Bangladesh University of Engineering and Technology

Department of Computer Science and Engineering

Course Code: CSE6413

Course Title: Network Science

Submitted To

Dr. Md. Saidur Rahman

Professor

Department of Computer Science and Engineering

Bangladesh University of Engineering and Technology

Submitted By

Rafi Khandokar, ID - 0422052078

Twitter network analysis through follow & retweet data

Abstract: Social networks are without a doubt one of the most discussed data sources for data scientists to investigate. Users of the microblogging website Twitter communicate with one another by sending "tweets." Twitter datasets are regarded as enormous resources for network analysis and are an excellent source of ideas. In this study, we have covered many analytical aspects of modeling a network situation using Twitter following relationship data. Follow relationship data explains how users engage with one another by counting the followers each user has on their Twitter accounts. This dataspace has been examined using statistical methods, and graphics have been used in a number of different distribution techniques.

Keywords: Network Analysis, Social network Data, Twitter user relationship, Follow and Followed, Graph distribution method, Twitter.

1. Introduction

Individuals may set up restrictions and channels for the flow of information through social media by using it. Information flow in social media platforms is influenced by a variety of interactive activities with individuals and material, including following, friending, subscribing, posting, and retweeting. An individual might find classic and non-traditional views on social media. He has the ability to inform people and serve as a resource for knowledge for them. Use venues and networks that are constructed around the type of social hierarchy that incorporates social norms and where conversations are about the activities of people with strong connections and good connections as an alternative to social media. By restricting contacts to a chosen set of like-minded persons, networks inside networks serve as an example of how to replicate the existing hierarchical structure of real life.

This report introduces a network analysis of the Twitter network, employing various social network ideas and their subsequent connection pattern measurements. Density and modularity are terms used to describe the degree to which users and clusters are interconnected. The degree to which connections are aggregated around just a few individuals in the network is used to represent networks based on information flow hierarchy. ("links," "ties," or "edges") are formed between social players, such as individuals and organizations. Online social networks enable users to build bonds or relationships among themselves while exchanging images, text, videos, and other digital artifacts. By emphasizing the social capital component of the network, or the relationships tying the entities together, the network viewpoint challenges a conventional approach to the study of social media. Emergent patterns or network motifs are formed from groups of these linkages or connections. Users engage with one another when they connect or share information on social networking sites, forming networks. Social networks on Twitter are made up of people and the relationships they make with one another through mentions and replies. When given the chance to engage freely, people and other social actors prefer to create smaller groups of linked individuals who are more interconnected with one another than with other, less connected members of their social network, according to network research. The fact that nodes in self-organized networks have the freedom to add and remove connections is one of their main characteristics. Milgram carried out various studies that served as the foundation for the social network research literature on small world theory. Milgram concluded that human society is made up of tiny clusters of closely linked individuals, regardless of the size of a social network. This led to a short average route length between any two persons. Strong links have the propensity to collapse into triangles, forming clusters of densely connected groups, according to Gravette, but weak ties are less likely to follow this tendency and often connect distinct clusters. In a network of pathways that would otherwise be lengthy, these patterns of interaction enable the development of shorter links. Watts discovered that high levels of local clustering or subgroups are connected to the short path length—short global separation—in many big networks (e.g., churches, Rotary Clubs, or neighborhood watch groups). One of the key discoveries was that "small world networks" had a high density of local clusters of persons in otherwise big and sparse networks. These clusters are linked by a small number of links, which weave the wider network into a more densely connected structure. The tiny world network structure was discovered in a variety of scale-free networks, including the Western United States' power grid, cooperation among film actors, metabolism, sexual interactions, "offline" friends, and relationships on Myspace.

The most well-known finding of this corpus of study is without a doubt the modest distances of connections between persons inside a network, sometimes referred to as the "six degrees of separation." But this study starts by asking unanswered concerns about Twitter networks: What types of network architectures might social media networks organically produce? What distinguishing structural traits do these network shapes have? How can social media networks be properly divided into usable categories using network metrics? Density, Clusters, and Information Flow: Density, Clusters, and Information Flow The intensity and bounds of information flow are reflected in the modularity, the amount to which users are interconnected, and the patterns of these relationships. Information is more likely to circulate among firmly linked individuals than among loosely connected persons. As clusters evolve as subgroups of networked individuals, connectedness between clusters impacts information flow across groups and can be an indicator of information sharing among diverse groupings. The rate of connectivity between users and clusters describes networks and can show relevant distinctions between them. Density, modularity, and isolation are three social network concepts that reflect these characteristics and help characterize the diversity of network architectures.

The interconnectedness of individuals in a network is captured by density. It can range from low density, where users are weakly connected, to high density, where users are strongly connected. The rate at which information moves through a network depends on how highly linked it is. Carley demonstrated how contact among people creates common knowledge, which in turn creates additional interaction. This result has significant ramifications for the stability of a group and its relationships with those outside its limits.

Clustering, also known as network transitivity, is a characteristic of many networks where two nodes linked to the same third node have a higher likelihood of also being connected to one another. Simply said, two of your friends have a higher chance of knowing one another than two strangers selected at random from the general population. Clusters arise as a network expands, whereas connections between these clusters become less dense. Clusters are subgroups of nodes inside a larger network that become increasingly interconnected and dense. Clusters are formed in social media when, for instance, connections are strategically made between blogs, individuals follow one another on Twitter, and they become Facebook friends. When users make these connections, they open up new channels for information to go along. The resultant groupings determine the social limits of information flow. Information can move freely inside these clusters, but it cannot move freely between clusters due to the limited connection between clusters. People living in interrelated clusters frequently have traits in common. Homophily is a situation where "interaction between similar persons happens at a faster rate than among dissimilar people," according to its definition. A crucial aspect of naturally existing social networks is homophily, also known as the saying "birds of a feather flock together," which illustrates how "distance in terms of social qualities translates into network distance." Simply set, homophily shows that people are more inclined to interact with others in their social networks who are frequently extremely similar to them. Therefore, homophily indicates that similar people will be more socially proximate to one another than different people. Researchers discovered that homophily is greatly influenced by demographic factors including age, sex, race, and education. They have studied homophily for over a century. Additionally, it was discovered that homophily is dependent on psychological traits including IQ, attitudes, and ambitions.

2. Experimental Details

This study makes use of a dataset that Kaist previously made available. Additionally, a Python-based application called NetworkX was utilized to extract the data's numerous statistics and visualizations.

2.1 Dataset

The dataset represents Twitter's follower network, which includes 1.4 billion directed follow edges between 41 million Twitter users. It is available at: http://an.kaist.ac.kr/traces/WWW2010.html. This is a snapshot of Twitter users in 2010.

2.2. Analysis Tool

The network has been explored using NetworkX. A Python module called NetworkX is used to investigate networks and graphs. Under the terms of the BSD-new license, NetworkX is free software. Large real-world graphs, such as those with more than 10 million nodes and 100 million edges, may be operated on using NetworkX. NetworkX is a moderately effective, extremely scalable, and highly portable framework for network and social network research due to its reliance on a pure-Python "dictionary of the dictionary" data structure.

3. Methodology

We choose to use networks and graph theory to define social systems. Users are seen as nodes, and the people's connections to the prominent users they follow are viewed as edges (in/out-degree). Scatter plots and varying degree distributions are then created. Fundamentally, by examining the relationships between entities, these elements provide better explanations of social processes. In its most basic form, our suggested method for examining the Twitter dataset for the provided connection resembles the illustration below:

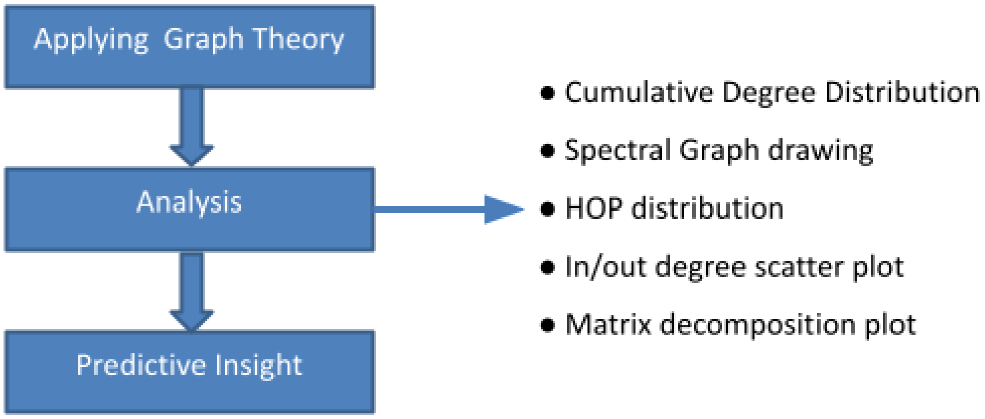

Figure: An explanatory image of step-by-step network analysis.

Since there are 342 networks in our Twitter dataset, showing the complete network will be confusing. By creating graphs for each network, we want to convey the scenario's structure in an intelligible manner. We may see the social network Twitter from many angles thanks to various distribution channels. In order to characterize the key characteristics of vertices and edges, regions, or the entire graph, network statistics assist remove extraneous information, making statistical techniques better for network research. In the next section, we'll go through the analytical elements that we plotted for a specific network such as

- Cumulative Degree Distribution
- Spectral Graph Drawing
- HOP Distribution
- In-Out Degree Scatter Plot
- Matrix Decomposition Plot

4. Results and Analysis

In the discussion of statistical methods, the ratio of nodes, edges, and their degrees is visually depicted. We will detail several strategies in accordance with our proposed plan in the sections that follow.

4.1 General Statistics

The dataset definitions listed below are essential for our research

Network format	Unipartite directed
Edge type	Unweighted, no multiple edges
Reciprocal	Contains reciprocal edges
Directed cycles	Contains directed cycles
Loops	Does not contain loops
Snapshot	Is a snapshot and likely to not contain all data
Connectedness	Only the largest connected component of the original data is included

4.1.1 The size in numbers

The dataset is quite substantial. In contrast to followers, who are represented by the number of edges, users are represented by the number of vertices. The edges are directed from follower to following.

Vertices	41,652,230
Edges	1,468,365,182
Loops	298
Wedges	123,435,589,564,841
Triangles	34,824,916,864
Clustering Coefficient	0.000 846 391

The number of wedges and triangles reflects the effect of the followers-to-followers ratio. There aren't many renowned individuals, but there are a lot of fans. As a result, they generate a lot of wedges but very few triangles. By 3544, wedge numbers outweigh triangular numbers. As a result, the clustering coefficient is quite low.

4.1.2 Degree distribution

The degree distribution represents the impact of the network's following follower relationships.

Maximum Degree	3,081,112
Maximum Outdegree	770,155
Maximum Indegree	2,997,469
Average Degree	70.506
Size of Largest Connected component	41,652,230
Diameter	23
50-percentile effective diameter	3.076 70

90-percentile effective diameter	3.973 28
Mean distance	3.549 32
Non-bipartivity	0.319 460

The highest level denotes the enormous amount of supporters someone had. It represents the total number of followers the person had, as well as those they followed. The typical degree is not that high. The biggest linked component, as shown, is the same size as the entire. Since every user in the snapshot is associated with at least one other user in some way. The diameter of 23 demonstrates how the small world effect operates in this situation. Only three hops are required to reach 50% of the users.

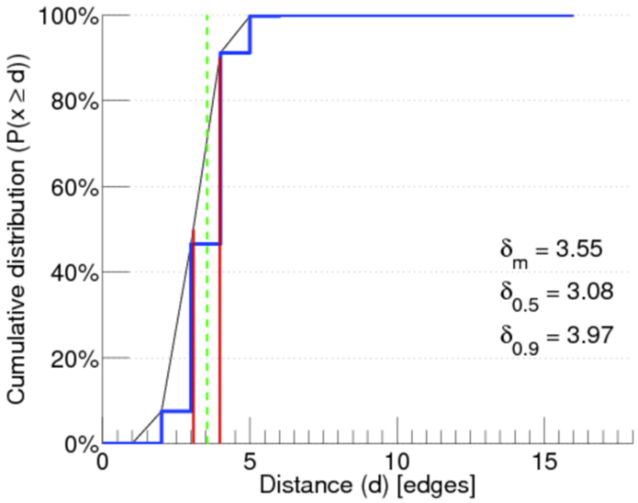

Figure: Distance graph

If a graph can be split into two sections with no connections between them, then the graph is bipartite. The non-bipartivity number demonstrates how distant from being bipartite the graph is.

4.2 Cumulative Degree Distribution

The cumulative degree distribution depicts the total number of nodes in each degree class. The distribution of degree values across all nodes defines the network as a whole. The degree distribution is often characterized as, P(k) = n_k /n. This indicates the likelihood that a certain node has degree k. The cumulative frequency of nodes (10^0-10^6) and node degree p(x) for both in and out-degree (d+ and d-) were used to create a graph. Node degree is displayed on the y-axis, while the cumulative frequency of the nodes is represented on the x-axis. The cumulative number of in-degree nodes and the inward connections at the vertices are displayed in the in-degree cumulative distribution.

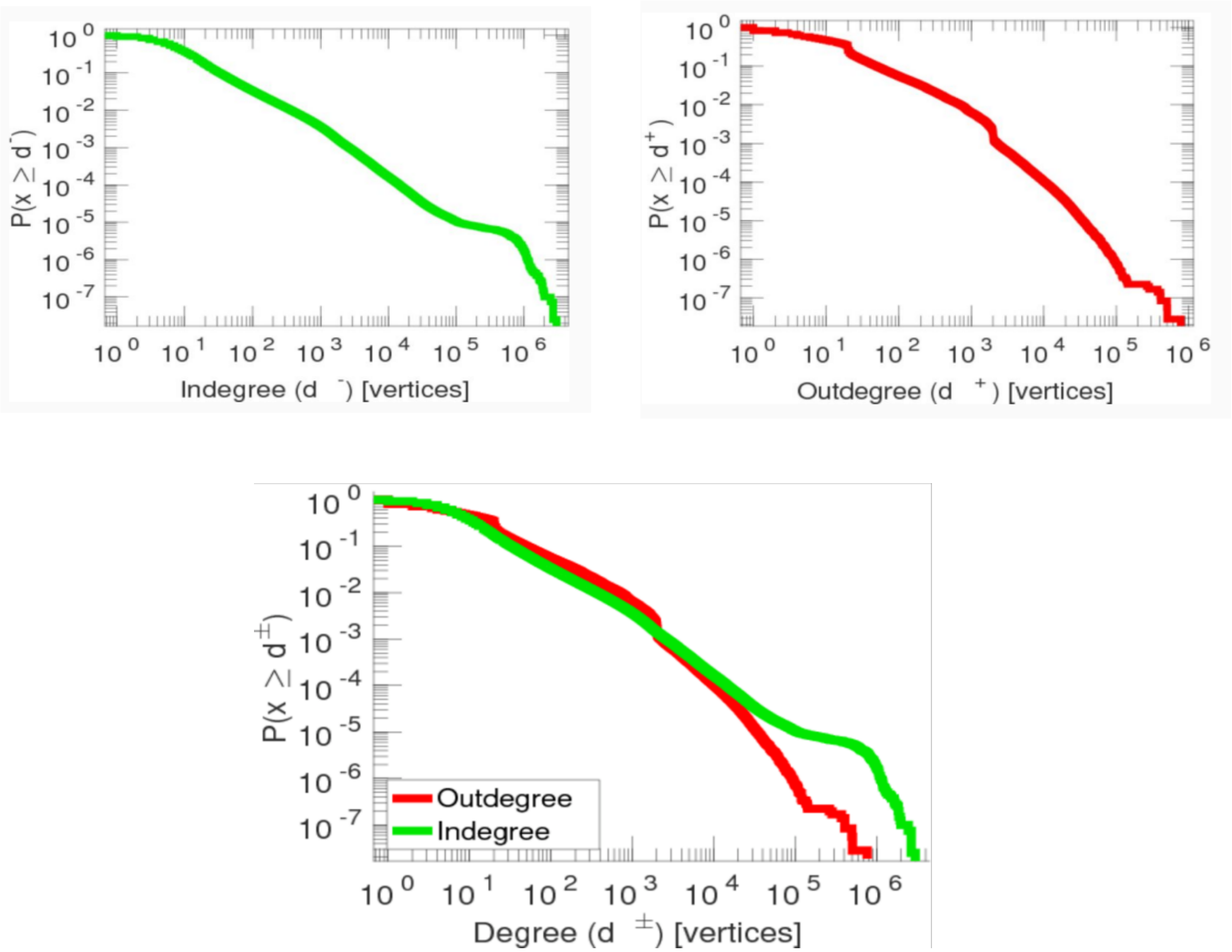

Figure: Cumulative degree Distribution graph of in-degree, out-degree, and both

The cumulative number of out-degree nodes and outward connections to other vertices are referred to as the "out-degree cumulative distribution." We plotted two of these graphs simultaneously to make them easier to grasp. The out-degree curve is more pronounced than the in-degree one because well-known users, politicians, and celebrities have a large number of followers but do not follow many individuals.

4.3 Spectral Graph Drawing

A spectral graph is used to depict a graph's criteria based on the eigenvalues and eigenvectors of matrices linked with that graph. Matrix types include Laplacian and adjacency matrices. Adjacency matrices were shown to be more dependable in our investigation for producing the scatter graph. We created a spectral graph of a graph based on its adjacency matrix to determine its layout. To be more specific, the eigenvector of linked matrices was employed as the plotted measurement. The term "eigenvector" refers to the degree of connectivity between a node and its neighboring nodes. The spectrum of the related graph is the set of the graph's eigenvalues. We are adding a picture of a section of the massive spectral graph produced by the whole network that contains certain eigenvalues derived from their adjacency matrices.

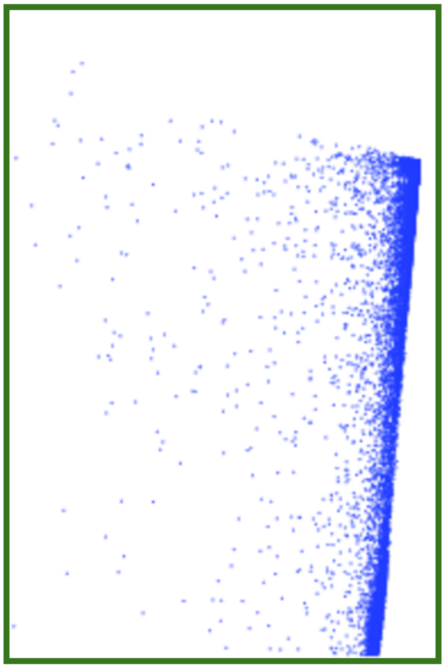

Figure: Spectral graph representing the eigenvectors

This spectral graph illustrates how people and their regular Twitter followers are arranged. Twitter follower relationships are shown in terms of a user's level of popularity. The graph demonstrates how people who have more than one account following them are more likely to have adjacency matrices transformed into eigenvalues and eigenvectors.

4.4 HOP Distribution Method

If a node can be specified using pairwise distance requirements, the HOP distribution is a graphical depiction of that node in a network. The fact that it measures the distance between two or more nodes gives it the nickname "better than degree distribution." For the purposes of our network, network analysis just requires a method that can calculate distances over a sizable network. To plot the graph, we utilized logistic and the tangent sigmoid distribution. The HOP count is regarded as a fundamental gauge of the network's distance. Understanding the approximate distance between a certain number of Twitter users who follow other users and those who are followed is the aim of this distribution. The frequency is in the X-axis and the degree distance is in the Y-axis in logistic and tangent sigmoid, respectively. The scaling equation for the HOP distance H(d) for logistic distribution was,

$$Logistic(\, -\log(\, \frac{1}{H}(\, d) - 1)\,)$$

Likewise, the scaling equation for the HOP distance H(d) for the Tangent Sigmoid was

$$Tangent\left(\tan \frac{H(\, d) \; + \dfrac{\pi}{2}}{\pi} \right)$$

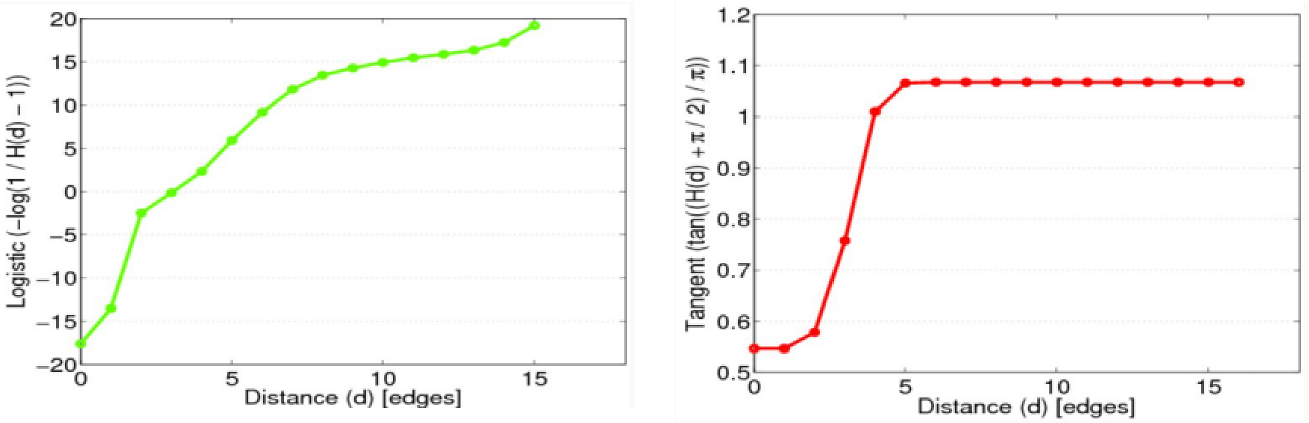

Figure: Logistic and Tangent sigmoid graph following the HOP distribution

4.5 In-Degree and Out-Degree Scatter Plot

The comparison of a large number of data points without respect to time is possible using scatter plot graphs. The in-degree nodes and out-degree nodes variables are taken into account for the comparison. The indicated node degrees' normal values are first displayed on a graph. Another graph is afterward drawn with enhanced values. Increased in and out-degree values allude to a performance that is at its best while maintaining overall controllability. Even though it was discovered after plotting that enhanced values are not significantly different from regular values, this feature is still used to highlight the network's dense connections.

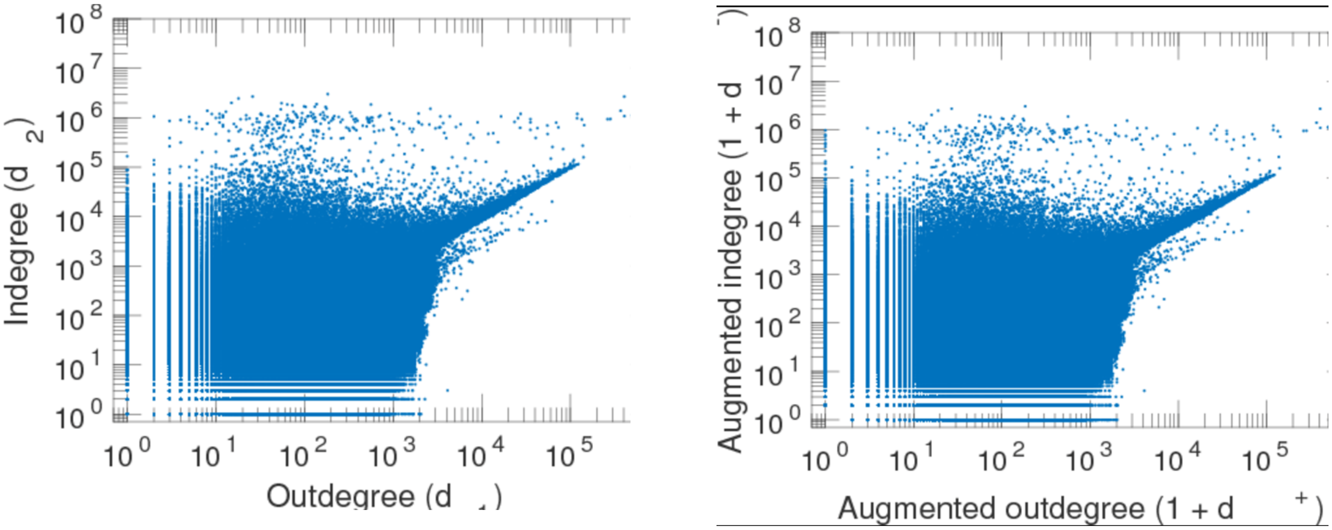

Figure: In-degree and Out-degree scatter plot with normal values(left) and augmented values(right)

5. Conclusion

We can construct twitter datasets for retweet datasets since we have covered five distribution or statistical approaches for analyzing Twitter networks. Our long-term objective is to provide a framework for both the Twitter follow and retweet datasets, allowing for more defined and efficient analysis. Twitter follow relationship data helps us understand the link between followers and users, while retweet data helps us understand the connection between users who follow and retweet each other.

Reference:

1. THE TWITTER EXPLORER: A FRAMEWORK FOR OBSERVING TWITTER THROUGH INTERACTIVE NETWORKS [https://arxiv.org/ftp/arxiv/papers/2003/2003.03599.pdf]
2. Social Network Analysis and Mining: Privacy and Security on Twitter [https://www.researchgate.net/profile/Zahraddeen-Gwarzo-2/publication/356172278_Social_Network_Analysis_and_Mining_Privacy_and_Security_on_Twitter/links/618e9584d7d1a f224be0d1d8/Social-Network-Analysis-and-Mining-Privacy-and-Security-on-Twitter.pdf]
3. What is Twitter, a Social Network or a News Media?[https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.947.64&rep=rep1&typ e=pdf]
4. dynamic Analysis of Twitter and Facebook Through Social Network Analysis [https://www.iaras.org/iaras/filedownloads/ijels/2021/002-0001(2021).pdf]