Graph Visualization Tools: A Comparative Analysis

Fariha Majeed[1], Dr. Saif-ur-Rahman[2]

[1,2]*Shaheed Zulfikar Ali Bhutto Institute of Science and Technology (SZABIST) Karachi, Pakistan*

[1]majeed.fariha@gmail.com

[2]saif.rahman@szabist.edu.pk

Abstract—Data visualization is becoming a necessity for big organizations as the social networking data is growing rapidly. It is becoming difficult to visualize data and perform complex comparisons. There have been large databases to store huge data but to study the behavior is becoming time consuming and sometimes impossible. One can analyze small sets of relations but as the relationship grows, it becomes difficult to make decisions. Data Visualization tools are used to overcome this issue; however, the algorithms used to perform the analysis requires high performance processors otherwise, the data size would degrade the performance. The research provides a comparative study on popular visualization tools that could be used in the analysis of large datasets. The comparison would comprise of statistics on their common features identified on the basis of market research and literature survey.

***Keywords*—Graph Visualization Tools, Live Journal, Information Analysis, Data Visualization**

I. INTRODUCTION

Data visualization is one the biggest feature that is empowering the companies to work on their data in order to generate profit. However, majority of the organizations are still hesitant to take advantage of modern tools in order to analyze their data. Data Visualization allows to access huge data which is not possible otherwise and the chances of making a good decision without the insights are less [1].

When the data is presented in a summarized statistics it may tend to miss important information and the meaningful data cannot be fully utilized by researchers. There are two primary goals for visualization; visualization for analytic also known as visual analytic where purpose is to explore and analyze data relationships and to find meaningful information through combination, second is visualization for communication where data is used for sharing and creating visualizations with respect to the skills and needs for predictive analysis. According to book of Ben Fry [2] there are seven stages for data visualization that are, acquire, when the data is acquired, second, parse, where the data is ordered in to categories, third, filter, where the irrelevant data is removed, fourth, mine, the application of statistics and providing mathematical context, fifth, represent, select a visual layout such as tree, graph, hierarchy etc., sixth ,refine, enhance the visuals and make it more vibrant and lastly, interact, work with different ways to manipulate the data and visualizing it with the help of features available.

This research is carried out in three parts. In the first part, the datasets will be loaded via visualization tools where data will be understood on the basis of few common features. The most common features used in this research are visualization, clusters, performance, usability, network metrics and etc. In the second part, the behavior of the datasets on the basis of the features identified in the first part will be studied. The results will be compared and analysis will be carried out for performance, execution time, visualization and etc. Finally, the results will show which tool is the best choice for the same dataset pertaining in the similar environment on the basis of the functionality.

The dilemma is to visualize the abstract form of graph which is large in size. The large graph tends to show bigger problems as the increase in number of elements degrades the performance or even reach the limits of viewing. However, it is also possible to display all the data in form of layout but the problem arises during the usability and viewing it as it is not possible to distinguish between nodes and edges. Further, it is believed that the better and detailed analysis is performed when the display is small because the visuals are of the large graph which shows the overall structure but it becomes harder to understand. Therefore, this study is focused on working with open source visualization tools for network analysis and visualization which is said to be an easier and broader way to access data. The research will be aimed to analyze graph datasets of one of the blogging website Livejournal that are readily available online. The dataset will be loaded through visualization tools for analysis on the basis of different criteria set in the same environment.

A social network is a collective structure comprised of entities called nodes which are connected through one or more common characteristics. It is believed that social network analysis is becoming mature. A graph is a set of vertices and a set of lines between pairs of vertices. A vertex is the smallest unit in a network. In SNA, it represents an actor (girl, organization, country). A line is a tie between two vertices in a network. In SNA, it can be any social relation. A loop is a special kind of line explicitly a line that connects a vertex to itself. A directed line is called an arc whereas an undirected line is an edge. A directed graph or digraph

contains one or more arcs. An undirected graph contains no arcs (all of its lines are edges). A simple undirected graph contains neither multiple edges nor loops. A simple directed graph contains no multiple arcs. [3]

II. A COMPARISION OF GRAPH VISUALIZATION ON TOOLS

The data in today's time cannot be ignored and working with it is becoming a necessity. Fortunately, nowadays obtaining the data is easy as there are various mechanisms for gathering and storing data and for their storage; databases are built accordingly to cater the need. While large organizations having huge amount of data opt for large databases to keep information in who, what, when, where, how and etc. format. Organizations store data for different reasons for example companies would collect data in order to understand customer behavior and counter the chances of customers turning away and a biological researcher would collect data in order to understand how a gene is interacting [4].

Visualizing such data requires tools like Tulip [5], Gephi [6], Pajek [7] and Cytoscape [1]. These are proposed tools for such usage as with immense datasets, the investigation becomes harder with the possible number of dimensions growing that are stored in databases [4].

A. State of the Art Graph Visualization Tools

As explained in the previous section, the growing need for analyzing social networks due to the increment in the new technologies and services, various SNA tools have been developed. Where these tools come in very handy during just not to analyze the network theoretically but also represent it graphically. The tools add different indicators that help in exploring the features of the network structure, the relationship and the position of the actor along with a comparison of various social networks [8].

1) Representation: When a graph is fully visualized, it is harder for humans to understand and interpret it due to the full view. Hence there are different algorithm and metrics to make the representation visible and easy to interpret the patterns, helping in making the best decisions and understand behaviors.

2) Visualization: Visualization of graph is one of the most important and wanted functionality among all [8]. Graph visualization aims at providing visual representation with different aspects and approaches.

Some of the state of the art graph visualization tools are defined in table 1 where the main objective of each tool is described. The functionality is divided in to three aspects: 1) It indicates the tool containing visualization option 2) Analysis and 3) The type of statistics it can provide.

Table 1. Overview for Graph Visualization

	Version	Objective	Functionality			Support		
			Visualization	*Analysis*	*Statistics*	*Manual*	*Help*	*Availability*
Gephi	0.8.2	Network Analysis and Management	Yes	Yes	Yes	No	Yes	Free
Cytoscape	3.1.0	Biological Research	Yes	Yes	Yes	Yes	Yes	Free
Tulip	4.6.0	analysis and visualization of relation data.	Yes	Yes	Yes	Yes	Yes	Free
Pajek	3.15	large data visualization	Yes	Yes	Yes	No	No	Free

The level of support is the last characteristic mentioned, check the software availability (open source or paid/commercial) and availability of manual and online help. Each tool is briefly defined as follow;

a) Gephi: In Gephi, the design of the node can be changed according to need and instead of having the regular form. Also, the layout algorithm can give real time graphics. For example the speed, gravity, repulsion, auto stabilization, inertia or size settings are adjusted in real time algorithm. Atlas Force is a force directed algorithm especially developed by the team. Various algorithms can be run at the same time without interfering with the rest of the work being run on the same interface. The software can also display labels that are associated with the nodes [6].

b) Pajek: The Pajek software emphasize on the analysis of large graphs, giving various powerful tools with k-core computation, eccentricity and others. In the previous versions, Tulip shared many similar ideas with this software. However, few visualization techniques outside graph are supported [9].

c) Cytoscape: Cytoscape is mainly used for visualization of networks in Biology. In many ways, it shares many ideas with the Tulip Framework. However, it is primarily focused on biological networks and can have scalability problems. For instance, loading and displaying a grid graph having 10000 nodes and 20000 edges requires 1.5 GB in Cytoscape where it only requires 98Mb with Tulip [9].

d) Tulip: Tulip is an information visualization framework dedicated to the analysis and visualization of relational data. The current framework enables the development of algorithms, visual encoding, interaction techniques, data models, and domain-specific visualizations. The software model facilitates the reuse of components and allows the developers to focus on programming their application. [9] The purpose of this article is to compare the mentioned tools functionality and do a comparative analysis for graph network visualization which is focused on the following features;

1) Representations
2) Visualization
3) Characterization via indicators
4) Clustering or Community Detection

B. Functionality of Graph Visualization tools

This section focuses on the different functionality provided by the visualization tools. These functionalities are firstly the representation of the network (directed or undirected), secondly the visualization of the network, thirdly the statistics based on nodes and edges, and finally the clustering or community detection.The tools provided with wide range of layouts to work around. The Figure 1 and 2 shows an example of a force graph layout algorithm and Force Atlas layout algorithm applied on a sample dataset from Live Journal. Among these algorithms, one can mention Fruchterman and Reingold [10], which are a well-used force-based algorithm for graph visualization and an example of it is shown in the Figure 2. The algorithm stimulates the graph as a system of mass particles where the nodes are considered as the mass particle and the edges are the spring between these particles. It is used widely but the performance stands still slow. Comparatively, the alternative algorithm is Hu [11] which is faster than Fruchterman [10]. It is a very fast algorithm as it combines force-directed model with a multilevel algorithm to minimize the complexity.

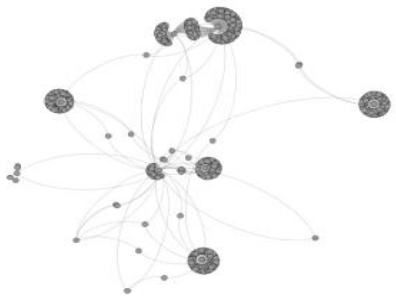

Fig. (1). Visualization of Live Journal using Gephi's ForceAtlas Layout

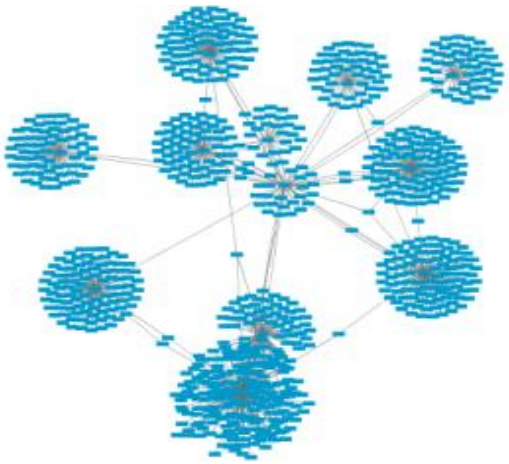

Fig. (2). Visualization of Live Journal using Cytoscape's ForceGraph Layout.

Both the layout algorithms are extensively used due to satisfying result which shows the graph structure clearly. They act as simulated physical systems which map the route distance between nodes in the network to Euclidean distance so to create natural representations. Both of the figures do not show any labels so far. Other layouts are different as they provide a view of the neighborhood of the node for example; circular layout, radial layout, label adjust layout and so on.

3) Characterization via indicators: Numerous quantifiable indicators have been defined on networks and the level of description of network level is compared through the proportion of nodes against the edges or through the evaluation of the graph properties for instance the randomness or small world distribution.

The network measures can be defined in two forms; metrics for networks and metrics for actors. The metrics for a network can be defined as Density. It represents the number of ties or connection which is expressed in number of ordered/unordered pairs. This measure shows that the graph is dense or sparse. Average Degree is the average number of links or connection per user or person. Diameter is the length of the longest and shortest path in the network/Number of Components. Average Distance is the average distance between all pair of nodes. The metrics for an actor are Degree/IN degree, Closeness Centrality and Between-ness Centrality.

4) Clustering or Community Detection: As discussed in section 2, it is beneficial to minimize the complexity of the visuals being used as it shows clarity and performance layout (rendering). Different techniques have been applied by the researchers to lower the difficulty in visualizing the graph and one of the techniques used is clustering.

The aim of clustering is to detect groups of nodes with dense connections within the groups and sparser connections between the groups. These groups are called clusters by statisticians and data mining professionals while sociologists prefer to use the word communities [12].

III. GEPHI VS PAJEK VS TULIP VS CYTOSCAPE

The objective of network visualization and analysis is to understand and distinguish patterns of social ties among the actors. The examination of social network analysis includes four parts; network definitions, manipulation, determine the structural components and visual review. The experiments performed through the tools are, Gephi 0.8.2 beta version, Tulip 4.6.0, Pajek 3.15, Cytoscape 3.1.0 and the dataset is in CSV file format of LiveJournal, a blogging community.

A. Network Visualization with Gephi

Analyzing the network seems to be the catch of every researcher eyes for analyzing data from a new perspective as Gephi is an easy and very powerful tool.

A network consists mainly of two components; actors list and relations list (connections between the actors). According to the norm, actors will be called vertices/nodes and relations will be called tiles/edges. The networks can be explored using visualization modules for various collaborative ways. It also supports import and export of different formats of data; Net, XML, CSV, SVG and etc. However, the interface is a bit cluttered and the redundant buttons in the tool makes it confusing. Further, the documentation for Gephi is not that through but the user guides gives an idea of how to work with it. The techniques have been developed with time to increase network clarity. Gephi further supports dynamic networks where real time change in data can be observed through third party databases or web services.

B. Network Visualization with Cytoscape

Cytoscape was initially designed for biological research but now it is used as a platform for general graphs which are considered as complex networks for analysis and visualization. It is open source software which was typically used for visualization of molecular interaction networks, biological patterns and other state data. Cytoscape like any other tool provides the basic features like integration, analysis and visualization along with new features by adding plugins. The plugins are used for profiling, layouts, file format support and connecting other data sources. The plugins can further be developed by user via open API which is based on Java. Cytoscape supports many standard network file format for export like XML, KGML, CSV, GraphML and etc. Also, networks can be exported as PDF, PNG, BMP or vector images and can be modified by other applications like Adobe Illustrator. Cytoscape also works as a web service client where it can directly connect to the external databases and import the network data. The work carried out on the tool is saved as session file hence the work is never lost and is recoverable. Layout networks are two dimensional where various layout algorithms are available for use like circular, edge weight, force directed, tree and organic.

The graph can be filtered by selecting nodes that are involved in the interactions and another network can be created from the results. The network can be customized using VisualStyles by using expression data that is mapped to color, label, borders etc. depending on user's desire. Clusters can be identified within the network where the regions are highly interconnected. Other than the features mentioned above, Cytoscape also provide the liberty of choosing a different language other than English.

C. Network Visualization with Tulip

Tulip is a social network analysis tool which is dedicated to the analysis and visualization of relation data. It is written in C++ where the framework allows the developer to customize algorithms, visual encoding, interaction techniques, data models, and domain-specific visualizations. Tulip provides the developer with complete library, support for designing interactive visualization applications that is used to resolve specific user problems which help the user to focus on programming the application. The format it supports is; tulip format (tlp), GraphViz (dot), GML, CSV and matrix. It can work on Windows Vista/XP/7/Linux/MAC OS [5].

Manipulation and Data Entry of Graph: Tulip has the ability to store and visualize large complete graph networks which makes it one of the few which allows the efficient way to navigate graph hierarchies or nested trees.

Graph Elements Storage: Tulip can visualize attributes of the graph elements like layouts, color, labels, size and etc. further, the features are available through the development of the algorithms.

Algorithms Application: Likewise, Tulip has been developed to easily work with graph algorithms like layout, clustering, indicators and etc. implemented in C++ plugins called Python. And also provide the widely used layouts in other popular tools.

Customized Tulip Plugins in Python: The developers can create new plugins with their choice of programming language.

D. Network Visualization with Pajek

Pajek is specially designed for network analysis and visualization of large data set size. It focuses primarily on three things; to aid in reducing large networks in to smaller networks which could be used for more refined methods; to give the user a powerful tool for visualization, and to work with proficient algorithms. The software is free to download and update continuously by developers. There is plenty of help and guidance available online. This is integral and is used by users that need attention in the analysis of networks [7].

1) Visualization techniques: The features available in Pajek for graph visualization are advanced. One of the features of draw window provides the user with many options for editing and manipulating graphics (design, size, color, rotation, etc.).

2) Data entry and manipulation: Pajek provides possibilities for managing the data structures where the networks can be transformed, directed graphs can be changed in to undirected graphs and vice versa, lines can be added or removed or the network can be reduced by decreasing the classes or by removing the parts. The program also contains basic network operations such as recording or dichotomization. There is no option to specify the relationships disappeared while it is possible to specify missing values for attributes (partitions and vectors). There are other changes that can be made in attributes and options

to create other data objects based on the attributes (hierarchies, clusters) [7].

3) Statistical Modeling: The statistical procedures in Pajek are few and basic. The properties expressed in attributes are available as partitions and vectors and is considered during statistical analysis like computations, cross tabulation and linear regression [7].

4) Descriptive Methods: Computing closeness centrality with Pajek is straightforward. The network has to be dichotomized before calculating the closeness. For directed graphs, the in- or out-closeness can be calculated as well as the closeness for the symmetrical network.

IV. EXPERIMENT

The findings of the experiment were based on table 1 which shows a comparative analysis of the four tools. This section provides the results and analysis conducted on the dataset used and the experiment for each tool was carried out in a similar environment. The specification of the machine are Intel Core i7 2.60 GHz processor, 8 GB RAM, Windows 7 64 bit Operating System and 500 GB hard drive.

A. Performance

During the experiment when the dataset was loaded in tools like Cytoscape, Tulip and Pajek on a system which has considerably low configurations, it has been observe that the tools crashed sometimes, mostly because of the load on the tools that causes memory leak that is exceeding memory limit or CPU utilization exceeded its limit. Whereas in comparison, Gephi proved to be working flawlessly on averagely configured computers and was able to perform analysis of large datasets. However, when it comes to performance of tool in terms of loading dataset of different sizes, Tulip, Gephi and Cytoscape loading time was increased with the increase in the size of dataset whereas Pajek proved to give a low consistent time for loading the dataset regardless of the size due to its non-graphical features unlike other tools. The other three had heavy graphics which were one of the main reasons for degradation of the tool performance. Another experiment proved that regardless of the heavy graphics, Gephi could visualize data in short time comparatively when its preview section was not used. The Figure 3 shows a significant comparison of how much time each tool took with dataset size 500 KB, 1 MB and 10 MB.

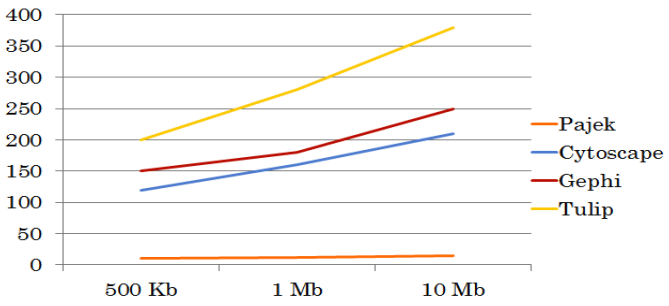

Fig. (3). Performance of Tool When Dataset Size was increased

B. Network Statistics

The quantifiable numbers shows a statistics according to metrics for networks and actors. Figure 4 shows the overall statistics loading time taken by each tool where tulip took the maximum time and Gephi took the minimum time. It was also analyzed that the level of stats was better in Cytoscape and each stats was easy to understand, whereas Pajek contained the less metrics on which the stats could be carried out.

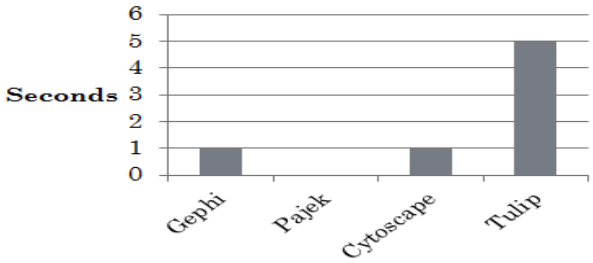

Fig. (4). Statistics Loading Time

C. Data Filter

Data can be filtered using different criteria that are available in the tools. The filter tool was easily found in Gephi where the graph could be filtered on the basis of degree rank, in degree, out degree, ego network and so on. Figure 5 shows the time taken to complete data filter for degree range.

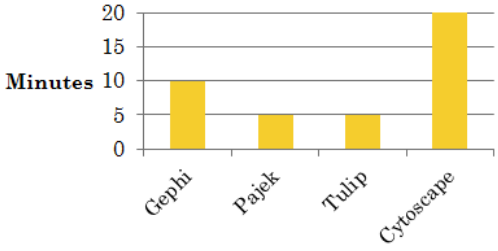

Fig. (5). Loading Time for DataFilter (Degree)

D. Layout Algorithm

The dataset could be visualized in different aspects using the layout features. Each tool had some common layouts whereas some of them were unique and was missing in the other tool. For instance, Fore Layout was available in Gephi which is a unique feature and not found in other tool. On the other hand, circular layout algorithm was found in all the tools with different type of representation. The Figure 6 shows a comparison of the tools in which each took amount of time to load the graph in circular layout. Pajek having the lowest type of graphic representation was the fastest one to give the layout whereas the other 3 took a significant amount of time. It was also noted that each time the dataset size was increased the algorithm took more time to show the circular layout and this was applied on all the layouts available on the tool. Lastly, it was analyzed that there were some layout that were not applicable due to the type of dataset being used; for instance, the dataset that was used in this research did not have a hierarchical layout hence, none of the tool was able to represent it.

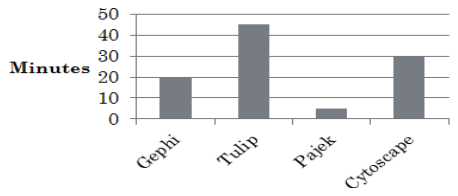

Fig. (6). Loading Time for Circular Layout

E. Loading Time of Dataset in Tools

The dataset was loaded in the tools for 3 times each to get the average time. Figure 7 shows the results where y axis represent time in seconds with 20 seconds interval and x axis represent the tools. The results show that Tulip took the maximum time to load the dataset whereas Pajek proved to be the fastest of all. One of the reasons of Pajek being fastest in loading is because it does not have a representation or overview of the graph. To visualize the graph, one needs to draw it through a separate option.

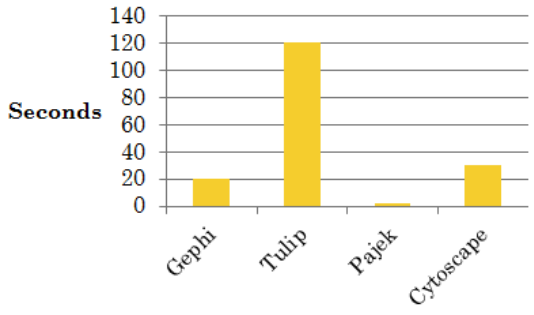

Fig. (7). Loading Time of Dataset

V. CONCLUSION AND FUTURE WORK

Network visualization has calculated different interesting results for example longest and shortest path, relationship, detecting groups, layout and representation. The study has provided a comparative analysis of four graph visualization tools where the common features were identified that are required for graph visualization. The research has concluded a comparative study upon four tools that are widely used in the research methodologies where it is scored on functionality, data manipulation, visualization, descriptive and statistical methods supported in the form on manual and help and user friendliness. The scores visible in the Figure 8 are ++ strong or very good, + is sufficient and is shortcomings. A brief comparison could be observed in Figure 8 where visualization and type of dataset imported functionality was found mature in Tulip, Cytoscape and Gephi. The indicators and attribute handling that are the characteristics functionality of a graph was mature in all the tools. When it comes to a tool with good graphics and maximum functionality, Gephi was observed to be on the highest performer, regardless of the heavy graphics. The analysis of large dataset was comparatively easier and visualization was clear. The tool is user friendly concluding the best overall performance. However, when it comes to low graphics and specific functionality, Pajek proved to be the lightest of all and easiest to use. The basic functionality provided by Pajek helps the user to stay focus on its result along with its log maintaining and the history tracking was easy. On the other hand, if a user wishes to develop and customize the dataset dynamically, Tulip proved to be best among all. With Tulip's workspace, the user can build their own plugins and modify and visualize the data accordingly. LiveJournal being a blogging website generated user views and analyzing such information can be very helpful for the organization.

	Functionality				Support		User Friendliness
	Data	Visualization	Descriptive Methods	Statistics	Manual	Help	
Gephi	++	++	++	++	++	++	++
Cytoscape	++	++	++	++	–	+	+
Tulip	++	++	+	+	++	++	+
Pajek	–	– –	+	–	++	–	+

\+ Mature Functionality – Not Available or Weak

Fig. (8). Results: Analysis of Features Comparison

The tools that are compared in this study had one common issue and that was loading the dataset in a low end configured environment which was addressed in this paper by identifying the tools what is best suited for purpose. It is recommended that the tools designed in future should be user friendly and provide a work space to users where they could visualize data with ease by reducing the white space from the visualization section, using low memory utilization algorithms. Lastly, this experiment has proved to show

positive and negative side of the tools and the changes suggested could help the tools to improve users' experience.

VI. ACKNOWLEDGMENT

Firstly, I would thank to Allah for his blessings upon me and for providing me with such capabilities. Secondly, I would like to show my gratitude towards my institution SZABIST and my advisor in this research, Dr. Saif Ur Rahman, whose timeless effort and constant guidance made this possible. Lastly, I would like to thank my family and friends for their throughout support and motivation.

REFERENCES

[1] J. Mcauley and J. Leskovec, "Discovering social circles in ego networks," *ACM Transactions on Knowledge Discovery from Data*, vol. 8, no. 4, 2014.

[2] B. Fry, *Visualizing data - exploring and explaining data with the processing environment*, O'Reilly, 2007.

[3] P. Boldi, B. Codenotti, M. Santini, and S. Vigna, "Ubicrawler: A scalable fully distributed web crawler," *Software: Practice & Experience*, vol. 34, no. 8, pp: 711-726, 2004.

[4] F. Gilbert and D. Auber, "From Databases to Graph Visualization," In *Proceedings of 14th International Conference on Information Visualization*, 2010, pp: 128.

[5] A. Lambert and D. Auber, "Graph analysis and visualization with tulip- python," In Proceedings *of 5th European meeting on Python in Science (EuroSciPy)*, 2012.

[6] M. Bastian, S. Heymann, and M. Jacomy, "Gephi: An open source software for exploring and manipulating networks," In *Proceedings of International AAAI Conference on Weblogs and Social Media*, 2009.

[7] M. Huisman and M. A. J. V. Duijn, "Stocnet: software for the statistical analysis of social networks", *Connections*, vol. 25, no. 1, pp: 7–26, 2003.

[8] D. Combe, C. Largeron, E. Egyed-Zsigmond, and M. Gery, "A comparative study of social network analysis tools," In *Proceedings of International Workshop on Web Intelligence Virtual Enterprise 2*, 2010.

[9] D. Auber, D. Archambault, R. Bourqui, A. Lambert, M. Mathiaut, P. Mary, M. Delest, J. Dubois and G. Melanon, "The tulip 3 framework: A scalable software library for information visualization applications based on relational data," Research Report RR-7860, Jan. 2012.

[10] T. M. J. Fruchterman and E. M. Rheingold, "Graph drawing by force-directed placement," *Software: Practice and Experience*, vol. 21, no. 11, pp: 1129–1164, 1991.

[11] Y. F. Hu, "Efficient and high quality force-directed graph drawing," *The Mathematica Journal*, vol. 10, pp: 37-71, 2006.

[12] I. Herman, G. Melançon and M. S. Marshall, "Graph visualization and navigation in information visualization-a survey," *IEEE Transactions on Visualization and Computer Graphics*, vol. 6, no. 1, pp: 24-43, 2000.