ISSN No. 0976-5697

Volume 8, No. 3, March – April 2017

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

Implementation of Enhanced Graph Layout Algorithm for Visualizing Social Network Data using NetworkX Library

Mandeep Kaur*
M.Tech Scholar, CSE Dept
SBBSIET, Padhiana
Jalandhar, India

Harpreet Kaur
Assistant Professor , CSE Dept
SBBSIET, Padhiana
Jalandhar, India

***Abstract*:** Networks are critical to modern society, and a thorough understanding of how they behave is crucial to their efficient operation. Fortunately, data on networks is plentiful; by visualizing this data, it is possible to greatly improve our understanding. Firstly, this paper discusses about social media and the use of Social Media Analytics by businesses. Second, how we plan to use available resources to analyze the sample case of a given social network. As a Case Study, this paper will demonstrate how by starting at a single twitter account we can build up a network graph of twitter followers and then visualize that network using the NetworkX library. In this seed account is used to collect twitter followers. After that, Process the collected twitter data to generate an output file of relationships between twitter accounts. At last, Visualize network data in a network graph using the NetworkX library. In this paper Snowball sampling algorithm is used.

Keywords: Network visualization, Parameter focusing, social media Analysis, Network data analysis, Data manipulation

I. INTRODUCTION

Social media enables users to generate content by sharing their knowledge, opinions and experiences on a variety of issues. Social media has changed the way customers engage with organizations, brands, products and services. It influences customer attitudes, perceptions and buying decisions. Social media provides organizations with many opportunities. It provides a new and powerful low cost marketing channel that can be harnessed to increase customer awareness of organizations and associated brands, products and services. Also, it enables organizations to improve their customer relationships through better engagement on a real time basis. [1]

We are currently in the midst of a networking revolution. Data communications networks such as the Internet now connect millions of computers; cellular phones have become commonplace, and personal communications networks are in the developmental stages. In parallel with the ever increasing network sizes has been a concomitant increase in the collection of network measurement data. Understanding this data is of crucial importance as we move to a modern, information-rich society

Unfortunately, tools for analyzing network data have not kept pace with the data volumes. More network measurement data is available today than ever before, yet it is useless until it is understood. [2] Traditional network analysis software and graphs cannot cope with the size of today's networks and their data collection capabilities. In 2016, it was estimated that there will be around 2.13 billion social network users around the globe; up from 1.4 billion in 2012.Social network penetration worldwide is ever-increasing. In 2012, 63.1 percent of internet users were also social network users and these figures were expected to grow.

Twitter gained 42 million users this year. The site welcomes 32 percent of Internet users age 18 to 24, and about 24 percent of that demographic has the app downloaded on a mobile device. About 86 percent of users access the site through mobile devices. Users spend an average of 17 minutes per day on the site, and 37 percent of users say they will buy products from a brand they follow on Twitter.

Second, As a Case Study, this paper will demonstrate how by starting at a single twitter account we can build up a network graph of twitter followers and then visualize that network using the NetworkX library.
The steps are:

- From initial seed account collect twitter followers.
- Process the collected twitter data to generate an output file of relationships between twitter accounts.
- Visualize network data in a network graph using the NetworkX library.[3]

There are various social network models which are as follows:

A. Using formal methods to show Social Networks:

One reason for using mathematical and graphical techniques in social network analysis is to represent the descriptions of networks compactly and systematically. A related reason for using (particularly mathematical) formal methods for representing social networks is that mathematical representations allow us to apply computers to the analysis of network data. The third, and final reason for using "formal" methods (mathematics and graphs) for representing social network data is that the techniques for graph processing and the rules of mathematics themselves suggest things that we might look for in our data. In the analysis of complete networks, a distinction can be made between

• Descriptive methods, also through graphical representations

• Analysis procedures, often based on a decomposition of the adjacency matrix

• Statistical models based on probability distributions

B. Using Graphs to Represent Social Relations:

Network analysis uses (primarily) one kind of graphic display that consists of points (or nodes) to represent actors and lines (or edges) to represent ties or relations. When sociologists borrowed this way of graphing things from the mathematicians, they renamed their graphs as"sociograms". There are a number of variations on the theme of sociograms, but they all share the common feature of using a labeled circle for each actor in the population we are describing, and line segments between pairs of actors to represent the observation that a tie exists between the two.

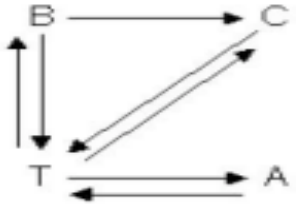

Figure 1.Using Graphs to Represent Social Relations

Visualization by displaying a sociogram as well as a summary of graph theoretical concepts provides a first description of social network data. For a small graph this may suffice, but usually the data and/or research questions are too complex for this relatively simple approach.

C. Using Matrices to Represent Social Relations:

The most common form of matrix in social network analysis is a very simple one composed of as many rows and columns as there are actors in our data set, and where the elements represent the ties between the actors. The simplest and most common matrix is binary. That is, if a tie is present, a one is entered in a cell; if there is no tie, a zero is entered. This kind of a matrix is the starting point for almost all network analysis, and is called an "adjacency matrix" because it represents who is next to, or adjacent to whom in the "social space" mapped by the relations that we have measured. By convention, in a directed graph, the sender of a tie is the row and the target of the tie is the column. Let's look at a simple example. The directed graph of friendship choices among Bob, Carol, Ted, and Alice looks like figure 1. Since the ties are measured at the nominal level (that is, the data are binary choice data), we can represent the same information in a matrix that looks like Table 1:

TABLE I. Using Matrices to Represent Social Relations

	Bob	Carol	Ted	Alice
Bob	---	1	0	0
Carol	1	---	1	0
Ted	1	1	---	1
Alice	0	0	1	---

D. Statistical Models for Social Network Analysis:

Statistical analysis of social networks spans over 60 years. Since the 1970s, one of the major directions in the field was to model probabilities of relational ties between interacting units (social actors), though in the beginning only very small groups of actors were considered. Extensive introduction to earlier methods is provided by Wasserman and Faust. Two of the most prominent current directions are Markov Random Fields (MRFs) introduced by Frank and Strauss and Exponential Random Graphical Models (ERGMs), also known as p. There are several useful properties of the stochastic models. Some of them are:

• The ability to explain important properties between entities that often occur in real life such as reciprocity, if i is related to j then j is more likely to be somehow related to i; and transitivity, if i knows j and j knows k, it is likely that i knows k.

• Inference methods for handling systematic errors in the measurement of links.

• General approaches for parameter estimation and model comparison using Markov Chain Monte Carlo methods.

• Taking into account individual variability and properties (covariates) of actors.

• Ability to handle groups of nodes with equivalent statistical properties.

There are several problems with existing models such as degeneracy analyzed by and scalability mentioned by several sources. The new specifications for the Exponential Random Graph Models proposed in attempt to find a solution for the unstable likelihood by proposing slightly different parameterization of the models than was used before.[3]

II. SOCIAL NETWORK PROPERTIES

There are some properties of social networks that are very important such as size, density, degree, reachability, distance, diameter, geodesic distance. Here we describe some more complicated properties which may be used in social network analysis.

***A. Maximum flow*:** One notion of how totally connected two actors are, asks how many different actors in the neighborhood of a source lead to pathways to a target. If I need to get a message to you, and there is only one other person to whom I can send this for retransmission, my connection is weak - even if the person I send it to may have many ways of reaching you. If, on the other hand, there are four people to whom I can send my message, each of whom has one or more ways of retransmitting my message to you, then my connection is stronger. This" flow" approach suggests that the strength of my tie to you is no stronger than the weakest link in the chain of connections, where weakness means a lack of alternatives.[4]

***B. Hubbell and Katz cohesion*:** The maximum flow approach focuses on the vulnerability or redundancy of connection between pairs of actors - kind of a" strength of the weakest link" argument. As an alternative approach, we might want to consider the strength of all links as defining the connection. If we are interested in how much two actors may influence on one another, or share a sense of common position, the full range of their connections should probably be considered. Even if we want to include all connections between two actors, it may not make a great deal of sense (in most cases) to consider a path of length 10 as important as a path of length 1. The Hubbell and Katz approaches count the total connections between actors (ties for undirected data, both sending and receiving ties for directed data). Each connection, however, is given a weight, according to its length. The greater the length, the weaker the connection.

C. Taylor's Influence: The Hubbell and Katz approach may make most sense when applied to symmetric data; because they pay no attention to the directions of connections (i.e. A's ties directed to B are just as important as B's ties to A in defining the distance or solidarity – closeness– between them). If we are more specifically interested in the influence of A on B in a directed graph, the Taylor influence approach provides an interesting alternative. The Taylor measure, like the others, uses all connections, and applies an attenuation factor. Rather than standardizing on the whole resulting matrix, however, a different approach is adopted. The column marginals for each actor are subtracted from the row marginals, and the result is then normed. Translated into English, we look at the balance between each actor's sending connections (row marginals) and their receiving connections (column marginals). Positive values then reflect a preponderance of sending over receiving to the other actor of the pair -or a balance of influence between the two-. D. Centrality and Power All sociologists would agree that power is a fundamental property of social structures. There is much less agreement about what power is, and how we can describe and analyze its causes and consequences. Table I summarizes some of the main approaches that social network analysis has developed to study power, and the closely related concept of centrality. [5]

TABLE II. Comparing three aspects of power in sociograms (degree, closeness, and betweenness)

Power Aspect Name	Definition	Influences
Degree	Number of ties for an actor	Having more oppurtunities and alternatives
Closeness	Length of paths to other actors	Direct bargaining and exchange with other actors
Betweenness	Lying between each other pairs of actors	Brokering contacts among actors to isolate them or prevent connections

III. RELATED WORK

Once graph data have been extracted from the respective social media platforms (the social networking sites, the microblogging sites, the image-sharing sites, the video sharing sites, the blog, wiki, the online encyclopedia, or the WWW and Internet), and the data have been processed (with graph metrics computed and groups identified), the next step is the graph layout. [6] The appearance of each graph will be necessarily unique because the data extracted from the social media platforms will generally be unique. The data feature that seems to most inform the graph layout is the number of nodes or vertices, or the data density vs. sparsity. (Some content networks may be less dynamic than other types of information.) It is helpful to remember that the data informing the graph layout in the graph pane is housed in the worksheets in the NodeXL Graph Template. Also, the data worksheets and the graph pane are interactive so if a cluster is highlighted in the pane, the highlighted records are indicated on the left, and vice versa. [7]

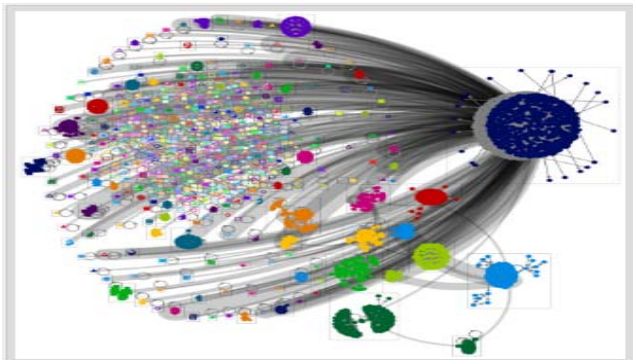

Figure 2 Hashtag Network on Twitter (basic network) [8]

IV. PROPOSED WORK

Following are the four major the steps that are used in proposed work:

Step1: Collect follower data from the Twitter API:

We have API keys to be able to query the Twitter API. While interacting with the Twitter API we learned that we need to cache data as we go along. This is because the API is rate limited and we find the script we write halts frequently when hitting a rate limit if we don't cache responses. The solution is to check for cached data before making an API call, if we get a cache miss then query the API and write the returned data to disk.

There are two directories for cached data. The directory 'following' contains a CSV file for each twitter account queried. The name of each file is the screen name of the twitter account and the content is a tab delimited list, each row contains the twitter id, screen name and account name of a follower, up to a maximum of 200 followers.

Step2. Process twitter data to generate an output file of relationships between twitter accounts:

The script below will process the data collected from the twitter API and generate an edge list. That is a list of relationships between twitter accounts. A weight value is included, this value is the total number of followers for the first twitter account, and this value is retrieved from the API.

The weight value can be used later to prune the network graph.

Step3. Visualizing the Network using the NetworkX library:

This gives the data we need to generate a network graph. These are the steps used to visualize the network graph:

1. Create a directed graph (net.DiGraph) containing all the edge data including metadata.

2. Remove nodes based on how connected they are to other nodes in the network (i.e. remove poorly connected nodes).

3. Remove edges that have less than a minimum number of followers.

4. Split nodes into two separate categories, 'TED' and 'non-TED' sets.

5. Render each nodeset.

6. Render edges between nodes.

7. Render node labels.

4. Output:

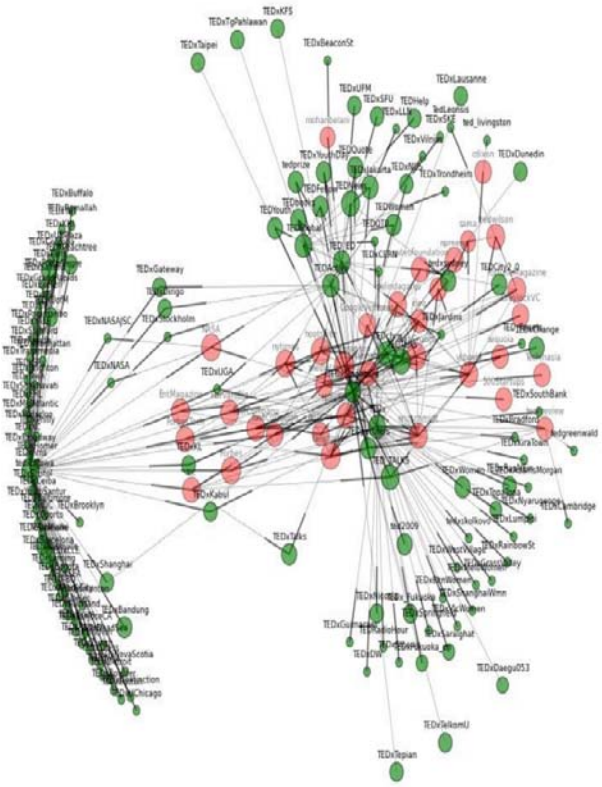

Figure 3 Output of proposed work

FLOWCHART OF PROPOSED ALGORITHM

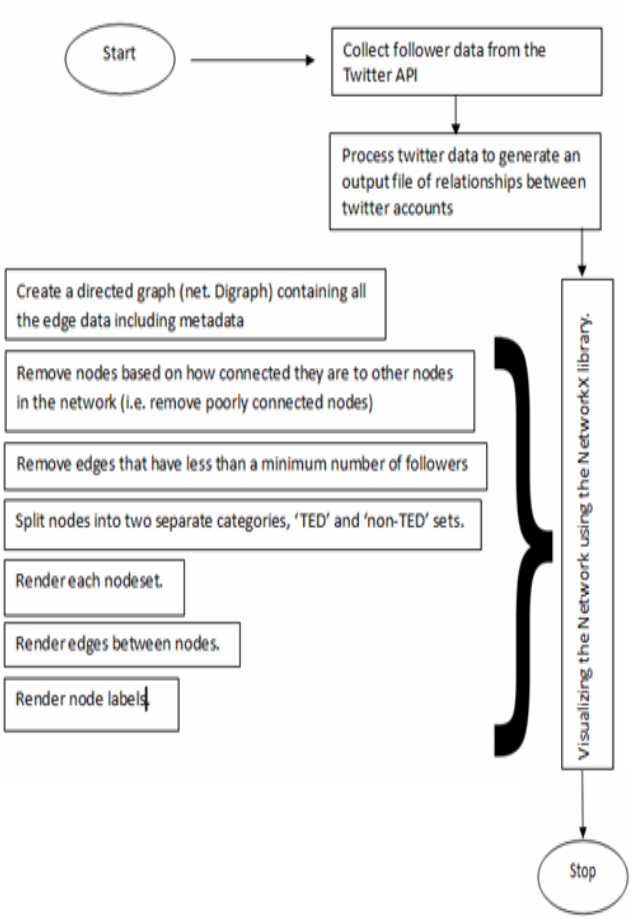

Figure 4 Flowchart of proposed algorithm

V. PROBLEM FORMULATION

Social network analysis (SNA) is the process of investigating social structures through the use of network and graph theories. It characterizes networked structures in terms of nodes (individual actors, people, or things within the network) and the ties, edges, or links (relationships or interactions) that connect them. Examples of social structures commonly visualized through social network analysis include social media networks, memes spread, friendship and acquaintance networks, collaboration graphs, kinship, disease transmission, and sexual relationships. These networks are often visualized through sociograms in which nodes are represented as points and ties are represented as lines. In the visualization of Social Network (Twitter), we use a following algorithm:

- From initial seed account collect twitter followers.
- Process the collected twitter data to generate an output file of relationships between twitter accounts.
- Visualize network data in a network graph using the NetworkX library.

OBJECTIVES OF PROPOSED WORK

The objectives of the proposed work are:

1. Study various Social Network Analytics Papers and Visualization Tools and Techniques.

2. Determine the method for efficient plotting of existing relationships with default seed account.

3. Implement the proposed algorithm to derive a visual representation of Social Network Data.

The approach behind the objectives of the proposed work is given below:

Social network analysis focuses on patterns of relationships between actors and examines the availability of resources and the exchange of resources between these actors. In this research, firstly I collect the initial seed account from the twitter a social network which I am using in this research. After that to process the collected twitter data to generate an output file of relationships between twitter accounts, which may be more than two accounts and then at last Visualize network data in a network graph using the NetworkX library.NetworkX library is a Python library for studying graphs and networks. NetworkX is free software released under the license.Follwing are the features of NetworkX library which are helpful in making this project:

- Classes for graphs and digraphs.
- Conversion of graphs to and from several formats.
- Ability to construct random graphs or construct them incrementally.
- Ability to find subgraphs, cliques, k-cores.
- Explore adjacency, degree, diameter, radius, centre, betweenness, etc.
- Draw networks in 2D and 3D

VI. RESULT AND DISCUSSION

In this research work, python tool is used. Python is a widely used high-level programming language used for general-purpose programming, created by Guido van Rossum and first released in 1991. An interpreted language, and a syntax which allows programmers to express concepts in fewer lines of code than possible in languages such as C++ or Java.

In this research three python based libraries one is NetworkX library second tweepy and pyplot. NetworkX library is a Python language software package for the creation, manipulation, and study of the structure, dynamics, and functions of complex networks. Tweepy is used to download your home timeline tweets and print each one of their texts to the console. Twitter requires all requests to use for authentication.

Pyplot is a collection of command style functions that make matplotlib work like MATLAB. Each pyplot function makes some change to a figure: e.g., creates a figure, creates a plotting area in a figure, plots some lines in a plotting area, decorates the plot with labels, etc. In matplotlib. pyplot various states are preserved across function calls, so that it keeps track of things like the current figure and plotting area, and the plotting functions are directed to the current axes (please note that “axes” here and in most places in the documentation refers to the *axes* part of a figure and not the strict mathematical term for more than one axis).’

Following is the graph showing the follower data of seed account up to three levels.

Figure 5. followers data upto three levels

These are the parameters that has been used in getting the followers from the seed account.

Table III. Parameters used

Parameter Name	Parameter Usage
CONSUMER_KEY	The consumer keys can be found on application's Details page located at https://dev.twitter.com/apps (under "OAuth settings")
CONSUMER_SECRET	The consumer keys can be found on application's Details page located at https://dev.twitter.com/apps (under "OAuth settings")
ACCESS_TOKEN	The access tokens can be found on application's Details page located at https://dev.twitter.com/apps (located under "Your access token")
ACCESS_TOKEN_SECRET	The access tokens can be found on application's Details page located at https://dev.twitter.com/ap

	ps (located under "Your access token")
INITIAL_USER_NAME	This is the account whose graph is to be plotted
DEPTH	How many tiers of data to be fetched

VII. CONCLUSION AND FUTURE SCOPE

Visualizing social networks is of immense help for social network researchers in understanding new ways to present and manage data and to effectively convert the data into meaningful information.

Social Network Analysis (SNA) is becoming an important tool for investigators, but all the necessary information is often distributed over a number of Web servers. Currently there are developing information system that helps managers and team leaders to monitor the status of a social network. This paper presented an overview of the basic concepts of social networks in data analysis including social network analysis metrics and performances. Different problems in social networks are discussed such as uncertainty, missing data and finding the shortest path in a social network. Community structure, detection and visualization in social network analysis were also discussed. The current implementation includes analyzing one social network connection map. As a future enhancement additional modules can be included to derive similar pattern from a group of heterogeneous social network.

VIII. ACKNOWLEDGEMENT

The satisfaction that accompanies that the successful completion of any task would be incomplete without the mention of people whose ceaseless cooperation made it possible, whose constant guidance and encouragement crown all efforts with success.

The project entitled “Implementation of enhanced graph layout algorithm for visualization of social network data using NetworkX library “is done under the guidance of my teacher “Er. Harpreet Kaur”. I am very thankful to my teacher for the inspiration and constructive suggestions that helpful us in the preparation of this project.’

IX. REFERENCES

[1] Marcin Mincera and Ewa Niewiadomska-Szynkiewicza, Institute of Control and Computation Engineering, Warsaw University of Technology, Warsaw, Poland Research and Academic Computer Network (NASK), Warsaw, Poland Application of Social Network Analysis to the Investigation of Interpersonal Connections, journal of telecommunications and information technology 2012.

[2] Caroline Haythornthwaite Graduate School of Library and Information Science University of Illinois, Urbana-Champaign Social Network Analysis: An Approach and Technique for the Study of Information Exchange LISR 18, 323-342 (1998)

[3] IEEE Transactions on Visualization and Computer Graphics, Vol. 1, No. 1, pages 16-21, March 1995. Visualizing Network Data Richard A. Becker Stephen G. Eick Allan R. Wilks.

[4] Different Aspects of Social Network Analysis Mohsen Jamali and Hassan Abolhassani Web Intelligence Research Laboratory Computer Engineering Department Sharif University of Technology, Tehran, Iran.

[5] Application of Social Network Analysis to the Investigation of Interpersonal Connections Marcin Mincera and Ewa Niewiadomska-Szynkiewicza,b a Institute of Control and Computation Engineering, Warsaw University of Technology, Warsaw, Poland b Research and Academic Computer Network (NASK), Warsaw, Poland in Interpersonal Connections, journal of telecommunications and information technology 2012.

[6] International Symposium on Social Science (ISSS 2015) A Survey on Social Network Visualization Jiang Du1 & Yafei Xian1, Jiayu Yang2 1College of Computer Science and Technology, Chongqing University of Posts and Telecommunications 2Chongqing iSoft Network Security Information Technology Co., Ltd.

[7] Social Network Analysis of Online Marketplaces Pushpa Kumar Kang Zhang Department of Computer Science, University of Texas at Dallas, Richardson, TX 75083 {pkumar, kzhang}@utdallas.edu.

[8] http://scalar.usc.edu/works/querying-social-media-with-nodexl/using-graph-layout-algorithms-in-nodexl.