How Mathematicians Connect: Visualizing the Mathematics Genealogy Project

Mark Horowitz, Amy Skinner, Julie Pignataro, Kendra K. Levine
iSchool of Drexel University
mdh47@drexel.edu, als92@drexel.edu, jap73@drexel.edu, kkl26@drexel.edu

Abstract

This paper examines the Mathematics Genealogy Project database using different information visualization methods and techniques. The authors analyze effectiveness and ease of use for creating visualizations in Many Eyes, Touchgraph, Microsoft Excel, and Treemap. Future directions are discussed as well as exploration of other academic genealogy visualizations that are available.

Keywords— **Information visualization, Social Networks, Academic genealogies, treemaps.**

1 Introduction

Information visualization has been defined as the use of computers to interactively amplify cognition, using visual representation. [1] Computers have made data analysis easier and the internet has made large data storage and transmission simpler, thereby creating information visualizations is more possible with a wider range of data readily available. Once any network (computer, data, or otherwise) becomes robust enough, it seems that a visualization of that network is not far behind. [2] Tools like IBMs Many Eyes [3] and TouchGraph [4] have made it easier for people to create their own network visualizations, illustrating abstract relationships that might not otherwise be visible. Social networks and professional communities have benefited from these technologies, as users can more clearly see relationships that might normally be obscured by the sheer volume of the data. Visualization tools have also made it possible to map different aspects of knowledge domains, adding to the academic community. [5, 6]

1.1 Mathematics Genealogy Project

The Mathematics Genealogy Project (MGP) [7], a service of North Dakota State University, contains information about those who have acquired advanced degrees in mathematical disciplines, including their advisors and dissertation titles. MGP's services are freely provided: a searchable database, data submission form, and genealogy posters (for purchase) which display a person's or educational department's genealogy history in the form of a family tree.

Our intent was to provide useful and insightful visualizations to the MGP using their database, which at the time we acquired it, contained 117,137 student records which go back to 1605. Mitch Keller, the Assistant Director of MGP, stated that he had tried to create visualizations previously that show the student-advisor relationship, but that the program he was using was overwhelmed by the quantity of data, and no visualizations were produced.

It was our goal to provide the MGP with visualizations that could help them grasp the breadth and depth of the data that they have, and hopefully help them gain insights into the connections among their data. Keeping with Keller's original intents, we focused on the student-advisor relationship. We choose to analyze the MGP data with Many Eyes, TouchGraph Navigator, TreeMap [8] and Microsoft Excel.

1.2 MGP Data

Keller provided our team data in tab-delimited form, which contained the following fields: id, last name, first name, middle name, degree, year, dissertation title, advisor 1, advisor 2, school, and miscellaneous field. All but the names and title were numeric codes. A separate school list was provided; it included name and country code, which was further decoded from a country list. We omitted any data sets which were incomplete in one way or another, such as missing the student ID or advisor code.

2 Visualization Methods

We decided to use networks and trees for our MGP visualizations because they would most clearly illustrate the relationships between students, advisors, and institutions. Through these types of visualizations, the underlying network patterns are easier to interpret and the data easier to mine. [9] We used Schneiderman's task by data type taxonomy to evaluate the effectiveness of the different methods. [10] Schneiderman's seven tasks that a user can apply to a visualization are: overview, zoom, filter, details-on-demand, relate, history, and extract. Expert users are more able to see the depth of the various relationships expressed in a visualization, though the graphic nature allows novices to see the most basic and rudimentary relationships as well.

Social networks are a common data set to create information visualizations because they contain rich information about networks that might not readily be seen another

way. Many visualizations of social networks focus on relationship rankings, cohesive subgroups of the network, and ego-centered networks where individuals serve as the primary node. [11] Mapping social networks is nothing new and has been in practice since the 1930s, though the field has boomed with the explosion of the internet. [12] Creating networks with the MGP data illustrates different trends, such as which institutions have the most students, or which advisors have spawned large numbers of graduate students and future faculty. To make these connections different tools were more effective to highlight each relationship.

2.1 Many Eyes

Many Eyes is a website sponsored by IBMs Collaborative User Experience (CUE) Visual Communication Lab, which enables users to freely generate a variety of information visualizations with their own data. Once users register with the site, they may upload their own tab-delineated data and create visualizations. [2, 13] The site is user-friendly and supports exploration by providing overview, zoom, pan and additional details on demand (typically by hovering over an item) depending on the type of visualization and the data [10]. There is also the ability to change the sort order from the visualization (instead of having to return to the data set) as well as use color to depict any differentiation or highlight a particular node.

Figure 1 shows experimentation with a subset from the data provided by MGP. The point of this visualization is to show the US schools that have graduated the most doctoral candidates in mathematics in the last 139 years. The top five schools are: Massachusetts Institute of Technology (2,212); Stanford University (2,059); Princeton University (1,348); Harvard University (1,275); and New York University (1,125).

A geographical visualization was also performed using Many Eyes with subsets of the MGP data to see global patterns. For the geographical visualization, a count of how many students per country was calculated from the data given, with 32 records removed (countries with zero doctoral degreed students) and 26 records of compound countries (i.e. Austria-Italy) also removed due to confusion as to which country the student(s) and/or school actually belonged (Figure 2).

While it is interesting to see the geographical spread of doctoral degrees, the significantly large number from the United States (over 55,000) skews the range (which is not an adjustable feature) and makes detecting greater differences or any patterns difficult. Perhaps if the range could be changed to reduce this effect, such as per hundred, the disparity would be decreased and any significant findings would be more noticeable.

The main constraint we encountered with Many Eyes comes from the 5 MB limit on data that can be visualized. This posed a problem with the data provided from MGP. Though the data was not extremely complex, the sheer quantity of data limited the production of useful visualizations that could use the entire data set. Also, the visualization can be saved, but not easily exported to other formats for inclusion in reports, presentations, etcetera. On the other hand, Many Eyes is an incredibly powerful tool for smaller data sets, and the number of available visualization options is a positive aspect.

2.2 TouchGraph

TouchGraph Navigator is a graphical visualization tool that shows how items (e.g. people) are related. This application accepts data in several formats and Microsoft's Excel spreadsheet was chosen for the MGP visualizations. An Excel format wizard is launched that enables a user to define the relationships between items from the spreadsheet(s) within the Excel workbook and select associated attributes (e.g. school, year) to show additional information. The visualization is generated (and one can watch the image appearing as it is being produced) after setting the parameters. On screen, settings (e.g. node appearance, size, edge width) can be changed to make the graph more visually aesthetic, and the cluster color scheme may also be changed. Windows to the left of the main visualization window display more in-depth information of the selected node and a list of the entire data set provided. Control over zooming and spacing between nodes is available as well as hiding or keeping nodes for information clarity. A nice feature for viewing relationship hierarchy is to hover over each node the edges change color with red linking the child node back to the parent node (Figure 3).

An original intent of the project was to visualize the student-advisor relationships in the hope of presenting an historical map of mathematicians. Unfortunately, our attempts of visualizing this network in its entirety fell short due to file size and display limitations of both Many Eyes and TouchGraph. The figure below (Figure 3) is a sample of only one of those relationships.

TouchGraph is a very powerful tool that has been used by Google, Live Journal, and Facebook to show social networking relationships. [12] It also allows easy exportation of visualizations in PNG format to facilitate information sharing. Since there is no question of whether TouchGraph can handle the quantity of data provided by the MGP, more time will be needed to explore how to optimize the database to create an entire network visualization.

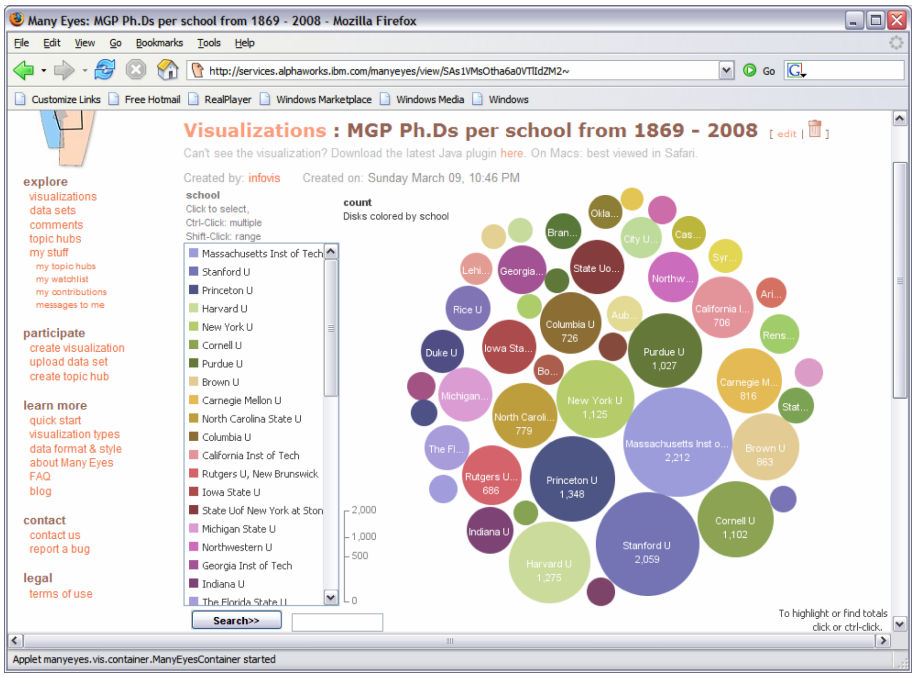

Figure 1: Many Eyes Visualization: Number of doctoral graduates in mathematics from 1869-2008 (US schools with less than 100 graduates were eliminated from this dataset for optimal visualization).

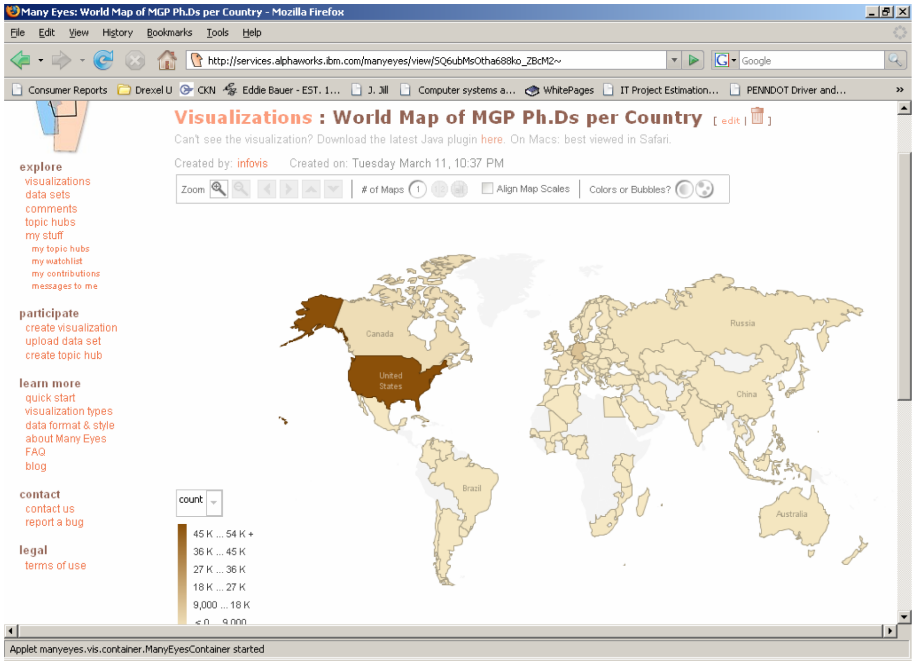

Figure 2: Many Eyes Visualization: Number of doctoral graduates in mathematics per country.

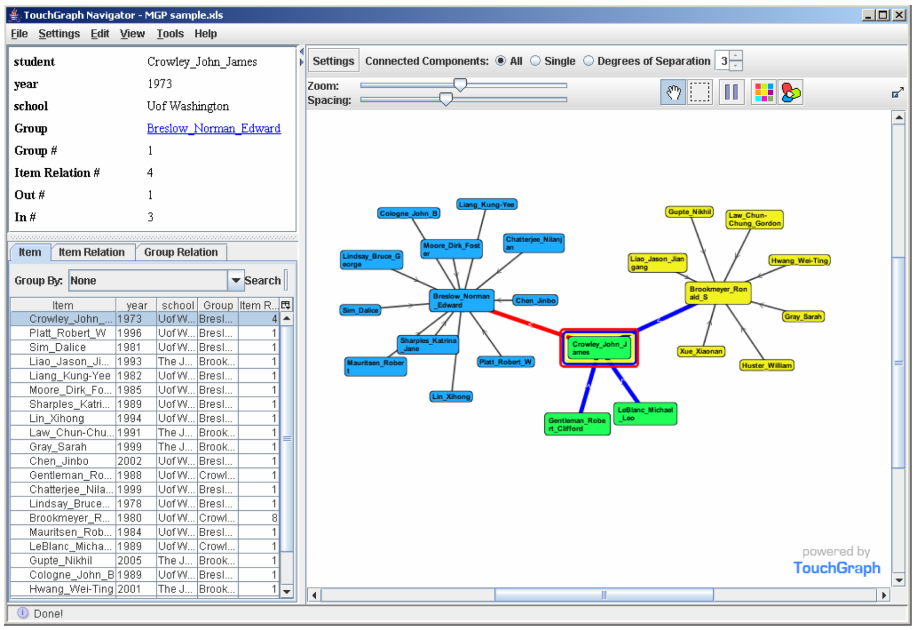

Figure 3: TouchGraph Navigator visualization showing a student-advisor relationship. Norman Edward Breslow (blue cluster to the left) was the advisor of John James Crowley, 1973 (green center cluster) who was the advisor of Ronald S Brookmeyer, 1980 (yellow cluster to the right). Clicking on a student provides additional details of graduation year and school.

2.3 Excel

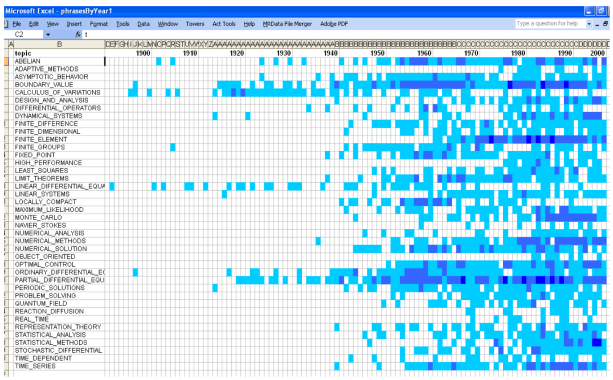

Figure 4: Microsoft Excel visualization showing the occurrence of the top 40 phrases used in dissertation topics from 1900-2008.

In order to explore the evolution of dissertation topics over time, we reorganized the data to extract two- and three-word phrases present in the dissertation titles. We then ranked them by frequency of occurrence. We arbitrarily chose the top 40 and determined the number of dissertations in which each occurred in each year from 1900 to 2008. We then used Microsoft Excel's conditional formatting feature to chromatically differentiate 4 different levels of phrase occurrence, 0 times, 1-3 times, 4-10 times, and more than 10. This permitted a very cognitively gentle visualization of phrase trends in the 20th century for these topics. We can see that boundary value presence did not pick up until the 1920s, reached a plateau in the mid 50s, only to decrease through the 70s and peak in the following two decades. Not surprisingly, the same pattern occurred for partial differential equations. Linear differential equations, interestingly, was relatively constant until the 1970s when it began to taper, probably because more and more mathematical models began to reflect multiple interacting variables. We expect that a better analysis of this data would result from the use of citation data related to the dissertations, which might more powerfully reflect the semantics of the dissertation topics.

2.4 TreeMap

Visualizations provided by the Mathematics Genealogy Project itself consist of only those on its posters, and they include ancestors of a single degree recipient or an academic department. Picturing the entire corpus of data has not been successful, nor would one reasonably expect that a network of roughly 100,000 nodes be easy to understand. Treemaps can reduce the complexity of the discussed visualizations because they highlight linear nodes, obscuring the structure of the data. [14] For this purpose we are ignor-

ing the second advisor data present for some of the degree recipients. TreeMap was designed following Schneiderman's rules for tasks, because it allows users to highlight areas of the tree for greater detail. A single click would produce a list of the ancestors present in the data, otherwise known as details on demand. Other parts of such a tool might include lists of schools and countries, and the ability to chromatically distinguish people by selecting a school or country. TreeMap, by definition, can show an entire tree of up to a million nodes, depending on display, on a single screen.

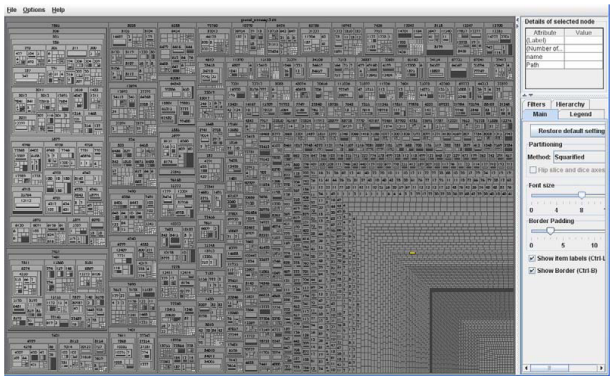

Figure 5: TreeMap Representation of the Entire Ancestor Structure

Although there are numerous open source TreeMap tools, we chose to use the software from the University of Maryland because of its flexibility and ease of interaction. For this purpose, we converted the MGP data, which is supplied in the form of tree nodes, one per degree recipient, to paths from most distant ancestor to each recipient without further descendents. We show the treemap created in Figure 5. Further work with the data and program will permit seeing colors, schools, countries, and names. Even without these additions, it is easy to see entire lineages within rectangles, such as that in the upper left hand corner, which corresponds to Hubert Anson Newton, who received his Ph.D. at Yale University in 1850, and has 12,018 descendants. This representation needs to be compared with the hypothetical one described above.

3 Conclusions and Future Direction

A benefit in trying to develop a more comprehensive visualization was the experimentation with different visualization types. The data received from the MGP did have us leaning towards the generation of a network view initially, but the experience of data manipulation led to testing other visualization types, visualization software and presenting other aspects (e.g. top schools, geographic dispersion) that may be interesting to see as well. On that note, there was more data preparation involved than we were expecting in order to create comprehensible visualizations. We found that we had to divide the data into smaller, more manageable sets which resulted in breaking up many relationships among the data. Therefore, despite the wide variety of visualization tools available, we found the necessary trimming down of the original database a large stumbling block.

The aspect of having to manipulate the data before creating visualizations is one of the useful features of the Many Eyes applications. Because the data becomes public domain on the web after it is uploaded, anyone can create visualizations from it (only the original user that uploaded the data can manipulate it) and this may create different perspectives of the data that the original provider of the data did not envision.

From the first three visualizations from Many Eyes and TouchGraph, it can be concluded that though the visualizations themselves were more pleasing to the eye than looking at a spreadsheet of the data, the information we gleaned from the visualizations could have easily been extracted from a spreadsheet type application. Figure 1 shows that the Massachusetts Institute of Technology had the most doctoral degreed students, but this data could have been done by organizing the database by the school field. The same applies to the country field (Figure 2) and it should also be noted that this was not surprising considering that the MGP is based in the United States and the website for registration is in English only. Perhaps if visualization could be prepared comparing the number of schools around the world that award doctoral degrees in mathematics with the number of qualified candidates in the MGP database, some meaningful information could be extracted.

3.1 Academic Genealogy Visualizations

Although there has been some work done on academic genealogy visualization [15, 16], the area contains more data than analysis or visualization. Therefore, we see the following steps taken:

- We envision a proposed standard for academic genealogy data. There are other sets of such data available, such as the Software Engingeering Academic Genealogy at North Carolina State University [17] and the Artificial Intelligence Genealogy Project from University of Texas, Austin. [18]

- The creation of a standard set of tools and visualizations would be useful for each of the communities which have compiled genealogy data. Further explorations in these collections of data will stimulate additional questions and potentially new methods of visualization. A challenge encountered in this project was the volume of data to visualize, using freely available, Web-based tools. We recognize that data such as

that used here may require more robust tools, some of which may need to be created for this purpose.

As technology progresses, it should become much easier for applications such as those described above to handle large sets of data such as that from the MGP. Additional visualization types that have potential in this area are a timeline, for example an advisor from 1901 followed through to today, and a tree view visualization would be worth exploring for displaying the student-advisor relationships comprehensively.

Acknowledgements

We graciously offer our thanks to Mitch Keller, Assistant Director of the Mathematics Genealogy Project, for providing us with the data used for this experiment. We also want to thank Alexander Shapiro, CEO of TouchGraph LLC for the trial software of TouchGraph Navigator, and Dr. Chaomei Chen, professor at Drexel University for his encouragement and insight.

References

[1] Zachary Pousman, John Stasko, and Michael Mateas. Casual information visualization: Depictions of data in everyday life. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1145–1152, 2007.

[2] Fernanda B. Viegas, Martin Wattenberg, Frank van Ham, Jesse Kriss, and Matt McKeon. Manyeyes: a site for visualization at internet scale. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1121–1128, 2007.

[3] Many Eyes. Website, 2008. http://services.alphaworks.ibm.com/manyeyes/page/About_Many_Eyes.html.

[4] TouchGraph. Website, 2008. http://www.touchgraph.com/.

[5] Remo A. Burkhard. Learning from architects: The difference between knowledge visualization and information visualization. *iv*, 00:519–524, 2004.

[6] Chaomei Chen. *Information Visualization: Beyond the Horizon*, chapter 5: Knowledge Domain Visualization. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[7] MGP data set. Website, 2008. http://genealogy.math.ndsu.nodak.edu/index.php.

[8] TreeMap. Website, 2008. http://www.cs.umd.edu/hcil/treemap/.

[9] John Galloway and Simeon J. Simoff. Network data mining: methods and techniques for discovering deep linkage between attributes. In *APCCM '06: Proceedings of the 3rd Asia-Pacific conference on Conceptual modelling*, pages 21–32, Darlinghurst, Australia, Australia, 2006. Australian Computer Society, Inc.

[10] Ben Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. In *VL '96: Proceedings of the 1996 IEEE Symposium on Visual Languages*, page 336, Washington, DC, USA, 1996. IEEE Computer Society.

[11] Adam Perer. Making sense of social networks. In *CHI '06: CHI '06 extended abstracts on Human factors in computing systems*, pages 1779–1782, New York, NY, USA, 2006. ACM.

[12] Jeffrey Heer and Danah Boyd. Vizster: Visualizing online social networks. In *INFOVIS '05: Proceedings of the Proceedings of the 2005 IEEE Symposium on Information Visualization*, page 5, Washington, DC, USA, 2005. IEEE Computer Society.

[13] Fernanda B. Viegas, Martin Wattenberg, Matt McKeon, Frank van Ham, and Jesse Kriss. Harry potter and the meat-filled freezer: A case study of spontaneous usage of visualization tools. In *HICSS '08: Proceedings of the Proceedings of the 41st Annual Hawaii International Conference on System Sciences*, page 159, Washington, DC, USA, 2008. IEEE Computer Society.

[14] Ivan Herman, Guy Melançon, and M. Scott Marshall. Graph visualization and navigation in information visualization: A survey. *IEEE Transactions on Visualization and Computer Graphics*, 6(1):24–43, 2000.

[15] Sooyoung Chang. Academic genealogy of american physicists. *AAPPS Bulletin*, 13(6), December 2003.

[16] Cheryl Davis Gary Marchionini, Paul Solomon and Terrell Russell. Information and library science mpact: A preliminary analysis. *Library & Information Science Research*, 28(4):480–500, Winter 2006.

[17] Software engineering academic genealogy project. Website. http://people.engr.ncsu.edu/txie/sefamily.htm.

[18] AI genealogy project. Website. http://aigp.csres.utexas.edu/ aigp/.