CSP301 Assignment 3

- -

Social Network Analytics

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

25-Nov-2012

Contents

1 The Social Network

We have been provided with live data from a hypothetical social media website. This hypothetical website has 2,500 users with approximately 65,000 edges between them. The edges between the nodes represents the connections between the users (akin to friends on Facebook). This data was provided to us in the form of the file *log-graph.out*.

The data of communication logs was provided to us in the form of files (a file per real-time day) which simulated the period of one year and contained 4 million conversations. We were asked to build a web based analytics dashboard to get a birds-eye view of the activity on the website. This analytics tool could be used by the administrators of the social networking website to analyse data flow patterns and introduce new features accordingly.

We were to submit a report consisting of all the interesting information that we could deduce from our analysis and visualizations.

2 Project Structure

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

We have created a dynamic Web User Interface using the **web2py** web framework. We have also utilized many features of the **D3** functional programming toolkit of Javascript. All client side scripting has been done in **Javascript** while the dynamization of the graphs has been by using Ajax for communication with the server side scripts written in web2py. Also central to this assignment was the usage of a Database Management System in order to store the humongous amounts of data being released everyday. For this we used the **SQLite** RDBMS which is natively supported in web2py and easily portable.

An instance of the Web UI may be run by importing the web2py application 'analytics' from our github repo and going to the URL "http(s)://web2py:port/analytics".

For analysis, we mainly used programs written in Python and various features of Excel. Also a lot of interesting insights about the data was made available to us by the visualizations which we created for the purpose of the dashboard. Many interesting results were obtained, a few of which are stated in the subsequent pages.

In this project we have used 6 main languages – Javascript, Ajax, HTML, CSS, Python and SQL.

3 Database Architecture

A very important part of this assignment on Big Data Visualization and Analysis was the use of a proper Database Architecture Model. We could not save all the data as it is, because it would be extremely inefficient in terms of both space as well as time to get the data from such a huge data repository everytime one wants to view a dynamic graph based on the data. Thus it was important to model an architecture which keeps all important information in the form of summary values. Also a proper organisation system was needed that could allow speedy access of data stored in the database. The RDBMS used by us in this assignment is SQLite 3.7.

Our database consists of the following 10 tables:

1. **db.nodes**: This table is populated with the list of 2500 nodes, their locations(under db.nodes.location), and the id of the cluster(db.nodes.clusterid) to which they belong. The id of node used in the log-graph file is put under db.nodes.uid but every node also has an auto-incremental id (db.nodes.id) which we have used throughout our code. So this table also provides us with a mapping between the uid and id.

2. **db.edges**: This table is basically the list of all edges that exist in the friendship graph. Every undirected edge $n1 \longleftrightarrow n2$ is represented as db.edges.node1 and db.edges.node2. However, in order to avoid data redundancy (which exists in the graph given to us), we have made sure that $db.edges.node1 < db.edges.node2$. The column db.edges.totalv stores the total amount of communication that has taken place on this edge.

3. **db.cledges**: In this table we store data about the undirected edges of the cluster-graph ('cledge'). db.cledges.cluster1 and db.cledges.cluster2 store the ids of the clusters the cledge connects. For this table, clusters play the same role as nodes in db.edges. In order to reduce data redundancy, we have always kept $db.cledges.cluster1 < db.cledges.cluster2$. db.cledges.totalv stores the total volume of communication that has occured across this cledge. db.edges.cledgeid stores the unique id associated with that cledge.

4. **db.chourwise**: This table stores information related to all the communication on the website in the last 10 days (more precisely 240 hours before the last entry in the latest log-comm file). db.chourwise.time stores the timestamp for the beginning of the hour in which the communication took place, db.chourwise.topic stores the name of the topic of communication, db.chourwise.cledge stores the id of the cledge on which the communication took place and db.chourwise.volume stores the volume of communication on that topic in that hour across the specific cledge. Cledges were chosen instead of edges in order to optimise the size. All features of the table are quite desirable so that we can query and get information about almost anything.

5. **db.cdaywise**: The structure of this table is similar to that of db.chourwise. However the db.cdaywise.volume stores the volume of communication on that topic on that *day* across the specified cledge. This table stores the information related to any communication within the past 9 months (assumed to be 30 days each).

6. **db.cweekwise**: In a similar manner, this table stores the communication data in a weekwise fashion since eternity.

7, 8, 9, 10. **db.locedges, db.lhourwise, db.ldaywise, db.lweekwise**: These tables are analogous to db.cledges, db.chourwise, db.cdaywise and db.cweekwise. Just that these tables store data considering the clustering to be purely location-based. Hence the clusterid's here are simply the name of the location the cluster represents. These tables play an important role in carrying out location based analysis of the communication data.

This structure of the data base allowed us to carry out our analysis about the social network effectively. The data for the recent past (10 days) was extremely precise. However, as we moved further into the past, our data resolution reduced and got more and more imprecise. This allows the analytics framework to make optimum usage of the available space and at the same time maintains a fine balance between data preciseness. Also to carry out queries quickly and effectively, we have used indexing in our database's tables.

While maintaining a database was important, it was also important that if at any time a file log is uploaded on the given URL then it is automatically downloaded and inserted into the database without any explicit commands being given to the system for processing. For this, we used a Python Script which pings the URL for new files once every hour (we did this using the Windows Task Scheduler) and if it finds that a new file has been uploaded on the site, it downloads it and inserts its contents into the database, all the time maintaining the time integrity of the data.

4 Visualization

4.1 Timeline Depiction

The aggregate communication activity of the Social Network for eternity has been displayed in the form of a timeline. The user of the dashboard can select the time frame as per his wishes to view the communication activity within that time frame. This information is provided to the user with the granularity of a day.

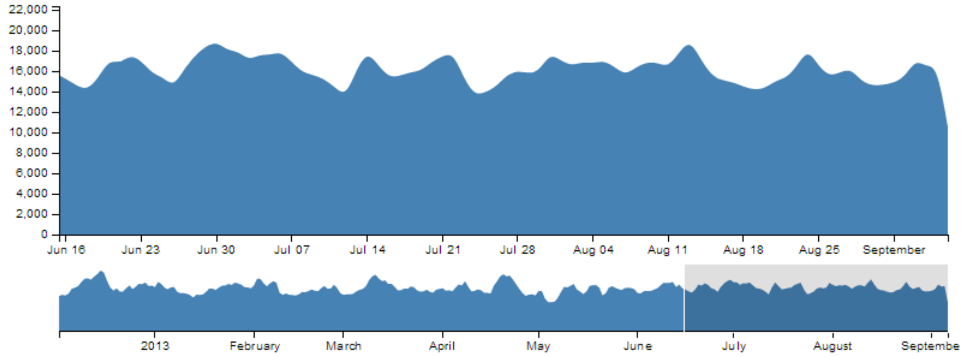


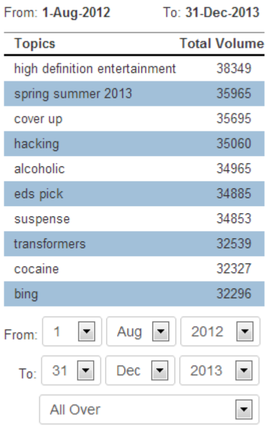
Fig 4.1.1: Timeline of Aggregate Communication Activity

Also, we have integrated into this graph functionality to view the communication on a specific topic at different resolutions - maybe hourwise (for 10 days), daywise (for 9 months) or weekwise (for eternity). This can provide us amazing insights about the time profile of a specific topic in the social network (discussed in detail later).

4.2 Top 10 Topics

The 'top 10 topics' tool gives the list of the topics on which most discussions were held. We have provided filters for locations and clusters so as to find the most talked about topics within specific locales. Also, one can choose the period of observation of topic trends as per his wishes. The top 10 topic functionality can be used by the administrators of the social network to provide a more customized browsing experience to the users.

A screenshot of the top 10 topics page is as below.



From: **1-Aug-2012** To: **31-Dec-2013**

Topics	Total Volume
high definition entertainment	38349
spring summer 2013	35965
cover up	35695
hacking	35060
alcoholic	34965
eds pick	34885
suspense	34853
transformers	32539
cocaine	32327
bing	32296

From: 1 Aug 2012

To: 31 Dec 2013

All Over

Fig 4.2.1: Top 10 topics of communication as on September 6, 2013

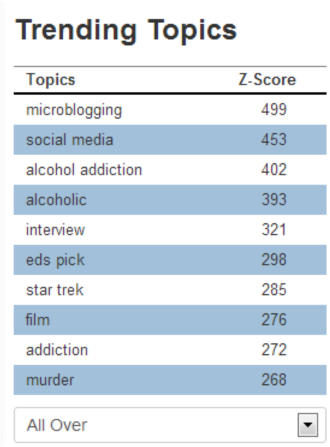
4.3 Trending Topics

Central to a social networking site is the concept of a Trending Topic - a topic on which there has been a significant rise in activity in the recent past. Websites like Twitter and Topix incorporate these statistics into their sites and are used to recommend topics for communication to the users.

The top 10 trending topics of communication have taken into account data from the past 3 days. The trends for the topics have been calculated using *z-score values* which indicates by how many standard deviations an observation or datum is above or below the mean.

We have also provided functionality to filter the trending topics by location and cluster. This could provide options for the administrators of the social networking website to provide topical suggestions to users on the basis of their locations or clusters i.e. connections.

A screenshot of the trending topics tool incorporated in our analytics website is shown below.



Trending Topics

Topics	Z-Score
microblogging	499
social media	453
alcohol addiction	402
alcoholic	393
interview	321
eds pick	298
star trek	285
film	276
addiction	272
murder	268

All Over

Fig 4.3.1: Trending Topics on Aug 14 2013

4.4 Cluster - Cluster Communication Mashup

A mashup was made in order to understand the distribution of volume of communication between the clusters and locations. Shown below is the cluster × cluster matrix for the communication betweeen nodes upto the hypothetical date September 6, 2013.

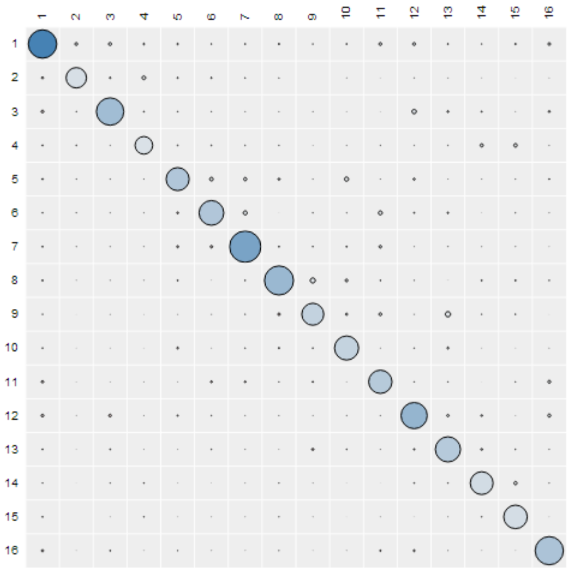


Fig 4.4.1: Cluster × Cluster Mashup

In this visualization tool, the sizes of the circles have been set according to the volume in the lower half (the half with red circles) while those in the upper half (the half with blue circles) have been sized according to their correlation coefficient values. The colour shades assigned to each circle follow the reverse order *i.e.* volumewise in the upper half and coefficient wise in the lower half. The correlation coefficient μ is defined as

$$\mu = \frac{V_C^{ij}}{V_C^i + V_C^j}$$

where V_C^{ij} represents the volume of communication between the i^{th} and j^{th} clusters and V_C^i represents the total cumulative communication in cluster i.

As can be seen from the image above, *clusters tend to communicate more within themselves than with other clusters.* Infact, it is found that the total volume of the intra-cluster communication for the analysed period is 3313954 while the net communication occurring over the period is 4712190. Thus it can be seen that 70.33% of all communication takes place within the clusters. This implies that most people tend to talk to the same people more often rather than new ones.

In the above mashup, we saw that the intra-cluster communication overshadowed the inter-cluster communication. In order to provide opportunity to analyse inter-cluster data flow, we have provided functionality to hide the intra-cluster communication. The inter-cluster communication pattern is as shown below.

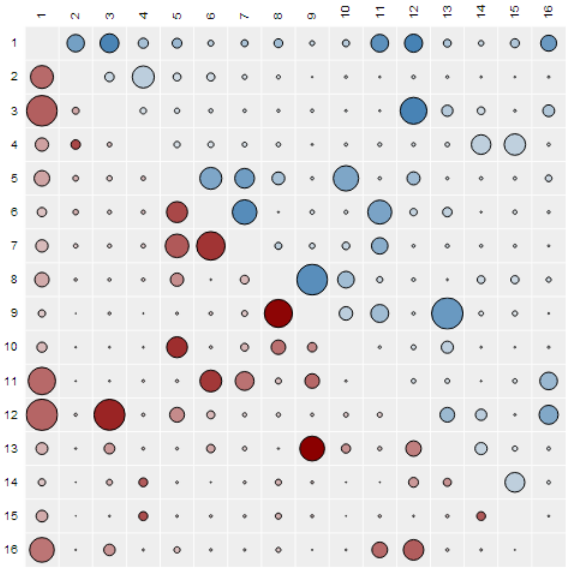


Fig 4.4.2: Inter-Cluster Communication

As can be seen from this graph, the volume of communication is high between the clusters which have larger circles in the cell corresponding to them in the grid. However, it is to be noted that this volume is still very low when compared to the intra-cluster communication.

The location × location matrix for the same period is as shown below.

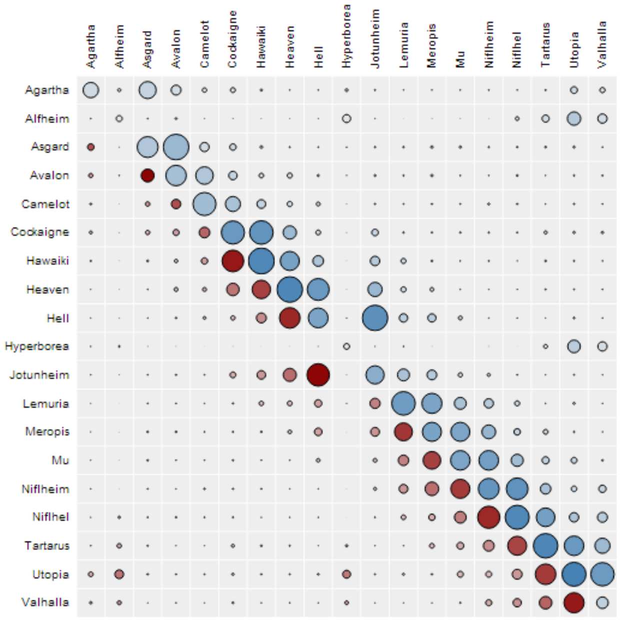


Fig 4.4.3: Location × Location Mashup

The image above clearly shows that within the data generation model *people tend to talk more often to the people in the same location or to those from the locations closer (lexicographically) to them.* Also, it was found that the total volume of the intra-location communication for the analysed period is 1454226 while the net communication occurring over the period is 4712190. Thus it can be seen that only 30.86% of all communication takes place within the same location. However if we consider the volume of communication occurring between people from the locations which are alphabetically closer to them, we see that the volume rises to 2627869 which is equivalent to 55.77% of the total communication. This implies that most communications take place between people from the same location or people from the locations closer to them alphabetically.

One notable exception to the rule stated above is Hyperborea which appears to be an extremely boring location and not too many communications take place in that location.

4.5 Stacked Area Graph

We have also provided a Stacked Area Graph which provides the user with data about how much communication on each topic has taken place within the past 10 days. This gives the administrator a fair idea of the relative popularity of the topics which have been in vogue in the recent past.

We have provided controls to the user that he can select a specific topic or multiple topics in order to view their relative performance on the social network. This feature is extremely useful as it would allow analysts to predict oncoming trends. For example, let us consider the scenario of the US Presidential elections. If one was to run the Stacked Area Chart on a real Social Network near the time of elections and select Barack Obama and Mitt Romney as his two topics for comparison, he could easily see trends as to who is currently on the minds of people. This could be used by the professional analysts to predict the results of the elections.

Also, there are controls to view the stacked area chart as a Stream or in the expanded view. The Stream view visualizes the communication occurring on the Social Network as a Stream and presents data in a fluid timescale format. This presents the administrators of the social network an option to view the data flow on each topic through their social network. The expanded view of the graph can be used to view the proportion of communication that each topic had on each day. Also, the tooltip for the graph has been programmed to reveal all communication stats about the selected topic for the specified day.

A screenshot of the Stacked Area Chart is as shown below:

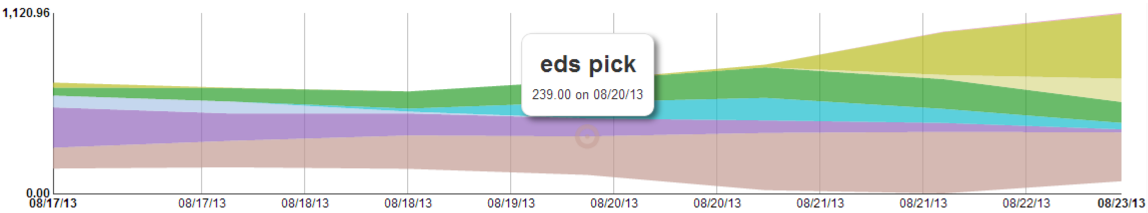


Fig 4.5.1: Stacked Area Chart in Stream View for 10 days preceding Aug 23 2013

4.6 Hierarchical Bar Chart

We have also included in our analytics website a Hierarchical Bar Chart. At the first level of hierarchy, it provides information about how much communication has taken place on topics active within the past week. At the second level of the hierarchy it provides a day-wise break up of all communication that has taken place on that topic in the past week.

At the third (and final) level of hierarchy, a location based break up of communication on the selected topic on the specified day is provided to the user.

A screenshot of the tool is as below (for the period of Aug 10 to Aug 16, 2013):

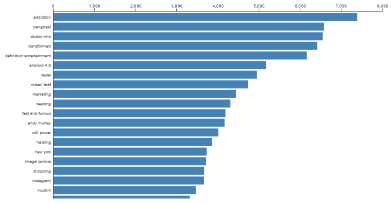


Fig 4.6.1: Topic Level Hierarchy

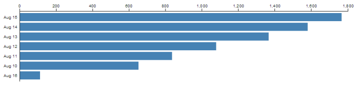


Fig 4.6.2: Day Level Hierarchy for Addiction

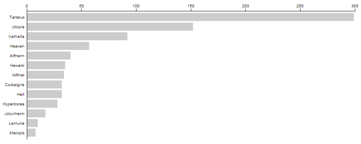


Fig 4.6.3: Location Level Hierarchy for Aug 11

As can be seen from the images above, at all levels of hierarchy, the y axis is sorted by the volume of communication which provides easy access to top topics and locations to the user. This graph also provides the administrators of the social network tools to analyse the data flow patterns on specific topics on specific days within different locations.

4.7 Motion Timeline Graph

In this feature of our analytics website, we have tried to depict how a selected topic has spread through the clusters in the social network over a selected time period. The x-axis has the list of clusters in the graph. The size of the circle corresponds to the volume of communication that occurred on that day in that cluster on the selected topic. The height of the circle above the x-axis gives the total volume of communication that has occurred on the selected topic within that cluster. This tool can be used to analyse the dynamics of the social network that how does a specific topic spread in the graph. Attributes such as the span of a topic on the graph can be well understood by this tool.

A screenshot of the tool is as shown below.

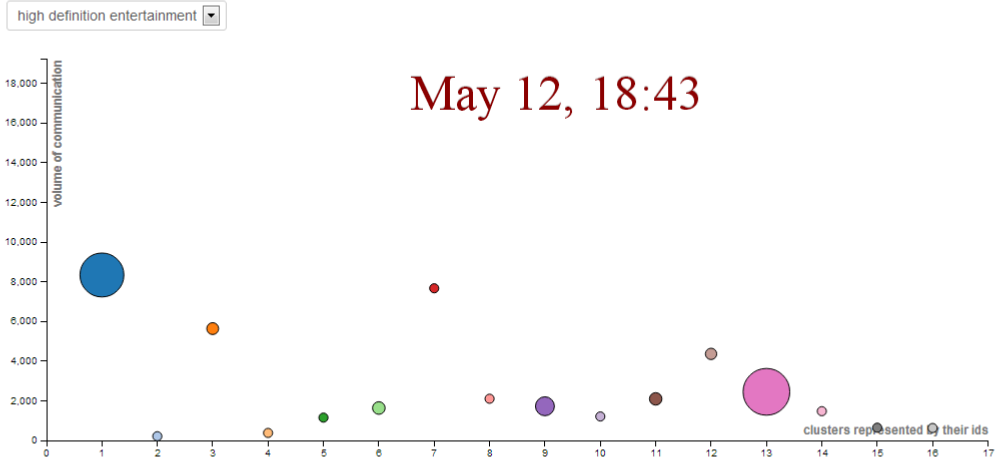


Fig 4.7.1: Screenshot of the Motion Timeline Graph

5 Analysis

5.1 Graph Structure

The social network graph has 2500 nodes (corresponding to 2500 users) and 50052 unique undirected edges (corresponding to connections between the users). A force directed layout of the graph can be seen as below.

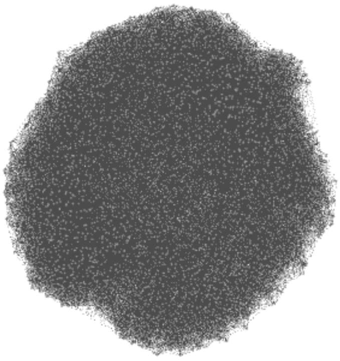


Fig 5.1.1: Force Directed Visualization of the Social Network Graph

As we can see, this does not lead to the formation of any distinct clusters in the graph and so, more sophisticated algorithms for finding clusters within the social network are required. A general analysis of the Graph was carried out and the results are depicted below. We are extremely thankful to the creators of the Software Tool **Gephi** for allowing us to carry out a lot of analysis quickly and with ease.

Degree Distribution

The degree distribution of the social network graph is as shown below.

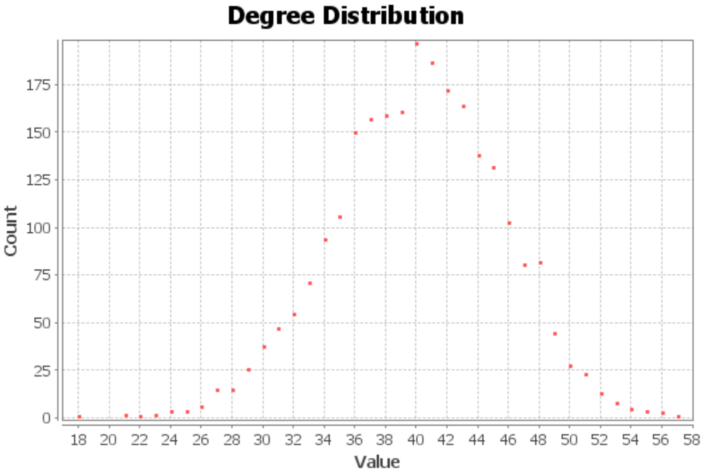


Fig 5.1.2: Degree Distribution in Social Network Graph

The mean degree of a node is 40.04. Also it can be easily seen that the degree distribution of nodes follows an almost Gaussian Bell Curve. An interesting thing to note here is that

$$40.04 \approx \frac{2 \times 50052}{2500}$$

i.e. there are $\frac{NK}{2}$ edges where N is the number of nodes and K is the mean degree of the nodes.

This seems to point to the *Watts-Strogatz random graph generation model.* This is also reinforced from the degree distribution curve we have plotted above. There is a pronounced peak at $k = K$ and decays exponentially for large $|k - K|$.

Graph Density

The density ρ of a graph $G = (V, E)$ measures how many edges are in set E compared to the maximum possible number of edges between vertices in set V. In fact the mathematical formula for calculating the density of an undirected graph is given by

$$\rho = \frac{2|E|}{|V|\left(|V| - 1\right)}$$

The density of the social network graph given to us is very low and is equal to 0.016. This mimics real life society to a huge extent because a person generally knows people within a particular community and not the entire world. Thus a low graph density is to be expected.

Graph Eccentricity

The **eccentricity** ϵ of a vertex v is the greatest geodesic distance between v and any other vertex. It can be thought of as how far a node is from the node most distant from it in the graph. The **radius** of a graph is the minimum eccentricity of any vertex. The **diameter** of a graph is the maximum eccentricity of any vertex in the graph. That is, the greatest distance between any pair of vertices.

The Eccentricity Distribution of the given social network is as shown in the figure below.

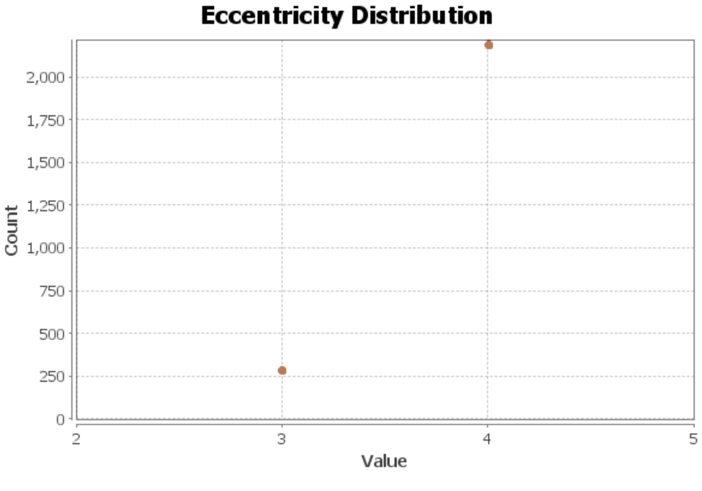


Fig 5.1.3: Eccentricity Distribution in Social Network Graph

It can be easily seen that all nodes have an eccentricity of either 3 or 4. Thus, the radius of the graph is 3 while the diameter of the network is 4. The mean path length between any two nodes in the graph is 2.66. The relatively small diameter is a pointer to the fact that this graph obeys the small world networking model wherein each node is reachable from every other within a small number of hops.

Betweenness Centrality

Betweenness centrality quantifies the number of times a node acts as a bridge along the shortest path between two other nodes. Vertices that have a high probability to occur on a randomly chosen shortest path between two randomly chosen vertices have a high betweenness centrality measure.

Mathematically, it is defined as

$$C_B(v) = \sum_{s \neq v \neq t \in V} \sigma_{st}(v)$$

where $\sigma_{st}(v)$ is the number of shortest paths that pass through v from node s to node t.

The betweenness centrality distribution for the social network graph provided to us is as shown below.

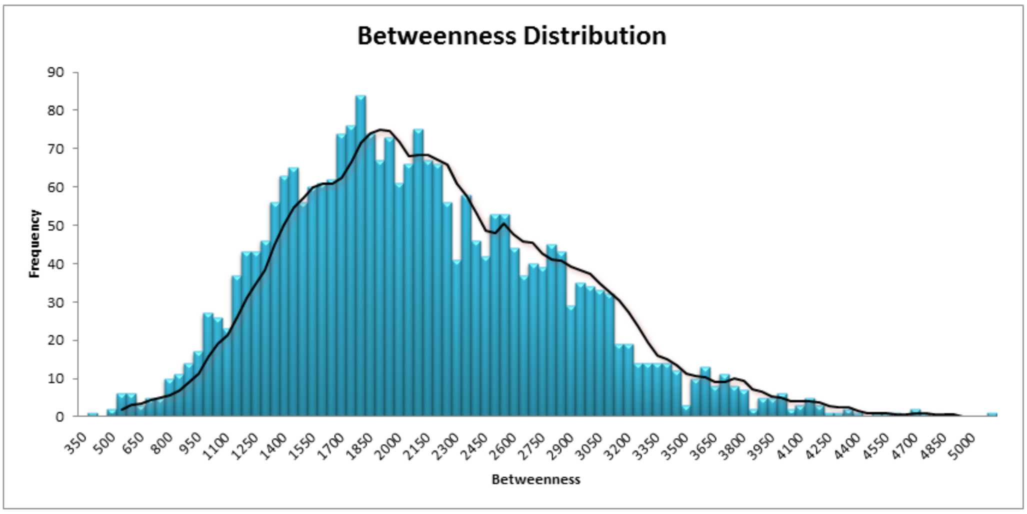


Fig 5.1.4: Betweenness Distribution in Social Network Graph

As can be seen from the graph above, the betweenness centrality measure follows a normal distribution - characteristic of a randomly generated Watts-Strogatz graph.

Closeness Centrality

This graph metric measures the average distance from a given starting node to all other nodes in the network. Closeness can be regarded as a measure of how fast it will take to spread information from s to all other nodes sequentially.

The closeness centrality distribution of the graph is as shown below.

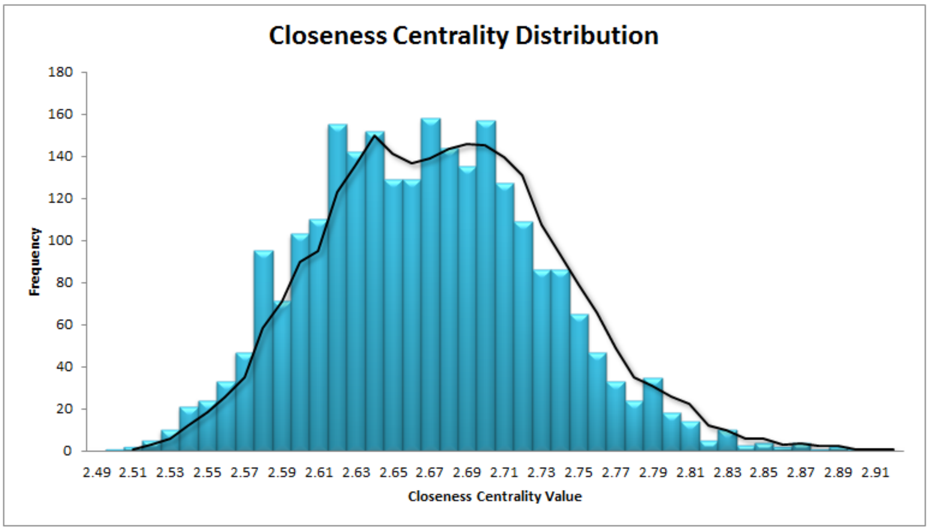


Fig 5.1.5: Closeness Distribution in Social Network Graph

As can be seen from the above graph, the closeness distribution of the network follows an almost perfectly Gaussian Distribution with a peak occurring near the 2.67 mark. Thus it means that spreading of information must be very fast in the social network given to us.

Clustering Coefficient & Triangles

We tried to find the number of triangles in the social network graph. It was found that the total number of triangles in the graph was 87990. The minimum number of triangles a node was involved in was 27 and the maximum number was found to be 200. A histogram showing the distribution of number of triangles a node is involved in, is shown below.

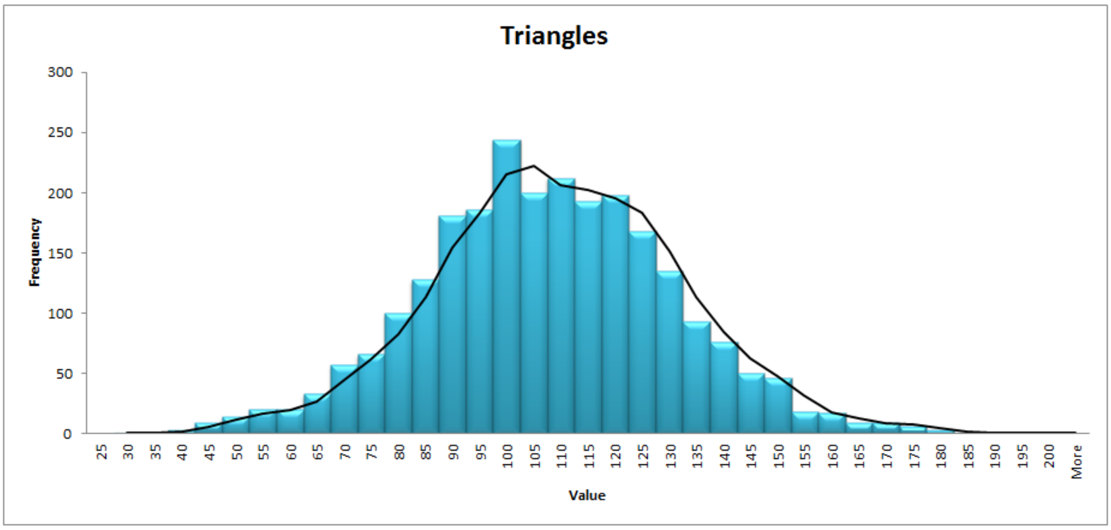


Fig 5.1.3: Triangle Distribution in Social Network Graph

In undirected networks, the Clustering Coefficient C_n of a node n is defined as

$$C_n = \frac{2 \times e_n}{k_n(k_n - 1)}$$

where k_n is the number of neighbors of n and e_n is the number of connected pairs between all neighbors of n.

It was found that the minimum Clustering Coefficient for the graph was 0.07 whereas the maximum Clustering Coefficient was 0.31. In fact a histogram showing the Clustering Coefficient Distribution of all nodes was plotted and is shown below.

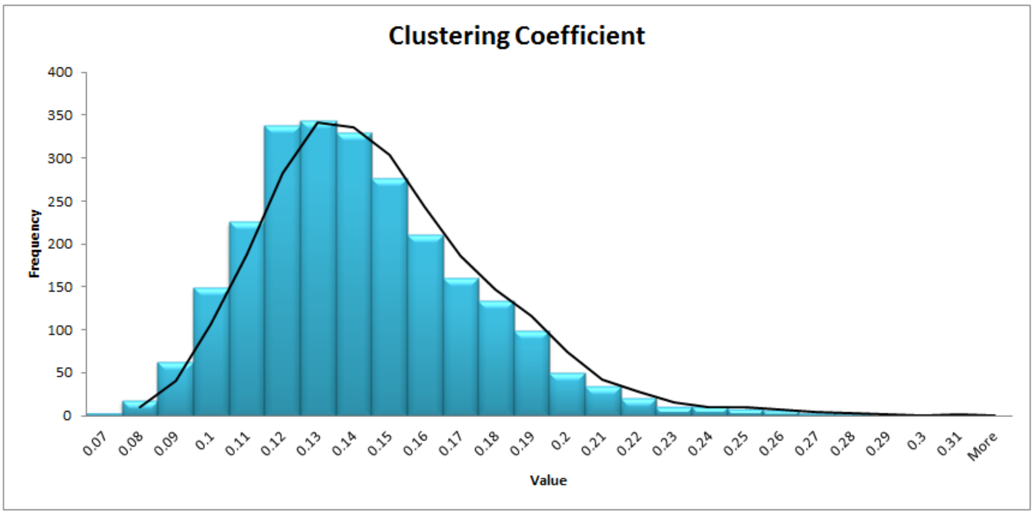


Fig 5.1.7: Clustering Coefficient Distribution in Social Network Graph

The global network clustering coefficient was found to be 0.138.

5.2 Topic Selection Mechanisms

5.2.1 Topic Popularity Distribution

Till September 6, 2013, there have been 289 topics of communication in the given social network. The Topic Popularity Distribution of the social network is as below.

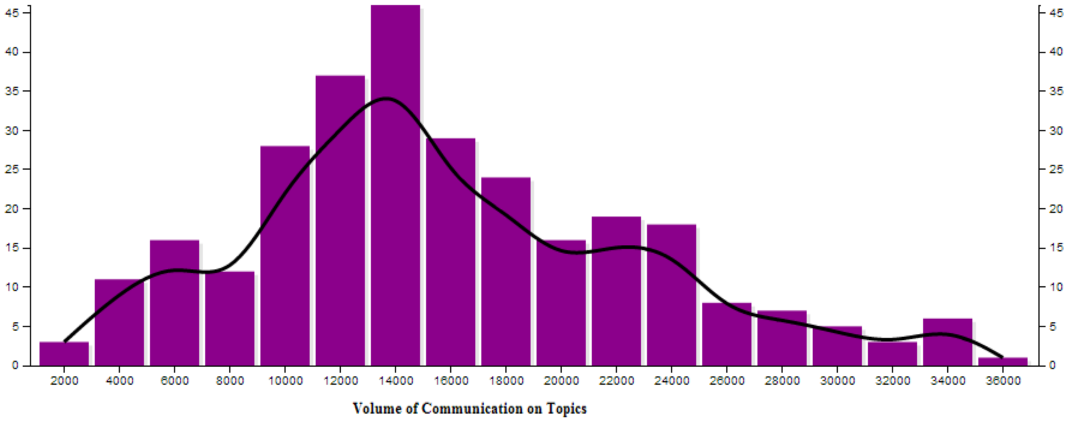


Fig 5.2.1: Topic Popularity Distribution

As can be easily seen from the trendline, the popularity distribution of the graph is almost NORMAL.

5.2.2 Topic Time Profiles

We tried to analyse the time profiles that various topics follow while it spreads through the social network using the timeline tool after setting it on the hourwise mode. We realised that the volume of communication on a particular topic was being controlled by two specific models.

1. **Stepped Volume Change**
   We realised that many topics show a stepwise increment and decrement in their volume. Generally, there are 3-4 such steps of increment and decrement which span a total period of 5-6 days.

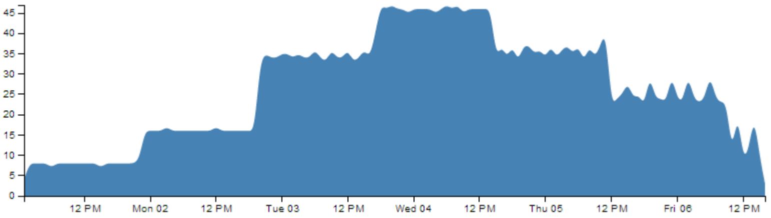


Fig 5.2.2: Communication Volume of Nissan Leaf from September 1 to September 6, 2013.

2. **Spiked Volume Change**
   Many topics show a spiked increase in volume over time. The volume of communication on the topic seems to be almost constant when there is a sudden spike in volume. The amplitude of the spike keeps on increasing as time progresses. This spike- based time profile continues for bot 5-6 dayswith around 10 spikes in one cycle.

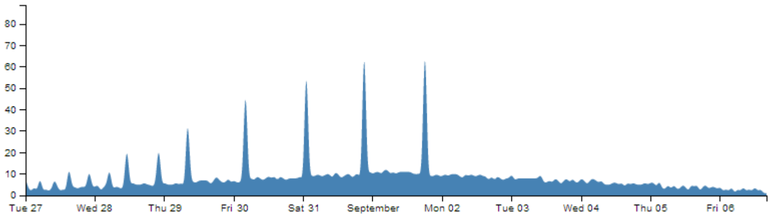


Fig 5.2.3: Communication Volume of Video Games from August 27 to September 6, 2013.

3. **Spike and Step Volume Change**
   It was also seen that many topics exhibited a volume increase model which combined the features of both the above mentioned models.

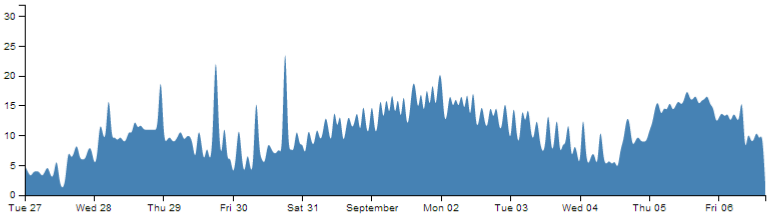


Fig 5.2.4: Communication Volume of Acting from August 27 to September 6, 2013.

From the analysis above, we are able to see that there are basically four states that a topic may be in:

1. It may be nascent with nobody talking about that topic at all.

2. It may be in the middle of a stepwise change in volume.

3. It may be in the middle of a spiked change in volume.

4. It may exhibit features which combine both spiked and stepwise changes in volume.

In the underlying model, this is probably controlled by assigning two boolean variables to each topic - one corresponding to stepwise change and the other to spiked change. The state of these variables determines what is the current volume of communication on that topic.

5.3 Cluster Analysis

Cluster Finding

Finding clusters in the social network was a difficult task and required much work from our side. We read about, and tried out, many clustering mechanisms and algorithms in order to obtain the optimal clustering. Some of the clustering algorithms that we tested included *Force Directed Clustering, TopGC, r-MCL clustering, Graclus, alpha-beta Clustering* etc. However, the clusters we obtained using these algorithms were not as good as we wanted them to be and also contained a few nodes that were common to multiple clusters. Subsequently, we decided to use some softwares like *Gephi and Pajek.* However, even these clusters were not satisfactory in our view.

Basically, we calculated a clustering ratio for the clusters - a measure of how well connected the cluster is in itself relative to the outside world, defined as the ratio of the number of neighbours of the node within the cluster to its degree, summed over all the nodes in the network.

Eventually, we came across a web tool hosted at *http://www.mapequation.org*. Using this tool we obtained 16 distinct clusters. After computing its clustering ratio and satisfying ourselves of its suitability, we decided to use these 16 as our clusters. The largest cluster we obtained was cluster 1 of size 346 whereas the smallest cluster obtained was cluster 10 of size 97. We found that this graph was a complete graph with $^{16}C_2 + 16 = 136$ edges between the nodes.

The online clustering tool implements a random walk clustering algorithm - a variant of the MCL clustering algorithm and provides us with truly differentiated communities within the social network.

A visualization of the clusters that we obtained is as follows

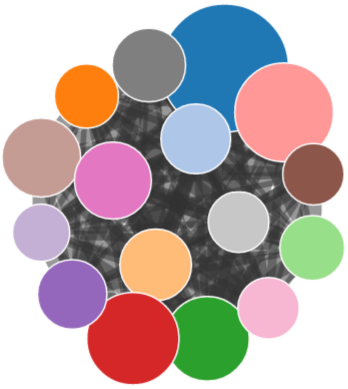


Fig 5.3.1: 16 clusters in the Social Network

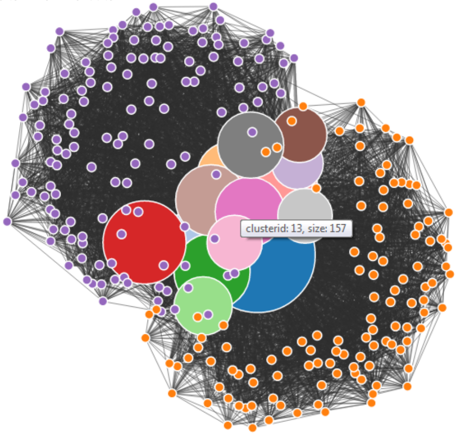
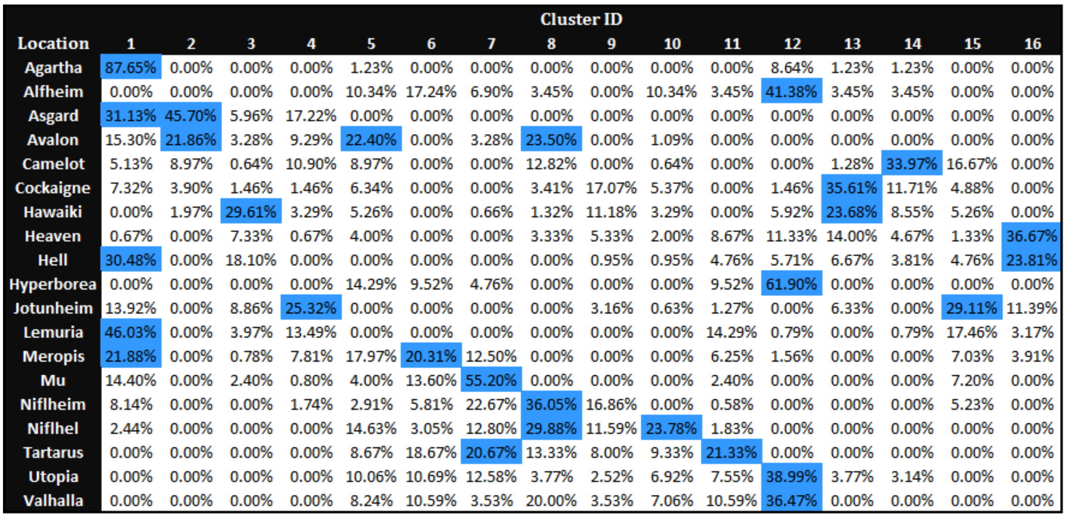


Fig 5.3.2: Members of 2 nodes along with the other 14 clusters

Cluster - Location Correlation

We wanted to find out if there existed any relation between the location and the cluster allocated to a user. For this, a table containing details about the location and cluster distribution of people was made.
Data for cluster distribution of people from a specific location is as follows:

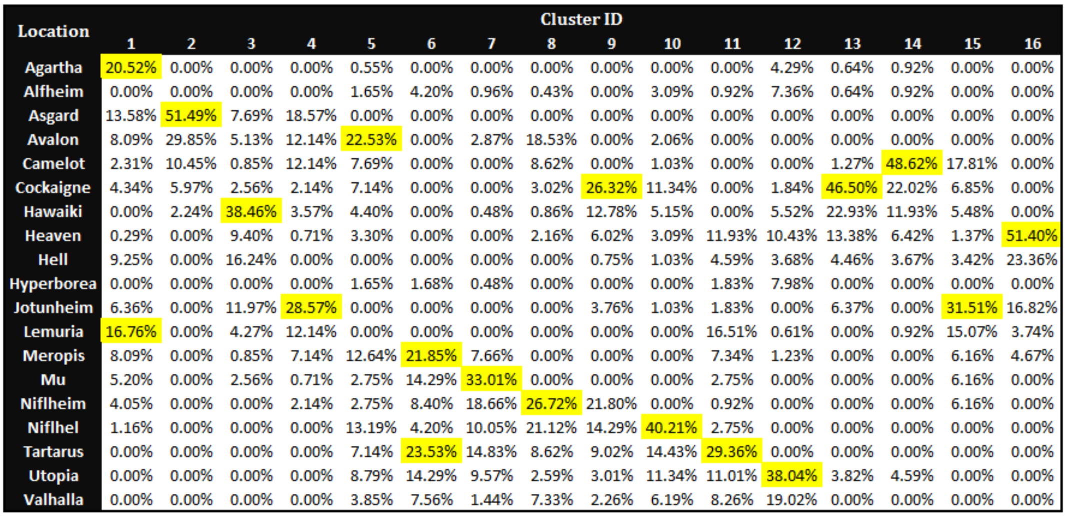


Location	Cluster ID															
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Agartha	87.65%	0.00%	0.00%	0.00%	1.23%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	8.64%	1.23%	1.23%	0.00%	0.00%
Alfheim	0.00%	0.00%	0.00%	0.00%	10.34%	17.24%	6.90%	3.45%	0.00%	10.34%	3.45%	41.38%	3.45%	3.45%	0.00%	0.00%
Asgard	31.13%	45.70%	5.96%	17.22%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Avalon	15.30%	21.86%	3.28%	9.29%	22.40%	0.00%	3.28%	23.50%	0.00%	1.09%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Camelot	5.13%	8.97%	0.64%	10.90%	8.97%	0.00%	0.00%	12.82%	0.00%	0.64%	0.00%	0.00%	1.28%	33.97%	16.67%	0.00%
Cockaigne	7.32%	3.90%	1.46%	1.46%	6.34%	0.00%	0.00%	3.41%	17.07%	5.37%	0.00%	1.46%	35.61%	11.71%	4.88%	0.00%
Hawaiki	0.00%	1.97%	29.61%	3.29%	5.26%	0.00%	0.66%	1.32%	11.18%	3.29%	0.00%	5.92%	23.68%	8.55%	5.26%	0.00%
Heaven	0.67%	0.00%	7.33%	0.67%	4.00%	0.00%	0.00%	3.33%	5.33%	2.00%	8.67%	11.33%	14.00%	4.67%	1.33%	36.67%
Hell	30.48%	0.00%	18.10%	0.00%	0.00%	0.00%	0.00%	0.00%	0.95%	0.95%	4.76%	5.71%	6.67%	3.81%	4.76%	23.81%
Hyperborea	0.00%	0.00%	0.00%	0.00%	14.29%	9.52%	4.76%	0.00%	0.00%	0.00%	9.52%	61.90%	0.00%	0.00%	0.00%	0.00%
Jotunheim	13.92%	0.00%	8.86%	25.32%	0.00%	0.00%	0.00%	0.00%	3.16%	0.63%	1.27%	0.00%	6.33%	0.00%	29.11%	11.39%
Lemuria	46.03%	0.00%	3.97%	13.49%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	14.29%	0.79%	0.00%	0.79%	17.46%	3.17%
Meropis	21.88%	0.00%	0.78%	7.81%	17.97%	20.31%	12.50%	0.00%	0.00%	0.00%	6.25%	1.56%	0.00%	0.00%	7.03%	3.91%
Mu	14.40%	0.00%	2.40%	0.80%	4.00%	13.60%	55.20%	0.00%	0.00%	0.00%	2.40%	0.00%	0.00%	0.00%	7.20%	0.00%
Niflheim	8.14%	0.00%	0.00%	1.74%	2.91%	5.81%	22.67%	36.05%	16.86%	0.00%	0.58%	0.00%	0.00%	0.00%	5.23%	0.00%
Niflhel	2.44%	0.00%	0.00%	0.00%	14.63%	3.05%	12.80%	29.88%	11.59%	23.78%	1.83%	0.00%	0.00%	0.00%	0.00%	0.00%
Tartarus	0.00%	0.00%	0.00%	0.00%	8.67%	18.67%	20.67%	13.33%	8.00%	9.33%	21.33%	0.00%	0.00%	0.00%	0.00%	0.00%
Utopia	0.00%	0.00%	0.00%	0.00%	10.06%	10.69%	12.58%	3.77%	2.52%	6.92%	7.55%	38.99%	3.77%	3.14%	0.00%	0.00%
Valhalla	0.00%	0.00%	0.00%	0.00%	8.24%	10.59%	3.53%	20.00%	3.53%	7.06%	10.59%	36.47%	0.00%	0.00%	0.00%	0.00%

Fig 5.3.3: Locationwise distribution of clusters

The cells in blue represent the cluster in which people from the specified location are present in huge numbers. For example, 87.65% of all people from Agartha are present in Cluster 1.

Data for location distribution of people from a specific cluster is as follows:



Location	Cluster ID															
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Agartha	20.52%	0.00%	0.00%	0.00%	0.55%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	4.29%	0.64%	0.92%	0.00%	0.00%
Alfheim	0.00%	0.00%	0.00%	0.00%	1.65%	4.20%	0.96%	0.43%	0.00%	3.09%	0.92%	7.36%	0.64%	0.92%	0.00%	0.00%
Asgard	13.58%	51.49%	7.69%	18.57%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Avalon	8.09%	29.85%	5.13%	12.14%	22.53%	0.00%	2.87%	18.53%	0.00%	2.06%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Camelot	2.31%	10.45%	0.85%	12.14%	7.69%	0.00%	0.00%	8.62%	0.00%	1.03%	0.00%	0.00%	1.27%	48.62%	17.81%	0.00%
Cockaigne	4.34%	5.97%	2.56%	2.14%	7.14%	0.00%	0.00%	3.02%	26.32%	11.34%	0.00%	1.84%	46.50%	22.02%	6.85%	0.00%
Hawaiki	0.00%	2.24%	38.46%	3.57%	4.40%	0.00%	0.48%	0.86%	12.78%	5.15%	0.00%	5.52%	22.93%	11.93%	5.48%	0.00%
Heaven	0.29%	0.00%	9.40%	0.71%	3.30%	0.00%	0.00%	2.16%	6.02%	3.09%	11.93%	10.43%	13.38%	6.42%	1.37%	51.40%
Hell	9.25%	0.00%	16.24%	0.00%	0.00%	0.00%	0.00%	0.00%	0.75%	1.03%	4.59%	3.68%	4.46%	3.67%	3.42%	23.36%
Hyperborea	0.00%	0.00%	0.00%	0.00%	1.65%	1.68%	0.48%	0.00%	0.00%	0.00%	1.83%	7.98%	0.00%	0.00%	0.00%	0.00%
Jotunheim	6.36%	0.00%	11.97%	28.57%	0.00%	0.00%	0.00%	0.00%	3.76%	1.03%	1.83%	0.00%	6.37%	0.00%	31.51%	16.82%
Lemuria	16.76%	0.00%	4.27%	12.14%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	16.51%	0.61%	0.00%	0.92%	15.07%	3.74%
Meropis	8.09%	0.00%	0.85%	7.14%	12.64%	21.85%	7.66%	0.00%	0.00%	0.00%	7.34%	1.23%	0.00%	0.00%	6.16%	4.67%
Mu	5.20%	0.00%	2.56%	0.71%	2.75%	14.29%	33.01%	0.00%	0.00%	0.00%	2.75%	0.00%	0.00%	0.00%	6.16%	0.00%
Niflheim	4.05%	0.00%	0.00%	2.14%	2.75%	8.40%	18.66%	26.72%	21.80%	0.00%	0.92%	0.00%	0.00%	0.00%	6.16%	0.00%
Niflhel	1.16%	0.00%	0.00%	0.00%	13.19%	4.20%	10.05%	21.12%	14.29%	40.21%	2.75%	0.00%	0.00%	0.00%	0.00%	0.00%
Tartarus	0.00%	0.00%	0.00%	0.00%	7.14%	23.53%	14.83%	8.62%	9.02%	14.43%	29.36%	0.00%	0.00%	0.00%	0.00%	0.00%
Utopia	0.00%	0.00%	0.00%	0.00%	8.79%	14.29%	9.57%	2.59%	3.01%	11.34%	11.01%	38.04%	3.82%	4.59%	0.00%	0.00%
Valhalla	0.00%	0.00%	0.00%	0.00%	3.85%	7.56%	1.44%	7.33%	2.26%	6.19%	8.26%	19.02%	0.00%	0.00%	0.00%	0.00%

Fig 5.3.4: Clusterwise distribution of locations

The cells in yellow represent the cluster in which people from the specified clusters are present in huge numbers. For example, 20.52% of the people in Cluster 1 are from Agartha.

As can be easily seen through these tables, the clusters and locations are highly correlated. Most people from a specific location lie within the same cluster. Also, most people in a specified cluster are from the same location.

We also did a χ^2-*test* in order to test the correlation between the locations and clusters. Our Null hypothesis for this was that there is no correlation between locations and clusters. The χ value obtained was 6052.951 and the degree of freedom was 270. This was much greater than the critical χ value of 380.77 for a p-value of 0.00001. Thus with a confidence of more than 99.99999%, we can reject the Null Hypothesis and say that there is a specific correlation between people's locations and their clusters.

This clearly depicts that users in the same cluster are likely to be from the same location. Also, users from the same location are generally within the same cluster. This is in consonance with our expectations for a real life social network and shows that the clusters obtained are optimal.

5.4 Topic Spread Mechanism

The spreading mechanism of a topic can be understood properly by analysing it on the Motion Graph and also taking into account the connections and general communication trends from the Mashup incorporated into our visualization. An amazing thing about the spreading of most topics in this Social Network is that a topic may be instantiated in any location, but once introduced in a location, it generally spreads through the network via locations which are lexicographically closer to itself. Thus, if a topic originates in Hawaiki, it spreads first to Heaven and Cockaigne and then gradually to Hell and Camelot and so on.

A notable exception to this rule, however, is the locale of Hyperborea which acts as a bottleneck in the spread of a topic through the social network. The general pattern is however followed for all other nodes in the graph.

When we come to clusters, it is seen that mostly a topic remains limited within the cluster of origin for most of the time. However, there is a very small trickle effect into the other clusters.
An interesting thing about the topic spread model was that despite a normal distribution of centralities, most topics took a lot of time to spread throughout the network. In order to analyse this, we tried to plot a histogram for the amount of communication occurring on each edge. The result obtained is shown below.

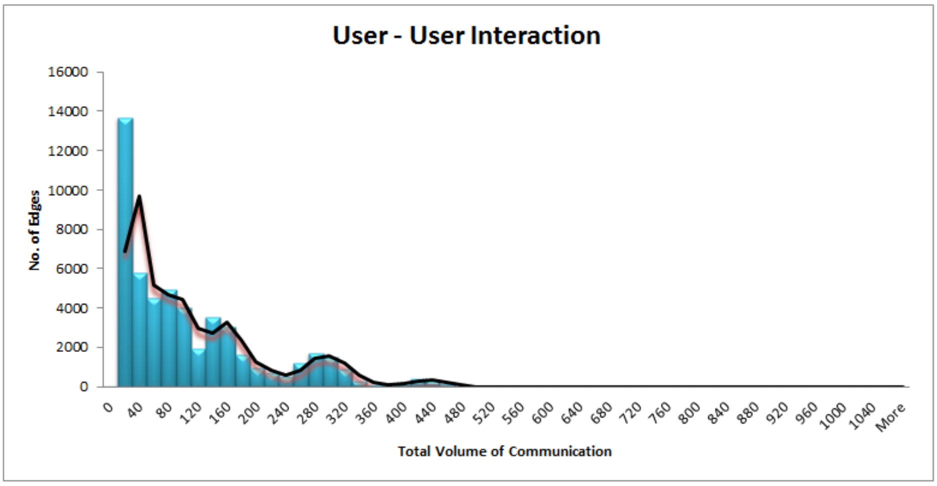


Fig 5.4.1: Histogram for communication on edges

As we can see, despite good connectivity in the graph, most edges have low activity. Thus, the spread of the topic in the social network is slow.

6 Underlying Model

The analysis and the visualizations that we have made were instrumental in uncovering the following insights about the model of generation:

1. The social network given to us may have been generated using a Watts-Strogatz Random Graph Generation Model.

2. The volume of communication of a particular topic is changed in two ways - spiked change or stepwise change. Various combinations of these two volume-change models are possible and are seen consistently in the social network graph.

3. A topic in the social network spreads by going from its place of origin to the places which are lexicographically closer to it. Hyperborea, however, acts as a bottleneck in this propagation of the topic through the network.

4. A topic tends to stay within the cluster that it originated in. There is a small trickle effect on other clusters with which it has relatively high communication.

Report developed in LaTeX