ANALYZING DATA VISUALIZATION METHODS FOR ACADEMIC SOCIAL NETWORKS

by

Marissa A. Mocenigo

A senior thesis submitted to the faculty of

Bryn Mawr College

in partial fulfillment of the requirements for the degree of

Bachelor of Arts

Department of Computer Science

Bryn Mawr College

Spring 2011

ABSTRACT

ANALYZING DATA VISUALIZATION METHODS FOR ACADEMIC SOCIAL NETWORKS

Marissa A. Mocenigo

Department of Computer Science

Bachelor of Arts

With the growing popularity of social networking sites, social network visualization is becoming increasingly more popular; these visualizations can help improve the user's experience or can benefit marketing agencies by condensing important information to an easier format; however, the advantages of social network analysis are not limited to sites such as Twitter or Facebook. The objective of this thesis is to examine methods of effectively visualizing social networks and applying them to a smaller scale network within an academic environment with the ultimate goal of providing a tool for students. In this case, the academic network is compromised of professors and their research connections, namely their co-authors. In particular, this data can be used to help students find relevant research and publications within the research area of their advisor. Ultimately, two methods are analyzed with academic network data; the first uses pre-existing software, Graphviz, and the second creates an independent interface that uses aspects of other visualizations to create an alternate perspective. Both emphasize different aspects of the data and provide varying viewpoints.

ACKNOWLEDGMENTS

I would like to acknowledge the Bryn Mawr College Computer Science Department for their knowledge and support; they have been an indispensable resource throughout my college career. In particular, thank you to my major advisor, Professor Dianna Xu, for the endless advice, inspiration, and friendship. Additionally, thank you to Professor Deepak Kumar for his guidance through the thesis process.

Furthermore, I owe much gratitude to my fellow computer science majors for all the hours in the lab and for the support and advice. I would not have survived the stress and demands of all my work if not for the support and aid of my fellow students.

Thank you, as well, to my friends and coworkers in Bryn Mawr College Computing Services for inspiring and fostering a love for computers.

I would like to extend a final thank you to my family and friends who have been nothing but supportive and encouraging for the past four years.

Contents

Table of Contents vi

List of Figures vii

1 Introduction 1

1.1 Motivation 1

2 Related Work 2

2.1 Data Visualization 2

2.2 Social Network Analysis 3

2.2.1 Network Clusters 3

2.2.2 Social Network Visualizations 4

3 GraphViz 6

3.1 The DOT Language 6

3.3 Applications to Other Social Networks 8

3.2 Applying GraphViz to an Academic Network 8

4 Developing a Comparative Method 11

4.1 Surveying Layouts 11

4.2 Implementation 14

5 Analysis of the Visualization Methods 18

6 Future Work and Conclusions 20

6.1 Future Work 20

6.1.1 Short-Term Improvements 20

6.1.2 Long-Term Improvements 20

6.2 Conclusion 21

Bibliography 22

A Graphviz Code 23

B Processing Code 28

List of Figures

2.1 The Billion Dollar Gram created by Dave McCandless 2

2.2 An example screenshot from TwitterPoster. 4

3.1 An example of a Graphviz representation using the DOT language. 7

3.2 An example from Proximity Graphs showing the connections of actor Kevin Bacon. 8

3.3 Graphviz representation of research connections for Professor Dianna Xu, using singly-linked, one-directional edges. 8

3.4 Graphviz representation of research connections for Professor Dianna Xu, using dually-linked edges. 9

3.5 Graphviz representation of research connections within the Bryn Mawr College Department of Computer Science. 9

4.1 Layout 1, included in survey. 12

4.2 Layout 2, included in survey. 12

4.3 Layout 3, included in survey. 13

4.4 Layout 4, included in survey. 14

4.5 Startup image depicting Computer Science departments and including drop-down menu example. 15

4.6 A visualization of professors within the Bryn Mawr College department. 16

4.7 A visualization of a Professor Dianna Xu's co-authors. 17

5.1 Disparity between prominence within department and strength of research connections. 18

1 Introduction

The human mind is disinclined to process large sets of numerical data. Graphs and charts have often been used to take raw data and transpose it to a visual representation that is easier for the mind to digest. In this manner, a person can see and interpret patterns that were previously hidden in complicated data sets. [2, 3] The field of data visualization seeks to explore the ways the human mind processes information and to translate it to a clear, efficient graphical representation. Visualization must keep two major factors in mind: function and aesthetics. The overall appearance of the graph cannot detract from the conveyance of information, but the layout should be intuitive and easy to interpret. [5] This field is becoming more and more prominent as our society's data output continues to grow; for instance, DNA sequencing generates large amounts of data for which there is little use. The applications of visualization are endless.

Simultaneously, our interactions on the web create a growing social network and, similarly, a growing source of data. Currently, social networking is a huge area in online marketing. [16] The popularity of social networking sites is phenomenal, and consequently, these sites are given access to an abundance of information by its users. Users submit everything from simple demographic information, from age and location, to highly personal information, such as interests and friends. With such a breadth of information from such a prolific source, data mining becomes incredibly powerful. Based on keywords taken from user's conversations and updates, researchers can predict the geographic spread of disease or the popularity of a product. [13]

Several research groups are currently exploring new ways of evaluating interpersonal connections on the web. Using Facebook, MySpace, or Twitter data, these groups use data visualization to show connections between people and the resources they use. These visualizations appeal to a wide audience; for example, advertisers could use connections between similar friends with similar interests to expand their target market. [16] Additionally, these connections act as a fingerprint for an individual and can be used to investigate cases of fraud. [14] Regardless of a person's presumed name, many of their friends or interests will be the same under multiple identities, making them easily traceable using an effective visualization.

1.1 Motivation

Many standards exist for representing connections in a social network. Our project seeks to investigate these individual methods and apply them to small-scale social networks in an academic environment such as Bryn Mawr College. The research connections that professors form mimic this social fingerprint; however, in this case, the data can be used to find relevant publications or related coursework. Access to this information benefits both students and other professors in pursuing their academic interests. My research seeks to create an optimal visualization of this data.

2 Related Work

2.1 Data Visualization

Data or information visualization specifically refers to visualization methods which use computer tools to analyze large data sets. [5]

The literature on visualization is expansive. Perhaps one of the most recognizable authors in the field of data visualization is Edward Tufte. He has published *The Visual Display of Quantitative Information*, *Envisioning Information*, and *Beautiful Evidence*. Tufte combines elements of statistics with graphic design to provide an in-depth look at creating visualizations.

Through statistics, Tufte reveals that major landmarks in human history and development follow statistical distributions that can be represented graphically. [4] He evaluates methods for presenting 3D data in a 2D graphic and emphasizing key pieces of information in layers. Tufte approaches visualization as a science but also as an art form. [3] He is largely responsible for the popularization of aesthetically beautiful charts and graphs.

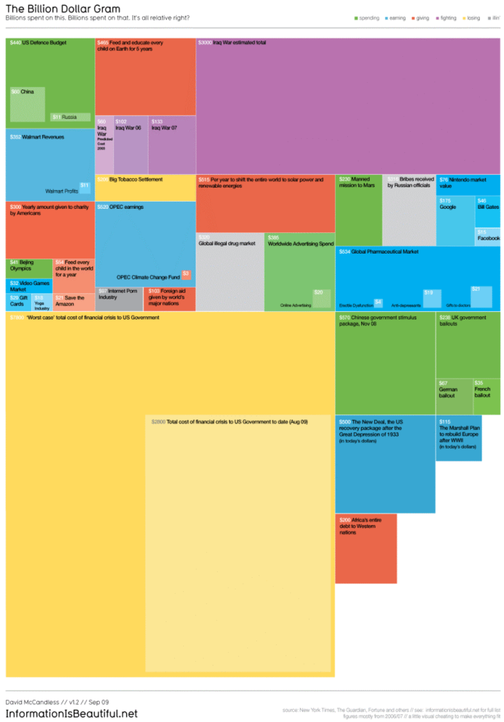


Figure 2.1: The Billion Dollar Gram created by Dave McCandless

Dave McCandless recently published *The Visual Miscellaneum*, which emphasizes the aesthetic aspect of data visualization. His research is founded on the belief that information should be beautiful. *The Visual Miscellaneum* is a collection of the various graphics he has created. One such visualization, which heavily influenced some of the later work on this project, can be seen in Figure 2.1. This visualization depicts how many billions of dollars are spent annually on various initiatives, from the US Defense budget to the yoga industry. [1]

Information Visualization: Beyond the Horizon provides a more analytical overview of many of the concepts presented initially by Tufte. Chen dedicates a chapter solely to the evaluation of networks. He gives the example of three network visualizations: SemNet, NicheWorks, and Narcissus. SemNet uses known connections to map data into three-dimensional space and highlights a need for simplicity in graphical models. NicheWorks, much like GraphViz, was developed at Bell Laboratories; its purpose is to algorithmically form a layout for a large number of nodes with the purpose of showing patterns within the data. NicheWorks alters the lengths of the edges between nodes to reflect additional properties within the graph. Narcissus uses three dimensions to display complex information. In 2D, the elements seem cluttered, but in 3D, there are multiple layers and large data sets can be compacted into a smaller frame. [5]

2.2 Social Network Analysis

Social network research provides a cornerstone for the evaluation of academic research networks. The connections that professors make and maintain reflect many similarities with more general social networks, such as those seen on Facebook.

2.2.1 Network Clusters

Many researchers are looking deeper at mathematical methods of evaluating these networks. "Community Structure in Large Networks: Natural Cluster Sizes and the Absence of Large Well-defined Clusters", an article by Jure Leskovec; Kevin J. Lang; Anirban Dasgupta; and Michael W. Mahoney, presents a different analytical perspective on representing social networks. This article examines an approximation algorithm for determining the magnitude of a connection between nodes in a community. Their results showed that large networks have a different structure than smaller social networks. [6]

On a small-scale of roughly 100 nodes, community groups are a central component, but once the view is expanded, these dissolve into the larger community. Additionally, the smaller communities can each be interpreted as part of a larger community; additionally, the network global minimum of the community profile plot often occurs within the smaller communities. Often, these communities are only connected to the remaining plot by a single edge. In larger communities, there is an inverse relation between the size and quality of a community. The prominent communities in the smaller plots begin to blend in with other communities and eventually disappear entirely. [6]

Newman's "Finding Community Structure in Networks Using the Eigenvectors of Matrices" continues a mathematical analysis of social networks. The article looks to apply the principles of visualization to a breadth of networks, from social to biological systems.
Additionally, they provide examples where insights into social networks help to interpret other systems. The communities in a social network mimic pathways in metabolic systems,

clustering of relevant articles on the web, and provide insight into networks that are hidden on a larger scale. [7]

Newman focuses on “modularity” as a method of evaluating communities in a social system. This method requires extensive computations, but results are consistently successful; however, Newman alters the modularity approach to use matrices. Traditional spectral partitioning methods do not work with social communities as the communities can be of arbitrary size. Consequently, he uses modularity instead to evaluate the division of networks. Community structure is determined by maximizing modularity over several divisions of a network. [7]

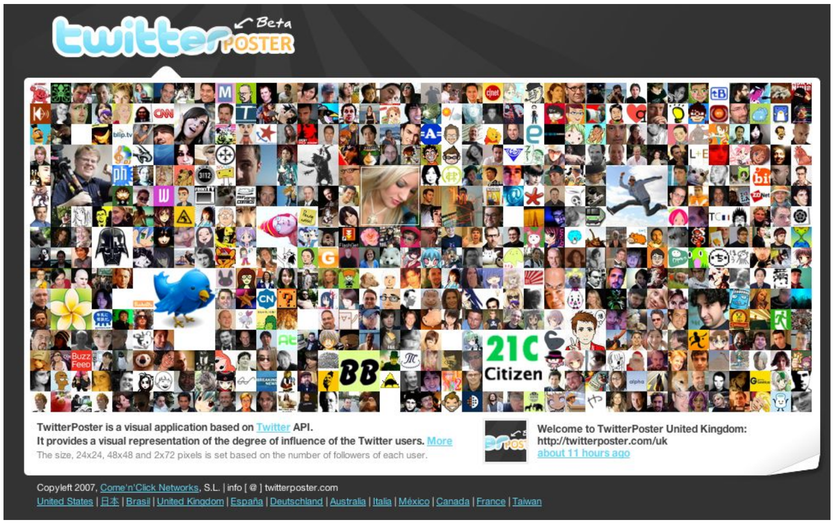
Network cluster research emphasizes the mathematical analysis of social networks; however, these clusters have an important function within social networks. Network clusters emphasize specific groups within a community. When evaluating a social network, even through a visualization, it is important to observe these distinct groupings as these clusters are particularly noticeable in a graphical representation.

2.2.2 Social Network Visualizations

Last.fm maintains a widget, called last.forward, which provides a graphing tool but tailored for use with social networking systems. Last.forward is open source and available freely online. [8] Similarly, last.fm also offers Friends Sociomap to show compatibility between the music interests of users. [9]

Based on the popularity of their online community, Digg has also implemented several visualization methods to process their network data. Their visualization software is made up of five major applications: Pics, Arc, BigSpy, Stack, and Swarm. Each application organizes data based on its popularity on the site. For instance, Swarm adds a ring to each story as its accessed; more popular stories have more rings, making them brighter and more noticeable. [9]

Twitter has inspired numerous visualization projects and offers some of the most relevant visualization research for social networks. One of the most popular Twitter visualizations is Twittervision, which plots tweets geographically. [10] Twitter visualizations analyze everything from geographic location to frequency of tweets to interconnected users. The data on Twitter is more accessible to the average user, due to fewer privacy settings, and consequently many open source and independent applications have arisen. [9]



TwitterPoster is a visual application based on Twitter API.
It provides a visual representation of the degree of influence of the Twitter users. More
The size, 24x24, 48x48 and 2x72 pixels is set based on the number of followers of each user.

Welcome to TwitterPoster United Kingdom:
http://twitterposter.com/uk
about 11 hours ago

Copyleft 2007, Come'n'Click Networks, S.L. | info [@] twitterposter.com

United States | 日本 | Brasil | United Kingdom | España | Deutschland | Australia | Italia | México | Canada | France | Taiwan

Figure 2.2: An example screenshot from TwitterPoster.

TwitterPoster has applications in other social networking realms; it uses data about individual user's to determine the most influential users. The number of their followers determines a user's influence. TwitterPoster generates a graphic to represent the most influential Twitter users at any given time; the larger a user's avatar, the more influential the user. [11] A similar application is We Feel Fine, which collects information from web blogs to determine the overall mood of Internet users. [9]

Beyond simple design applications, the visualization of social networks is gaining popularity in academic circles. Paul Torres explores the Habbo Hotel chat community as part of a social network visualization. He explains that "within the field of information visualization, the best way to represent large networks is still very much an open question." [12] Torres uses visualizations to analyze social constructs within the chat community. He wanted to show how and why different users formed cliques and groups. However, he found a major problem with data collection and processing. Social networks change very quickly, and Torres was unable to obtain any raw data from the site for several months, due to security reasons. [12]

At the University of California at Berkeley, Vizster is an ongoing research project that evaluates different methods of visualizing friendships in a social network. In this instance, the researchers think of each person as the vertex of a graph with each "friendship" as an edge connecting the nodes. They began designing a visualization geared toward site users that would educate them about their online community. In the past, many social science researchers who were evaluating online communities believed that a visualization design should focus first on the big picture and only present details on command; however, the researchers at Berkeley found that many users preferred to see relevant profile information up front. [13]

3 Graphviz

AT&T Labs researchers are currently working on a visualization software project called Graphviz. Overall, their research seeks to create graphical networks for a broad range of inputs. The software is built with Java and runs on Windows, Linux, Mac OS, and even has an iPhone app; in fact, Graphviz is used on the AT&T Labs website to show the connections between researchers in the Labs. [14]

The software itself is very versatile, and one of its primary uses is in telephone networking. In this case, the social network is built around a consumer's phone calls. This series of phone calls creates a unique fingerprint for a user. [14] If he or she attempts to use a false name to evade payment, their unique fingerprint can be used to track them and identify high-risk consumers.

3.1 The DOT Language

Graphviz takes input in the DOT language and outputs a tree-like graph. For example, Professor Dianna Xu has co-authored a paper with both Marcelo Siqueira and Luis Gustavo Nonato. Professor Xu will appear as the top node with connections to both researchers; however, the graph should also show that Siqueira and Nonato are connected. In the DOT language, this would be represented as follows:

```
strict digraph example {
        node[shape=”ellipse”]

        prof[label=”Dianna Xu”];
        coauth1[label=”Marcelo Siqueira”];
        coauth2[label=”Luis Gustavo Nonato”];

        prof->coauth1;
        prof->coauth2;
        coauth1->coauth2[arrowhead=”none”];
}
```

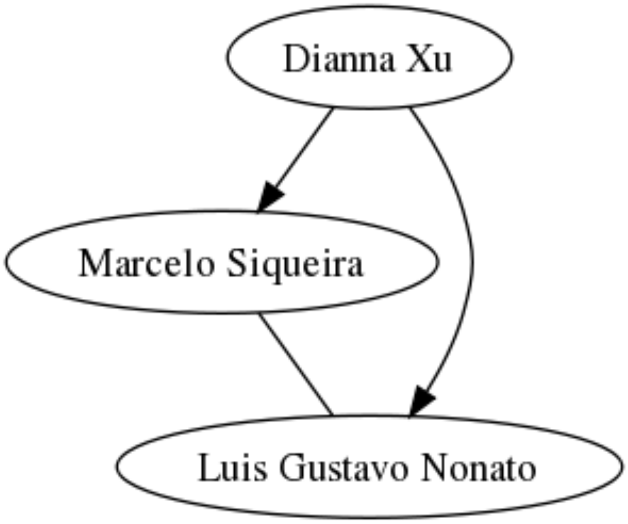


Figure 3.1: An example of a Graphviz representation using the DOT language.

This example creates a visual representation of these connections, as seen in Figure 3.1, but it shows only a couple nodes. This is a very simplistic layout, but it is also a very straightforward representation of the data. The language allows for simple customization of the appearance. The overall appearance of the nodes can be altered; in this case, the nodes have been set to ellipses. Each individual node can be changed; here, they have merely been labeled with the professor or researchers name. Additionally, the edges can be customized; two have been left with default directional arrows, and one has been sent to an unadorned line. In this case, the directed arrows are used to show the hierarchy between the professor and co-authors, and the undirected arrows show connections between co-authors only. Many of these properties were used in the customization of the academic data.

3.2 Applications to Other Social Networks

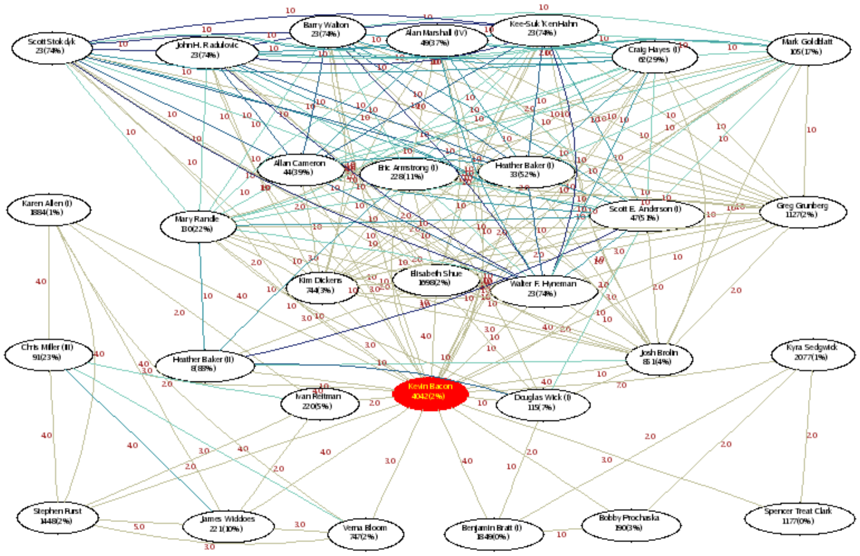


Figure 3.2: An example from Proximity Graphs showing the connections of actor Kevin Bacon.

Proximity Graphs, also at AT&T Labs, and the Hammond Jazz Inventory use Graphviz to represent unique social networks. Proximity Graphs uses information from the Internet Movie Database to show the links between user-inputted actors, as seen in Figure 3.2. A number labeling the edge is used to indicate the strength of the connection to another actor. Additionally, the Proximity Graphs have been extended to include information from the Co-authorship Network to connect authors. The shortest distance between two authors is shown with connection to the most frequently referenced authors. [15] Alternately, the Hammond Jazz Inventory shows the related recordings of listed musicians. [14]

3.3 Applying GraphViz to an Academic Network

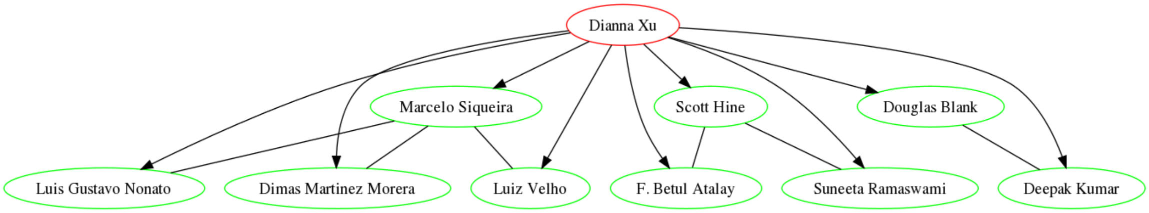


Figure 3.3: Graphviz representation of research connections for Professor Dianna Xu, using singly-linked, one-directional edges.

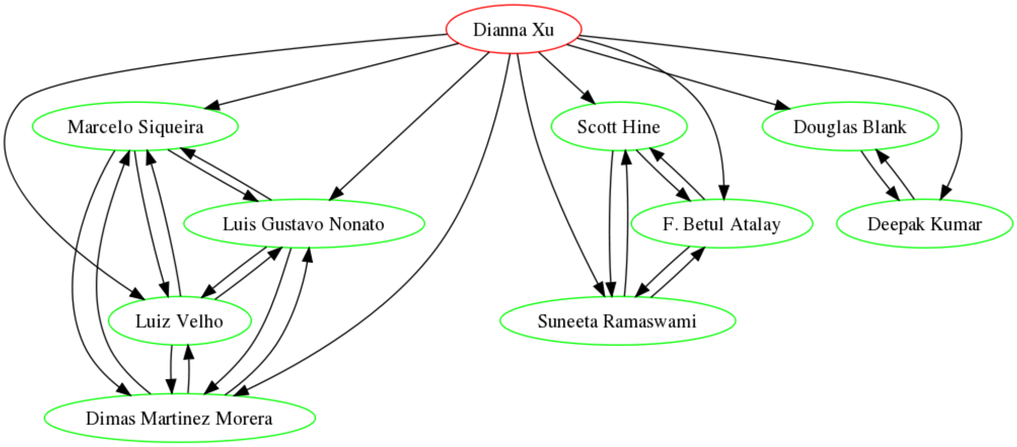


Figure 3.4: Graphviz representation of research connections for Professor Dianna Xu, using dually-linked edges.

Data was collected from professors' selected publications; the names of all unique co-authors within the most recent ten papers were used, then an edge was added between the professor and coauthor. Patterns and clusters become more obvious in larger data sets. Figure 3.2 and 3.3 show a Graphviz representation of all Professor Dianna Xu's co-authors. Figure 3.2 emulates the same arrow representations as the example in Figure 3.2. The professor's connections are shown with directed arrows as their connections are distinct; however, interchange between the co-authors is undirected. The dually-linked nodes seen in Figure 3.3 are used to show the exchange of information between multiple co-authors and replaces the undirected arrows seen in previous representations, but it also enhances the obvious grouping of certain researchers. Initially, publication titles were also included in the graphical representation; however, this greatly skewed the graphs, and the connections were occluded by the layout.

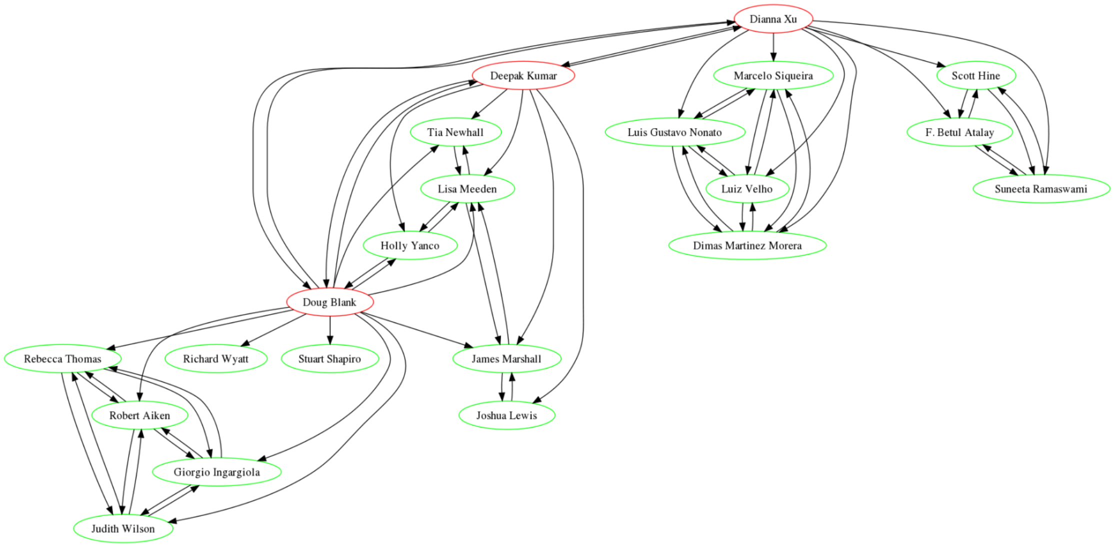


Figure 3.5: Graphviz representation of research connections within the Bryn Mawr College Department of Computer Science

The clusters observed in Figure 3.3 become even more evident when the view is expanded to the entire department, as seen in Figure 3.5. Here, it becomes evident which connections are shared between professors.

4 Developing a Comparative Method

In implementing a new visualization method for the data, the layout sought to emphasize the influence of each professor's research connections. As with the data used in GraphViz, each professor's co-authors were collected from their most recent ten publications. Largely, a co-authors influence was determined by the frequency with which the researcher was listed as a co-author. The more often a co-author appeared, the more influential that co-author would be on the professor's work. Influence could also be affected by how recently the paper was published; however, this was eliminated from this implementation as only the most recent ten publications were used.

4.1 Surveying Layouts

The aesthetic aspect of the layout was based on research from previous visualizations; however, this did not account for the usability of the visualization. Consequently, a survey was sent out to twenty-five students to gauge the usability of various layouts. Over the course of a month, only eight students responded; however, this small sample group represented a diverse range of computer literacy. Most identified as having an "average" or "fair" of computers. Two students listed that they had exclusively programming experience, one student had exclusively hardware experience, one had both, one had very little general computer experience, and the others identified as average. This was entirely self-identified; regardless, there seemed to be no correlation between a user's responses and their computer literacy.

Users were shown four layouts individually and asked three questions: "Imagine this is the visualization for one professor's fellow researchers. Is it easy to understand this visualization? What would you observe about the researchers from this graphic alone?"; "Could this visualization be enhanced with further information or text? What additional information would you like to see?"; and "Please enter a few words or phrases that you feel describe this image."

The goal of these questions was to see, first, what impressions users had about the data presented. Were they able to understand the graphic quickly and with little prior knowledge? What additional information would help them understand the visualization? How could it be improved? In most instances, the results were very similar from user to user.

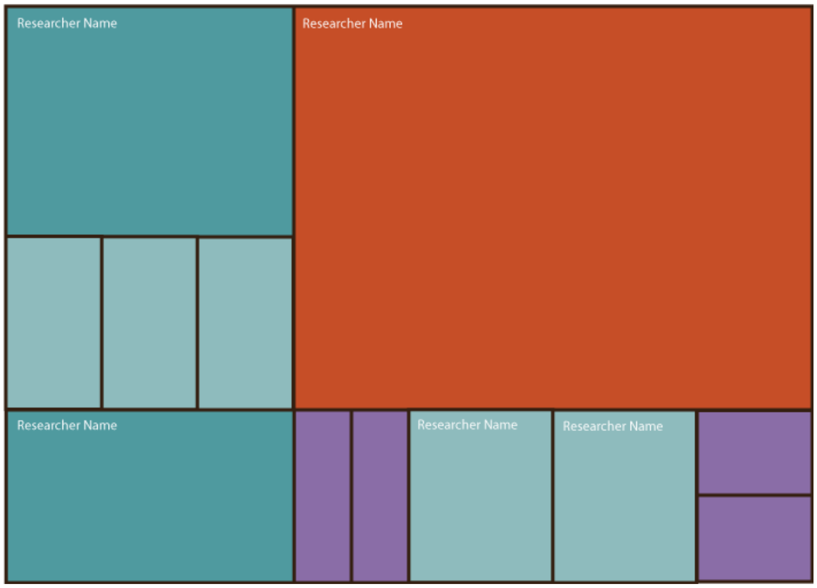


Figure 4.1: Layout 1, included in survey.

The first layout, seen in Figure 4.1, meant to create an artistic, modern representation of the data; however, this layout was very poorly received. Of the students surveyed, 62.5% voted that this layout was their least favorite. It was described as "abstract", "confusing", "vague", and "cluttered". Some users interpreted that a researcher's influence was linked to the size of the corresponding rectangle while others associated influence with color. Overall, it was stated that the graphic would be greatly improved with more textual information but that this would detract from the aesthetic aspect of the visualization.

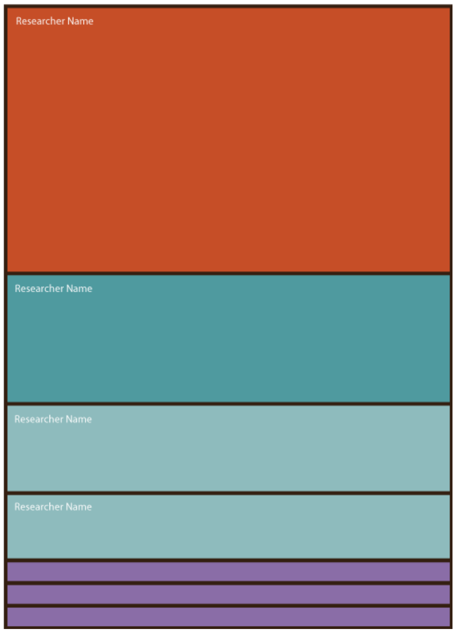


Figure 4.2: Layout 2, included in survey.

Figure 4.2 shows the second layout, which 50.0% of surveyed students voted to be their favorite layout; however, no users ranked this layout as their least favorite. This layout provided an unusual, unique representation while still maintaining elements of familiar graphs. Ultimately, this was clearly preferred by all users and was described as "ordered", "understandable", "informative", "logical", and "organized". Here, users were able to quickly pick out influential researchers, and many suggested that no more than the researcher's name would be necessary to enhance the graphic.

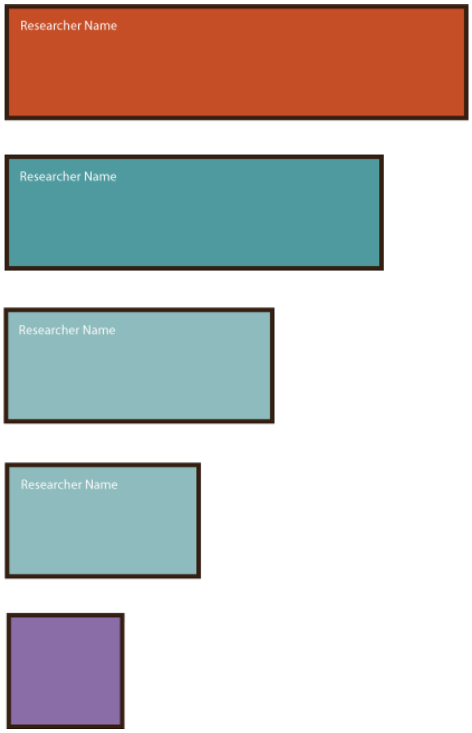


Figure 4.3: Layout 3, included in survey.

The third layout, Figure 4.3, was the most polarizing layout. Many users were able to quickly find the relevant data and found the bar graph appearance very relatable. The familiarity of the graph appealed to many users, but the simplicity and boring design deterred others. Ultimately, only 12.5% of users ranked this as their favorite, and 12.5% ranked it as their least favorite. It was described as "communicative", "ordered", "direct", "too traditional", "vague", "informative", and "less appealing".

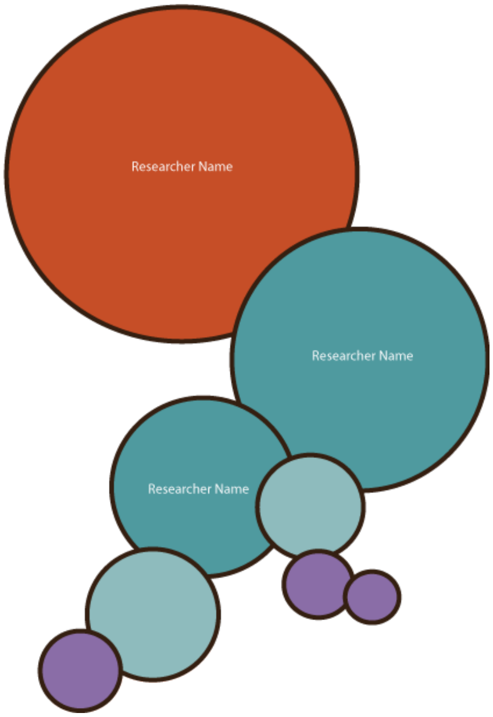


Figure 4.4: Layout 4, included in survey.

Layout 4, Figure 4.4, was a close favorite with the second layout; 37.5% of users voted this to be their favorite layout, and 25.0% voted this to be their least favorite. The overlap of the ellipses provides an opportunity to incorporate more data components into the visualization, and several users commented that they liked the originality of its appearance. When asked to describe it, users wrote "fun but not practical", "interesting", "thought-provoking", "pretty", and "unnecessarily complicated." Regardless of how much users liked the layout, few could quickly understand the various features within the graphics; they agreed that the shapes and overlay had great potential to represent additional data, beyond researcher's influence, but most felt it would simply make the visualization more confusing and obscure the purpose.

4.2 Implementation

Based on the survey feedback, the second layout, seen in Figure 4.2, was chosen for the implementation. Ultimately, it was designed in Processing, as the language is easy-to-use and allows for easy integration with Java.

Two classes were created in Java to store and organize data: a Professor class and a Researcher class. The Professor class stores the professor's first and last name, their department, and a Vector of Researcher class objects. The Professor class was designed such that it could hold information for researchers listed as co-authors on publications or referenced within the publication. The data for referenced authors versus co-authors is contained in two separate Vectors but sorted similarly.

The graphical interface is implemented entirely in Processing. As the data sets used were so small and cluttered information was frequently altered and removed, the data is currently hard-coded. With the use of the aforementioned classes, this could easily be changed later.

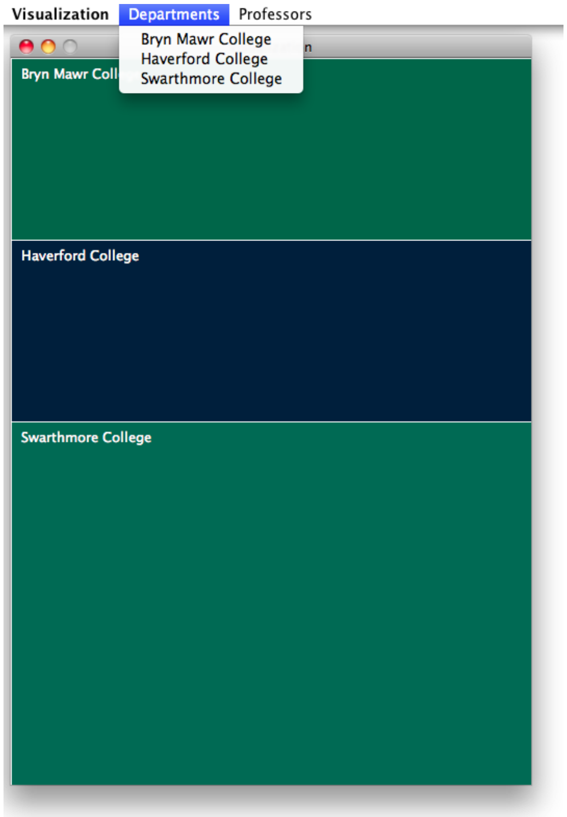


Figure 4.5: Startup image depicting Computer Science departments and including drop-down menu example.

The first image generates an image depicting the computer science departments at Bryn Mawr, Haverford, and Swarthmore Colleges. Each college is represented by a colored rectangle, which is proportional to the number of professors in each department; the larger a college's department, the larger the rectangle.

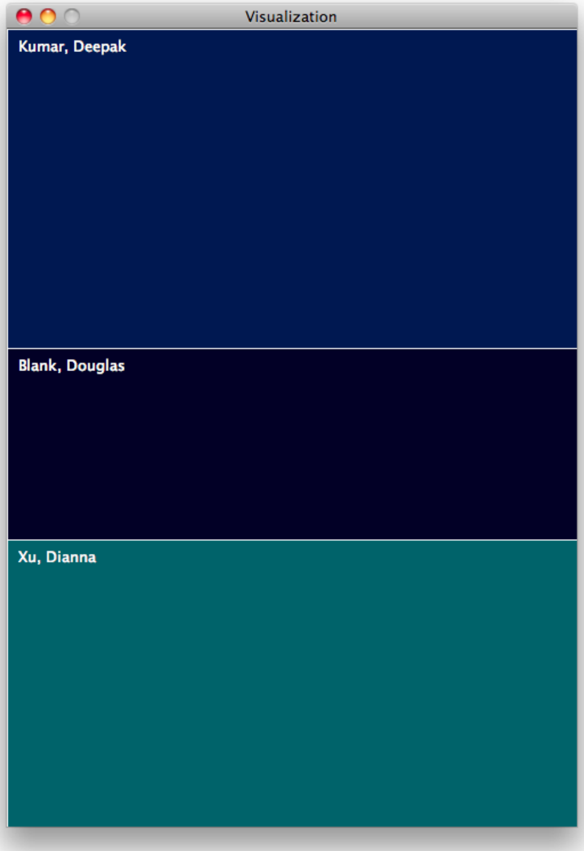


Figure 4.6: A visualization of professors within the Bryn Mawr College department.

The user can then navigate the application from the options in the menu bar. There are two drop-down menus: departments and professors. When a college's department is selected, the view will change to show the professors in the department, and the rectangles will reflect the proportion of researchers affiliated with each professor, as seen in Figure 4.6.

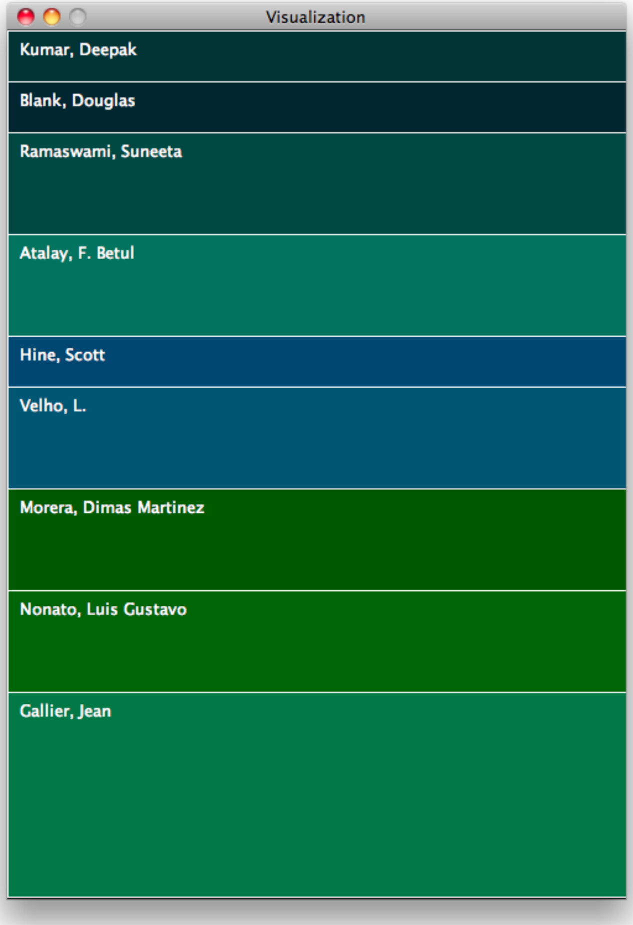


Figure 4.7: A visualization of a Professor Dianna Xu's co-authors.

The visualization for an individual professor most prominently uses the concept of influence, as seen in TwitterPoster. In TwitterPoster, influence is determined by the number of followers a user has. Here, the influence of a co-author is determined by the number of times the co-author appears in a professor's selected publications. The rectangles are drawn proportional to the co-author's influence.

The colors of the visualization could be altered to reflect influence or perhaps another variable; however, currently the colors are random. In the survey, a color palette was used from ColourLovers.com to try to create a fun, interesting vibe; however, the randomly generated green-blue range was initially used as a placeholder and kept as a personal preference.

5 Analysis of the Visualization Methods

The visualization, as created by Graphviz and using nodes and edges to represent connections prominently, shows clusters of researchers and professors that tend to work together. This visualization would be useful for a user trying to find researchers with similar publications; for example, if a user is working on research with Professor Dianna Xu and finds a publication by Marcelo Siqueira particularly useful, the user may want to peruse publications by other researchers within that cluster. Alternately, there is no guarantee that the chosen researcher's own work will be relevant; he or she could have contributed minimally to the paper.

This is counteracted in the second visualization method. The emphasis on influence guarantees that a researcher's work is primarily more similar to that of the professor. If they are publishing four or five papers together, it is likely that this is a prominent area of interest. The simple node and edge method used in Graphviz does not compensate for influence. The Proximity Graphs use a floating point number to denote the strength of an edge; however, this does not become clear in the visual representation, and the user is left to interpret a mess of numbers.

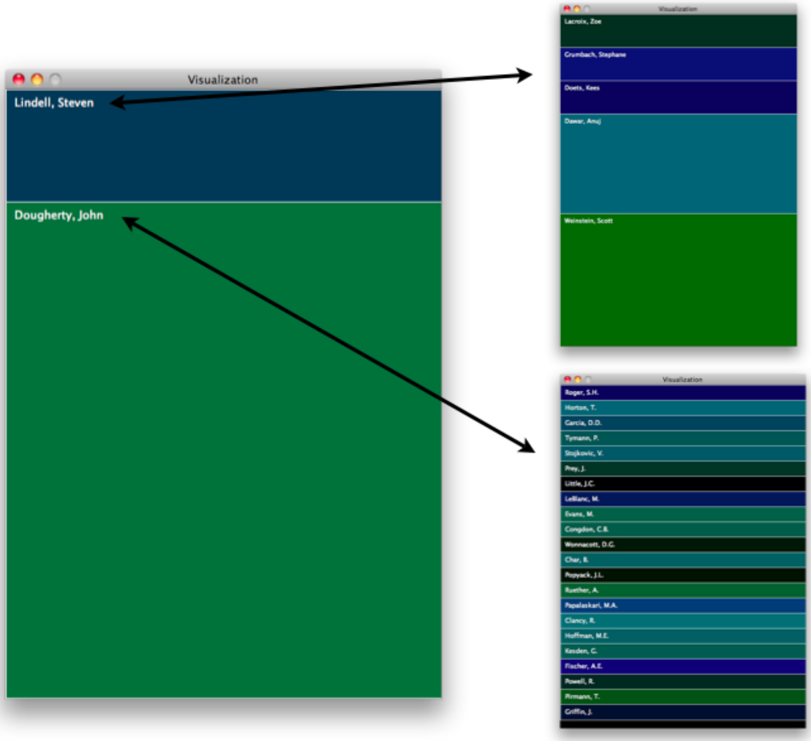


Figure 5.1: Disparity between prominence within department and strength of research connections.

The major issue with the second method of evaluation is the disparity between the information depicted in each view. The prominence of a professor within a department does not necessarily mean that they have developed a consistent working relationship with each or any researchers. As seen in Figure 5.1, Professor John Dougherty dominates the visualization for Haverford College's computer science department.

Alternately, the Processing implementation reveals interesting insight about a professor's researchers. It is easy to quickly identify researchers with whom a professor works frequently. A student could use this information to find additional publications by knowing which authors would be most relevant to continued work. Graphviz has a similar advantage in that it shows which professors share researchers and which researchers tend to work together. Unfortunately, in Graphviz, there is no way to tell which researchers are most influential, and in the Processing implementation, there is no way to determine which researchers often work together; the two interfaces are largely incompatible.

6 Future Work and Conclusions

6.1 Future Work

The current implementation has much room for improvement. The current features provide basic insight into visualizing an academic network but largely fails to provide an ideal interface.

6.1.1Short-Term Improvements

1. An alteration to the data input could greatly improve the efficiency of the program. Currently, the data is hard-coded as the data sets were very small and were frequently changed, and Graphviz and the Processing layouts required different inputs. However, it would be quick and easy to read in the data from a text file. A text file could be created for each professor including their name, department, and the names of their co-authors.

2. In the survey, it was frequently stated that the visualizations could be improved by including the titles of the shared publications, as many users like to see relevant information immediately. [13] This was not implemented as the titles were often so long that their inclusion entirely altered the appearance of the visualization. Further research would be required to find a method by which to include the publication titles; however, this would greatly improve the ease of use in the long-term.

3. Additional information about each researchers could be added via color-coding. For example, different colors could be used to show the dates of the publications so that more recent publications would stand out more.

6.1.2Long-Term Improvements

1. This project reveals a few important features of academic networks. Graphviz emphasizes the clusters of researcher networks, and the Processing visualization displays the influence of each researcher; however, there is no easy way to combine these two elements as the interfaces are vastly different.

2. Once a better method of processing data is implemented, the program can be expanded to include larger networks. For example, the departments could be expanded to include a number of larger colleges. A comparison could then be made between research at a small liberal arts school and a larger technical university. Additionally, the data could be expanded to include other departments. This was a major issue initially as many professors outside the sciences did not post their publications, but with this data, the difference between humanities, natural sciences, and social sciences could be more closely evaluated.

3. As many publications are formatted in some bibliography style, it would be relatively straightforward to implement a web crawler to collect this information. The design of a web crawler could greatly expand the breadth of data included in this visualization, both for the Processing and Graphviz implementation.

4. An additional feature could be added to output the DOT file for Graphviz directly from the Processing implementation. The relevant data is stored within the Java classes. When a user requests a particular professor, the program could simultaneously write the data to an external file. In order to compile this automatically, however, the program would need to interface with a C script. [14]

6.2 Conclusion

Many methods exist for visualizing and analyzing social networks; however, these have many applications outside the friend networks aggregated from social networking websites. Instead, these methods can have applications to other networks, such as those developed in an academic environment. Many professors include publications online, and this information can help student researchers. The long list of publications included on most professors' pages can be confusing and difficult to analyze at a glance.

Graphviz is an open source software package designed to analyze node-edge structures and output a graphical representation. The software is very versatile and easily adapted to emphasize certain aspects of the data; for example, network clusters, such as researchers who worked on the same projects, were easy to identify when all connections were included. However, the simple node-edge structure does not reflect the strength of a professor's connection to each researcher. As seen in Proximity Graphs, this capability does exist, but it largely uses numerical values to interpret the weight of an edge and obscures the interpretation.

Alternately, a method was developed in Processing to represent the data and build off examples of other social network visualizations. This visualization made it very easy to find important researchers among a long list of co-authors and revealed interesting patterns within the different levels of data; however, some functionality was lost to aesthetics. The connections between groups of researchers were lost. This could be remedied in the long run by the use of color to indicate researchers who are frequently cited together or to include mouse-over text with the names of each researcher's cited publications.

Ultimately, both methods have room for improvement but provide a solid base for further work. The different methods highlight different aspects of the same data and reflect ongoing research in social network visualization. Graphviz has already shown to have extensive applications within visualization; however, the Processing implementation is also a versatile interface. The implementation can be expanded to other social networking data; as with Twitter poster, the co-authors could be replaced with Twitter users and followers to represent the influence of a particular Twitter user or group. The interface offers an adaptable method for data comparison that can have many future applications.

Bibliography

[1] David McCandless. *The Visual Miscellaneum*. Harper Design, 2009.

[2] Edward Tufte. *The Visual Display of Quantitative Information*. Graphics Press, 2001.

[3] Edward. Tufte. *Envisioning Information*. Graphics Press, 1990.

[4] Edward Tufte. *Visual Explanations: Images and Quantities, Evidence and Narrative*. Graphics Press, 1997.

[5] Chaomei Chen. *Information Visualization: Beyond the Horizon*. Springer, 2006.

[6] Jure Leskovec, Kevin J. Lang, Anirban Dasgupta, and Michael W. Mahoney. Community Structure in Large Networks: Natural Cluster Sizes and the Absence of Large Well-defined Clusters. *CoRR*, 2008.

[7] M.E.J. Newman. Finding Community Structure in Networks Using the Eigenvectors of Matrices. *PHYS.REV.E*, 2006.

[8] Thomas Knoll. Last.forward. http://build.last.fm/item/42

[9] Sarah Perez. The Best Tools for Visualization. http://www.readwriteweb.com/archives/the_best_tools_for_visualization.php

[10] David Troy. Twitter Vision. http://twittervision.com/

[11] TwitterPoster. http://twitterposter.com/

[12] Paul Torres. Visualizing Social Networks: A social network visualization of groups in the online chat community of Habbo Hotel. May 2004.

[13] Jeffrey Heer and Danah Boyd. Vizster: Visualizing Online Social Networks. *IEEE InfoVis*, 2005.

[14] John Ellson, Emden Gansner, Yifun Hu, and Arif Bilgin. Graphviz – Graph Visualization Software. http://graphviz.org/

[15] Chris Volinsky. Proximity Graphs. http://public.research.att.com/~volinsky/cgi-bin/prox/prox.pl

[16] David Kempe, Jon Kleinberg, and Eva Tardos. Maximizing the Spread of Influence through a Social Network. *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2003.