RESEARCHERID
THOMSON REUTERS

[Rani, 5(7): July 2018] ISSN 2348 – 8034
DOI- 10.5281/zenodo.1305005 Impact Factor- 5.070

GLOBAL JOURNAL OF ENGINEERING SCIENCE AND RESEARCHES

SOCIAL NETWORK ANALYSIS – SURVEY ON DATA COLLECTION METHODS, DATA FORMATS AND ANALYSIS TOOLS

Soja Rani S[*1] & Tinu N S[2]

[*1&2]Assistant professor, Dept of CSE, New Horizon College of Engg., Bangalore, India

ABSTRACT

Now days, social network analysis has a great theoretical and practical significance. It is an interdisciplinary subject which spreads over various areas such as, social psychology, anthropology, economics, demography, communication studies, geography, biology, history, sociolinguistics, information science, organizational studies, political science, development studies, and computer science. The popular online social networks like Twitter, Facebook, etc hold big amounts of data about the people and their likes, dislikes, behaviors, relations etc. Graph based mining tools are required to analyze these huge social networks. A number of such analysis and visualization tools are available which have their own functionalities and characteristics. Choosing an appropriate tool for a particular task is difficult to decide. This paper puts an effort to do a survey on the different data collection mechanisms, relational data measurements, network data representations together with the different social network dataset formats and social network analysis and visualization tools.

Keywords: *Social Network Analysis, Web Mining, Social Network Data, Network Analysis tools, SNA Data Formats, Online Social Networks.*

I. INTRODUCTION

Nowadays, the wide use of Internet around the world allows connecting a lot of people. The explosion of Web 2.0 (blogs, wikis, content sharing sites, social networks, etc.) opens up new perspectives for sharing and managing information. This development of networks made social network mining and social network analysis a necessity in order to provide understanding of the network and to detect communities in the network. This field has applications in business marketing, e-commerce, recommendation systems, etc. A social structure formed with actors, which can be individuals, organizations, or any entities can be called as a social network. It depicts how different social relations link or connect these actors. The type relations considered can vary from being a casual acquaintance to close kinship bonds [1]. Email traffic, epidemic transmission, and terrorist activity, etc can all be modeled as social networks. Any complex systems will be having an intrinsic network wiring mechanism. Social network analysis is the mapping and measuring of relationships between these actors. The nodes otherwise called actors in the social network represent the people or the entities considered, while the links represents the relationships between these entities. Social network analysis which is the application of graph theory provides both a visual analysis as well as a mathematical model of the relationships between individuals. One striking thing about social structures is their substructure in terms of groupings or cliques which can provide lot of information regarding the relationships and connections that exist in the network. The number of actors, size of the network, and connections among the dyads, triads and subgroups in the social network provides us a lot of information about the likely behavior of the network. How fast will information or an epidemic move across the network? Up to what level is the subgroups and social structures overlapping with one another? [1] These different facts about sub-group structure become very relevant in predicting the behavior of the network as a whole.

Social network data consist of various elements. Following the words of Wasserman and Faust [2] can be defined in six words: a structure of ties between nodes. Other than the relations or ties, additional information can also be included in the network in the form of actor attribute variables or multiple relations can exist between a pair of node. As in any other analysis data collection is the most crucial in developing a social network and social network data

can be collected in various ways. Commonly used approach is the use of questionnaires, but other methods like interviews, observations, and secondary sources are also frequently used in data collection methods for a social network. [3].

II. SOCIAL NETWORK DATA

A. Population, Samples, Boundaries and Snow Sampling

Traditional data involves information on actors and their attributes where as network data involves information on actors and the relationships among them. This difference in the type of data makes the methodology used for research design, sampling technique, measurement, and handling of the resulting data also different. Social network data involves two types of variables –

- **Structural variables** correspond to relationship measured on a pair of actors/nodes or dyads. Social network dataset primarily focus on these data. They measure the particular kind of relationship that exists between ant two actors in network. For example, structural variables can measure any type of relations that shows in the social network which can vary from business transactions between corporations to friendship/kinship between people, or even trade between nations. The actors involved in the relationship measured are usually the same type of actors.
- **Composition variables** correspond to the attributes of the actors. They are otherwise called actor attribute variables which are defined at the individual actor level. For example, it can be the different attribute of an individual like gender, race, ethnicity, geographic location, etc.

A **population** is a whole group of individuals or entities having common characteristics. It includes entire entities or individuals that are considered in that study inference procedure. Size of population can vary from big to small based on the type of application over which the study is carried on. Network analysts study diverse populations ranging from text symbols or sound verbalizations at one end to the countries in the world, even though the most common among populations are individuals itself. In any case the entities of the population to be studied should fall within some boundary [4].

Social network analysts very rarely draw **samples** from the population to perform the analysis in their work. Network analysts perform census which includes each entity in the population rather than samples in the analysis. If one actor is selected, the all other nodes or actors to whom that ego has ties are also to be included for the study which makes network approaches to conduct analysis on the whole populations rather than on sample**.** Due to the rich and unrestricted types and number of human relations, fixing proper boundaries for network analysis is a difficult task. The two approached used are: **Realist and Nominalist**

- **Realist:** These are actor perceived boundaries. The entities in the network define the boundaries based on their perceptions and experiences. For example, a network of school committee board individuals is the population who can make some difference on the rules and conditions on which teachers are hired each year.
- **Nominalist:** Researcher fixed boundary based on the criteria selected by researcher. This approach doesn't ensure the existence of relationships between the entities in the population even though researcher can find groups with the required traits.

Snowball Sampling is sampling mechanism used to create a network that is intermediate between an ego network – formed of a ego, the focal actor, and set of alters/actors who have ties/relations to ego together with the information regarding the ties/relations between the ego and these alters/actors - and a whole network - have information regarding all actors/nodes and ties/relations in the network. For example, take a random sample of students and ask to provide information regarding their close friends and further ask those friends to provide information regarding their close friends and so on so the network will keep building.

Procedure:

1. From the given population select a small sample randomly.
2. Enquire each entity or element of the random sample about network alters.

3. Contact those alters and enquire information on those alters.
4. Contact the alters of the alters.
5. Continue….

B. Scales of measurement in relations

1. Binary (nominal) measures of relations: Nominal measurement provides information regarding the relation present (coded one), and relation being absent (coded zero).
2. Multiple-category nominal measures of relations: This measure is used when there are many choices for the relation information and not only the binary nominal which tells you the existence and nonexistence of relation. The relations are coded by its type, and not by its strength. The multiple category nominal measure is multiple choices where as binary nominal which proves true-false data.
3. Grouped ordinal measures of relations: This type of measurement is used to provide additional information regarding the rank-ordering which also gives the intensity of relationships For example in a study the participants may be asked to rate others in the network as "liked" "disliked" or "neutral." This information can be encoded using a three-point scale which can be coded as -1, 0, and +1 to reflect “liked”, “neutral” and “disliked”.
4. Interval measures of relations: This measurement scale is used when we want to encode still more quantitative information in the social networks. Unlike the ordinal measures, in interval scale, the distance between numbers or units on the scale is equal over all levels of the scale which means the difference between a "1" and a "2" is same as the difference between "3" and "4”.

III. SOCIAL NETWORK DATA-COLLECTION

There are different proven methods in which social network data can be collected. The most commonly used techniques include questionnaires, interviews, observations, archival records, experiments, etc.

1. Questionnaires

This is the commonly used method especially when actors are people and the relations are the ones that can directly reported the actors. The questionnaire contains a set of questions about the participant’s relations to the other actors in the network. Questionnaires are also used when the nodes in the study are a collective entity, like an organization, corporation or an association, but an individual person forms the spokesperson to represent the collective entity. Questionnaire can be of three different formats - Roster vs. free recall, Free vs. fixed choice or Ratings Vs. complete rankings.

Roster vs. free recall -The questionnaire is of roster form when the participant is presented with a complete list of actors in the network called the roster. Preparing the roster is a challenging work and can be done only when the researcher has information regarding the members in the set of actors.

Free vs. fixed choice- In fixed choice format the participants are asked are asked to make a fixed number of choices like, to name a specific number of "best friends". When the actors are not given constraints on the number of nominations to make, the questionnaire format becomes free choice.

Ratings Vs complete rankings-In ratings format, each participant is required to assign a value or rating to each tie. In Complete rankings format, each participant is required to rank their ties to all other actors.

2. Interview

This is used in case studies where questionnaires are not the feasible method of data collection. It involves methods where data is gathered either through face-to-face or over the telephone interviews.

3. Observation

Data collection method widely used in studies of smaller groups where data is collected by observing interactions among actors during face to face interactions.

4. Archival Records

Method by which data is obtained by examining measurements taken from records of interactions.

5. Data Collection from Online Social Networks

Data collection from online social networks include the extraction of public and private data of users, groups and pages, which contain posts, tweets, likes, comments, photos, videos etc. A number of tools and APIs are available for extracting data from various online social networks like Facebook, Twitter, YouTube and few more. Facebook graph API, Twitter API, Netvizz[6] and NodeXL[7] are such tools for extracting social network data and further analyzing most of the criminal and terrorist activities using online social networks.

IV. REPRESENTATION OF SOCIAL NETWROK

1. Networks as Graphs – Sociograms: This is a graphical representation where nodes represent actors and lines or edges represent relations or ties. Sociologists borrowed this concept from graph theory and renamed the graph as sociograms [5].

2. Networks as Matrices – Sociomatrix: Most common and simplest social network representation is the adjacency matrix. It is a square array of measurements with the rows and columns same as the number of actors in the data set. Matrix is binary when the network data is nominal and otherwise if the data is measured in ordinal and interval level.1 represents the presence of a relationship and 0 the absence of the relation in the social network.

3. Networks as Edgelist: Edgelist is most efficient form of social network representation considering data storage. Here every row contains two values where first value is the source node and the second value is the target node and presence of this row in the edge list indicates the existence of a relational tie from source node to target node. The method only gives information regarding existing ties so the total number of actors in the network should also be appended to complete the network representation

4. Network as Adjacency List: It is a set of rows of text, just like an edge list. In a twist, the first node listed in each row is the node from which ties emanate, and all the other nodes listed in that row are tied to that first node. This format is less common than the other three.

V. SNA DATASET FILE FORMATS

In order to apply the knowledge of social network analysis, you need real world data sets to work on. Fortunately, there are different organizations and individuals that have gathered network data sets for different situations and topics which are publicly available to access and download. Social networks can with directed edged or undirected edges. Undirected edges are used to represent symmetric relationships like friendship where as directed edges are used to represent asymmetric relationships like an email network. Examples of social networks include friendship network – where nodes are people and a link will be present between two nodes if those two people are friends with each other, road network – where nodes are cities and a link will be present between two nodes if there is a road between these two cities, email network – where nodes will be people and a directed edge from one node to another node exists if the first person has send an email to the second person, citation network – where nodes represent papers and a directed edge from node A to node B exists if paper A cites or uses some information from paper B, collaboration/co-authorship network – where nodes will be people or scientists and the directed edge from one node to other represents who all collaborated for a work, etc. Data sets are available in different formats. The most commonly used formats are csv, gml, Pajek net, GraphML, GEXF, etc.

1. csv

comma separated values: These files will be available in the extension .txt or .csv. The file can be edge list format or as adjacency list format. Edge list provide you the list of edges in the network.

2. gml

Graph modeling language: Due to its simplicity and flexibility in labeling and assigning attributes to nodes and edges, this is one of the most commonly used network file formats. The file starts with a graph keyword. Node keyword describes a node and edge keyword represents the edge. Using node keyword we can mention the identifier of the node and with edge keyword we can mention the source and target node identifiers. A network with n nodes and m edges will have n node keywords and m edge keywords in the gml file. GML supports [6]:– directed and undirected graphs – node and edge labels – graphical placement of nodes (coordinates) – other annotations

3. Pajek net

This file will be available in the extension .net or .paj. The file starts with *vertices keyword together with the number of vertices in the graph followed by each nodes and the labels used for them. The *arcs keyword indicated the edges in the graph with each row mentioning the source node and the target node. This format is not often handled in the other implementations except the Pajek program, which allows edge representation with a matrix or an edge list or arc list (for directed graphs)

4. GraphML

ML here stands for XML. It uses a hierarchical structure and makes use of various HTML tags. The file starts with an xml tag followed by the graphml tag and inside graphml tag there will be the graph tag followed by a number of node tags and edge tags. It supports – directed, undirected, and mixed graphs, hyper graphs, hierarchical graphs, graphical representations, and application specific attribute data [8].

5. GEXF

Graph Exchange XML Format: This format was put forward by gephi which is open source tool for visualizing and analyzing networks. This is also inspired from XML uses various XML tags. The format is similar to GraphML with added flexibility is assigning attributes to the nodes and edges. It supports dynamic graphs, application specific attribute data, through the use of users XML namespaces, – hierarchical structure (nodes can contain nodes) – visualization and positioning information such as 3D coordinates, colors, shapes [8].

The most useful network dataset repositories over internet include SNAP provided by Stanford University, UCI network data repository maintained by University of California, KONECT repository maintained by Koblenz University and many more.

VI. SOCIAL NETWORK ANALYSIS AND VISUALIZATION TOOLS

Various tools and libraries are commonly used by researchers for social network analysis and visualization of social. We are considering few commonly used tools for comparison with respect to their functionality, platform, license type and file-formats respectively, based on the features like network visualization, computation of node level and network-level measures and handling of large networks with many nodes.

1. UCINET: [8] this is a software package for the analysis of social network data. It was developed by Lin Freeman, Martin Everett and Steve Borgatti. Software allows for the computational aspects of analysis, various centrality and power measures, many clustering and community algorithms, as well as hypothesis testing. It comes with the NetDraw network visualization tool. NetDraw allows for graphic representation of networks including relations and attributes. It allows for various data Import and export formats and also handles large datasets. The software was developed for Windows platform. Support various file formats including dl, .net, .vna, .csv and raw matrices. The software limitations include platform dependent, not scalable and lack of dynamic network analysis.

2. Gephi: [9] is an open source, interactive visualization and exploration platform for all kinds of networks, dynamic and hierarchical graphs. This most commonly used software supports social network analysis and visualization, provides various layouts for visualization, various measures of centrality, clustering and modularity, ranking and partitioning of network, timeline feature and plug-in support available. It runs on Windows, Linux and Mac OS X. Gephi is a tool for people who have to explore and understand graphs. It supports various file formats like .dot,

.gml,, .gdf, .graphml, .net, .dot, .dl, .gexf and .csv. Software supports both one mode and two mode network types. Gephi is not that suitable for very large data sets and multi-relational networks.

3. ORA: [10] ORA are a network analysis tool that detects risks or vulnerabilities of an organization's design structure. This software is developed for Windows platform. ORA functionalities include analysis and visualization of two mode and multi-relational networks, various visualization layouts, dynamic network analysis, measures of centrality, power and clustering, report generation of analysis. It supports .xml, .dynetml and .zip file formats. It is not suitable for large data sets and is also platform dependent.

4. NodeXL: [11] is an open source social network analysis and visualization tool for MS-Excel. This software has functionalities to import data directly from few online social networks, export data in various formats, visualization layouts, and different measures for analysis. It was developed to work with Windows platform. It supports .dl, .net, .graphml, and matrix formats. Software is Incapable of handling and visualizing large datasets.

5. JUNG: [12] Java Universal Network/Graph Framework–is a open source software library that provides a common and extendible language for the modeling, analysis, and visualization of data that can be represented as a graph or network. It is written in Java, which allows JUNG-based applications to make use of the extensive built-in capabilities of the Java API, as well as those of other existing third-party Java libraries. It includes various algorithms for centrality, clustering, flow etc., and is capable of handling very large datasets. Software is platform independent. It supports .net and .graphml. Software lacks an interactive interface and is not appropriate for dynamic network analysis.

6. Pajek: [13] A widely used Software for drawing networks, Pajek also has analytical capabilities, and can be used to compute most centrality measures, identify structural holes, block model, and so on. Functionalities include social network analysis and visualization for large networks, two-mode, multi-relational and temporal network analysis, algorithms for centrality and graph layouts. Software is adaptable for both Windows and Linux platforms. It supports .net and .dl formats.

7. NetworkX: [14] is an open source Python language software package for the creation, analysis, manipulation, structural and functional study social networks. Loading and storing of most of the standard data formats, network structure analysis, network model building, visualizing the network, etc are possible with this tool. Networkx has many features like language data structures for graphs, dIGraphs, and multIGraphs. Nodes can be "anything" (e.g. text, images), Edges can hold arbitrary data (e.g. weights, time-series), Standard graph algorithms, Network structure and analysis measures etc. It can be used both in Windows and Linux platforms. It supports .gml, .graphml and .net formats.

8. Igraph: [15] is a free software package for creating and manipulating graphs. The efficient implementation of IGraph allows it to handle graphs with millions of nodes and edges. IGraph can be installed as libraries for C, R, Python and Ruby. This tool supports, two - mode networks, lot of measures for centrality, clustering and layouts. It can be used both in Windows and Linux platforms. It supports .gml, .graphml, and .net formats. The tool does not allow direct importing from online social networks.

VII. APPLICATIONS OF SOCIAL NETWORK ANALYSIS

Following are some applications of social network analysis: [16][17]

- Identify new scientific trends becoming commercially viable, e.g. RFID, Genome sequencing, tissue engineering
- Analyze expert network, Co-authorship networks, co-citation networks, patent networks
- Measurement of success
- Ranking of trends, of authors, of companies commercializing trend
- Analyzing page importance Page Rank (Related to recursive in-degree computation), Authorities/Hubs
- Discovering Communities: Finding near-cliques

- Analyzing Trust: Propagating Trust
- Using propagated trust to fight spam: In Email, in Web page ranking

VIII. CONCLUSION

Social Network Analysis has always been popular to give analytical inferences about the society which have diverse areas of successful applications. Development of computer technology played a major role to social networking a part of everyone‘s daily routine. And it took little time to encounter very large social networks which are continuously growing rapidly. Social network analysis methods provide some useful tools for addressing many aspects of large social networks. Tools or software is very much required for graph visualization as well as to map the data from one format to another. Libraries like Networkx or IGraph are very useful network analysis and for tasks involving large number of nodes and for different operations among them and clustering. Stand alone software are easy to use and easy to learn so for beginner Pajek and Gephi is suitable software. For complex dataset and research purpose we can use Networkx and IGraph software. For one mode or two mode network analysis we can use any of software tools but for multi-relational network graph, we have only Pajek software tools. Most of the software can compute centrality, clustering coefficient, network diameter, page rank, density. But if we want to compute some specific feature we choose different software which can do that specific work.

REFERENCES

1. *A.Hanneman and M.Riddle,“Introduction to social network methods,” online at http://www.faculty.ucr.edu/ hanneman/nettext/, 2005.*
2. *S. Wasserman and K. Faust, Social Network Analysis: Methods and Applications. Cambridge University Press, Novemver 1994. [3] R. L. Breiger, The Analysis of Social Networks. London: Sage Publications Ltd, 2004, pages 505-526 in Handbook of Data Analysis, edited by Melissa Hardy and Alan Bryman.*
3. *R. L. Breiger, The Analysis of Social Networks. London: Sage Publications Ltd, 2004, pages 505-526 in Handbook of Data Analysis, edited by Melissa Hardy and Alan Bryman.*
4. *Pankaj Choudhary, Upasna Singh, “A Survey on Social Network Analysis for Counter Terrorism” International Journal of Computer Applications, 2015*
5. *David Combe1 ∗, Christine Largeron1, El˝od Egyed-Zsigmond2 and Mathias Géry1 A comparative study of social network analysis tools*
6. *Everett, Martin G., and Borgatti, Stephen P., (2014) "Networks containing negative ties", Social Networks, Vol. 38, pp111-120.*
7. *Katz, Leo, (1953) "A new status index derived from sociometric analysis", Psychometrika Vol. 18, No. 1, pp39-43*
8. *Borgatti, Stephen P., Martin G. Everett, & Linton C. Freeman, (2002) "Ucinet for Windows: Software for social network analysis".*
9. *Bastian, M., Heymann, S., & Jacomy, M., (2009) “Gephi: an open source software for exploring and manipulating networks”. ICWSM, 8, 361-362.*
10. *Carley, K. M., & Reminga, J., (2004) “ORA: Organization risk analyzer (No. CMU-ISRI-04-106)”, Carnegie-Mellon University, Pittsburgh, Institute of Software Research Internet.*
11. *Networkx http://Networkx.lanl.gov/index.html*
12. *O’Madadhain, J., Fisher, D., White, S., & Boey, Y., (2003) “The jung (java universal network/graph) framework”, University of California, Irvine, California.*
13. *Batagelj, V., & Mrvar, A., (1998) “Pajek-program for large network analysis”, Connections, Vol. 21, No. 2, pp47-57.*
14. *Hagberg, A., Schult, D., Swart, P., Conway, D., SéguinCharbonneau, L., Ellison, C. and Torrents, J. (2004). “Networkx. High productivity software for complex networks”, link: https://networkx. lanl. gov/wiki.*
15. *Csardi, G., & Nepusz, T. (2006). “The igraph software package for complex network research. InterJournal, Complex Systems, 1695(5).*
16. *Social Networks Overview: Current Trends and Research Challenges” November 2010 Coordinated by the —NextMEDIA CSA.*
17. *Business Application of Social Network Analysis BASNA-2013 www.basna.in*