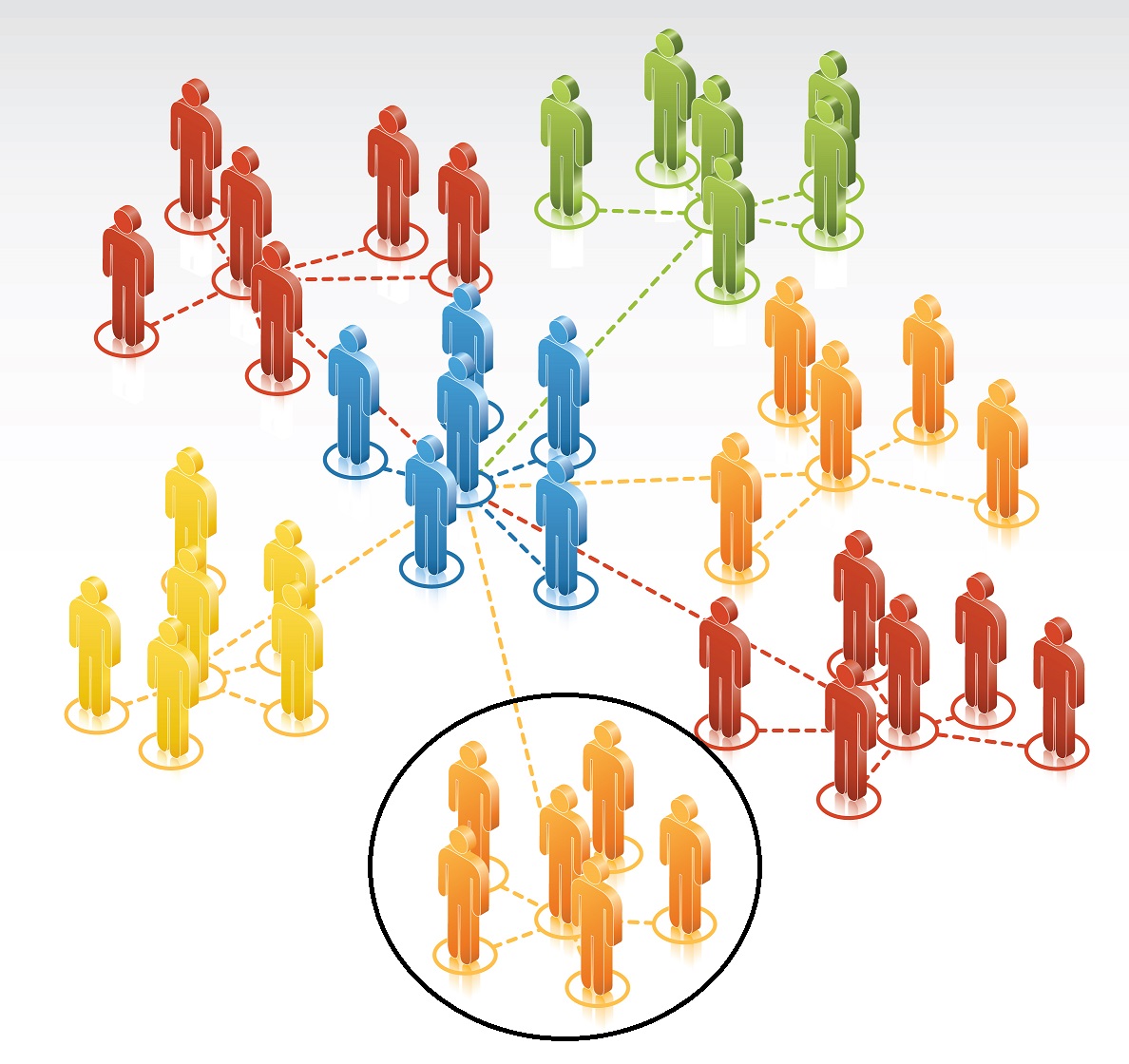
**DATA MINING 2 REPORT**

**SOCIAL NETWORK ANALYSIS**

****

**Aarti Ranganathan  
Ambrose Wong**

**Swaraj Mohapatra**

**Roshan Shetty**

**ABSTRACT**

Social network analysis helps in discerning information about the structure, evolution of networks and processes that occur on them. Overlap of communities in networks can be explained with the help of link communities. Modeling social networks helps in capturing both the regularities in the formation of the network ties and the variability that can be associated to noise. The project deals with all the above facets with the help of Gephi as well as R packages like linkcomm and statnet. A brief introduction to the project proceeds with an overview of network analysis with Gephi, GML and Exponential Random Graph Models (ERGM). The datasets used in the analysis are then introduced with subsequent sections explaining the key highlights of the analysis. Relevant link communities are then found in a Facebook social network. ERGM is then used to build several models of a Facebook network and the best model is chosen from them.

Table of Contents

[1. Introduction 2](#_Toc417838743)

[2. Social Network Analysis (SNA) 2](#_Toc417838744)

[3. Exponential Random Graph Models for Social Network 3](#_Toc417838745)

[4. Software 4](#_Toc417838746)

[5. Graph Modeling Language 6](#_Toc417838747)

[6. Datasets 7](#_Toc417838748)

[7. Facebook Network Analysis 7](#_Toc417838749)

[8. Word Adjacencies 9](#_Toc417838750)

[9. Books about US Politics 12](#_Toc417838751)

[10. Political Blogs 14](#_Toc417838752)

[11. Unsupervised learning 16](#_Toc417838753)

[12. Random graph models 16](#_Toc417838754)

[13. Conclusions 18](#_Toc417838755)

[14. References 18](#_Toc417838756)

# Introduction

In today's developed world everything is connected: people, information, events and places, and online social media has made it even easier. A practical way of making sense of the fuzzy connections is to analyze them as networks. With the help of Gephi, large amount of social network data is analyzed in an interactive way. The report demonstrates the hands-on analysis of real-world data sets and focuses on a range of tasks: from identifying important nodes in the network, to detecting communities, to tracing information diffusion and opinion formation. The idea is to be able to extract information from a seemingly fuzzy network and recognize communities and cliques. We further used the same techniques and applied them on other networks such as word adjacencies, books about US politics, political blogs and Facebook data.

Using these datasets, we portrayed gender distribution in a Facebook network, commonly used adjectives in a book, conflicting trends in customer demands, and political dysfunction through blogs. These exploratory analysis was combined with unsupervised learning to detect communities in a Facebook dataset. These detection algorithms used clustering type algorithms where each node was selected into a certain group. However, this was a major drawback because people could belong to multiple communities. Hence linked community analysis was also performed where people could belong to multiple clusters.

Finally, with the help of ERGM, various models are built for Facebook data and the GWESP based model with degree 1 term is chosen based on MCMC diagnostics – which reveals the error distribution. Although the model has the second best AIC and BIC values, it is still chosen as its noise distribution is normal.

# Social Network Analysis (SNA)

With the advent of social media and increases in technology, people are connected to each other in much more sophisticated ways than in the past. To understand this phenomenon, we have to understand the system that connects them together. In today’s terminology, these systems are called social networks and it includes attributes or information on how people are organized. Network analysis on these social networks allows for the extraction of interesting information about the structure, evolution of networks and processes that occur within them. Certain important concepts of social network analysis are as follows:

* To analyze the relationship among the actors (nodes) in the network.
* To analyze the information flow among the different actors (nodes) in the network.
* To analyze the attributes such as the social, economic, and political structures of the network and to conceptualize them as patterns of relationship among the actors (nodes).

Social networking has gained innumerous popularity with the advent of Facebook, google plus and other online networking sites. There are more than one billion users on Facebook alone. Social networks formed explicitly on these internet sites and implicitly through other social interactions can be used to understand the importance of a single person in a network, the formation of groups of people in a network, and the interconnection amongst these groups. At the same time, the user interaction can indicate the participation of the user in a particular process and how certain processes such as a disease or rumor spreads in the network. We present an example in the next figure.

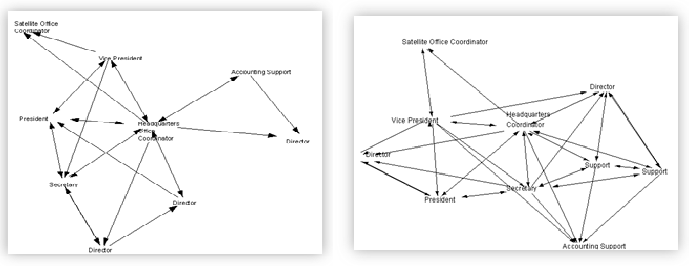


Fig 1.1: Network dynamics before and after the introduction of content management system

Figure 1.1 depicts the flow of information from one node to another and how this flow of information can be affected by external forces. In addition to network dynamics, businesses can use SNA to analyze and improve their communications within the organization and also with their customers and clients. Social medial websites such as Facebook uses SNA to recommend potential friends based network connections amongst your friends.

# Exponential Random Graph Models for Social Network

The observed network is regarded as one realization from a set of an infinite number of networks with similar important characteristics (such as number of nodes). In other words, the outcome is determined by an unknown stochastic process. Since the stochastic process is unknown, our goal is to formulate a model that is plausible and theoretically principled at the same time [5]. The network can be described by structural characteristics such as parameters and the range of possible networks can be described using a probability distribution on the set of nodes.

**3.1. Exponential random graph model: dependence assumptions and parameter constraints**

Exponential random graphs have the following form:

……………………………….. (1)

Where (i) the summation is over all conﬁgurations A; (ii) is a parameter corresponding to the conﬁguration A (and is non-zero only if all pairs of variables in A are assumed to be conditionally dependent); (iii) = is the network statistic corresponding to conﬁguration A; =1 if the conﬁguration is observed in the network y, and is 0 otherwise; (iv) is a normalizing quantity which ensures that (1) is a proper probability distribution

**A special form: Markov Random Graph**

A Markov random graph model for a non-directed network with edge, two-star, three-star and triangle effects is:

where and are the numbers of two-stars and three-stars, respectively, in the network y and is the number of triangles in y. The Markov Random graph will be primarily used in the estimation technique: *Markov Chain Monte Carlo Estimation Technique.* The central approach is to simulate a distribution of random graphs from a starting set of parameter values and iteratively reﬁne the parameter values by comparing the distribution of graphs against the observed graph until the parameter estimates stabilize.

# Software

**GEPHI**

Gephi is an open-source [network analysis](http://en.wikipedia.org/wiki/Network_theory) and [visualization](http://en.wikipedia.org/wiki/Network_visualization) software package written in Java on the [NetBeans](http://en.wikipedia.org/wiki/NetBeans) platform. It was initially developed by students of the [University of Technology of Compiègne](http://en.wikipedia.org/wiki/University_of_Technology_of_Compi%C3%A8gne) (UTC) in France. It has been used worldwide in research, journalism and other interesting applications. For example, it was used to visualize the global connectivity of New York Times content and to examine the Twitter network traffic during social unrest along with more traditional network analysis topics.

The power of Gephi lies in its interactive nature and its ability to visualize and explore data from all types of networks. The user interacts with the graph and manipulates the structure, shapes and colors to reveal hidden properties. It uses a 3D rendering engine to display large networks in real-time and to speed up exploration. In summary, it’s a flexible and multi-task architecture that brings new possibilities in complex data sets by producing valuable visual results.

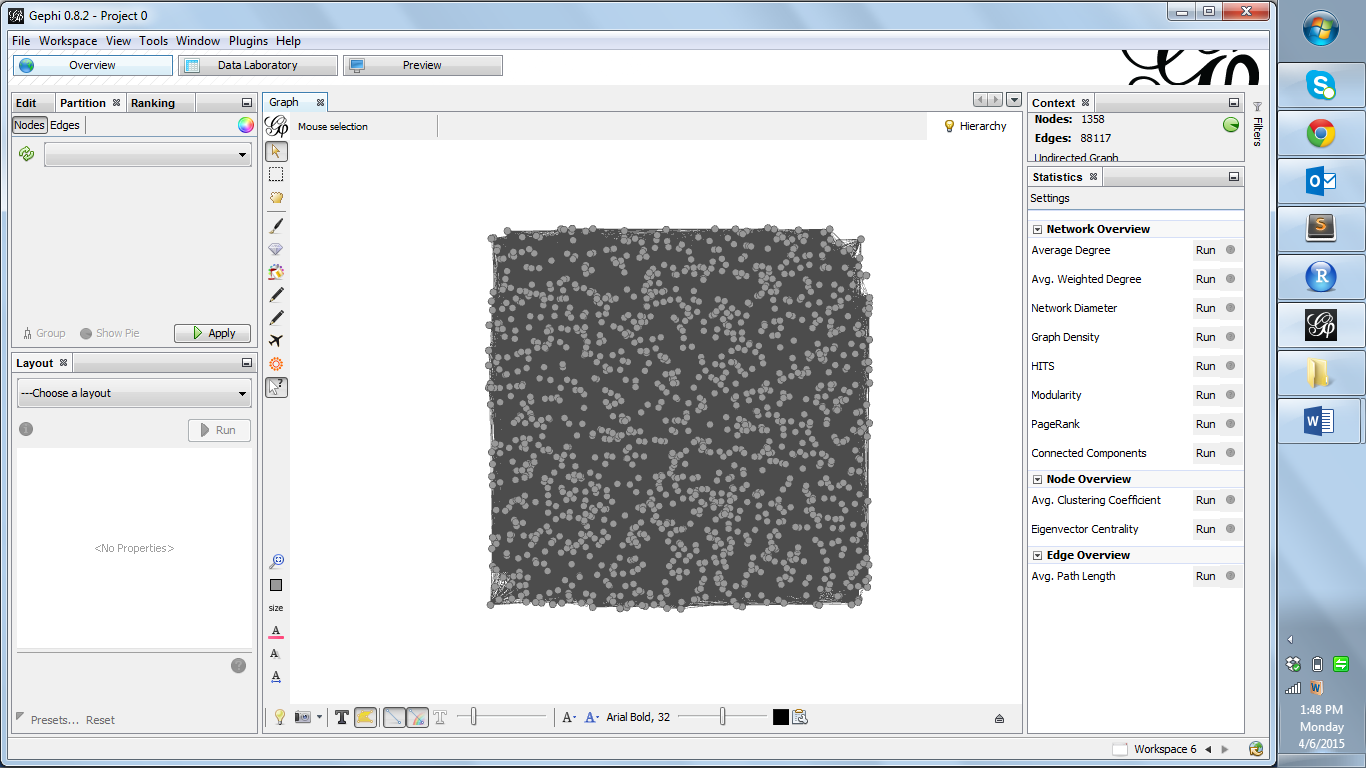


Fig 4.1 : Snapshot of the Gephi software tool

**R**

Statnet comprises a suite of R packages for the analysis of social networks and other relational data. We primarily use the following 3 packages in it: **network, sna** and **ergm.** The **network** and **sna** packages are used for querying the network data and network exploration while the **ergm** package allowed the user to ﬁt exponential-family random graph (ERG) models to network data sets. Each of the models can be compared using the MLE criterion to arrive at the best model. The goodness of fit of the set of networks simulated from the model can be done by comparing it with the observed network using **simulate** and **gof** functions.

The package igraph and Linkcomm in R was also used extensively. Igraph was used for clustering analysis and for converting between data types. And linkcomm was used to perform linked community analysis.

# Graph Modeling Language

Graph Modeling Language (GML) is a hierarchical key value list that uses ASCII-based file format for describing graphs. It has been also named as Graph Meta Language. Its key features are portability, simple syntax, extensibility and flexibility. Graphs can be annotated with arbitrary data structures called attributes and the idea for a common file format was born at the GD'95 after many discussions. GML is the standard file format in the Graphlet graph editor system and it has been adapted by several other systems for drawing graphs. A simple graph example in GML format is given below:

**graph** [

comment "This is a sample graph"

directed 1

id 42

label "Hello, I am a graph"

**node** [

id 1

label "node 1"

thisIsASampleAttribute 42

]

**node** [

id 2

label "node 2"

thisIsASampleAttribute 43

]

**edge** [

source 1

target 2

label "Edge from node 1 to node 2"

]

]

# Datasets

We analyzed four datasets in Gephi. The datasets differ from each other on various levels and the idea is to be able to extract information from seemingly fuzzy and different networks. The datasets include User Facebook data, word adjacencies, US political blogs and political books. In addition to Gephi, we also used R to analyze Facebook data. Note that the Facebook data was extracted using an API.

# Facebook network Analysis

Facebook data from one of the users of the group was used for analysis.The attributes extracted from the facebook data include:

|  |  |  |
| --- | --- | --- |
| Sl # | Attribute | Attribute Explanation |
| 1. | Nodes | The name of the people on the facebook user’s network |
| 2. | Id | Id is a variable that uniquely identifies the nodes on the network |
| 3. | Label | Label is an attribute that defines how a node is labelled on the network |
| 4. | Sex | Sex of the facebook user’s friend |
| 5. | Agerank | The user’s friends’s age rank. May be 13-17, 18-20 or 21+ |
| 6. | Locale | A two-letter language code and a two-letter country code representing the users’ friends’ [locale](http://www.facebook.com/translations/FacebookLocales.xml). |
| 7. | Degree | The total number of edges connected to a node in the network |

Table 7.1: Facebook data descriptions

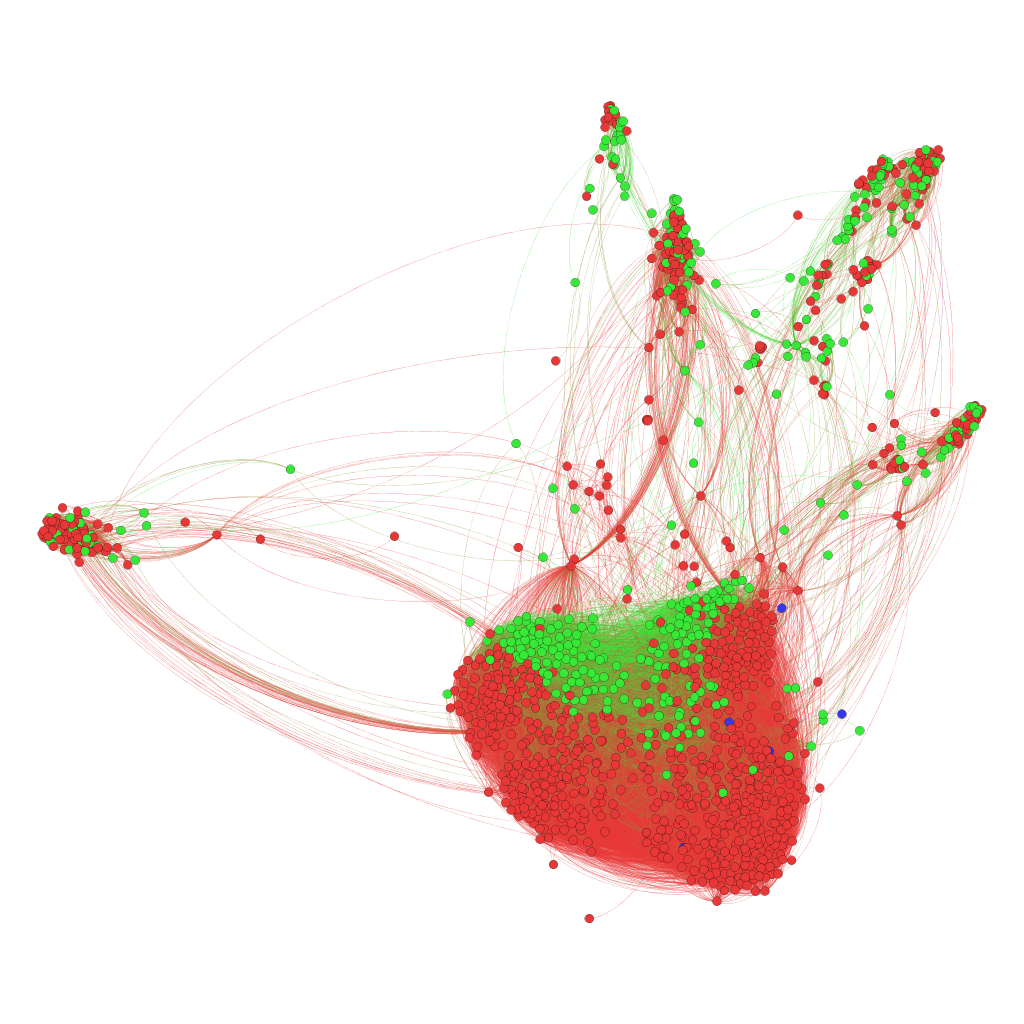


Fig 7.1: Facebook Network with color coding based on gender

The network in figure 7.1 shows that the user has quite a few communities and cliques in his Facebook network. Note that the nodes represent friends of the user and they are colored based on gender. With red nodes representing males and green nodes representing females. People who have not updated their sex on Facebook are represented as blue nodes. The links between the nodes indicate connections between the friends of the user and are called edges. In the above graph, the edges are undirected and each have the same weight. This implies that if you are my friend than I’m also your friend.

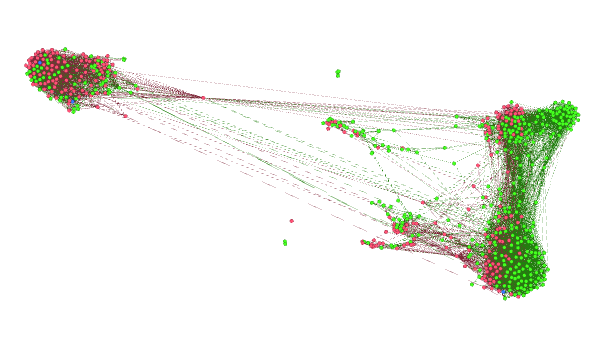


Fig 7.2: Facebook Network with an important node highlighted.

Figure 7.2 shows a node which acts as a significant connection between the two groups of Facebook friends. This person plays a vital role and influences the common posts that the different user groups can find in their notifications. Studying this node further can help analyze the user behavior.

Based on the Facebook network, we can understand the distribution of the user’s friends. The people who play important roles in the network based on the number of mutual friends, friends who play the role of intermittent connections between different communities. This information affects the newsfeed of the user, information that will be accessible by the user and mutual friends’ suggestions.

# Word Adjacencies

We explored the common adjectives and nouns in the novel "David Copperfield" by Charles Dickens and created an adjacency network. The attribute in the dataset include:

|  |  |  |
| --- | --- | --- |
| Sl.no | Attribute Name | Attribute Explanation |
| 1. | Nodes | The noun or adjective commonly used in the book |
| 2. | Id | Id is a variable that uniquely identifies the nodes on the network |
| 3. | Label | Label is the attribute which defines how the node is labelled on the network |
| 4. | Value | Indicates if the work is a noun or an adjective. 0: Adjective ,1: Noun |

Table 8.1: Word adjacency data: Node descriptions

|  |  |  |
| --- | --- | --- |
| Sl.no | Attribute Name | Attribute Explanation |
| 1. | Source | Indicates the node that acts as the source for the edge |
| 2. | Target | Indicates the node that acts as the target for the edge |
| 3. | Type | Specifies the type of the edge: Directed |
| 4. | Id | Id is a variable that uniquely identifies the edge on the network |
| 5. | Weight | The weight of an edge. In this dataset, each edge has an equal weight. |

Table 8.2: Word adjacency data: Edge descriptions

Since the dataset is directed, every node has an in-degree and an out-degree where in-degree is the number of words that have the current word as its predecessor and out-degree is the number of words that have the current word as its successor. The edge has a source and target indicating which node is the predecessor and which node is the successor for that connection.

The image on the left in Figure 8.1 indicates the distribution of the commonly used nouns and adjectives in the book. The nodes in red are adjectives and the nodes in blue are nouns. The image on the left gives more perspective as the size of the nodes depend on the in-degree and the out-degree. The nodes that is connected to many other nodes is indicated with a size bigger than the node that is less widely used. We can study these nodes further.

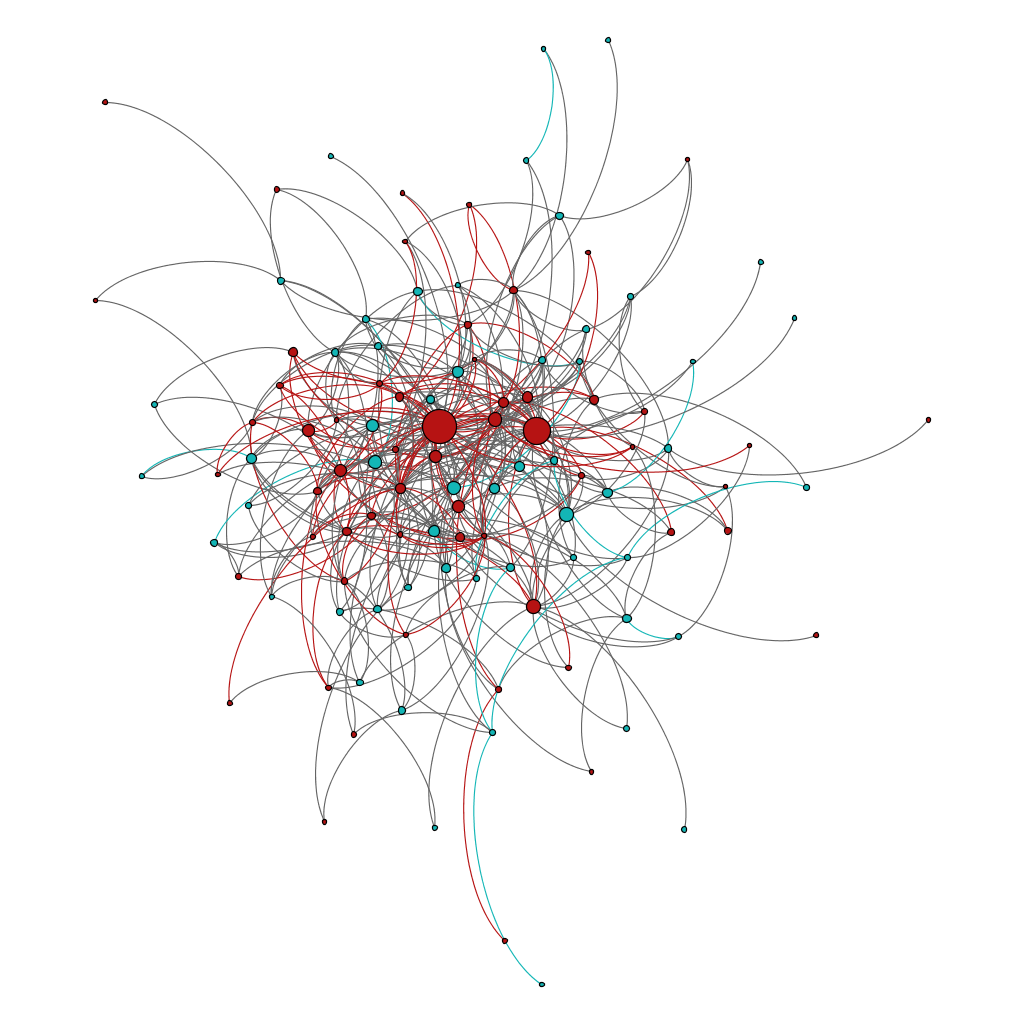
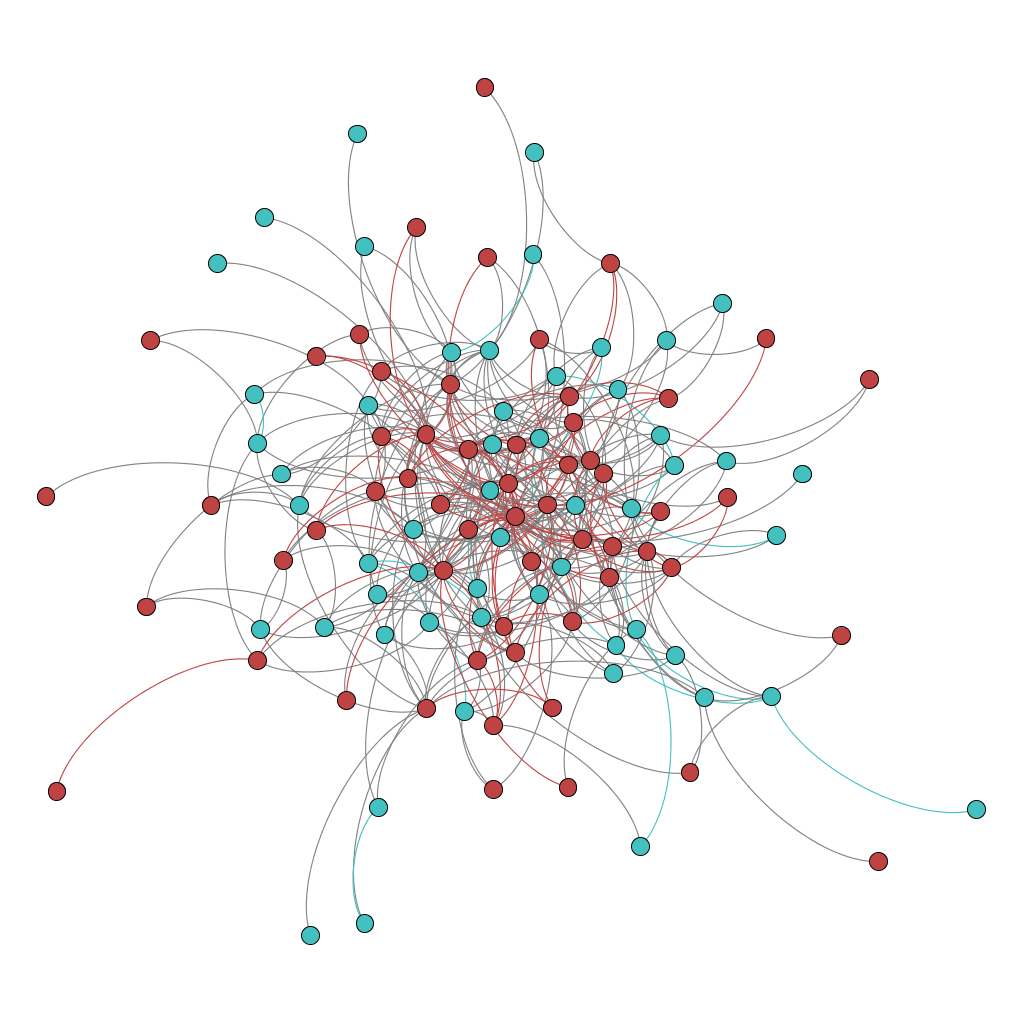


Fig 8.1: Word adjacency network with color based on type.

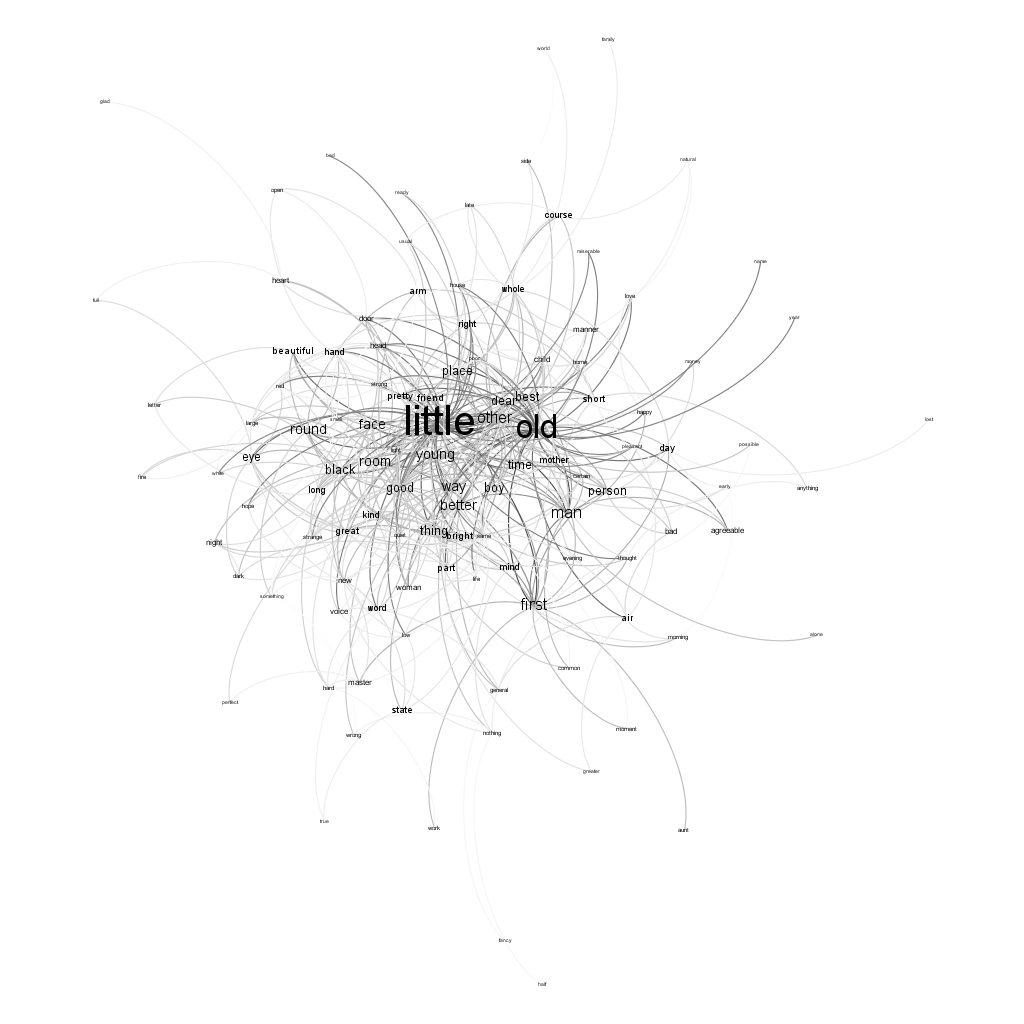


Fig 8.2: Word adjacency network: High degree adjectives

Figure 8.2 shows that the adjectives “little” and “old” are the two commonly used adjectives that were highlighted in figure 8.1. By analyzing the word adjacency dataset we showed that Gephi is powerful in text analysis and can be used in language study and many other applications apart from social networks.

# Books about US Politics

This dataset consists of a network of books about US politics published around the time of the 2004 presidential election and sold by the online bookseller Amazon.com. Edges between books represent frequent co-purchasing of books by the same buyers. The attributes in the dataset include:

|  |  |  |
| --- | --- | --- |
| Sl.no | Attribute | Attribute Explanation |
| 1. | Nodes | The name of the book |
| 2. | Id | Id is a variable that uniquely identifies the nodes on the network |
| 3. | Label | Label is the attribute which defines how the node is labelled on the network |
| 4. | Value | Indicates the category of the book: liberal(l), neutral(n) or conservative(c) |

Table 9.1: Political Books data: Node descriptions

|  |  |  |
| --- | --- | --- |
| Sl.no | Attribute | Attribute Explanation |
| 1. | Source | Indicates the node that acts as the source for the edge |
| 2. | Target | Indicates the node that acts as the target for the edge |
| 3. | Type | Specifies the type of the edge: UnDirected |
| 4. | Id | Id is a variable that uniquely identifies the edge on the network |
| 5. | Weight | The weight of every edge. In this dataset, every edge has an equal weight of 1 |

Table 9.2: Political Books data: Edge descriptions

Fig 9.1 divides the political books sold by Amazon in a particular period into three categories. The nodes in blue are books on conservatives, the nodes in green are books on liberals and the nodes in red are neurtal. The network indicates the tendency of a buyer to buy the same kind of the book is high but there are a few inter-linking books that indicate a different trend.

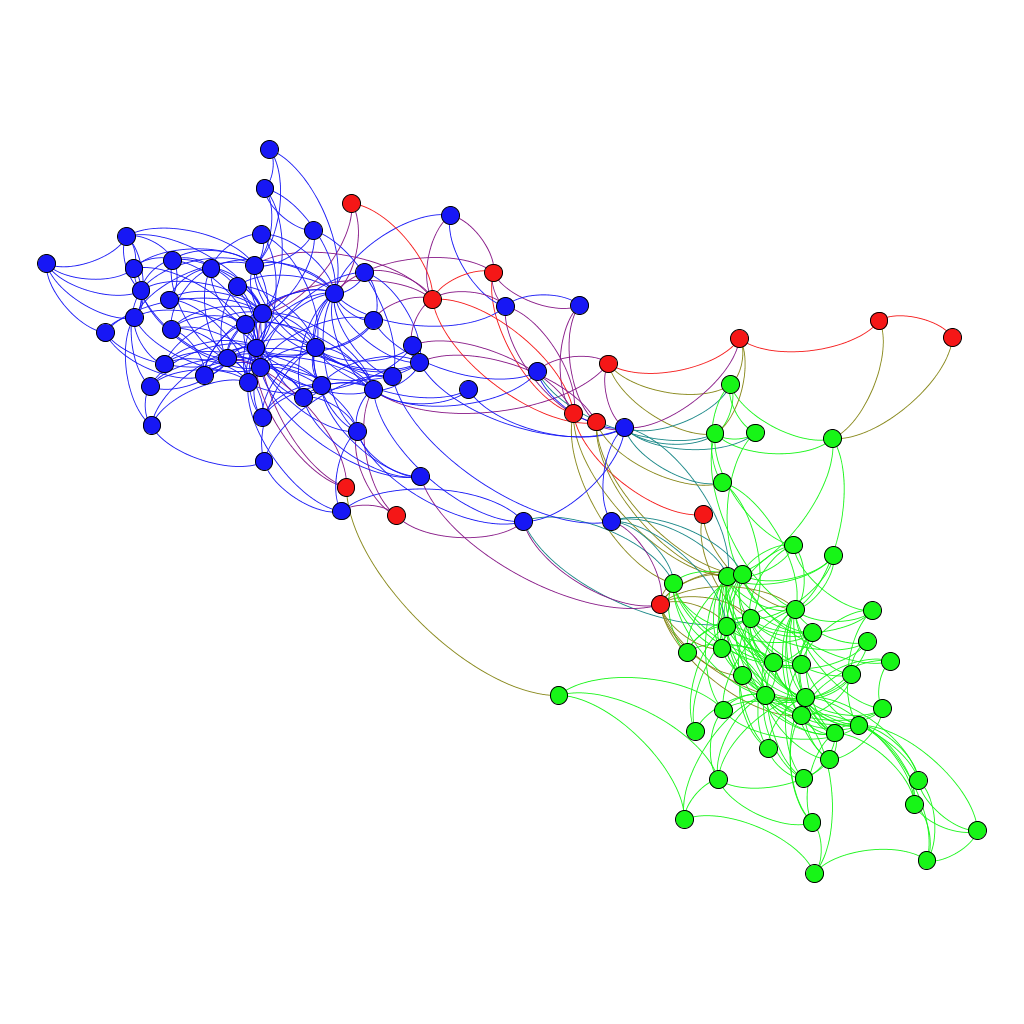


Fig 9.1: Political books network

Fig 9.2 shows the books that a user purchased along with “Surprise, Security and the Americal Experience”. We see that buyers have a equal tendency to purchase a conservative or a liberal book along with this book.

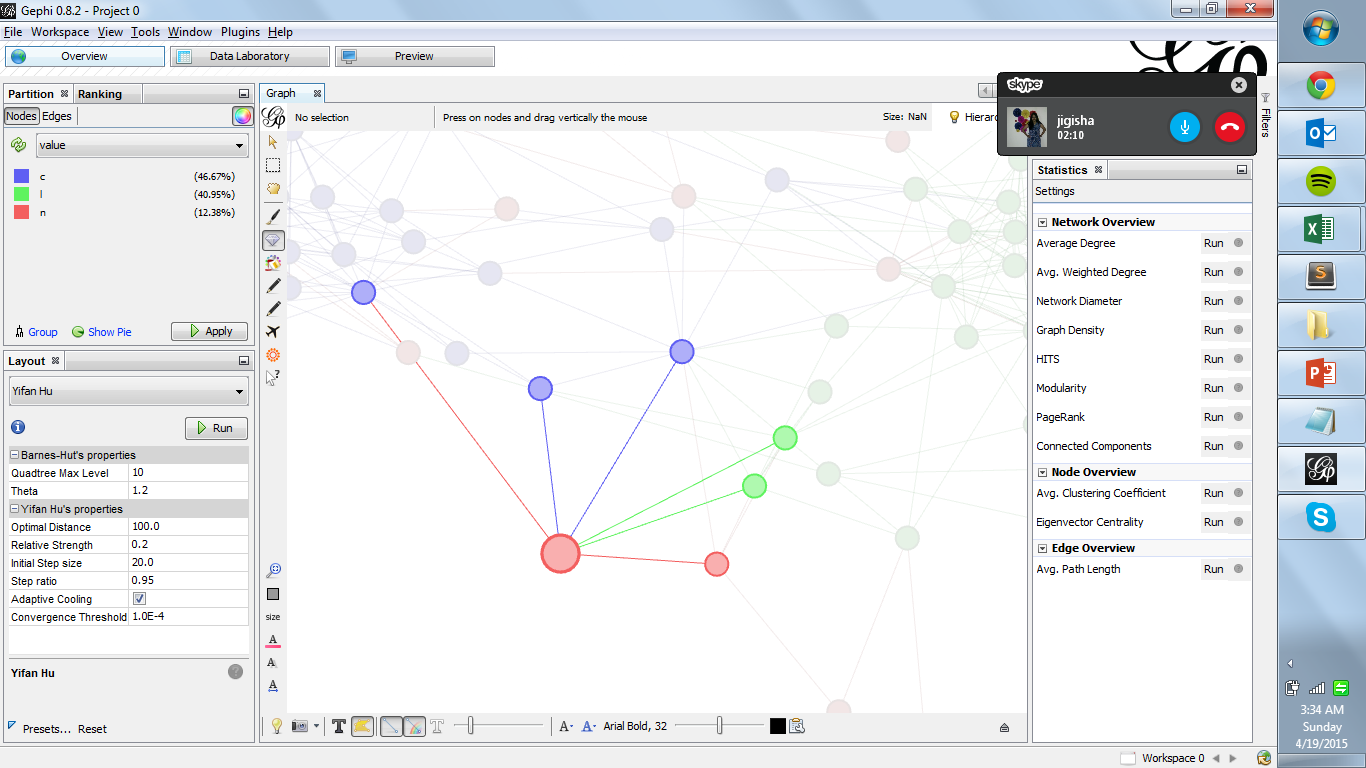
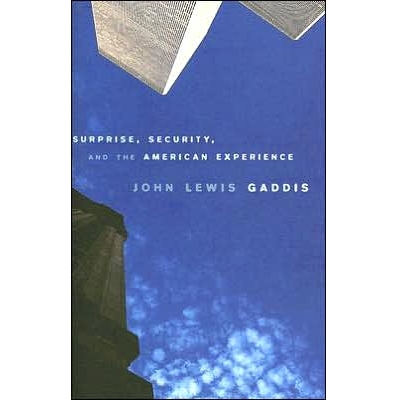


Fig 9.2: Network of the book “Surprise, Security and the Americal Experience

# Political Blogs

A directed network of hyperlinks between weblogs on US politics, recorded in 2005 by Adamic and Glance. Using the source and target attributes of the edges, we analyze the following data:

|  |  |  |
| --- | --- | --- |
| Sl.no | Attribute | Attribute Explanation |
| 1. | Nodes | The URL of the blog |
| 2. | Id | Id is a variable that uniquely identifies the nodes on the network |
| 3. | Label | Label is the attribute which defines how the node is labelled on the network |
| 4. | Value | 0: left or liberal political parties, 1: right or conservative parties |
| 5. | Source | Specifies the source of the blog. Example: Blogarama |

Table 10.1: Political Blogs data: Node descriptions

|  |  |  |
| --- | --- | --- |
| Sl.no | Attribute | Attribute Explanation |
| 1. | Source | Indicates the node that acts as the source for the edge |
| 2. | Target | Indicates the node that acts as the target for the edge |
| 3. | Type | Specifies the type of the edge: Directed |
| 4. | Id | Id is a variable that uniquely identifies the edge on the network |
| 5. | Weight | The weight of every edge. In this dataset, every edge has an equal weight of 1 |

Table 10.2: Political Blogs data: Edge descriptions

The Political blogs considered in the dataset are written around the 2004 presidential election. These blogs distinguished based on political leaning. The two kinds of blogs include left or liberal and right or conservative. Analysis of Gephi network indicated the type and source of the blogs. The web crawls from one political blog to another are also considered and these are depicted by the edges in the graph. In Fig 10.1, the left or liberal blogs are indicated in red and the and right or conservative are shown in blue. We can see that a person is more likely to read a similar kind of a blog rather than switching between libeal and conservative. The graph also shows certain blogs which link these two groups. Such blogs will be considered for further analysis.

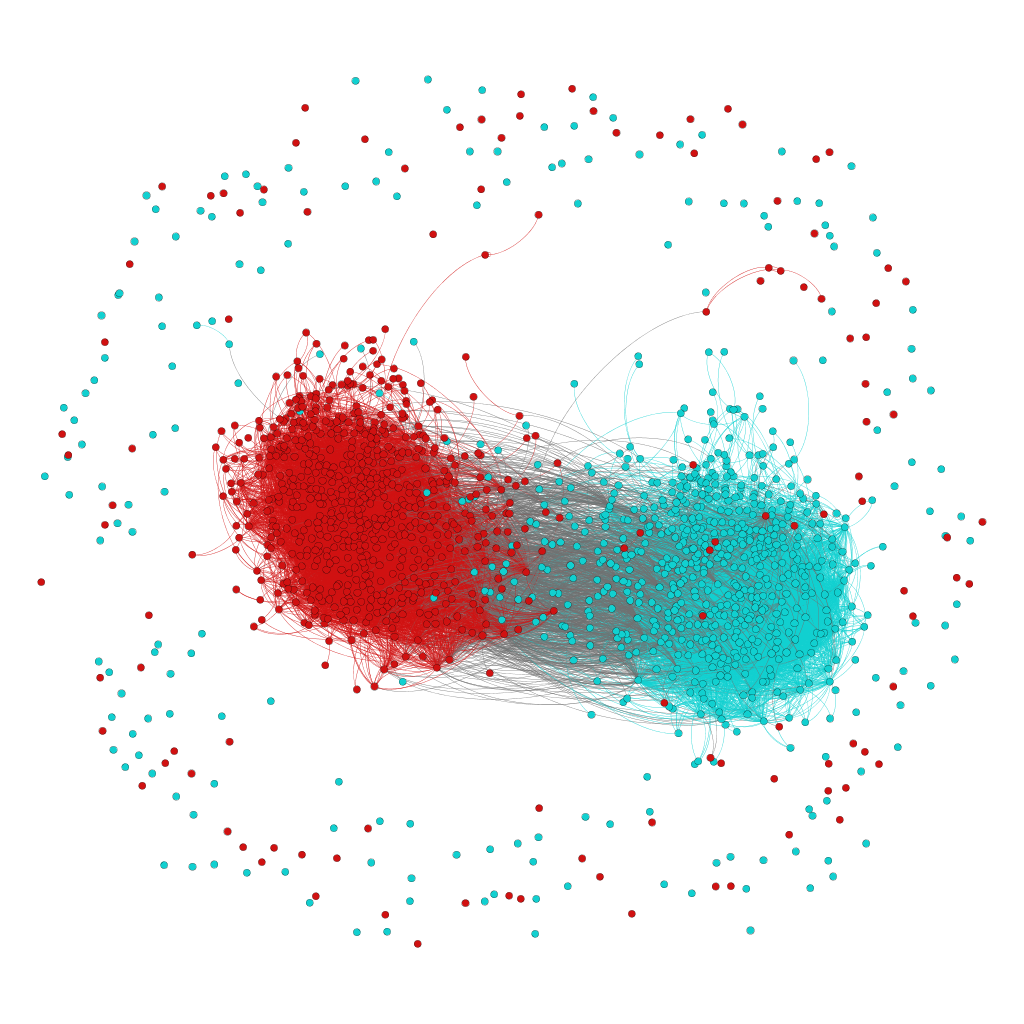


Fig 10.1: Political blogs network with color based on value

In Figure 10.2, the size of the nodes are shown with respect to its degree. The highlighted nodes indicate that users have traversed to this blog from other political blogs more often compared to other blogs.

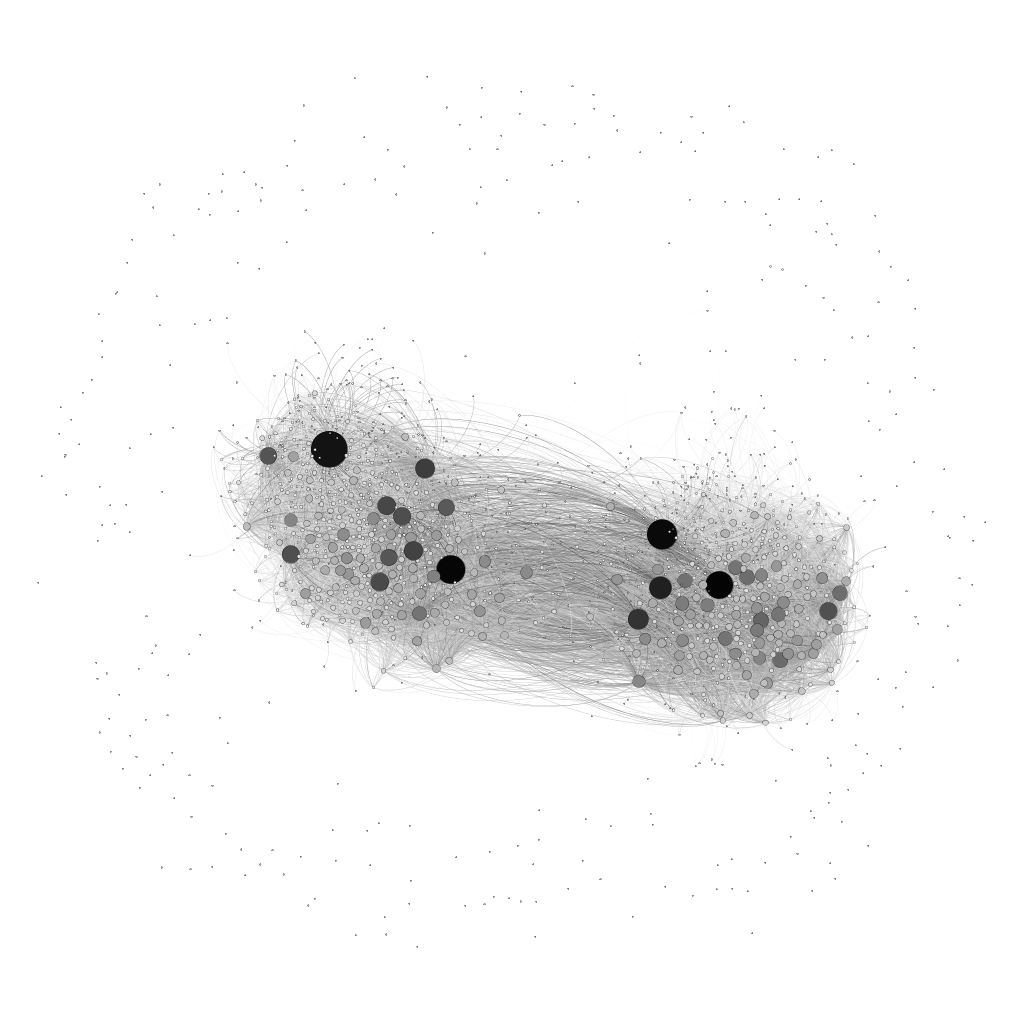


Fig 10.2: Political blogs based on size of node

# Unsupervised learning

Unsupervised learning plays a major role in the detection of communities in social networks. A common used measure is the modularity of a network which based its calculation on the total number of edges in the group and the expected number of edges in the group. However, this method is computationally very expensive because it considers all edges. Hence, there are algorithms developed that will optimally find the solution using a smaller number of sets. In Gephi, the Louvain method is used and in R, many other algorithms are used [6]. We demonstrated the modularity in Gephi during the class demonstrations. This was also done in R for a Facebook dataset using the Fast Greedy Community algorithm.

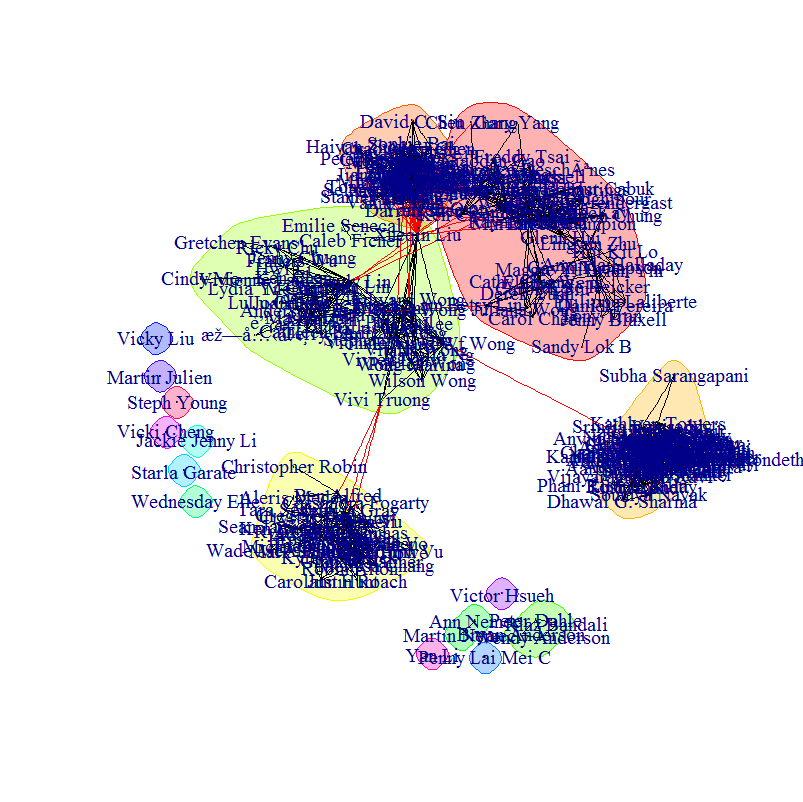


Fig 11.1: Community detection of Facebook data in R

An inherent difficulty with this type of analysis is that nodes in social networks can belong to multiple communities, but this type of clustering only allows each node to belong to one community. Using another package in R: Linkcomm, we can place different nodes into multiple communities. This analysis is shown in figure 11.2.

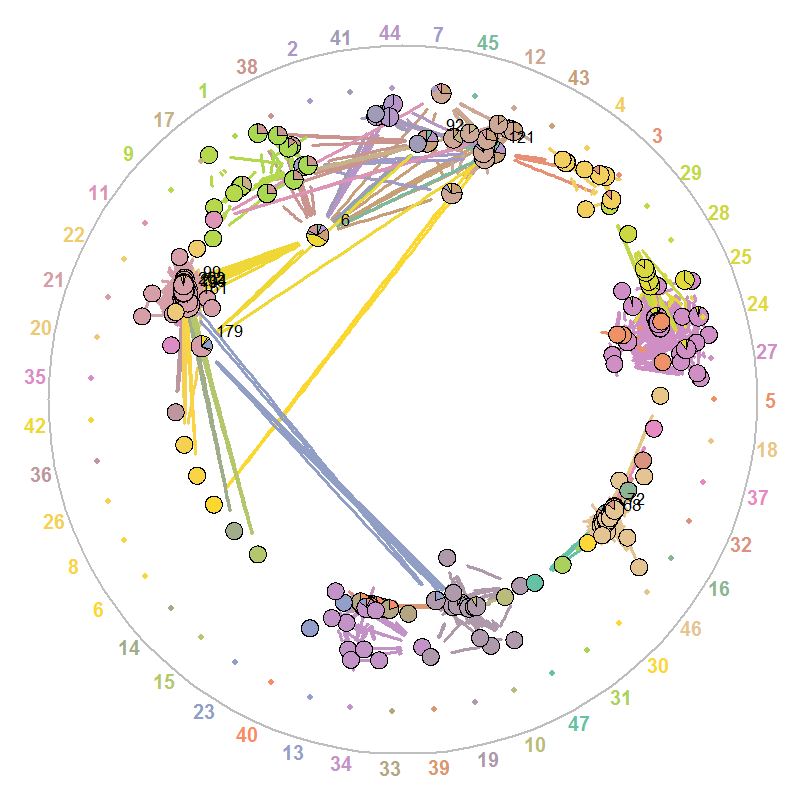
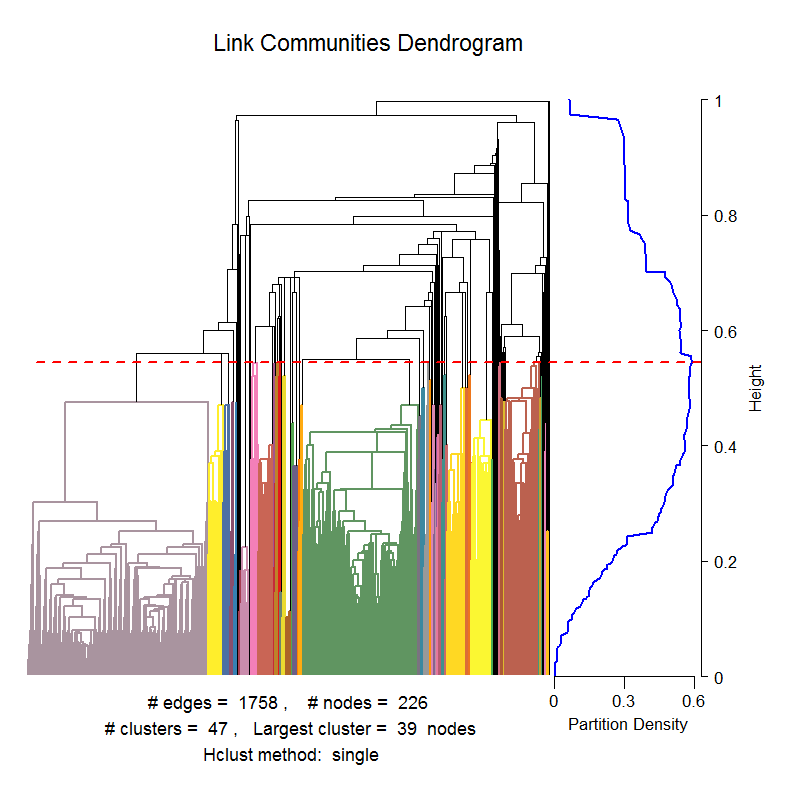


Fig 11.2: Community detection of Facebook data in R

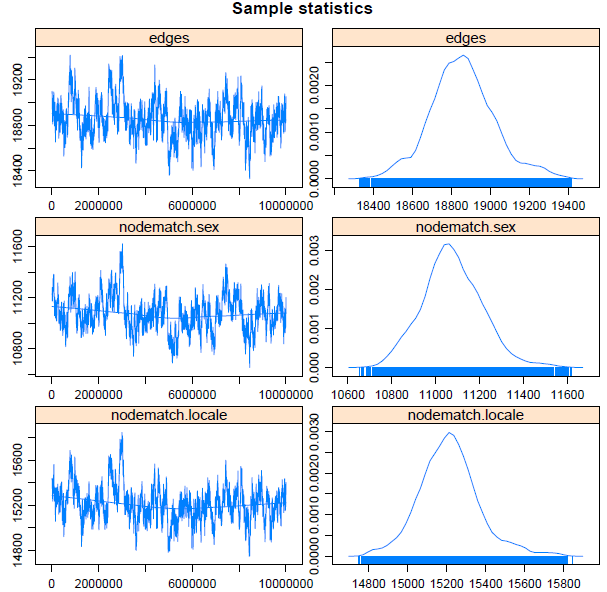
# Random graph models

The **statnet** package in R has been used to do build ERGM models from the Facebook dataset. The models have been created under several parameter constraints leading to different models.

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Likelihood | AIC | BIC |
| Erdos-Renyi (just edges) | -130019.2 | 260040 | 260050 |
| Edge based assortative model (edges + node matching on sex and location) | -129572.7 | 259151 | 259182 |
| Edge and triangle based model | -36095.6 | 72195 | 72216 |
| GWESP based model – No degree 1 term (edges + sex and locale nodes + GWESP term | -134193.3 | 268395 | 268435 |
| GWESP based model – With degree 1 term | -50388.23 | 100786 | 100837 |

Table 12.1: ERGM Model Comparison

A comparison of these models have been done to decide the best model. The edge and triangle based model has the highest likelihood and lowest AIC and BIC values. The GWESP based model with degree 1 term has the next best values. The MCMC diagnostics of the model which show how the error for each of the terms are modeled show that the GWESP based model with degree 1 term has the terms normally distributed while the edge and triangle based model does not. This mean that the GWESP model captures the network better and hence it is chosen.



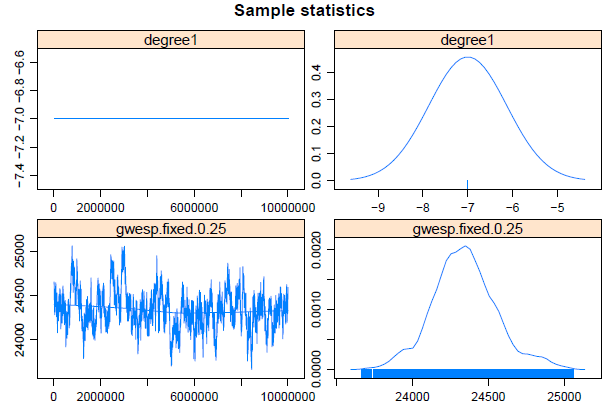
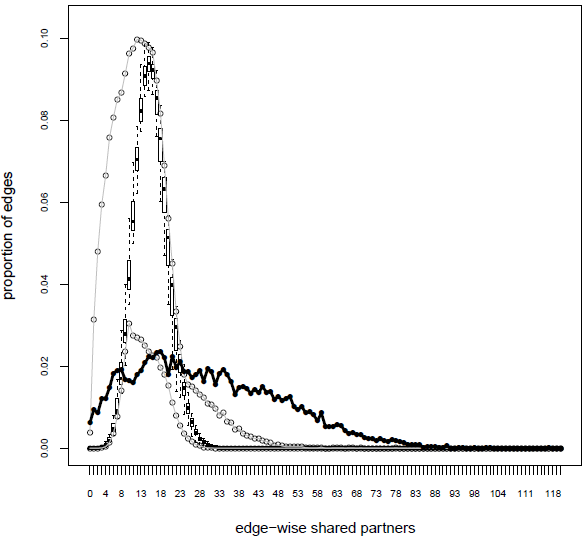
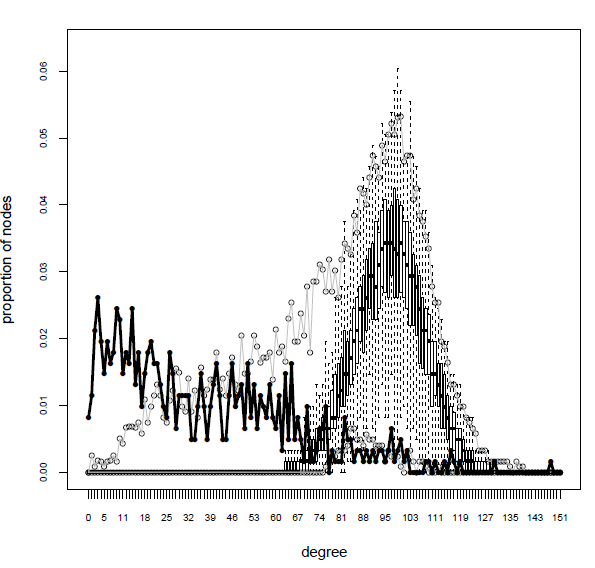


Fig 12.1: MCMC Diagnostics for GWESP based model with degree 1 term

The goodness-of-fit diagnostics of the model shows how the observed and distribution of simulated networks compare with each other. The figures below contain the observed network’s statistics plotted in bold while the simulated networks box-plot distribution is plotted in gray. The observed network is contained in the simulated distribution of networks for the most part.

 Fig 12.2: Goodness-of-Fit Statistics for GWESP based model with degree 1 term

# Conclusions

Network analysis of various dataset: A Facebook network, words in a book, political blogs, and books on politics were analyzed. Exploratory analysis was initially done explore the graphs and then clustering detection algorithms was used to find groups within networks. Finally, linked community analysis was used to detect multiple communities that certain nodes could belong to. In addition, random graphs were used to compare our Facebook networks. In particular, a GWESP based ERGM model with degree 1 term is the best model on the Facebook data as its noise distribution is normal.

# References

1. L. A. Adamic and N. Glance, "The political blogosphere and the 2004 US Election", in Proceedings of the WWW-2005 Workshop on the Weblogging Ecosystem (2005). Thanks to Lada Adamic for permission to post these data on this web site.
2. <http://en.wikipedia.org/wiki/Gephi>
3. M. E. J. Newman, Phys. Rev. E 74, 036104 (2006).
4. <http://www-personal.umich.edu/~mejn/netdata/>
5. Garry Robins∗, Pip Pattison, Yuval Kalish, Dean Lusher, “An introduction to exponential random graph (p\*) models for social networks”
6. <https://sites.google.com/site/findcommunities/>
7. <http://gephi.github.io/users/supported-graph-formats/gml-format/>
8. Golbeck, Jennifer, “Analyzing the Social Web”, 2013