**NETWORK ANALYSIS OF STACK OVERFLOW**

# Team Members:

# 16BCE0891 (Nakul Lakhotia )

# 16BCE0995 ( Raghav Bhatia )

# 16BCE0815(Prakhar Paliwar)

**Report submitted for the Final Project Review of**

**Course Code: CSE3021 – Social and Information Networks Slot: A2 + TA2**

**Professor: Rajkumar R.**

# Introduction:

In the past decade, social network analysis has come out to be a very efficient way of mapping and depicting the behavior of various web users on the basis of different types of interactions between them. These interactions are depicted graphically with nodes as users and edges as the respective interactions between those users. SNA as abbreviated has opened doors for discovering interactions on the internet. On the basis of observations made, giants like Facebook, Twitter and Google develop algorithms to enhance user experience by adding more features to promote different types of such interactions. SNA has been a critical logical research center in administration, human science, and social insurance for quite a long time. Be that as it may, conventional SNA depended vigorously on manual strategies, for example, polls and meetings, to build social systems. The outcomes are generally static and the extension has been restricted. Today, laborers every now and again communicate carefully. Due to the constraint of important information procurement, particularly from scholastic, more deliberate continuous largescale explores are as yet holding up to be done to use the sufficient information that are made by individuals' communications, such as email, call logs, content informing, report stores, furthermore, web 2.0 instruments in associations. It is exceptionally hard to lead huge scale cross methodology or multi-modality examination, e.g., looking at how individual system structures influence income. This hole is dangerous, on the grounds that the writing on authoritative systems experiences the same shortages that a significant part of the informal community writing does. It needs to concentrate on little, static systems, on the grounds that electronic follows dwell in heterogeneous spots. The scope of this project is on one of such huge systems, Stack Overflow where various users interact with each other through by answering questions and commenting on questions and answers. The website is directed towards asking and answering programming related technical questions. There is a programming editor available for each user. User can put forward his doubts with his code included in the question. Also the users answering the question can also include their code for correct reference of the questioner so that the doubt can be cleared and hence programming can be made interactive. Answers are given up votes by the registered users on the basis of how much they find it relevant and accurate for the question asked. The answers are displayed in the decreasing order of up votes on the web page. The data obtained is in the form of source and destination nodes, each pair defining a unique edge. With more than ten million edges, we consider the first thousand interactions in all three types of interactions available. The aim of the research is to show graphically the interactions and calculate various metrics of the network thus formed. We aim to calculate the weighted degree distribution, clustering coefficient distribution, average clustering coefficient and the average graph density. Later, we conclude with the practical life significance of these metrics and depict how they can be used by the websites to know about their network.

# Literature Review Summary Table:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| ***Authors and Year (Reference)*** | ***Title (Study)*** | ***Concept***  ***/ Theoretic al model/ Framewo rk*** | ***Methodolog y used/ Implementa tion*** | ***Dataset details/ Analysis*** | ***Relevant Finding*** | ***Limitatio ns/ Future Research***  ***/ Gaps identified*** |
| *John Tang,Micro Musolesi, Cecilia Mascolo,Latora* | *Tempor al Distance Metrics for Social Network Analysis* | *Temporal network analysis* | *Temporal distances study using metrics and numerical formulas* | *INFOCO M*  *datasets* | *new temporal distance based metrics and have shown how they can be applied effectively to characteris e the temporal dynamics social*  *networks.* | *Complex explanatio n* |
| *Ashwin Paranjape , Austin R. Benson*  *, Jure Leskovec* | *Motifs in Tempor al Network s* | *Temporal Network Analysis of various online discussion platforms* | *Different algorithms for motif counts* | *Stack Overflow Dump* | *Finding temporal motif as a tool for analysing temporal networks* | *Overview of stackOverf low datasets using graphs but not detailed explanatio*  *n.* |
| *Bogdan Batrinca*  *, Philip C. Treleaven* | *Social media analytic s: a*  *survey* | *Overview of different methods*  *and tools* | *Algorithms and tools to analyse different*  *social* | *Different platforms for websites,*  *eg* | *Explainin g all the terminolo gies and*  *their* | *Not working in depth of any*  *particular* |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | *of techniqu es, tools*  *,*  *platform*  *s* | *to carry out social network analysis* | *netowrks like facebook,wi kimedia* | *Twitter Data Grants for Twitter* | *practical importanc e based on different outputs* | *case* |
| *Mirjam Wattenhofer , Roger Wattenhofer , Zack Zhu* | *The YouTub e Social Network* | *Graphical representa tion of subscriber s relationsh ips as well as popularit y of*  *Youtube* | *Tools to graph the values provided by Google* | *Google datasets* | *measurem ents to classify for pre- filtering potential YouTube partners for manual*  *selection.* | *Comprehe nsive as compared to other already present work* |
| *Jure Leskovec, Daniel Huttenlocher, Jon Kleinberg* | *Signed Network s in Social Media* | *Examine the interplay between positive and negative links in social*  *media* | *Theories of signed netwroks* | *Epinions datasets, Slashdot datasets, Wikipedi a voting datasets* | *Number of common neighbour s vs number of plus edges graphs for the three*  *datasets* | *Less mathemat ical approach* |

1. **Objective of the project:**

The project aims at analyzing the 3 primary interactions on stack overflow, namely answer to questions, comments to questions and comments to answers. We have calculated 3 metrics for each project. The metrics are weighted degree distribution, average path length and average clustering coefficient. Each metric denotes a practical characteristic of the network and based on the mutual discussions of the website development teams, the developers can enhance the website to provide better features to have their users utilize the maximum of what they are offered. For instance, if they find that any particular interaction among the 3 of them doesn’t come up with expected traffic, they can add more easy to use features and promote that interaction. Before social network analysis came into picture, such improvements were made on the basis of polls conducted by the organization that owns the website or on the basis of user e-mails or error reports provided to the development team. This process had its own

disadvantages. Firstly, conducting a poll for the whole website for a feature would just give the user a chance to select a feature enhancement from the choices provided by the poll conductor. Secondly, if the feature to be added had to be selected from various e- mails, reports and comments on their social network handles, it would take a lot of time to analyze all of them. Thirdly, even if an opinion of the user was noted to add a feature upon, without the actual visualization of all the users as a whole community, it could not be accurately predicted if the feature would lead to more interaction among all the users on the website who were not participating in such interactions till then. With our project, not only the 3 metrics would be clearly depicted for the 3 interactions but also the real life significance of the metric can help the developers to know about their user community better and work to provide them with the best user experience that too while promoting the use of the website to the people who prefer their competitors above them.

# Innovation component in the project:

Usually projects based on social network analysis revolve around the mathematical notions based on formulas and graphing of networks to depict how it looks. Ours take into account these parts of the network with a real-life significance of such metrics clearly described for the interactions. The results obtained using such an approach for any social network can even help the owners to design strategies for betterment of their product and the users too. Doing this, they can leave behind their competitors. Hence, this project can help to technically enhancing the website ultimately leading to more monetary profits.

# Work done and implementation:

* 1. **Methodology adapted:**

The technical methodology of the project is to create graphical visualization of the three different interactions and calculate the required metrics from the data available about the network. This is done using a software called Gephi. As there were more than 100,000 edges in each interaction, we have incorporated the first 1000 for each interaction to calculate the different metrics. The metrics calculated are weighted degree distribution, average path length and average clustering coefficient. The minimum hardware requirements of Gephi are a 500 Megahertz CPU and 128 megabytes of RAM. Also, because Gephi uses an OpenGL 3D engine to speed up graph visualization, a compatible graphics card is required. Gephi requires Java to be installed on the system, more specifically Java version 7 or 8 for its latest release. Gephi can be connected with a MySQL, SQLite or PostgreSQL database to import a graph edge list.

# Dataset used:

1. Where from you are taking your dataset?

The dataset is available on Stanford University’s website in the SNAP section. The dataset is in the form of node ids and time stamps.

1. Is your project based on any other reference project? No, our project is not based on any other reference project.

# Tools used:

The entire numerical and graphical analysis of the project is done using Gephi 0.9.2. We selected this tool because of the following reasons:

1. All the features of Gephi are available free of cost.
2. The graph visualization and edge extraction is very easy. Each edge can be extracted individually by just placing the mouse pointer on the edge.
3. All interactions made by each node can be viewed by placing the mouse pointer on the node.
4. Distribution graphs for all metrics can be generated and viewed as a web page or an image.

# Screenshot and Demo:

Given below are the directed network and weighted degree distribution graphs of the three interactions.

# Answers to questions

In this scenario, the source node is the user answering a question asked by the destination node across an edge.

Figure A.1 shows the graphical visualization of the first 1000 interactions of the data obtained.

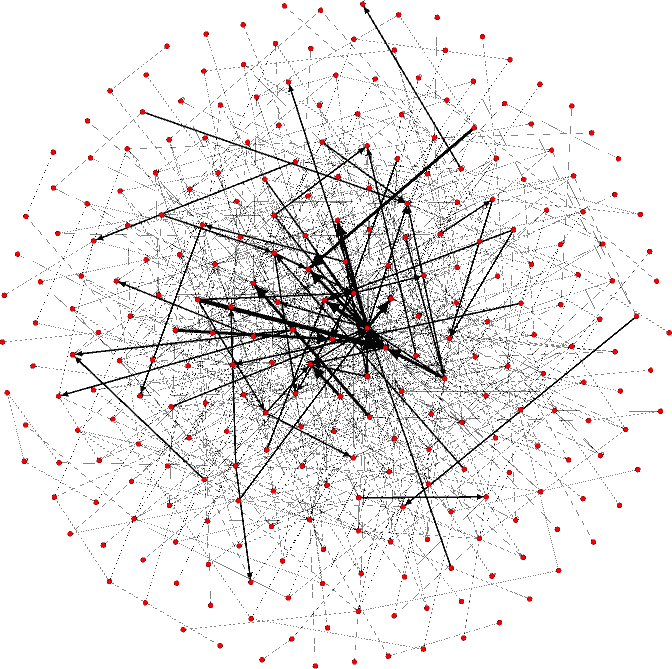


Figure A.1

Figure A.2 shows the average weighted distribution of the network in figure A.1 respectively.

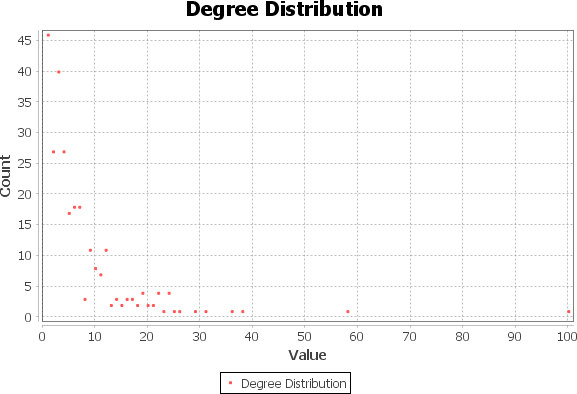


Figure A.2

The average clustering coefficient of this network is calculated to be 0.076 and average graph density is equal to 0.012.

# Comments to questions

In this scenario, the source node is the user commenting on a question asked by the destination node across an edge.

Figure B.1 shows the graphical visualization of the first 1000 interactions of the data obtained.

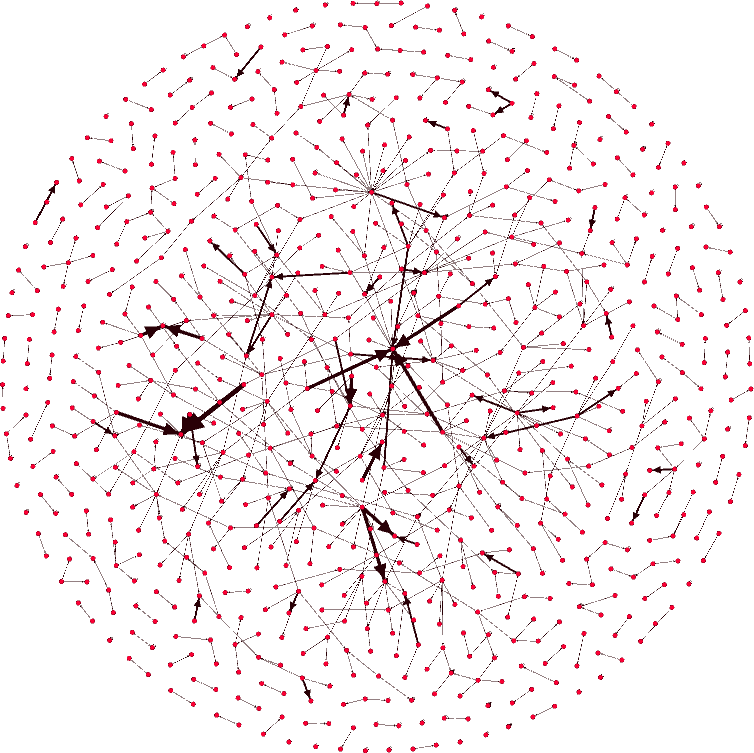


Figure B.1

Figure B.2 shows the weighted degree distribution of the network in figure B.1 respectively.

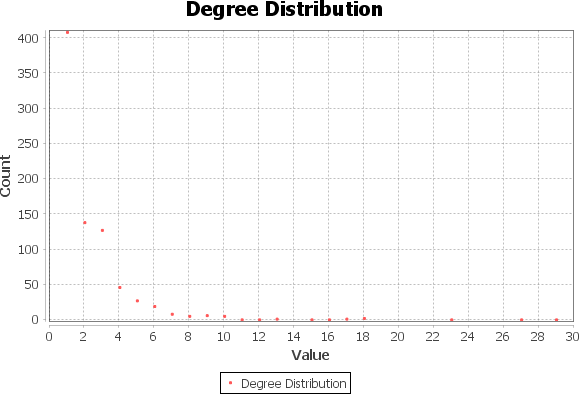


Figure B.2

The average clustering coefficient is equal to 0.070 whereas the average graph density is calculated to be 0.001.

# Comments to answers

In this scenario, the source node is the user commenting on an answer given by the destination node across an edge.

Figure C.1 shows the graphical visualization of the first 1000 interactions of the data obtained.

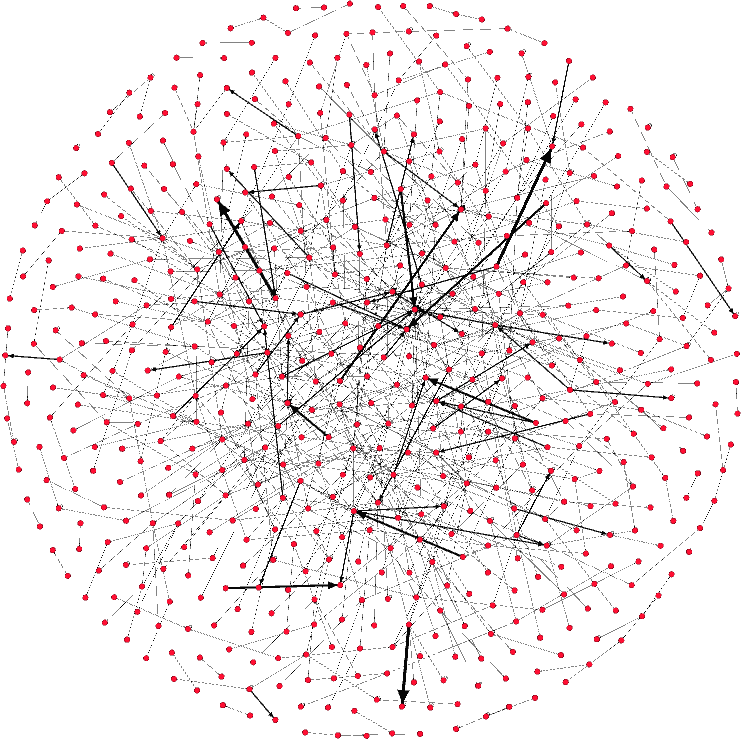


Figure C.1

Figure C.2 shows the weighted degree distribution of the network in figure C.1 respectively.

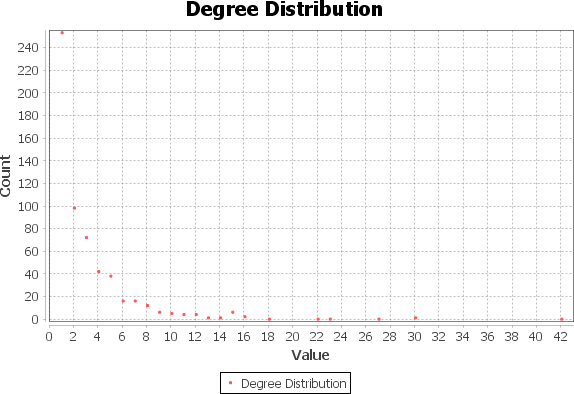


Figure C.2

The average clustering coefficient for the above network is equal to 0.050 and the average graph density equals 0.002.

# Results and discussion

In this section, we discuss about the three metrics that we have studied for the three interactions and analyze the interactions on the basis of these metrics to come up with a practical inference on the same.

# Average Weighted Degree

In Table 1, the maximum value for the average weighted degree is for the answers to questions interaction. This is also a clear outcome of the purpose of the website. The bigger weighted degree value denotes how much the users are involved in answering questions. This is more than double to interactions made by comments by each user on

questions and answers separately. This infers that for a user on an average, comments would be less than the answers given for the questions.

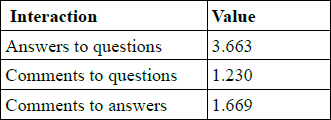


Table 1

# Average Clustering Coefficient

Higher value of local clustering coefficient for a node denotes how complete the neighborhood of that node is in terms of the interactions between the neighbors. By this we mean, if a node has a high clustering coefficient for an interaction, it implies that its neighbors are also involved in similar interaction among each other. Average or global clustering coefficient takes into account clustering coefficients of all the nodes present in the graph and computes the average of them that is used as a metric for the whole graph and not a single node. As shown in table 2, answers to questions and comments to questions have comparable values while comments to answers have a lesser value. This roughly implies that more interactions of the respective types are made between neighbors of each node for the first

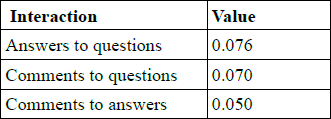
two graphs.

Table 2

# Average Graph Density

Mathematically, the average graph density is defined as the ratio of number of edges in a graph to the total possible edges. This denotes how complete the graph is. This leads us to infer from table 3 that answers to questions generate a more dense graph. It implies that maximum number of interactions from the total possible interactions are made in this case as compared to the other two.

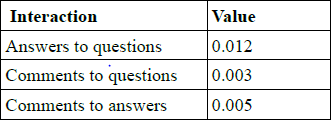


Table 3

# Conclusion-

The possible outcomes of various metrics values for different networks can be used to gather very useful data. Careful study of such metrics for various websites can help regulate things over world wide web. Network protocols that promote such interactions that yield useful data can further be developed using attained information. The data set used for this research paper has provided us with the following outcomes:

1. Answering the questions is the most widely practiced interaction on Stack Overflow.
2. The activity on the network between neighbors of different users in the graph is comparatively similar for answering the questions and commenting on the questions.
3. The flow of data is the most dense for answering the questions.
4. Stack Overflow was made for programmers to clear out their programming related queries by reaching to people with their right answers and thus it is clearly being achieved.

# 7. References:

1. Data set obtained from https://snap.stanford.edu/data/sxstackoverflow.html.
2. https://gephi.org/users/ for Gephi instructions and download.
3. https://[www.packtpub.com/mapt/book/big\_d](http://www.packtpub.com/mapt/book/big_d) ata\_and\_business\_intelligence/97817839874 05/5/ch05lvl1sec56/finding-the- averagedegree-and-average-weighted-degree-of-agraph, step by step tutorial for Gephi use.
4. <http://www.the-vital-edge.com/what-is-network-density/>for average graph density significance.
5. <https://networkscience.wordpress.com/2013/09/08/defining-the-clustering-coefficient/> for clustering coefficient significance.
6. [http://matthieu-totet.fr/Koumin/2013/12/16/understand-degree-weighted-degree- betweeness-centrality/](http://matthieu-totet.fr/Koumin/2013/12/16/understand-degree-weighted-degree-betweeness-centrality/) for weighted degree significance.