Journal of Contemporary Eastern Asia, Volume 14, No.2: 87-96
http://dx.doi.org/10.17477/jcea.2015.14.2.087

Catalyzing social media scholarship with open tools and data

Marc A. Smith[1]

Social media comprises a vast and consequential landscape that has been poorly mapped and understood. Hundreds of millions of people have eagerly moved many of the conversations and discussions that compose civil society into these services and platforms. There is a need to document and analyze these social spaces for many academic and commercial purposes. The Social Media Research Foundation has engaged a strategy to cultivate better research into the structure and dynamics of social media. The foundation is dedicated to the creation of open tools, open data, and open scholarship related to social media. It has implemented a free and open network collection, analysis, and visualization tool called NodeXL to facilitate social media network research. Using NodeXL a group of researchers has collectively authored a publicly available archive, called the NodeXL Graph Gallery, composed of network data sets and visualizations from users around the world. This site has enabled the aggregation of tens of thousands of network datasets and images. Use of the archive has led to scholarly research results that are based on the wide range and scope of social media data sets available.

Introduction

Social media research can be catalyzed by improving the quality of research tools, expanding the scope and accessibility of research data, and openly sharing research results. The Social Media Research Foundation was founded to realize the goal of "Open Tools, Open Data, and Open Scholarship" related to social media. The Foundation is a not-for-profit organization dedicated to creating research tools and facilitating better understanding of the landscape of social media.

Challenge

Many tools for network data collection, analysis, and visualization are freely available. However, many require programming skills that are not widely found, particularly in the social science, humanities, and business management community. Researchers and practitioners in these fields have been blocked from directly accessing and analyzing social media network data because of the significant skills required to do so using programming languages. Programmers must often master a variety of tools and technologies, from programming languages to databases and application programming interfaces in order to collect and gain insights into connected structures.

[1] Social Media Research Foundation, USA.

The obstacle this presents to many researchers in non-technical fields has been a significant barrier to entry.

Inspiration

Firefox is a free and open web browser for HTML, the XML file format for describing a web page. The Mozilla Foundation makes this software globally available to enable and ensure that the widest possible audience has the tools needed to access the wealth of information stored in the World Wide Web. The Social Media Research Foundation shares the goal of providing a freely available public tool for access to a different but equally important type of data format: the network. While Firefox is widely known as a "web browser", in practice it rarely displays a "web". In practice, Firefox is a "page" browser, leaving a need for a real "web" browser that can simplify access to insights into connected structures. To avoid confusion we might describe the software as a "net browser" to distinguish it from web browsers.

Goals

The Social Media Research Foundation seeks to create easy to use tools for social media data collection, analysis, visualization and publication. NodeXL, the primary project of the foundation, is a tool that aspires to become the "point-and-shoot digital camera" for social media crowds. Using NodeXL most users with basic spreadsheet skills can collect, store, analyze, visualize and publish network data extracted from a variety of social media data sources.

A community of NodeXL users has emerged, some of whom have contributed data sets, visualizations, and guides to data analysis for other users. The creation of an archive of data sets enables public analysis of a wide range of social media network data. As this global collection of users has developed, the result has been an increasingly diverse set of geographic and topical data. Cultural, regional, and topical differences have emerged that illustrate the ways the same technology can be used in diverse ways.

This large social media network archive has enabled the publication of freely available research findings based on the results of analysis using these tools and data. The archive often provides data sets that can be contrasted to illustrate the diversity of structures and forms present in social media systems.

NodeXL is designed for use by anyone interested in networks, social networks and particularly social media networks. NodeXL users are often scholars, researchers, students, managers, and analysts who are interested in understanding the shape, structure, and key positions within a connected structure. Since societies are connected structures, people interested in understanding organizations, groups, enterprises, and markets are often interested in networks. The NodeXL project is focused on creating a network analysis tool built for ease of use, automation, and reporting that highlights key features of interest in a connected structure. Integration with Microsoft's widely used Excel program is intended to make NodeXL simple to use. For example, if you can make a pie chart in Excel, you can easily make a network chart using NodeXL.

Research examples

There are already a number of examples in which NodeXL has been successfully employed by users in a variety of disciplines and professions.

Professor Diane Cline at George Washington University is a scholar of antiquity with a specialty in the life of Alexander the Great. She has been able to quickly master NodeXL and use it to create some of the first maps of Alexander's social network. While he may not have had a Facebook page, Alexander the Great did have a web of connections and relationships that are now easier for scholars to understand by visualizing and analyzing them in NodeXL.

http://www.academia.edu/2153390/The_Social_network_of_Alexander_the_Great_Social_Network_Analysis_in_Ancient_History

Business professor Scott Dempwolf at the University of Maryland uses NodeXL to map the connections formed when people author patents together. These networks reveal clusters of innovations that define a regional economic specialty.

http://portal.sliderocket.com/ATWBE/Using-SNA-to-find-and-manage-RICs
http://portal.sliderocket.com/ATWBE/Using-SNA-to-find-and-manage-RICs

Lee Rainie, director of the Pew Research Internet Project, partnered with the Social Media Research Foundation to use NodeXL to map a range of Twitter social media networks. The research team documented the existence of six distinct patterns of network structures that regularly occur in Twitter topic streams. These six network types can help people understand the conversations in which they participate and recognize the patterns of conversations that they want to emulate.

http://www.pewinternet.org/2014/02/20/mapping-twitter-topic-networks-from-polarized-crowds-to-community-clusters/
http://www.pewinternet.org/2014/02/20/mapping-twitter-topic-networks-from-polarized-crowds-to-community-clusters/

Next steps for social media network analysis

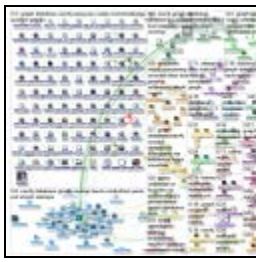
The Social Media Research Foundation is also focusing on new and improved network data importers from social media services like Twitter, Facebook, YouTube, Sina Weibo, VKontackte and others. Many sources of social media network data exist on the Internet and a goal of the NodeXL project is to make it easy for users to extract those networks without requiring any programming skills.

Work remains on simplifying the use of NodeXL. While it was designed for simplicity and end user ease of use, it can still be a challenge for new users to create their first network image captures. Future design changes will focus on reducing and removing these obstacles. The project will develop simpler interfaces that will lead users step by step through the collection, analysis, visualization and publication of network analysis.

Updates to the web interface for the NodeXL Graph Gallery will enable more interaction with network datasets without the need to download and run the desktop client NodeXL application. This will allow much of the data consumption task to be performed using almost any computing platform.

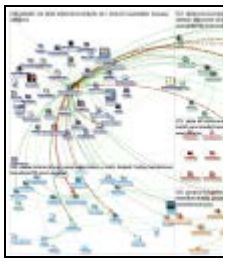
Documenting the range of variation of social media networks in Twitter allows us to understand more clearly where users are in the social media landscape. For example, consider the landscape around X on Twitter:

This map can be contrasted with social media network maps for related topics as in the examples below:



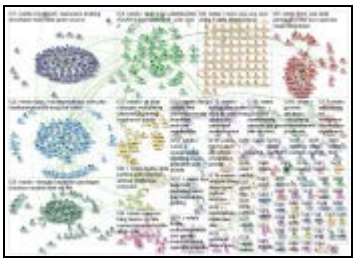
graph database Twitter ...

data science OR #datasc...



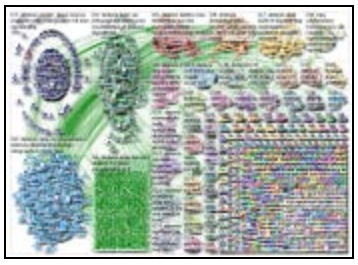
datacommunitydc

datascientist Twitter N...



rstats

#bigdata



dataviz OR datavis

These different patterns can be compared with the six basic patterns of network structures the Social Media Research Foundation discovered (see page 7). Social media researchers and analysts can use these maps to identify the key people, groups, topics and links in each topic stream. Each report links to a summary of the content in the network as a whole and well as the details of each sub-group's discussion.

An example of how such network visualizations may be used would be identifying the key person in, or "mayor" of, the discussion around "datavis" or "#bigdata" in Twitter. Users with high centrality are good candidates for the role of "mayor". NodeXL social media network maps can also identify sub-groups with different interests in topics. Analysis of frequently used URLs, hashtags and vocabulary in each cluster can reveal differences and similarities among sub-groups. The overall structures of networks aid in the understanding and analysis of the nature of groups and discussions. The use of NodeXL allows researchers to examine network characteristics such as why some groups are polarized and divided, while others are unified and why some groups are fragmented or clustered, while others have distinct hub-and-spoke shapes.

Open Tools

In common with organizations such as the Mozilla Foundation, known for its Firefox browser, the Social Media Research Foundation believes that some tools are better if they are open and free. The open NodeXL application enables more people to better understand the structures and dynamics of social media, which is used by hundreds of millions of people on a daily basis. NodeXL was designed to function as a kind of "point-and-shoot digital camera" for capturing images of crowds in cyberspace. The foundation seeks to promote and enable the creation of as many pictures of social media networks as possible to better document the characteristics and activities of these populations. To this end, the foundation wants to make NodeXL freely and easily available to as many people as possible. An open approach is also important for facilitating collaboration for both academic and commercial purposes. Open data and open scholarship are important to ensure that researchers in every disciple can access and share data widely. With a quarter of humanity living in cyberspace, it is time to properly document, map and study this new terrain. Open tools help make that possible.

Future directions

The Social Media Research Foundation seeks to create a reality in which social media analytics is not just for analysts. Many of us spend a lot of our time using social media: it's where our people are. But there is far more social media data than any human can consume, so we need to

prioritize and filter our feeds. Analytic tools will become mainstream as people seek tools to help manage torrents of posts and messages into a focused image that reveals the key people, groups, topics, and bridges. Social media analytics might "disappear" at that point, becoming a normal part of our interfaces with social data. Visualization will be a big part of that — a necessary method of bringing analytic insights to people with limited quantitative training or skills. So "Data Science" may soon follow the path traveled by "Desktop Publishing" — software will simplify complex processes so that casual users can do 80% of what they need for themselves.

While software development and database skills are and will remain very useful, the foundation is focused on tools like NodeXL that allow 80% of what we now call "data science" to be done by casual end-users in the same way that "desktop publishing" enabled anyone with some text to create professional looking results. As technical skills decline in importance, insight into social processes and structure, the work of the social sciences, will increase in importance.

Initial results

The NodeXL application has been downloaded more than 300,000 times. More than 50 universities have courses featuring the use of NodeXL. Hundreds of research papers have been published which contain one or more network maps and reports generated with the NodeXL application . Research is being conducted using the archive of data collectively constructed in the NodeXL Graph Gallery. Research into network structures has expanded into disciplines with lower levels of technical skills. Employing the easy to use NodeXL application, researchers in the Digital Humanities have been able to extract and analyze social networks based on data on the lives of historical figures. Business scholars have mapped the networks of relationships created when people author patents together. This research reveals the “innovation clusters” that form in regions or particular cities which attract a particular kind of technical skills or industry.

Recent research has made extensive use of the NodeXL application’s automation capabilities. Using automated data collection and analysis, a large collection of networks related to a wide variety of topics can be easily created. Many network image captures generated daily over long periods of time can produce a data set that depicts the variation in social media structures. Analysis of these network structures revealed that six patterns could be found appearing repeatedly.

The NodeXL application can collect and categorize a large set of network visualizations and data sets. The NodeXL project (http://nodexl.codeplex.com) from the Social Media Research Foundation (http://www.smrfoundation.org) is a free and open add-in to the familiar Microsoft Excel spreadsheet that integrates features for collecting, analyzing, visualizing and publishing social media networks (http://nodexlgraphgallery.org). It provides a set of simple to use features that process the short texts often found in social media content like Tweets. NodeXL supports direct access to a range of social media data sources including Facebook, YouTube, flickr, email, Twitter, the World Wide Web, Wikis, blogs and data formats like GraphML.

NodeXL builds and segments a social graph of connections among people based on the connections formed when users reply and mention one another [3]. NodeXL first builds a visualization of the connections among the users who mentioned a selected keyword or search term, which can include arbitrary strings, hashtags, URLs, or user names [4]. Network visualizations are further segmented to highlight each algorithmically discovered cluster. Each cluster is displayed in its own region and the most commonly used hashtags mentioned in each cluster are displayed above it [5, 6]. Each user is located in a cluster and is represented by their profile photo, which is

scaled in size in proportion to their count of followers. Each user photo has lines connecting each to all others on the map that they reply to or mention.

Sub-collections of messages are created within the social clusters found within these social networks. Sub-groups within the same social networks often use distinctive hashtags, URLs, word pairs and user names. Contrasting summaries of the text used within each group offers a quick way to distinguish the interests, resources and leaders of each cluster in the network.

Social network visualizations of connections among Twitter users illustrate the macro structure that emerges from the micro level of relationships created by each user. All users displayed in these maps are there because they posted a tweet that contained the search term. Currently, the Twitter API returns up to a maximum of 18,000 tweets for a single query, which may span as long as ten days or as little as a few minutes or seconds of content depending on the speed and volume of content created that contains that term. In practice Twitter often returns far fewer than 18,000 tweets, either because there are fewer than that number in the time period or it imposes limits. Populations of users derived from these query results vary from a maximum of 18000 to as few as zero. NodeXL displays a list of all the users in a data set with summary data about their history in Twitter and their network metrics.

1. Attributes of social media network patterns

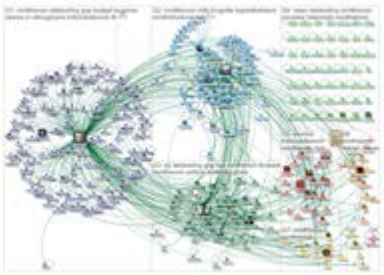
Six basic social media network patterns can be found in services like Twitter. We analyzed a large collection of network maps and reports. Over time, these six structures reappeared frequently. There may be additional network structures in social media but these patterns seem stable and regularly occur in many data sets. Each has distinctive network properties.

Polarized Crowd: Polarized discussions feature two big and dense groups that have little connection between them. The topics being discussed are often highly divisive and heated political subjects. In fact, there is usually little conversation between these groups despite the fact they are focused on the same topic. Polarized Crowds on Twitter are not arguing. They are ignoring one another while pointing to different web resources and using different hashtags.

Tight Crowd: These discussions are characterized by highly interconnected people with few isolated participants. Many conferences, professional topics, hobby groups, and other subjects that attract communities take this Tight Crowd form.

Brand Clusters: When well-known products or services or popular subjects like celebrities are discussed on Twitter, there is often comment from disconnected participants (the "isolates" who are participating in a conversation cluster are on the left side of the picture on the left). Well-known brands and other popular subjects can attract large fragmented Twitter populations who Tweet about it but not to each other. The larger the population talking about a brand the less likely it is that people are connected to one another. Brand mentioning participants focus on a topic, but tend not connect to each other.

Community Clusters: Some popular topics may develop multiple smaller groups, which often form around a few hubs each with its own audience, influencers, and sources of information. These Community Clusters conversations look like bazaars which host multiple centers of activity. Global news stories often attract coverage from many news outlets, each with its own following. That creates a collection of medium-sized groups – and a fair number of isolates (the left side of the picture above).



Broadcast Network: Twitter commentary around breaking news stories and the output of well-known media outlets and pundits has a distinctive hub and spoke structure in which many people

repeat what prominent news and media organizations tweet. The members of the Broadcast Network audience are often connected only to the hub news source, without connecting to one another. In some cases there are smaller subgroups of densely connected people – think of them as subject groupies – who do discuss the news with one another.

Support Network: Customer complaints for a major business are often handled by a Twitter service account that attempts to resolve and manage customer issues around their products and services. This produces a hub and spoke structure that is different from the Broadcast pattern. In the Support Network structure, the hub account replies-to many otherwise disconnected users, creating an outward hub. In contrast, in the Broadcast pattern, the hub gets replied to or re-tweeted by many disconnected people, creating an inward hub.

Discussion

Conclusion

The work of the Social Media Research Foundation is rooted in the idea that open and free tools such as NodeXL can expand the population of researchers and practitioners able to engage in social media network data collection and analysis. As researchers from a wider variety of disciplines gain the ability to collect, and find insights into, connected structures, our understanding of social media will improve.

Acknowledgements

Our thanks to the members of the NodeXL project team and the Social Media Research Foundation and the Pew Internet Research Center. Special thanks to Lee Rainie, Itai Himelboim, and Ben Shneiderman, and Jana Diesner.

References

[1] Smith, M., Lee Rainie, Ben Shneiderman, Itai Himelboim. 2014. Mapping Twitter Topic Networks: From Polarized Crowds to Community Clusters, Pew Internet Research Center. http://www.pewinternet.org/2014/02/20/mapping-twitter-topic-networks-from-polarized-crowds-to-community-clusters/

[2] Smith, M., Shneiderman, B., Milic-Frayling, N., Rodrigues, E.M., Barash, V., Dunne, C., Capone, T., Perer, A. & Gleave, E. (2009), "Analyzing (social media) networks with NodeXL", In C&T '09: Proc. fourth international conference on Communities and Technologies. New York, NY, USA., pp. 255-264. ACM.

[3] Clauset, A., Newman, M. E. J., & Moore, C. (2004). Finding community structure in very large networks. Physical Review E, 70(6), 066111.

[4] Harel, D., & Koren, Y. (2001). A fast multi-scale method for drawing large graphs. In 8th International Symposium on Graph Drawing, 1984 Lecture Notes in Computer Science, 183-196.
[5] Eduarda Mendes Rodrigues, Natasa Milic-Frayling, Marc Smith, Ben Shneiderman, Derek Hansen, Group-in-a-box Layout for Multi-faceted Analysis of Communities. IEEE Third International Conference on Social Computing, October 9-11, 2011. Boston, MA