefficient ranged from 0.63 to 0.74. The variation in Pearson's Correlation Coefficient can be seen in the folowing Illustration.

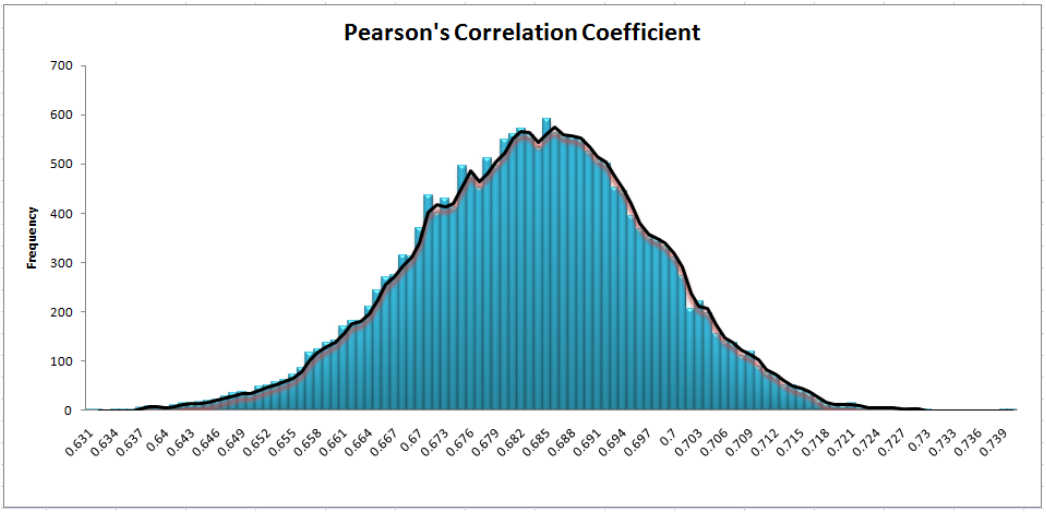

Fig 4.2.3 : The graph depicts a typical Gaussian curve for random distribution.

polblogs lies to the extreme right of the above Gaussian distribution. This shows that in polblogs there is an almost completely linear dependence between the number of

triangles formed by a node and its degree as can also be seen from the following scatter plot.

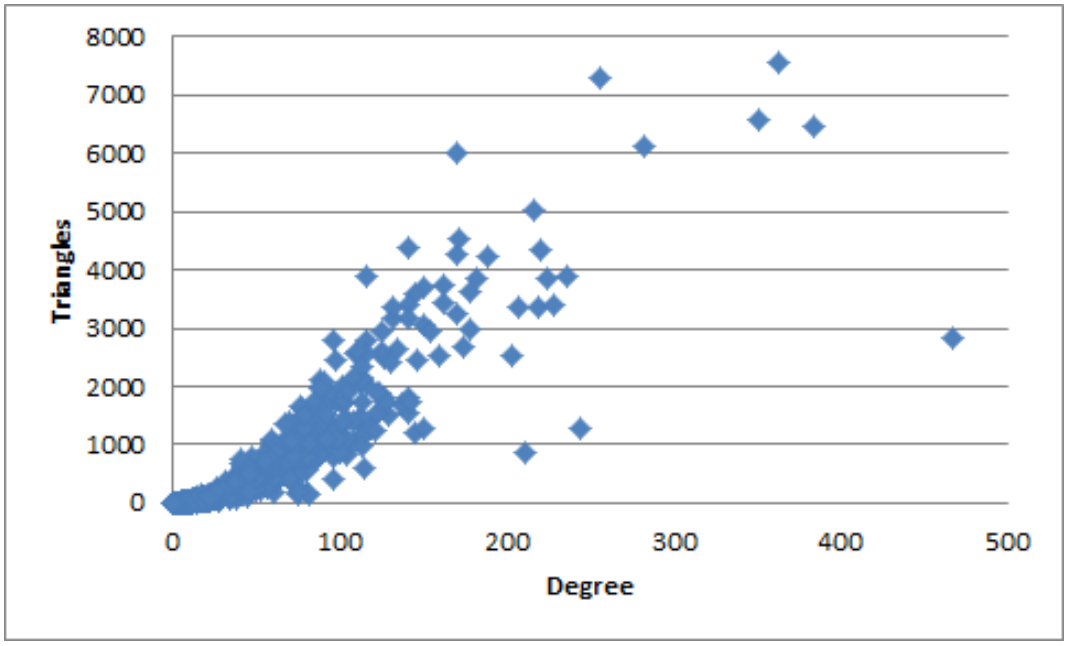

Fig 4.2.4 : Scatter plot between number of triangles and degree of a node.

A similar analysis was also done on social network graphs generated using scale-free graph generation algorithms. However the analysis generated was also quite far away from the analysis for polblogs. An analysis of the same can be seen on our github repo at:
https://github.com/CSDesign/secret-bear/blob/master/csp301/randSocblogsGraphAnalysis.xlsx

5 Visualisation

5.1 Features

- **Zooming** : *Zoom* into or out of graph using the mousewheel or right clicking and dragging.
- **Panning** : *Pan* into the graph by clicking and dragging.
- **Tool Tip** : Keeping cursor on any node will pop a *Tip* containing important information about the particular node.
- **Force Simulator** : Change sliders present on the right jPanel and observe changes in the behaviour of the graph.
- **Connectivity Filter** : Click on any node and change the slider of connectiviy filter. This shows all nodes in the graph which have a path of the specified length from the selected node.
- **Search Field** : Type into the search box and the nodes whose labels contain the search item get dynamically highlighted.
- **Neighbour Highlighting** : Hover over any node and its neighbour nodes and edges get highlighted.
- **Highlighting High Degree Nodes** : Nodes are given different Stroke colors according to their degree. Nodes with degree higher than 3 times the median degree have a darker stroke color.
- **Distinct Colors and Shapes** : Different colors and shapes are given to nodes according to their affiliation.
- **Node Info** : Hover over any node and all its information becomes visible on the right panel.
- **Overview** : In case you zoom in and dont want to zoom out an overview of the graph is provided on right panel.
- **Forced Directed Layout** : Forced Directed layout has been implemented using Prefuse's Runge Kutta Integrator.
- **Edge Renderer** : Edge Renderer has been used to control the size and color of edges and their arrows heads.
- **Tranparency** : Mostly nodes do not cover each other compeletly but in case of strongly connected components of polblogs i.e. *graphGamma*, the bigger node covers some nodes. This was solved by increasing transparency of the node when we hover over it.
- **Relative Sizing** : The node size in *graphGamma* has been set according to the size of strongly connected component within it.
- **Click action** : On clicking any of the nodes of *graphGamma*, a new window opens up with the subgraph contained within that node.
- **Cliques** : The graphDelta visualisation shows all cliques in the polblogs dataset as triangles. The conservative cliques point to the right while the liberal point to the left. When we click on any of these cliques, a new window opens up with the subgraph contained within that node.

5.2 Screenshots

5.2.1 polbooks

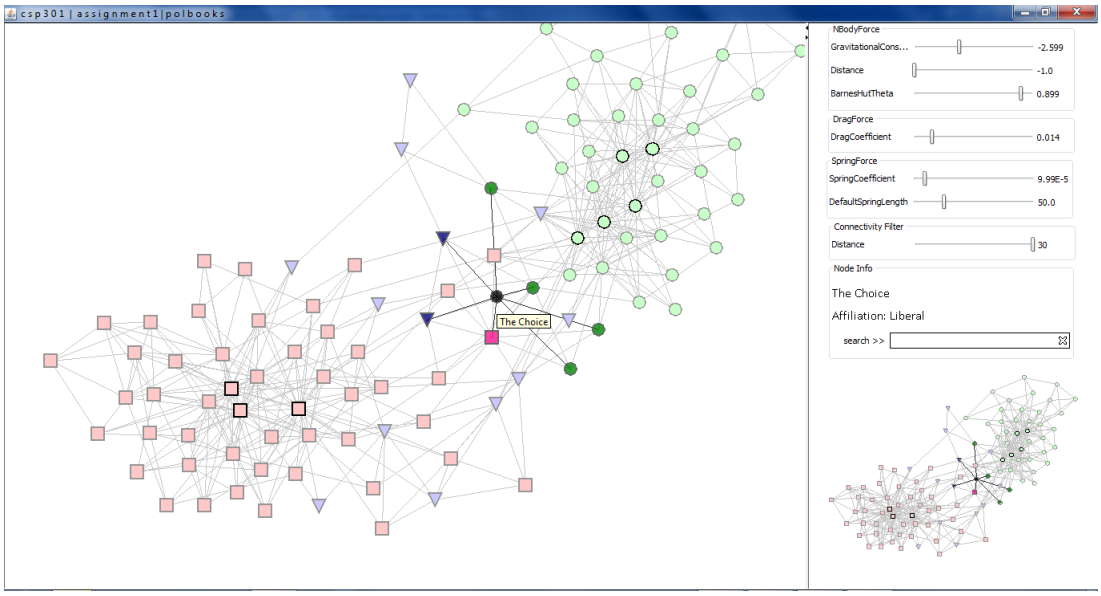

Fig 5.1 : Visualisation of graphBeta.

5.2.2 polblogs

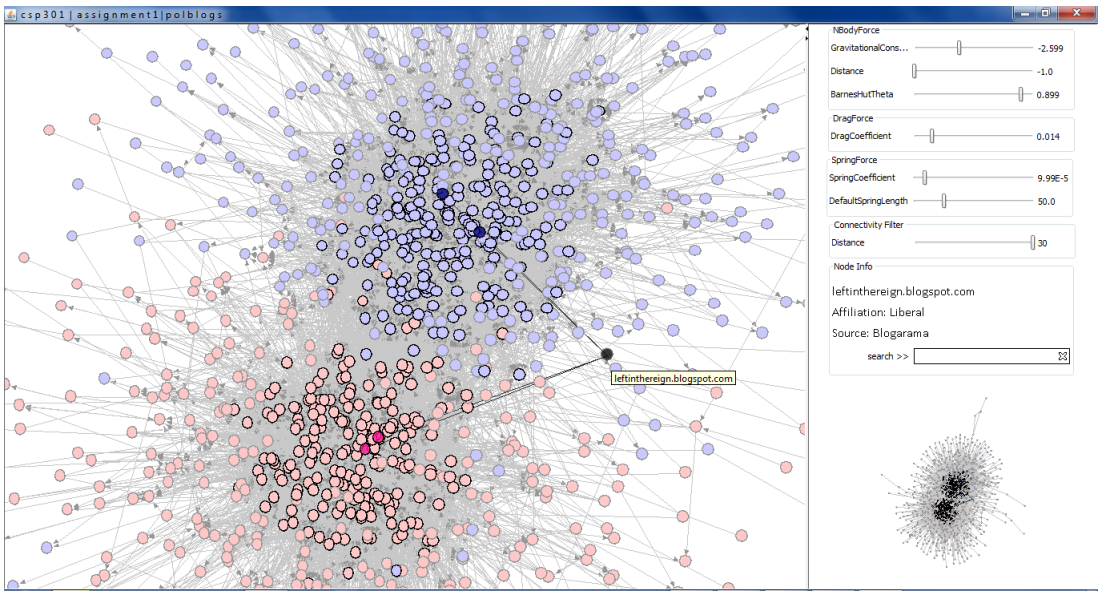

Fig 5.2 : Visualisation of graphAlpha.

5.2.3 Strongly Connected Components

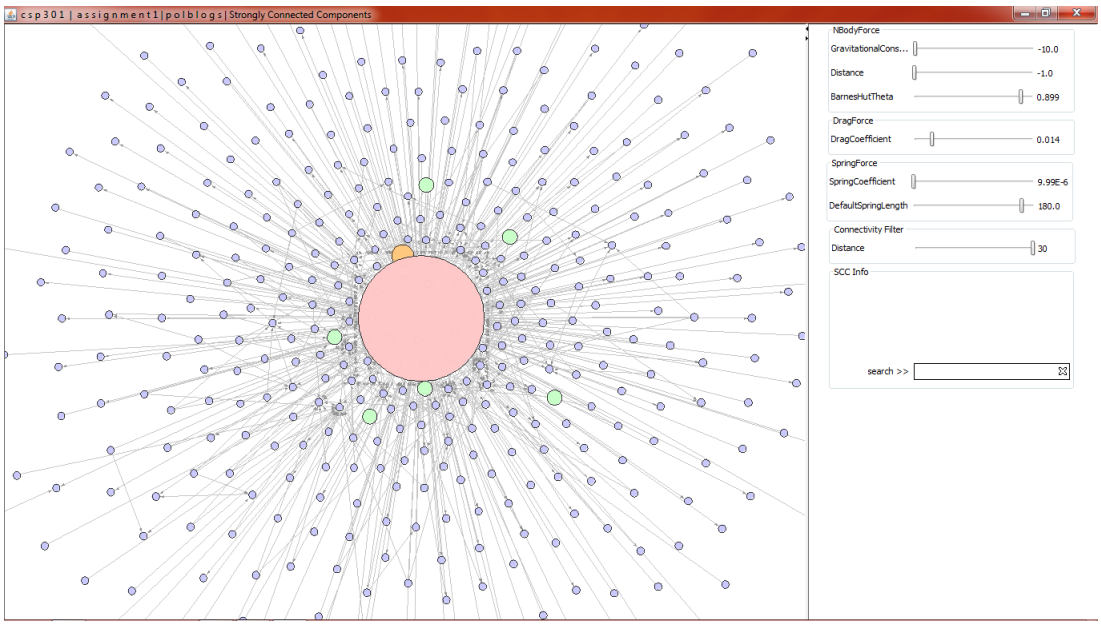

Fig 5.3 : Visualisation of graphGamma.

5.2.4 Cliques

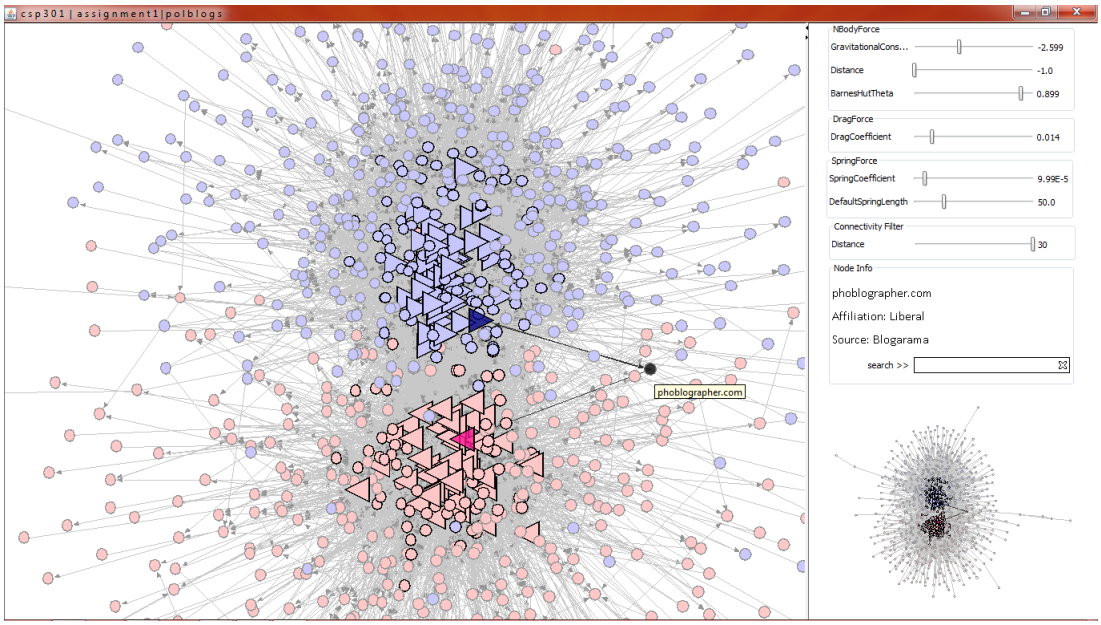

Fig 5.4 : Visualisation of graphDelta.

• • •

Report developed in LaTeX