Department of Electrical and Computer Engineering
North South University

Senior Design Project

Ordering Different Networks Based on Statistical Tests

Submitted by

Syed Rhythm Ahir Hussain- ID # 1521234042
Md. Nazmul Islam - ID # 1610629042
Aminul Haq - ID # 1712246642
Mohammad Sohail - ID # 1510108642
Abdur Raufus Saleheen - ID # 1610472642

Faculty Advisor:

Dr. Mohammad Ashrafuzzaman Khan
Associate Professor
Electrical and Computer Engineering Department

Spring, 2019

Agreement Form

We take great pleasure in submitting our senior design project report on "**Ordering Different Networks Based on Statistical Tests**". This report is prepared as a requirement of the Capstone Design Project CSE/EEE/ETE 499 A & B which is a two semester long senior design course. This course involves multidisciplinary teams of students who build and test custom designed systems, components or engineering processes. We would like to request you to accept this report as a partial fulfillment of Bachelor of Science degree under Electrical and Computer Engineering Department of North South University.

Declared by:

……………………………………………
Name: Syed Rhythm Ahir Hussain
ID: 1521234042

……………………………………………
Name: Md. Nazmul Islam
ID: 1610629042

……………………………………………
Name: Aminul Haq
ID: 1712246642

……………………………………………
Name: Mohammad Sohail
ID: 1510108642

……………………………………………
Name: Abdur Raufus Saleheen
ID: 1610472642

Approved by:

……………………………
Professor & Chair
Dr. K. M. A. Salam
Department of Electrical and Computer Engineering North South, Dhaka, Bangladesh.

Approved by:

……………………………
Supervisor
Dr. Mohammad Ashrafuzzaman Khan
Associate Professor, Department of Electrical and Computer Engineering North South University, Dhaka, Bangladesh.

Declaration

This is to declare that no part of this report or the project has been previously submitted elsewhere for the fulfillment of any other degree or program. Proper acknowledgement has been provided for any material that has been taken from previously published sources in the bibliography section of this report.

....
Syed Rhythm Ahir Hussain
ECE Department (CSE)
North South University, Bangladesh

....
Md. Nazmul Islam
ECE Department (CSE)
North South University, Bangladesh

....
Aminul Haq
ECE Department (CSE)
North South University, Bangladesh

....
Mohammad Sohail
ECE Department (CSE)
North South University, Bangladesh

....
Abdur Raufus Saleheen
ECE Department (CSE)
North South University, Bangladesh

Acknowledgement

First of all, we would like thank Almighty for all the fate related to our studies and secondly to express our profound gratitude to our honorable course instructor, **Dr. Mohammad Ashrafuzzaman Khan,** for his constant and meticulous supervision, valuable suggestions, his patience and encouragement to complete the project work.

We would also like to thank the ECE department of North South University for providing us with the opportunity to have an industrial level design experience as part of our curriculum for the undergraduate program.

Finally, we would like to thank our families, friends, classmates and everybody who supported us and provided with guidance for the completion of this project.

Abstract

The main purpose of the research team is to find the differences, hypothesize the causes and rank different networks according to their properties based on statistical tests. The Internet network is built by the forces of regional politics, money, business opportunities, geography etc. The Facebook graph, on the other hand, is built by the force of friendship.

In this research, the research team tried to figure out a common pattern among the different networks so that we can identify the type of network.

Analysis of different kind of social network graph can enable us to gain valuable insights into how these networks are organized. How the networks are created and how the factors of the social networks are working among different social networks, are analyzed here.

The research team began the analysis by constructing a random graph and finding different factors (e.g. Degree, eigenvalue, page rank, betweenness etc. Then the research team worked on different social networks such as Facebook network, AS network, Google network, Twitter network, Amazon network, Communication network (Email- enron), Crime network etc. The research team found out the same factors for these networks and tried to find out a pattern among the factors of network which will help the research team to identify an anonymous type of social network graph.

TABLE OF CONTENTS

Chapter 4: Software Details

Chapter 5: Working Sheet

Chapter 6: Project Summary

Appendix

Chapter 1
Introduction

1.1 Introduction

Social network is a" group of internet-based application that is built on the ideological and technological foundation of web, and allow the creation and exchange of user generated content". Understanding such a network provides us useful insights of ways regarding how social communities are formed and interactions thereafter. Therefore it is required to analyze different factors in order to understand the networks, so that a pattern can be visualized among the networks

Social network is the structure, which shows the relations between individuals and organizations. It indicates the ways in which they are connected through various social similarities ranging from casual acquaintance to close familiar bonds.

In this research, the research team tried to figure out a common pattern among the social network graphs so that having a view of any dataset of any anonymous social network graph can be identified.

A considerable amount of work has been accomplished on structure of the networks with appreciable accuracy. The same cannot be told about the structure of networks of different social media due to limited number of datasets. We intend to work further on the network structure of social media. Our main purpose is to find the differences and hypothesize the causes. We want to rank these networks according to different network properties based on statistical tests.

Keywords— Graph visualization, random graph generation, Degree analysis, eigenvector and centrality, Jonckheere-Terpstra test.

1.2 Background

At the beginning of this research project, the research team started their work by analyzing a simple random graph. After the visualization of the graph and finding out different factors such as degree, eigenvalue, PageRank, betweenness of the random graph, the research team tried to find out the same factors for random social network graphs e.g. Facebook network, AS network, Google network, Twitter network, Amazon network, Communication network(Email- enron), Crime network etc.

Finding out degree, eigenvalue, PageRank, betweenness for these different social network graphs, the research team tried to figure out a pattern among the factors of these different social network graph. In order to find a pattern the research team did a rank test among the different type of social network graphs. For performing rank test, initially five different types of social networks each having minimum four different datasets are collected. After finding out eigenvalue, degree, PageRank, betweenness for each of the four datasets of each type of social network, the team found out the average for each factors of each dataset.

In order to perform a ranking, the research team used JT test to perform a ranking among the different social network graphs. In JT test, average of eigenvalues for each dataset of each network is used. Similarly, the average degree, the average betweenness, the average PageRank is used in JT test to perform a ranking among the different social networks.

1.3 Objective

A considerable amount of work has been accomplished on structure of the networks with appreciable accuracy. The same cannot be told about the structure of networks of different social media due to limited number of datasets. We intend to work further on the network structure of social media. Our main purpose is to find the differences and hypothesize the causes. We want to rank these networks according to different network properties based on statistical tests.

Keywords— Graph visualization, random graph generation, Degree analysis, eigenvector and centrality, Jonckheere-Terpstra test.

This model can be used to identify which type of network we are dealing. When we have ordered list of a huge number of networks, we can predict the type of an unknown network by this statistical tests. It will eventually help us to detect whether a unknown network is crime network or not. If the network data behavior is very much similar with our crime network data behavior we can predict that network can be a crime network. We can find influential nodes among them which has the least Probability value. This will help us to take actions against any conspiracy beforehand. Cyber security can be improved by controlling the flow of fake news and propagandas. We can even use this model to create a model by visualization and train the data of different network dataset.

1.4 Motivation

The fire triangle represents the three elements a fire needs to burn: oxygen, heat, and a fuel. Similarly, fake news requires different items to succeed. Tools and services for manipulating and spreading the message across relevant social media networks. Of course, for these tools to be of any use, social networks have to exist as a platform for spreading propaganda. With people spending more time on these sites as a way to get the latest news and information. However, People with criminal motives do not stop simply after posting propaganda, they actually try turning it into something that the target audience consumes. Our motive is to control those propagandas in other we can handle them from being violent.

1.5 Our Approach

The basic feature of our project is to predict a network type of an unknown network by comparing with our model along with finding the hierarchy of influential nodes of that network to reduce cybercrime or fake propagandas. Propaganda has been around for centuries, and the internet is only the latest means of communication to be abused to spread lies and misinformation.

Chapter 2
Project Overview

2.1 Workflow

For this project we used the following properties of graph. These are computed using NetworkX package in Python. NetworkX is a package of Python for manipulation, creation, and it also helps to figure out the complex networks. All the following properties are built in in this package in Python.

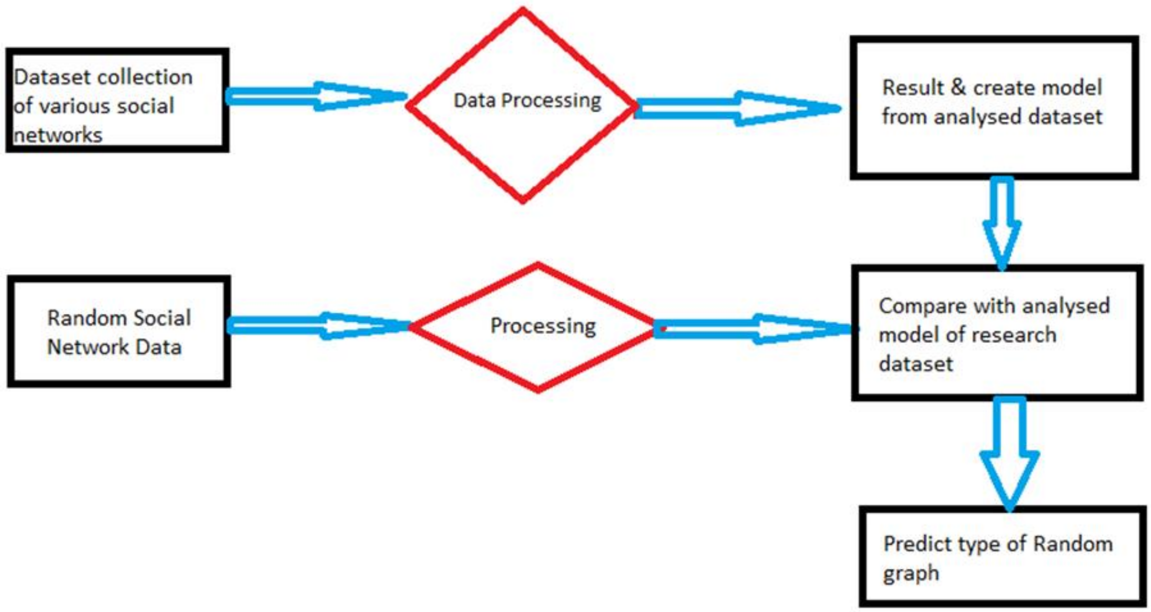


Diagram: Workflow

2.2 Graph Properties

Degree Centrality

This algorithm computes the degree centrality for nodes. The degree centrality for a node is the number of nodes it is connected to. The degree centrality values are normalized by dividing by the maximum possible degree in a simple graph n-1 where n is the number of nodes in G. Suppose, we have a graph of a, b, c nodes. Here is a connected to b and c. Then degree centrality of is a 2. If b is connected to only a then degree centrality of b is 1. Same way, all degree centrality of each nodes is calculated. For multi-graph or graphs with self- loops the maximum degree might be higher than n-1 and values of degree centrality greater than 1 are possible though we do not use it. The following network shows the in degree of each node.

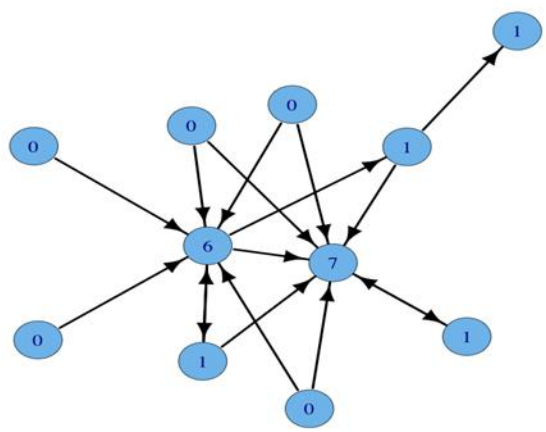


Diagram: Degree Centrality

Betweenness Centrality

It computes the shortest-path betweenness centrality for nodes. Betweenness Centrality gives the value of the number of times a node acts as a bridge along the shortest path between two other nodes.

Suppose we have a graph of a, b, c nodes. Here is a connected to c through b. so here betweenness centrality of b is 1. This way each node's betweenness centrality is calculated.

In the case of different networks the distance from other units is not the only important property of a unit. The more important thing is which units lie on the shortest paths among pairs of other units. Such units have control over the flow of information in the network. Betweenness centrality is useful as a measure of the potential of a vertex for control of communication.

Here A and B works as bridges of many nodes/vertexes.

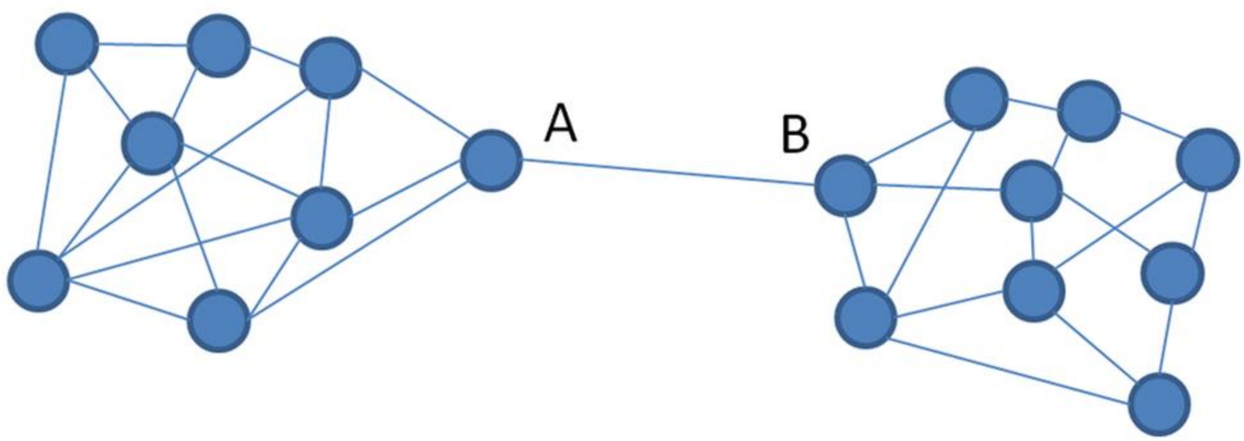


Diagram: Betweenness Centrality

Eigenvector Centrality:

Eigenvector centrality measures a node's importance while giving consideration to the importance of its neighbors. For example, a node with 300 relatively unpopular friends on Facebook would have lower eigenvector centrality than someone with 300 very popular friends (like Ronaldo).

Eigenvector centrality, regarded as a ranking measure, is a remarkably old method. Early pioneers of this technique are Wassily W. Leontief (The Structure of American Economy, 1919-1929. Harvard University Press, 1941) and John R. Seeley (The net of reciprocal influence: A problem in treating sociometric data.

Here the graph shows the Eigen value of each vertex.

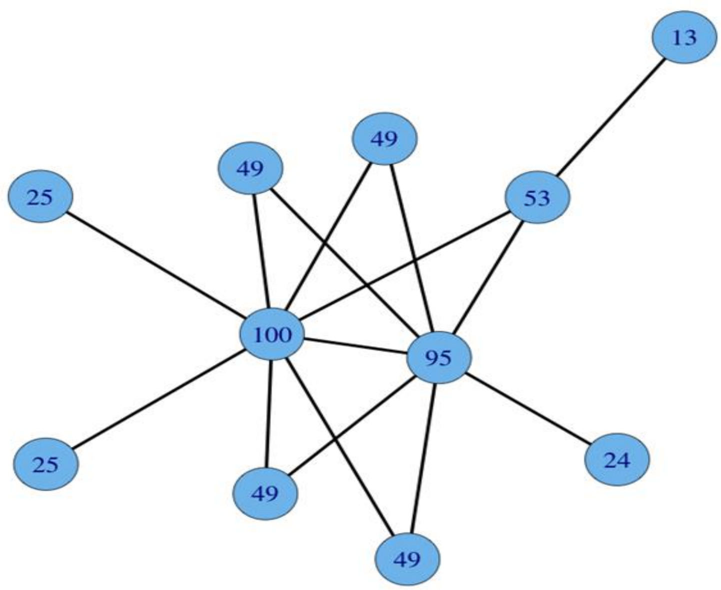


Diagram: Eigenvector Centrality

PageRank:

We used it for the PageRank of the nodes in the graph. It computes a ranking of the nodes in the graph based on the structure of the in degree. PageRank is an algorithm used by Google Search to rank websites in their search engine results. PageRank was named after Larry Page, one of the founders of Google. PageRank is a way of measuring the importance of website pages. In our project we used it to find the PageRank value of each vertex.

2.3 Dataset

Names	Set-1	Set-2	Set-3	Set-4
Facebook Source: kaggle.com	Name: Artist Edges Type: Graph Number of nodes: 50515 Number of edges: 819306 Average degree: 32.4381	Name: Company Type: Graph Number of nodes: 14113 Number of edges: 52310 Average degree: 7.4130	Name: Government Type: Graph Number of nodes: 7057 Number of edges: 89455 Average degree: 25.3521	Name: Politician Type: Graph Number of nodes: 5908 Number of edges: 41729 Average degree: 14.1263
Crime Source:	Name: 9_11 Type: Graph Number of nodes: 61 Number of edges: 132 Average degree: 4.3279	Name: 2015 FIFA Type: Graph Number of nodes: 450 Number of edges: 5022 Average degree: 22.3200	Name: BALIBOMBING2002_2000 Type: Graph Number of nodes: 23 Number of edges: 97 Average degree: 8.4348	Name: Mali_Terrorist_Network Type: Graph Number of nodes: 36 Number of edges: 67 Average degree: 3.7222

P2P Source: snap.stanford.edu	Name: p2p-Gnutella04 Type: Graph Number of nodes: 10876 Number of edges: 39994 Average degree: 7.3545	Name: p2p-Gnutella05 Type: Graph Number of nodes: 8846 Number of edges: 31839 Average degree: 7.1985	Name: p2p-Gnutella06 Type: Graph Number of nodes: 8717 Number of edges: 31525 Average degree: 7.2330	Name: p2p-Gnutella08 Type: Graph Number of nodes: 6301 Number of edges: 20777 Average degree: 6.5948
Amazon Source: snap.stanford.edu	Name: Amazon0302 Type: Graph Number of nodes: 232780 Number of edges: 767456 Average degree: 6.5938	Name: com-amazon.all.dedup.cmty Type: Graph Number of nodes: 222365 Number of edges: 222345 Average degree: 1.9998	Name: com-amazon.top5000.cmty.txt Type: Graph Number of nodes: 16716 Number of edges: 15344 Average degree: 1.8358	Name: com-amazon.ungraph.txt Type: Graph Number of nodes: 334863 Number of edges: 925872 Average degree: 5.5299

Email_eu Source: snap.stanford.edu	Name: email-Eu-core-temporal-Dept1.txt Type: Graph Number of nodes: 309 Number of edges: 1938 Average degree: 12.5437	Name: email-Eu-core-temporal-Dept2 Type: Graph Number of nodes: 162 Number of edges: 1045 Average degree: 12.9012	Name: email-Eu-core-temporal-Dept3 Type: Graph Number of nodes: 89 Number of edges: 973 Average degree: 21.8652	Name: email-Eu-core-temporal-Dept4 Type: Graph Number of nodes: 142 Number of edges: 833 Average degree: 11.7324

2.4 Features Extractions

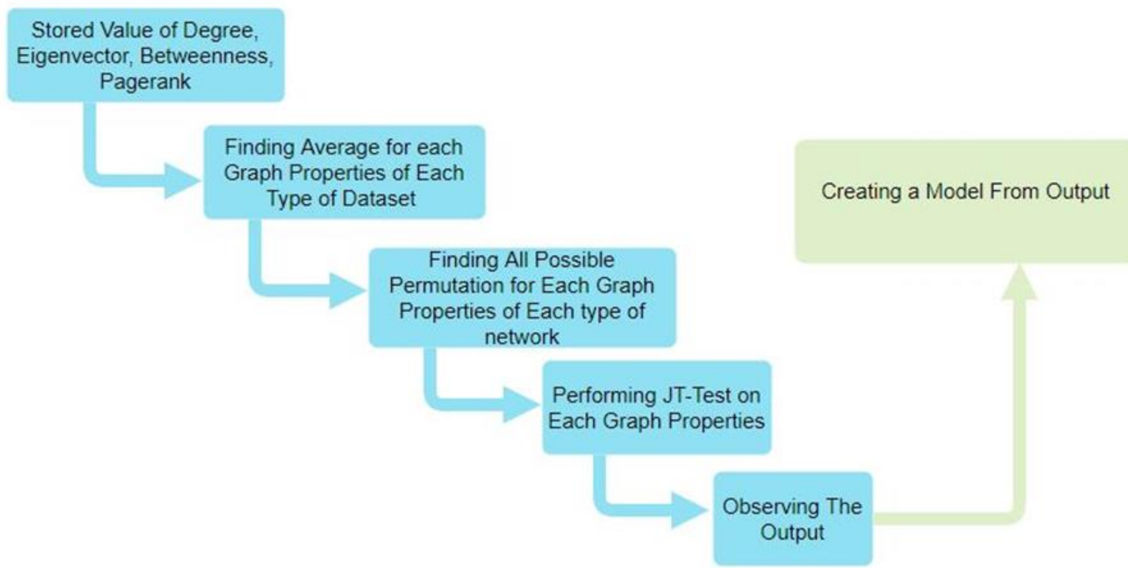


Diagram: Feature Extractions

2.5 Jonckheere's trend test

In statistics, the Jonckheere trend test (sometimes called the Jonckheere–Terpstra test) is a test for an ordered alternative hypothesis within an independent samples (between-participants) design. It is similar to the Kruskal–Wallis test in that the null hypothesis is that several independent samples are from the same population. However, with the Kruskal–Wallis test there is no a priori ordering of the populations from which the samples are drawn. When there is an a priori ordering, the Jonckheere test has more statistical power than the Kruskal–Wallis test. The test was developed by A. R. Jonckheere, who was a psychologist and statistician at University College London.

We considered 5 types of Networks Graphs. Facebook, Crime, Peer to Peer, Amazon & Email_eu. We calculated Betweenness, Average Degree, PageRank & EigenVector for each data type total 20 dataset. We find the average value of them. With those average values we run them on JT test.
The Jonckheere-Terpstra test is a rank-based nonparametric test that can be used to determine if there is a statistically significant trend between an ordinal independent variable and a continuous or ordinal dependent variable.
The JT test gives us the P value and J value which helps us to determine the closeness and rank among nodes. If the P value is very near to 0(zero) that represents that network has better relations among the nodes.
The standardized test statistic is computed as

$$Z = \frac{J - E(J)}{\sqrt{Var(J)}}$$

Here

$$E(J) = \frac{N^2 - \sum_{j=1}^{k} n_j^2}{4}$$

And

$$Var(J) = N^2(2N+3) - \sum_{j=1}^{k} n_j^2(2n_j+3)$$

Chapter 3
Network Visualization

3.1 Introduction

Network virtualization is the process of combining hardware and software network resources and network functionality into a single, software-based administrative entity, a virtual network. Network virtualization involves platform virtualization, often combined with resource virtualization.

3.2 Necessary Graph Properties

Graph properties which are need to be visualized

- Degree
- EigenVector
- PageRank
- Betweenness
- Other Components

3.3 Network Visualization Process

Network Visualization enables us to test point-to-point performance of our graph under real-world network conditions. By visualizing each types of graphs, we can introduce highly probable effects such as latency, packet loss, and link faults over our network. We can create more meaningful results by configuring multiple loads of data generated histograms. We can compare groups on a single type data with the same unique set of network effects. That can even help us in finding influential nodes.

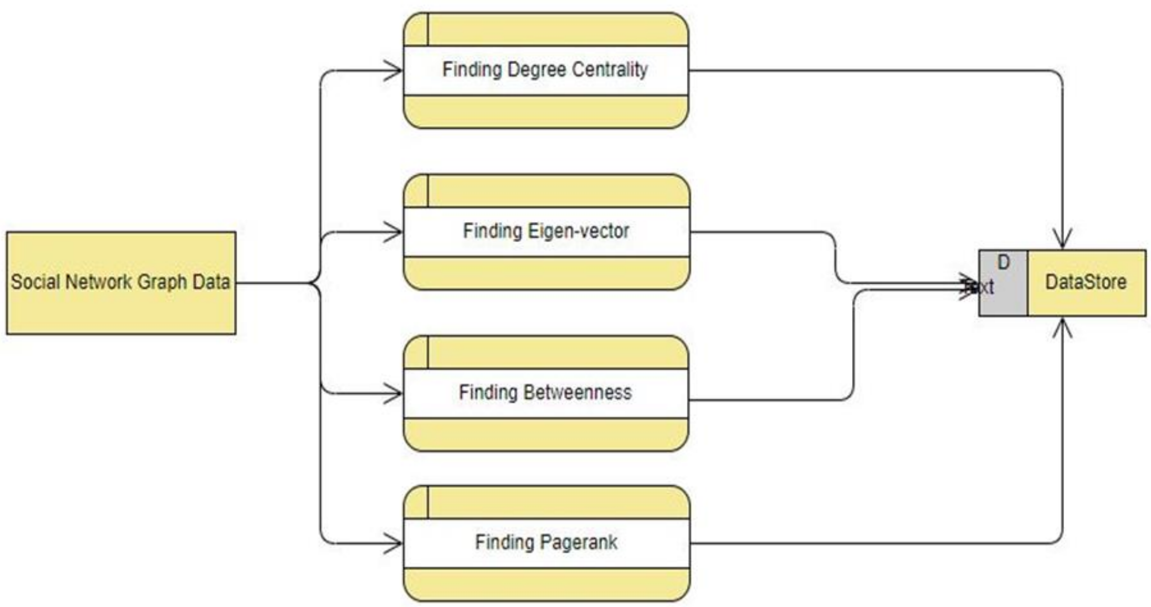


Diagram: Processing Overview

3.4 Types of Network Visualized

Graph properties which are need to be visualized

- Facebook
- Crime
- P2P
- Amazon
- Email_EU

Facebook

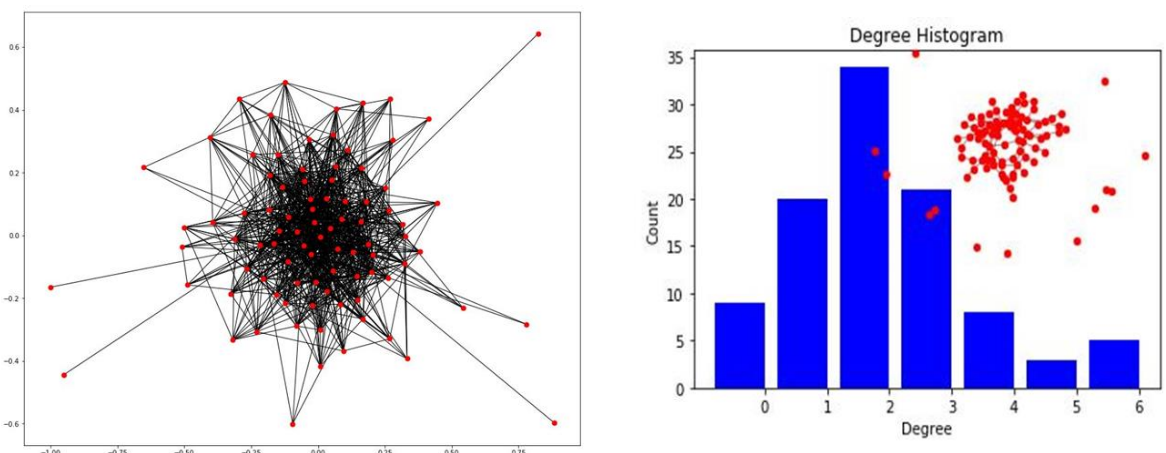


Diagram: Network Visualization of Facebook Data

Crime

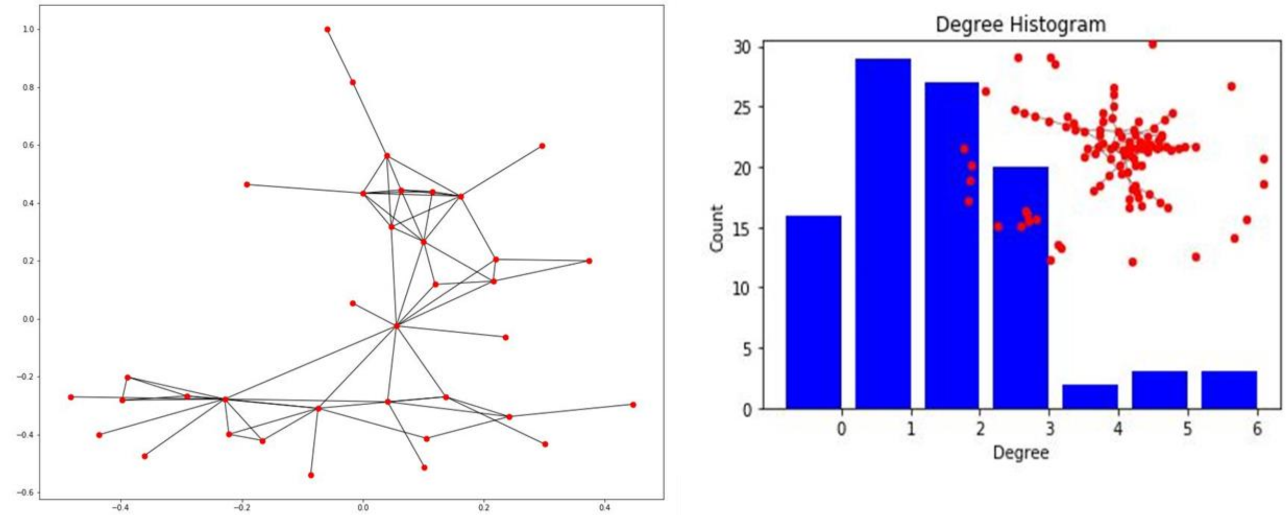


Diagram: Network Visualization of Crime Data

P2P

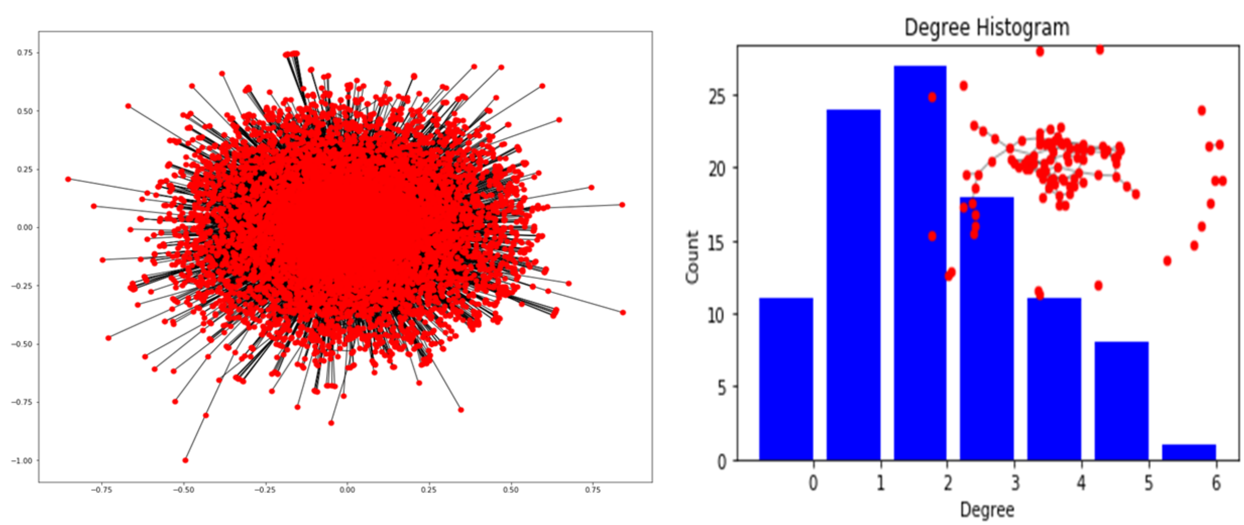


Diagram: Network Visualization of P2P Data

Amazon

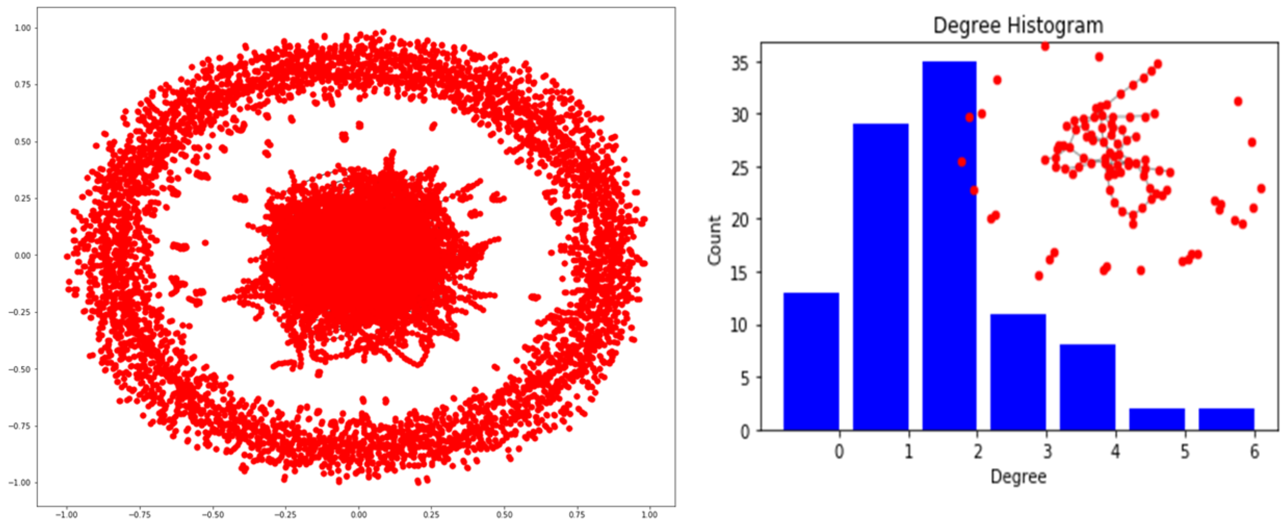


Diagram: Network Visualization of Amazon Data

Email_EU

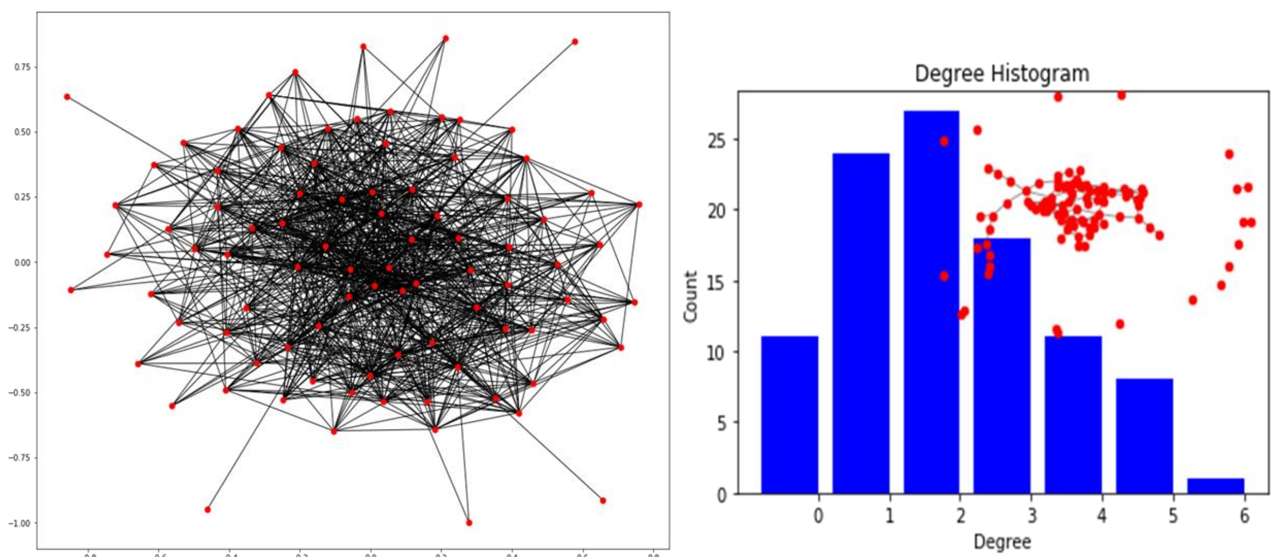


Diagram: Network Visualization of Email(Europe) Data

Chapter 4
Software Details

4.1 Software Components

As we want to design an AI based interactive blind assistant device, we need to have some knowledge in different types of programming languages by which we can easily build a connection between our server and Android app. To send the voice command we are using our android app, Arduino nano module had been programmed by the Arduino IDE software.

RSudio:

RStudio is an integrated development environment (IDE) for R, a programming language for statistical computing and graphics. The RStudio IDE is developed by RStudio, Inc., a commercial enterprise founded by JJ Allaire, creator of the programming language ColdFusion. Hadley Wickham is the Chief Scientist at RStudio . RStudio, Inc. has no formal connection to the R Foundation, a not for profit organization located in Vienna Austria, which is responsible for overseeing development of the R environment for statistical computing.

RStudio is available in two formats: RStudio Desktop, where the program is run locally as a regular desktop application; and RStudio Server, which allows accessing RStudio using a web browser while it is running on a remote Linux server.

RStudio Desktop and RStudio Server are both available in free and fee-based (commercial) editions. OS support depends on the format/edition of the IDE.

Prepackaged distributions of RStudio Desktop are available for Windows, macOS, and

Linux. RStudio Server and Server Pro run on Debian, Ubuntu, Red Hat Linux, CentOS, openSUSE and SLES.

RStudio is partly written in the C++ programming language and uses the Qt framework for its graphical user interface. The bigger percentage of the code is written in Java. JavaScript is also amongst the languages used.

Diagram: RStudio user interface

IntelliJ IDEA:

IntelliJ IDEA is a Java integrated development environment (IDE) for developing computer software. It is developed by JetBrains (formerly known as IntelliJ), and is available as an Apache 2 Licensed community edition, and in a proprietary commercial edition. Both can be used for commercial development.

Diagram: InteliJ IDEA user interface

PyCharm:

PyCharm is an integrated development environment (IDE) used in computer programming, specifically for the Python language. It is developed by the Czech company JetBrains. It provides code analysis, a graphical debugger, an integrated unit tester, integration with version control systems (VCSes), and supports web development with Django. PyCharm provides smart code completion, code inspections, on-the-fly error highlighting and quick-fixes, along with automated code refactoring and rich navigation capabilities. PyCharm's smart code editor provides first-class support for Python, JavaScript, CoffeeScript, TypeScript, CSS, popular template languages and more. Take advantage of language-aware code completion, error detection, and on-the-fly code fixes.

Diagram: PyCharm IDE

Use smart search to jump to any class, file or symbol, or even any IDE action or tool window. It only takes one click to switch to the declaration, super method, test, usages, implementation, and more.

For running the server side's codes in python programming language and processing the whole system we used PyCharm. As it is graphical user interface (GUI) based software we can easily handle our functionalities through this.

Anaconda:

Anaconda is a free and open-source distribution of the Python and R programming languages, which aims to simplify package management and deployment. Package versions are managed by the package management system conda. The Anaconda distribution is used by over 15 million users and includes more than 1500 popular data-science packages suitable for Windows, Linux, and MacOS.

Anaconda distribution comes with more than 1,500 packages as well as the Conda package and virtual environment manager. It also includes a GUI, Anaconda Navigator, as a graphical alternative to the command line interface (CLI).

Diagram: Anaconda user interface

Conda analyzes your current environment, everything you have installed, any version limitations you specify (e.g. you only want tensorflow >= 2.0) and figures out how to install compatible dependencies.

Chapter 5
Working Sheet

5.1 Proposed Workflow of our work in 499A:

January	Work-Flow	Check Point
Week-1		
Week-2	Detailed Discussion & Draft construction	
Week-3	Related Work Study & Final Design Construction	
Week-4	Design implementation & Basic Model build	

February	Work-Flow	Check Point
Week-1	Data set construction	
Week-2	Data set construction and final processing	
Week-3	Understanding Graph Properties	
Week-4	Understanding Graph Properties	

March	Work-Flow	Check Point
Week-1	Understanding Graph Properties	
Week-2	Phase-2: Classify different Networks and Finding Properties	
Week-3	Network Visualization	
Week-4	Network Visualization	

5.2 Proposed Workflow of our work in 499B:

May	Work-Flow	Check Point
Week-1	Different Types of Network Data Collection	
Week-2	Finding similarity among Different Types of Network Data Collection	
Week-3	Finding Betweenness, Degree, PageRank & Eigenector of every type of Network	
Week-4	Finding Betweenness, Degree, PageRank & Eigenector of every type of Network	

June	Work-Flow	Check Point
Week-1	Design of the Networks	
Week-2	Basic Functionalities for assistance	
Week-3	Studying About Different tests to find ordering network graph	
Week-4	Studying about Jonckheere-Terpstra test	

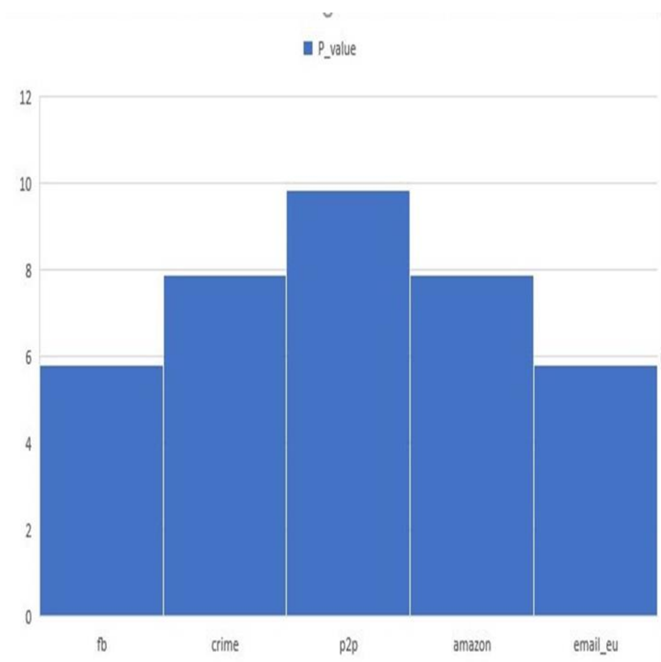
July	Work-Flow	Check Point
Week-1	Implementing Jonckheere-Terpstra test	
Week-2	Implementing Jonckheere-Terpstra test	
Week-3	Result Observation	
Week-4	Representations of Results	

Chapter 6
Project Summary

6.1 Result and Discussion

We considered 5 types of Networks Graphs. Facebook, Crime, Peer to Peer, Amazon & Email_eu. We calculated Betweenness, Average Degree, PageRank and EigenVector for each data type total 20 dataset. We find the average value of them. With those average values we run them on JT test.

The Jonckheere-Terpstra test is a rank-based nonparametric test that can be used to determine if there is a statistically significant trend between an ordinal independent variable and a continuous or ordinal dependent variable.



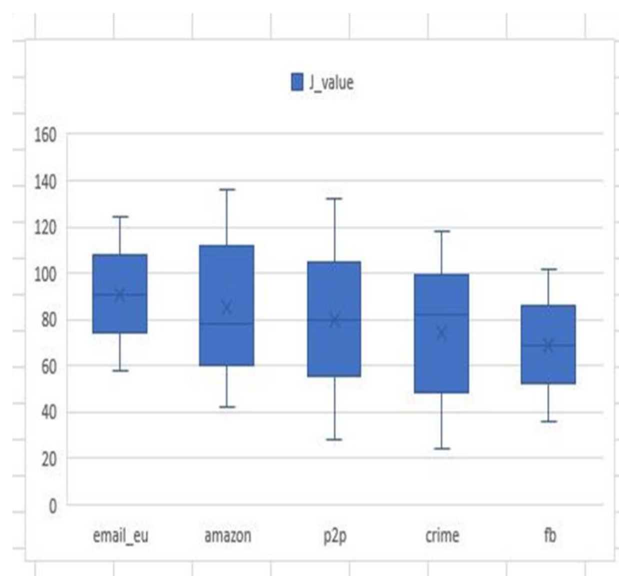


Diagram: JT Test Histogram of Graph PageRank

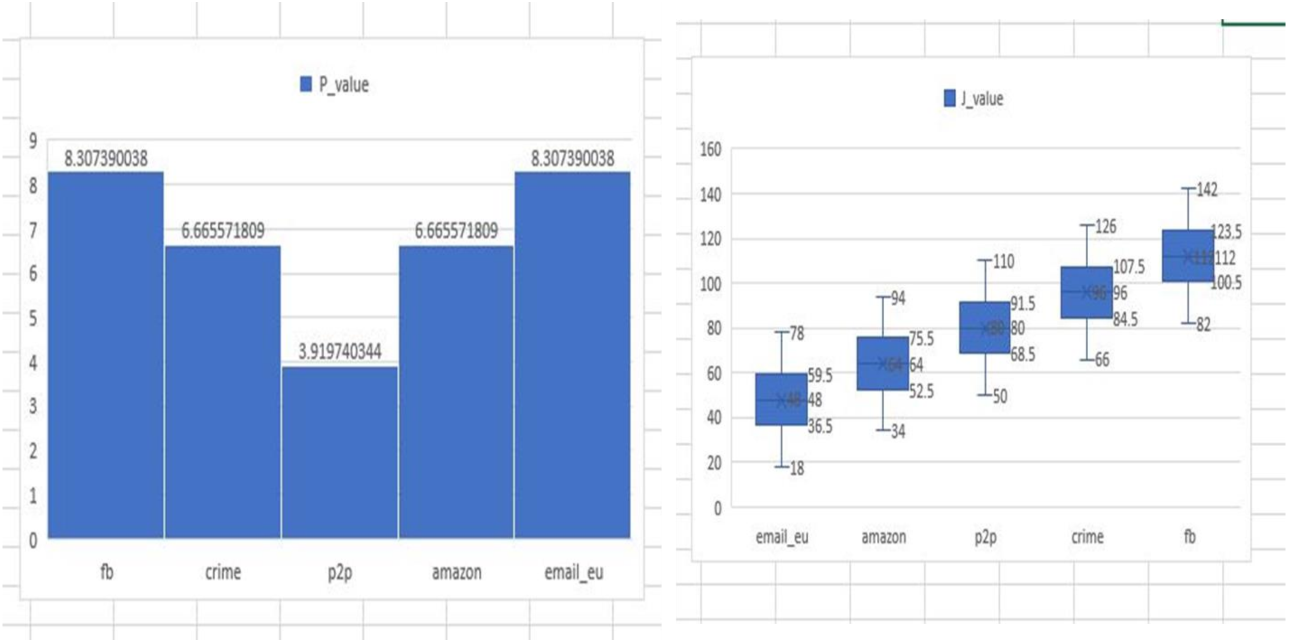


Diagram: JT Test Histogram of Graph Average Degree

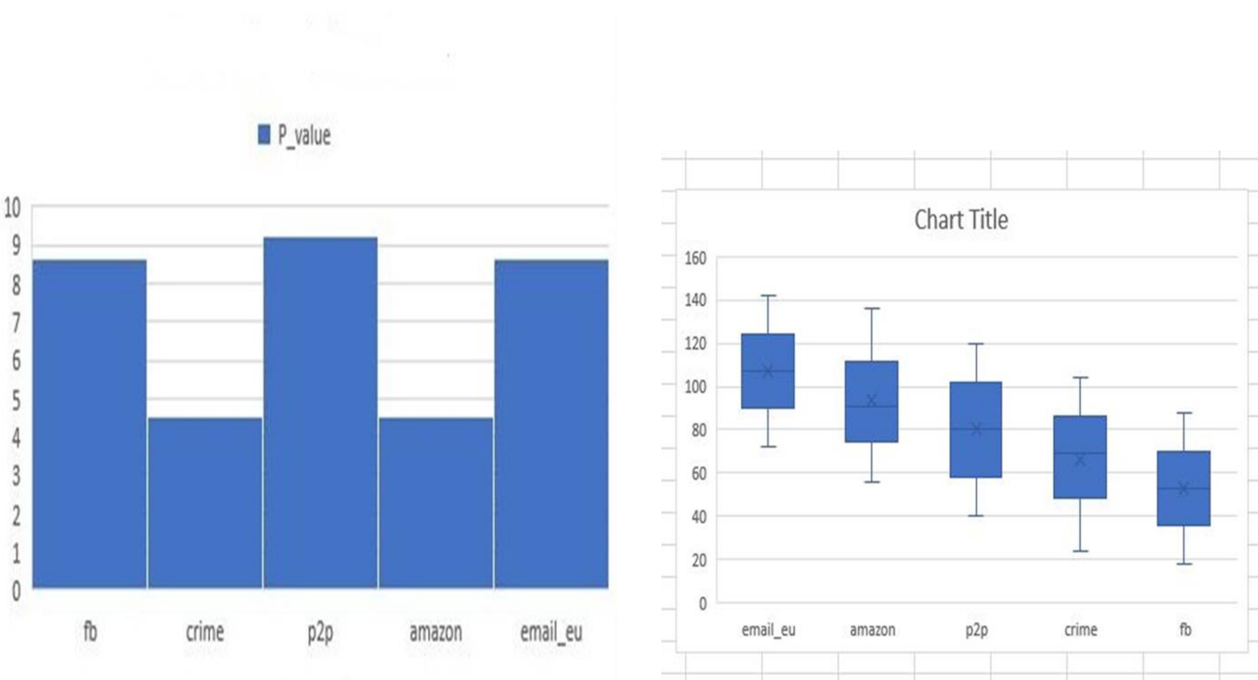


Diagram: JT Test Histogram of Graph Betweenness

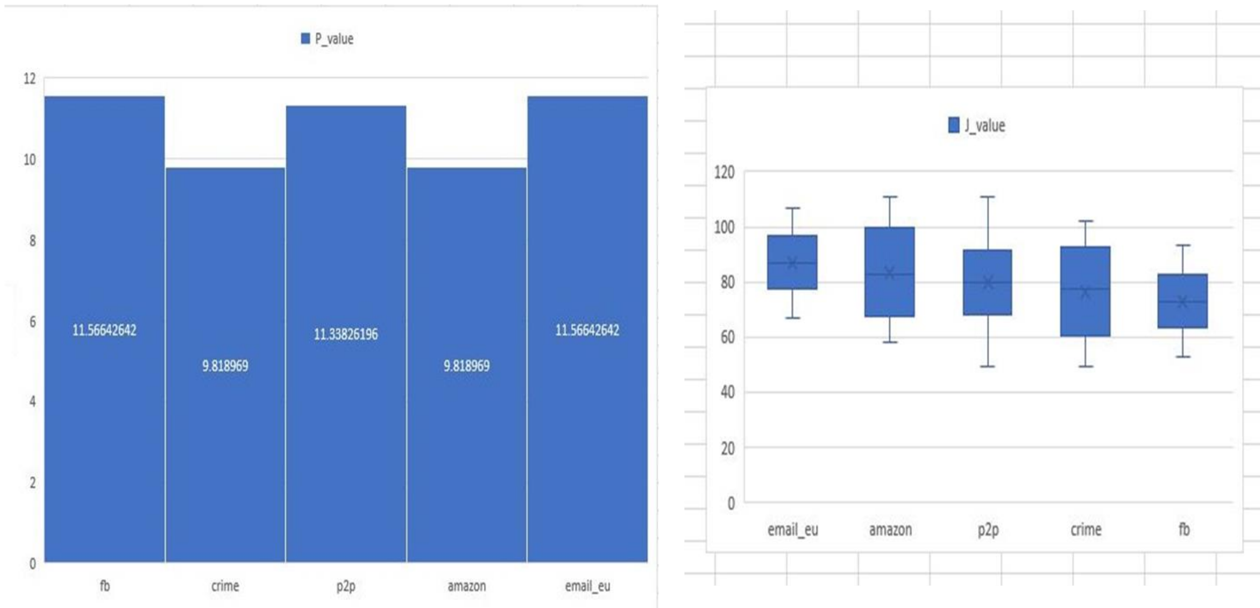


Diagram: JT Test Histogram of Graph EigenVector

The JT test gives us the P value and J value which helps us to determine the closeness and rank among nodes. If the P value is very near to 0(zero) that represents that network has better relations among the nodes. We used a histogram to identify which type of network gives us more accurate value when we compare them with other which is P value less than 0.1.

6.2 Economic Overview:

After generating output from JT test, we took the best 30% of the output data, based on the lowest of the P values. This table shows the percentage of a network at a particular position after ordering. We observed that, some networks tend to be on some particular position. For example, in the degree centrality table, 33.33% of time p2p networks comes in the first position and amazon or crime networks tends to be in the last position. P2p never takes the last position here. Which actually matches with our hypothesis. Because p2p is more dense network than amazon. P2P's node's degrees should be definitely higher than amazon. Since in Peer-to-Peer one user connects with many other users, but on amazon's network there are only edges between buyer and seller. We see similar trends across the networks in all the tables. But also we observed some ambiguity. But we believe, if we process more data from other different networks we will get more consistent results and a more stable ordering will be found. All the tables with position percentages are given below.

Degree centrality

Position	Amazon	Crime	Facebook	Email_Eu	P2P
1	25.0%	25%	8.33%	8.33%	33.33%
2	29.18%	29.18%	8.33%	8.33%	25.0%
3	20.86%	20.86%	12.5%	12.5%	33.33%
4	12.5%	12.5%	33.33%	33.33%	8.33%
5	12.5%	12.5%	37.5%	37.5%	0.0%

EigenVector

Position	Amazon	Crime	Facebook	Email_Eu	P2P
1	16.66%	16.66%	16.66%	16.66%	33.33%
2	25.0%	25.0%	25.0%	25.0%	0.0%
3	25.0%	25.0%	25.0%	25.0%	0.0%
4	16.64%	16.64%	16.64%	16.64%	33.33%
5	16.64%	16.64%	16.64%	16.64%	33.33%

Betweenness

Position	Amazon	Crime	Facebook	Email_Eu	P2P
1	25.0%	25.0%	25.0%	8.33%	8.33%
2	20.83%	20.83%	20.83%	25.0%	25.0%
3	16.64%	16.64%	16.66%	16..66%	16.66%
4	16.64%	16.64%	16.66%	25.0%	25.0%
5	20.83%	20.83%	20.83%	25.0%	25.0%

Pagerank

Position	Amazon	Crime	Facebook	Eamil_Eu	P2P
1	16.66%	16.66%	37.5%	29.16%	0.0%
2	16.66%	16.66%	20.83%	25.0%	20.83%
3	20.83%	20.83%	25.0%	25.0%	12.5%
4	20.83%	20.83%	8.33%	12.5%	33.33%
5	25%	25.0%	8.33%	8.33%	33.33%

6.3 Future Ideas:

- **Recognizing Individual Persons:** Our model can detect different type of nodes and person. It can detect multiple persons as well. In future we want to add such a feature in which our model can recognize the person whom was detected before. At the same time, it will be able to give that person's identity through the mobile app.
- **Control Propaganda on Bangladeshi Environment:** We have a plan integrate Bangladeshi Data on our model in which the device will be able to detect the person we are looking for.

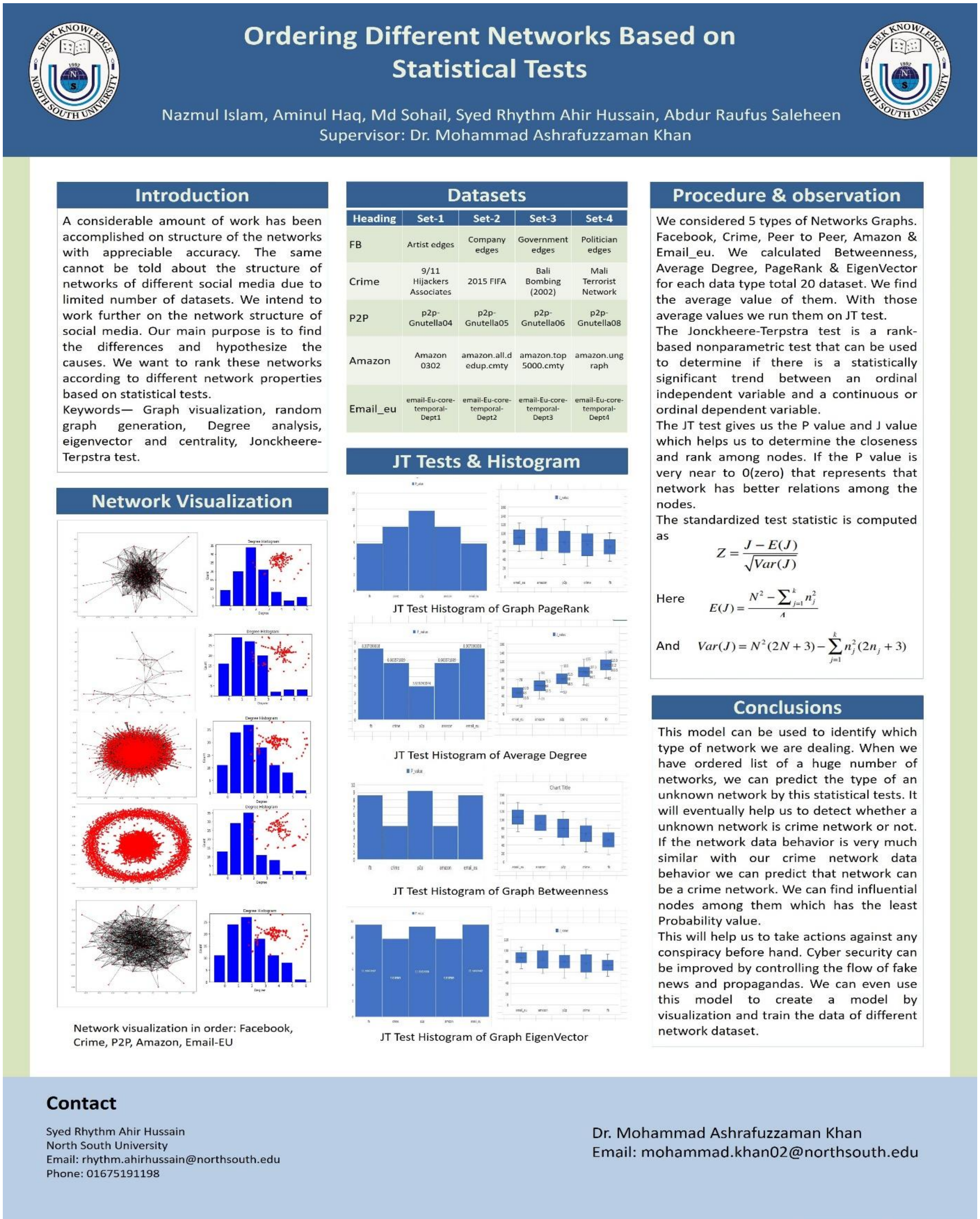
  In country like Bangladesh people have the tendency to believe what they see in social network and act emotional to it. That is why we think that it will be most suitable our environment.

6.4 Conclusion:

This model can be used to identify which type of network we are dealing. When we have ordered list of a huge number of networks, we can predict the type of an unknown network by this statistical tests. It will eventually help us to detect whether a unknown network is crime network or not. If the network data behavior is very much similar with our crime network data behavior we can predict that network can be a crime network. We can find influential nodes among them which has the least Probability value.

This will help us to take actions against any conspiracy beforehand. Cyber security can be improved by controlling the flow of fake news and propagandas. We can even use this model to create a model by visualization and train the data of different network dataset.

6.5 Poster:


Ordering Different Networks Based on Statistical Tests
Nazmul Islam, Aminul Haq, Md Sohail, Syed Rhythm Ahir Hussain, Abdur Raufus Saleheen
Supervisor: Dr. Mohammad Ashrafuzzaman Khan
Introduction
A considerable amount of work has been accomplished on structure of the networks with appreciable accuracy. The same cannot be told about the structure of networks of different social media due to limited number of datasets. We intend to work further on the network structure of social media. Our main purpose is to find the differences and hypothesize the causes. We want to rank these networks according to different network properties based on statistical tests.
Keywords— Graph visualization, random graph generation, Degree analysis, eigenvector and centrality, Jonckheere-Terpstra test.
Network Visualization
Network visualization in order: Facebook, Crime, P2P, Amazon, Email-EU
Datasets
Heading | Set-1 | Set-2 | Set-3 | Set-4
FB | Artist edges | Company edges | Government edges | Politician edges
Crime | 9/11 Hijackers Associates | 2015 FIFA | Bali Bombing (2002) | Mali Terrorist Network
P2P | p2p-Gnutella04 | p2p-Gnutella05 | p2p-Gnutella06 | p2p-Gnutella08
Amazon | Amazon 0302 | amazon.all.dedup.cmty | amazon.top5000.cmty | amazon.ungraph
Email_eu | email-Eu-core-temporal-Dept1 | email-Eu-core-temporal-Dept2 | email-Eu-core-temporal-Dept3 | email-Eu-core-temporal-Dept4
JT Tests & Histogram
JT Test Histogram of Graph PageRank
JT Test Histogram of Average Degree
JT Test Histogram of Graph Betweenness
JT Test Histogram of Graph EigenVector
Procedure & observation
We considered 5 types of Networks Graphs. Facebook, Crime, Peer to Peer, Amazon & Email_eu. We calculated Betweenness, Average Degree, PageRank & EigenVector for each data type total 20 dataset. We find the average value of them. With those average values we run them on JT test.
The Jonckheere-Terpstra test is a rank-based nonparametric test that can be used to determine if there is a statistically significant trend between an ordinal independent variable and a continuous or ordinal dependent variable.
The JT test gives us the P value and J value which helps us to determine the closeness and rank among nodes. If the P value is very near to 0(zero) that represents that network has better relations among the nodes.
The standardized test statistic is computed as
$Z = \frac{J - E(J)}{\sqrt{Var(J)}}$
Here $E(J) = \frac{N^2 - \sum_{j=1}^{k} n_j^2}{4}$
And $Var(J) = N^2(2N+3) - \sum_{j=1}^{k} n_j^2(2n_j+3)$
Conclusions
This model can be used to identify which type of network we are dealing. When we have ordered list of a huge number of networks, we can predict the type of an unknown network by this statistical tests. It will eventually help us to detect whether a unknown network is crime network or not. If the network data behavior is very much similar with our crime network data behavior we can predict that network can be a crime network. We can find influential nodes among them which has the least Probability value.
This will help us to take actions against any conspiracy before hand. Cyber security can be improved by controlling the flow of fake news and propagandas. We can even use this model to create a model by visualization and train the data of different network dataset.
Contact
Syed Rhythm Ahir Hussain
North South University
Email: rhythm.ahirhussain@northsouth.edu
Phone: 01675191198
Dr. Mohammad Ashrafuzzaman Khan
Email: mohammad.khan02@northsouth.edu

Appendix

A. Reference

References and Resources:

http://snap.stanford.edu/data/

https://www.kaggle.com/kkanda/analyzing-uci-crime-and-communities-dataset

https://www.rdocumentation.org/packages/SAGx/versions/1.46.0/topics/JT.test

https://sci2s.ugr.es/sci2s.ugr.es/software/javanpts

http://archive.ics.uci.edu/ml/datasets/Communities%20and%20Crime%20Unnormalized

https://www.rdocumentation.org/

https://networkx.github.io/documentation/stable/?fbclid=IwAR1NgPeZL6G0YtFu7g8KNkt9f9zD

132ugU5F2GDfdMqyJDyhXBUIg8Z7wj8

B. CODES

RStudio Codes for Finding Graph Properties and JT test:

betweeness_&_degree.R:

```
library("igraph")

library("poweRlaw")

library("ggplot2")


# Just loading my data

edge_list <- read.csv("Amazon0302.csv")

g <- graph.data.frame(edge_list)


write.csv(betweenness(g), file = "betweeness_Amazon0302.csv")


write.csv(degree(g), file = "degree_Amazon0302.csv")
```

JT_Test.R:

```
library("SAGx")

library("igraph")

library("poweRlaw")

library("ggplot2")
```

```
#G <- as.matrix(c(99,114,116,127,146,111, 125,143,148,157,133,139, 149, 160, 184))

G <- as.matrix(c(1266.43314, 1891.08269, 4412.3894, 31989.28866,
          9.17391, 61.11111, 72.37705, 603.06889,
          11752.64857, 18116.70758, 19105.3258, 24967.38829,
          142.10986, 670945.77595, 1583160.9857, 47552794845.82855,
          13.14198, 17.30282, 51.41424, 70.74157))

# create the class labels
g <- c(rep(1, 4),rep(2, 4),rep(3, 4), rep(4, 4), rep(5, 4))
# The groups have the medians
tapply(G, g, median)
# JT.test indicates that this trend is significant at the 5% level
JT.test(data = G, class = g, labs = c("GRP 1", "GRP 2", "GRP 3", "GRP 4", "GRP 5"), alternative = "two-
sided")
```