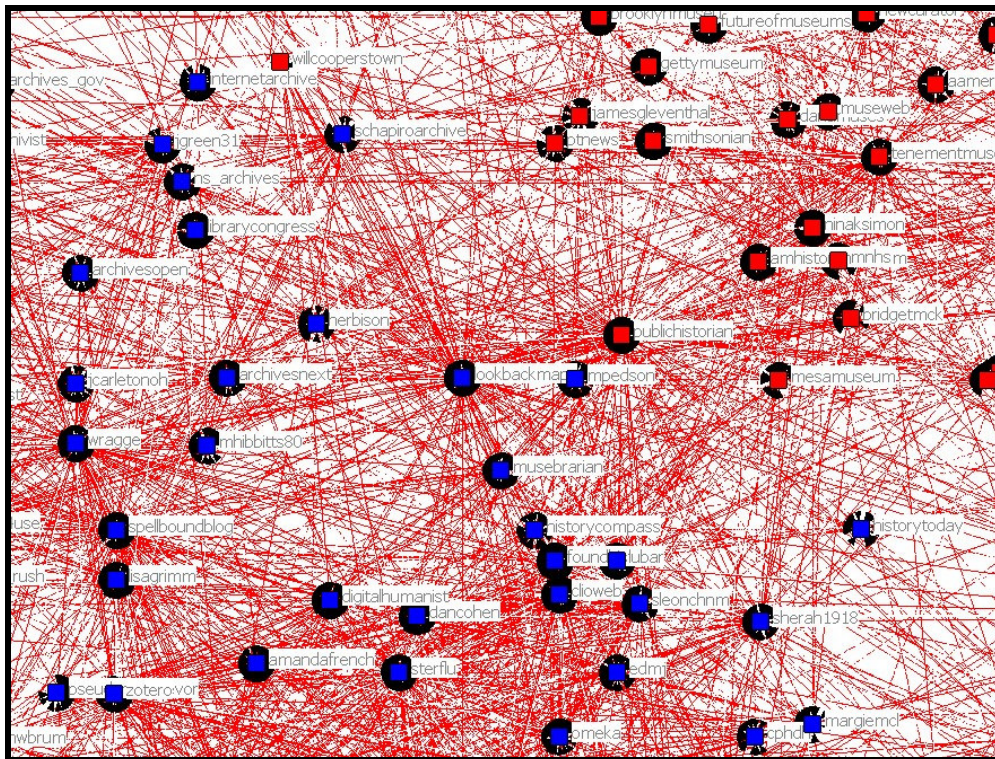




intelligent information design

December 10, 2009

Ph. 415-593-5508 jon@jumpslide.com



An analysis of the Twitter network of user @LookBackMaps reveals a tight knit community of people interested in making historical collections more accessible, interactive, and interesting to the general public. Social Network Analysis and mapping reveals two distinct factions of key players, those in the digital humanities, and those in the museum world. Tools used included custom Javascripts, the Twitter API, Microsoft Excel, and Analytic Technologies' NetDraw. This article examines the methodology used to collect 1st and 2nd degree network nodes related to @LookBackMaps, and what kind of information was gleaned from the analysis of this data.

Recently, I needed to review options and create documentation for social network analysis software for a client. Because I find it really difficult to explore such systems in the abstract, I decided to analyze the social network of my Twitter connections for @LookBackMaps.

Of course, I'm a technologist and not a mathematician, but I felt that the use of network visualization software would give me some tools to look deeper into this unique network without an intricate understanding of the math behind various algorithms of network theory.

Observations about my network

More than a year ago, I started work on a project called LookBackMaps. Simply put, LookBackMaps.net is a mashup that displays geotagged historical photographs on a Google Map. While still a prototype, it's a simple, yet robust way of visually organizing, exploring and engaging in history and historical photographs. Over the course of the last year, the idea has evolved into finding new ways to connect to disparate, stand alone historical archives, and building community around history. When I started, I didn't know anything about digital humanities in academia, or the movement afoot amongst archivists, libraries and public history professionals to make historical archives more accessible in a Web 2.0 environment. Through word of mouth and contacts within my social network, I met several influential people involved in this ongoing sea change. Utilizing Twitter, I began to follow people and organizations that are active in these discussions, joining in the conversation, and researching technical innovations in the world of web archives. Before I knew it, I found myself in an interesting and open community of like-minded people engaged in the digital humanities.

While there are a lot of academics involved in this work, I was surprised to see how open they are to contributions and innovation from outside the university. Yet I found there was also somewhat of a disconnect between the open source efforts of innovative academics and those of commercial technologists of the Silicon Valley network. I suppose that makes sense given the fact that the aims of both groups are very different, even while there is a mutual respect and even admiration that is a two way street between the two. Somewhere in the middle lie projects like LookBackMaps, which focus on the general public, not academics, and seek to provide free access to publicly available data in an open and collaborative environment. Additionally, there are a growing number of museums, libraries, and archives that are searching for ways to open their collections to more collaborative efforts, adding their voice to the conversation.

This is the very unique community I found myself participating in with Twitter and my @LookBackMaps screen name. I wanted to dig deeper and build a map of first and second degree connections to see what this network looked like, how were people connected, what kinds of groups or factions existed, who had the most influence?

Discovering Factions

Upon analysis of my first and second degree connections within my Twitter network, I found two rather distinct groups, or factions, which I ascribed to digital humanities and museums2.0. The statistical data

highlighted the existence of these groups, and I could easily see the difference in content as well as connections¹.

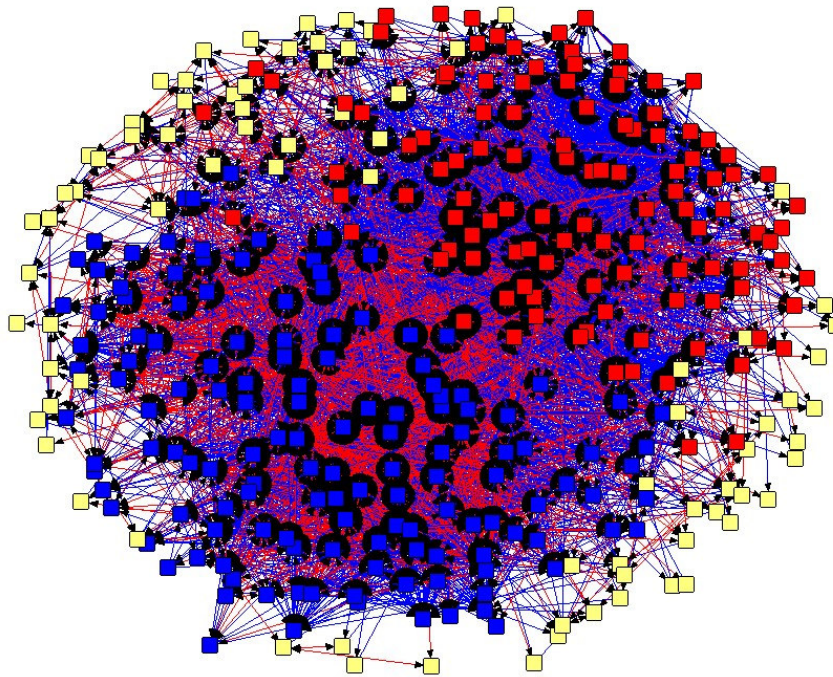


Figure 1. Blue nodes are mostly digital humanities and red nodes are mostly museum2.0, while yellow are outlying others. Red links reflect reciprocal relationships.

The roughly defined digital humanities network is exemplary of a rather small, niche community that uses Twitter as a unique and open tool for networking, idea exploration, sharing, and innovation. From a qualitative perspective, I've noticed that many **tweets** (the 140 character messages posted to Twitter) within this network concerned conference information and planning, meeting follow-ups, and logistics. Often, short back and forth discussion may revolve around an article that someone shared². For someone like me, who initially had no idea about these conferences and may not stumble across the articles under discussion in my own reading, the open collaboration and sharing was a window into the actual work and projects of other professionals in this field. This correspondence, open to the public, may have only a year ago happened via email or on closed email lists, or in many cases, not at all. Within the network, increased openness and visibility leads to better collaboration on projects, and openness to people outside of the network leads to opportunities that may have otherwise not been possible.

¹ Note that the factions are based on an algorithm which groups them by maximizing the connection within factions and minimizing connections between factions (See paragraph [Analysis>Subgroups>Factions](#) at http://faculty.ucr.edu/~hanneman/nettext/C4_netdraw.html). It is sorting them strictly by their relation to one another, not at all on their content or relevancy based on retweets or the like.

² www.digitalhumanitiesnow.org is a great example of a recent innovation using the Twitter API and algorithms to highlight articles of importance from the digital humanities community. Digital Humanities Now was created by [Dan Cohen](#), assisted by [Jeremy Boggs](#), and is a production of the [Center for History and New Media](#) at [George Mason University](#).

Figure 1, above, shows the network without screen names, and gives a good look at the nature of the relationships and factions. You can see that the blue nodes have a high rate of reciprocal ties (reciprocal ties are marked with red lines between nodes, non-reciprocal ties are blue lines) suggesting more internal interaction than the red nodes. The yellow nodes are outliers in this analysis, and may not fit in with the red or blue faction.

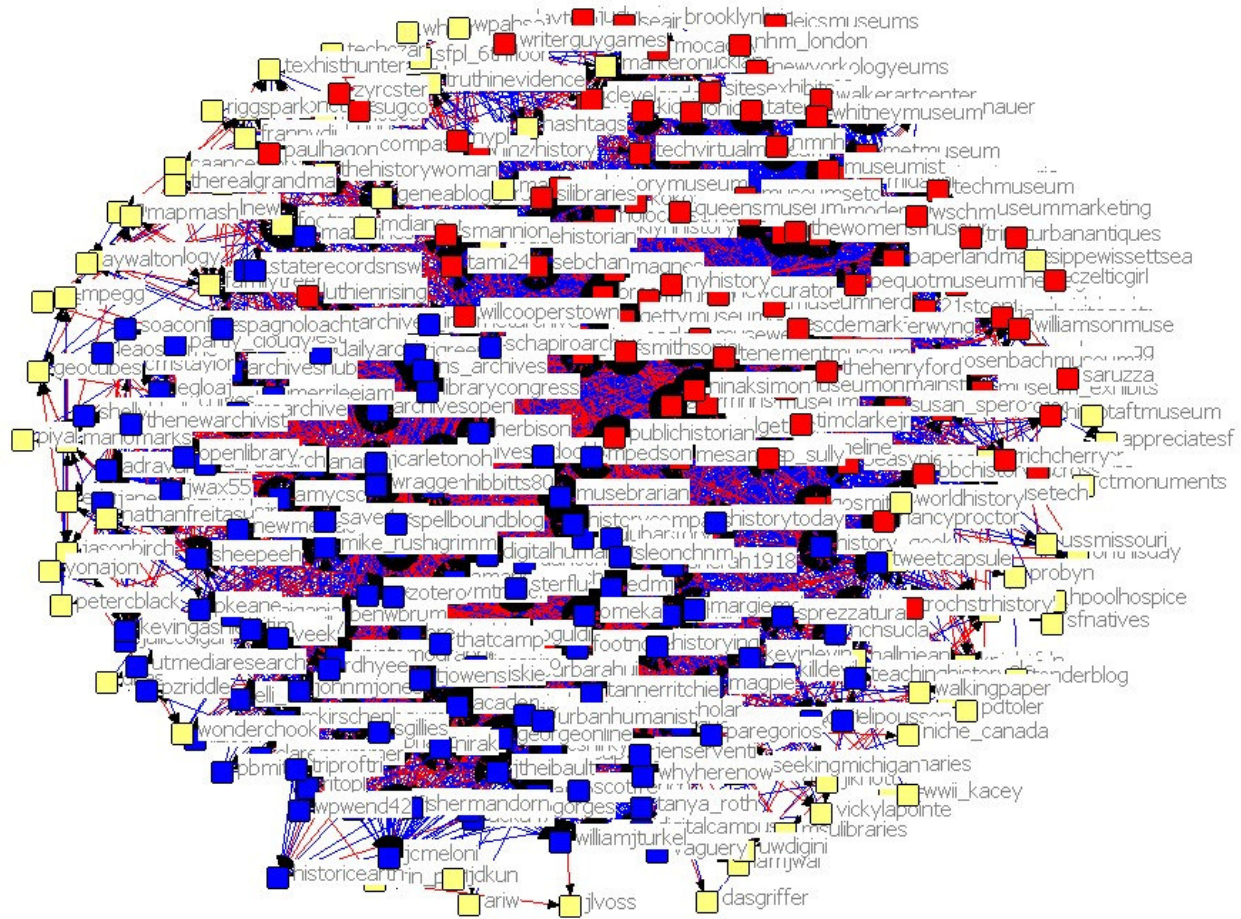


Figure 2. Most popular first and second degree contacts from LookBackMaps, with screen names.

Figure 2 shows us the network with the screen names, which gives me a better idea of what these factions might have in common. Generally, these groups can be broken down into digital humanities (blue), museums2.0 (red), and other (yellow).

The high rate of reciprocal ties is very interesting in the digital humanities faction. Is it because of the interdisciplinary nature of digital humanities, an early adoption of Twitter amongst these academics and professionals, interaction at common conferences, or all of the above? Most likely some combination, but in any case, the reciprocal connections suggests the possibility for open dialogue and discussion via Twitter within this faction of my network.

Conversely, why is there such a lack of reciprocal connections within the museum faction of my network? Zooming in to this sector, I notice that there are many more institutions than people represented here, so we find that many nodes here are followed by many users but are not following at

anywhere near the same rate. It may speak to a Web 1.0 model within a Web 2.0 environment, in which the medium is used more for broadcast of information rather than an interactive experience, at least as far as Twitter is concerned. The individuals that fall into the museum faction do have a stronger presence (and in this mapping, more central presence) and more reciprocal ties than the institutions.

Possibilities for observing other networks

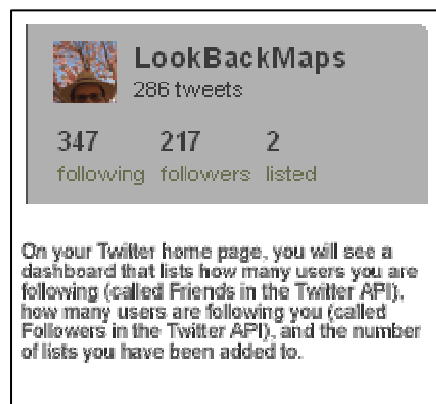
I've been able to explore my own network and the relationships between nodes in a way that would be impossible without the open architecture and access of Twitter and the Twitter API. I believe this gives us an incredible opportunity to observe the evolution of networks and community over time, if we utilize the right tools.

For instance, you could measure the growth and efficiency of networks by monitoring Twitter connections of conference participants before and after a conference. Of course, this could be done with a survey that measures and weighs relationships before and after an event or period of time, but the larger the event, the harder it is to depend on traditional measures such as surveys. An obvious obstacle is that in many cases, you will have many participants that don't use Twitter at all. However, if you can identify Twitter users before a conference and take a snapshot of their network before an event, and then again perhaps a month after, you would have a very telling measure of the impact of that event on networking. Using Twitter could also allow you to include users who are following a conference from afar and communicating with onsite participants through the use of hashtags.

This may not be appropriate for all events, but I think that within the community of digital humanities, where there is a high rate of adoption of communication and social networking technologies like Twitter, it may prove to be a very good measure. THATCamp is a particular conference within the digital humanities that comes to mind as one in which this could be a particularly useful metric.

Future additions to Twitter and the Twitter API may make following networks even more useful, such as the recent addition of **Lists** as a feature. Lists allow any user to create a list of users that they would like to follow, and other users can in turn choose to follow those lists. You could create a list that represents a network you'd like to map over time and compare.

It's also possible that the use of Twitter as a tool for network analysis and visualization may diminish over time as users either stop using the service or add so many Friends that there is too much noise for the data to be very useful. Filtering out Friends with a certain number of Friends is a good way to focus the scope at this point, and I've found that users with a relatively small (200-500) number of Friends returns more useful relationship data.



Collecting the data, and how to do it yourself

Because Twitter offers a public application programming interface (API), I knew that it was possible to build a script to collect the data that I was looking for using scripts run on a website. But first I had to figure out what data I was looking for.

At this point it's necessary to define some of the terms that Twitter uses. Each user has a unique id, which is a number correlating with a **screen name**. These users are the nodes we are interested in collecting, and to make it useful, we need to return screen names, such as @LookBackMaps or @jumpSLIDE or @jonvoss. It's important to note that Twitter, unlike **Facebook**, is an open network by default. Though you can choose to "protect your tweets," and only let people whom you approve follow your tweets, the vast majority of users do not. Consequently, each user can connect to other users, and these are one-way connections, called **Friends** by the Twitter API. The term is a little misleading, as "friend" implies a reciprocal relationship (and in Facebook, that is the case), but it is not necessarily reciprocated. When another user **follows** you, they are counted as a **Follower**, and referred that way in the Twitter API. You can see how this appears in your Twitter dashboard above.

A simplified example is shown in figure 3, where @msulibraries has three Friends and no Followers. @Librarycongress has no Friends and two Followers, and @lookbackmaps has one Friend and one Follower. @internetarchive has two followers (@lookbackmaps and @msulibraries) and one friend (@librarycongress).

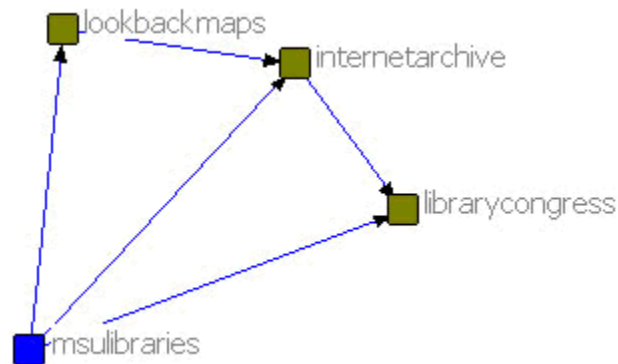


Figure 3. Msulibraries' Friends

Another example, in figure 4, shows four users who are both followed by and following one another.

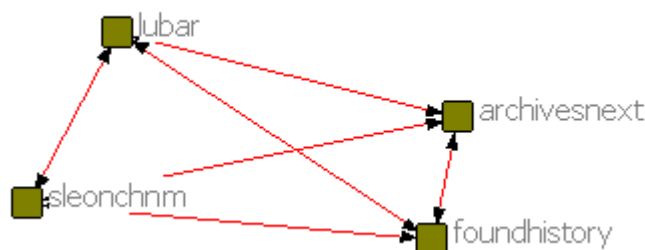


Figure 4. A tight-knit group.

One of the main questions I wanted to answer in my analysis was: who has the most influence in my Twitter network? To do this, I needed to look at the list of users I was following, and in turn, the lists of

users they were following. For my purposes, Followers were irrelevant, as I was looking for who people chose to follow.

I worked with a developer to build a script that would utilize Javascript and the Twitter API to call for a list of my Friends, and then from that list, call for a list of each of my Friends' Friends³. Again, because the Twitter platform is open, as long as a user has not opted to "protect their tweets," which also hides their Friend and Follower data, the data was easily accessible through calls to their API. I had the data returned in a format that would easily import into excel, and in turn into the network visualization software:

lookbackmaps	_Genealogy
lookbackmaps	1968_Project
lookbackmaps	21stCenturyAbe
lookbackmaps	5easypieces
lookbackmaps	AAMers

Figure 5. Data table for import into Excel and NetDraw

I found that I had to do some filtering while I collected this list of users. There are a number of users in Twitter that have enormous numbers of Friends and Followers, and collecting the lists of their Friends not only took a lot of time, but only provided noise to the data. Examples would include @MrTweet, @WeFollow, and @BarackObama. We created an option in the script to exclude users with more than a certain number of Friends, which I kept at 3,000, though I did make some exceptions.

It should be noted that the Twitter API allows up to 150 calls per hour, and will return 100 Friends per call. It does not return an error when the number of calls is exceeded. Also, we found that during business hours of United States Eastern Time, the API calls would often time out without error, presumably due to traffic on the Twitter servers. We set delays in the script calls to avoid these timeouts, and let it run over the course of several days.

Crunching numbers

Once I had all the data I wanted, I imported the data into Excel in the format shown above. I started with 341 Friends on Twitter, or 1st degree relationships. From there, I gathered 82,244 2nd degree relationships, or Friends of Friends⁴.

The picture of my network started to take shape in Excel. The first step was to create a list of all the individual 2nd degree friends, and then to count the frequency that each was named a Friend. The chart looks like this:

³ Please contact me directly to access this script for use, it's free for personal and non-commercial use.

⁴ I had to be sure to use Excel 2007 on my PC (Dell Vostro Core2Duo @2.2GHz, 3GB RAM, XP Pro) in regular mode, not "compatibility mode," (i.e. using xlsx extensions vs. xls extensions) as the latter is limited to 65,536 rows of data, and I was working with 136,424 rows.

FRIEND	Freq	Friend
LookBackMaps	147	341
dancohen	122	251
foundhistory	112	338
publichistorian	109	827
digitalhumanist	105	328
clioweb	87	371
wefollow	84	0
brooklynmuseum	84	558
GettyMuseum	84	0
ninaksimn	83	1146
smithsonian	82	65
MrTweet	78	0
archivesnext	78	403
lubar	77	228
amandafrench	75	212
librarycongress	74	0
spellboundblog	74	301
musweb	72	670
Musebrarian	72	426
lisagrimm	70	372
MuseumModernArt	68	0
anarchivist	68	0
amhistorymuseum	67	317
BarackObama	67	0

Figure 6. An excerpt of my list of 1st and 2nd degree Friends.

This chart begins to answer the question of who has the most influence within my 1st and 2nd degree network. The first column is the screen name, followed by the number of people within my network that have connected to that screen name. The third column shows how many Friends were returned from that screen name when the collecting script was run. Zeros in that column can mean either 1) that user was excluded from the script due to a high number of Friends, or 2) I am not following them. In this excerpt there are examples of both.

As you can imagine, the “long tail” is very much alive in this analysis. Of the 82,585 1st and 2nd degree friends, 80% of them had a follower frequency of 1. 96.9% of the total users had a frequency of 5 or less. The graph in figure 7, below, illustrates this.

Looking at the top 100 or so users by frequency gives me a lot of information. First of all, I can see that my network is largely in the digital humanities and museum sectors. When we plug this data into the network visualization software, that will become even clearer. Another thing that jumps out at me are many users that my network values but I had not otherwise discovered. For example, in figure 6, I can see that @anarchivist is a user that has a frequency of 68. @MuseumModernArt is another user with a high frequency, but when I look at their Twitter page, I can see that they are followed by 47,264 people,

so it's not quite the same thing. Chances are that both are going to have relevant information for me, but the former is likely to be more relevant from a collaborative sense.

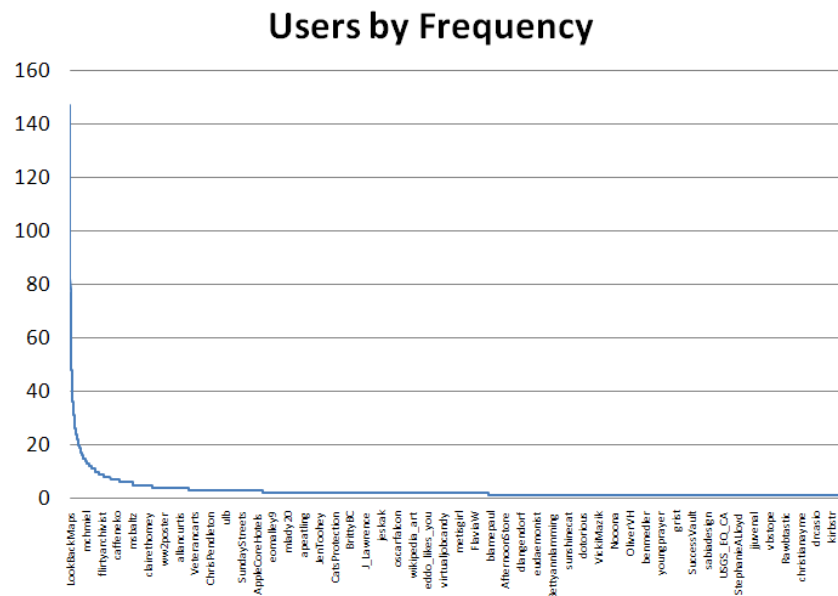


Figure 7. Users (only sampling of screen names appear in x axis) by Frequency, the long tail.

Another thing that I noticed, and can be seen in figure 6, is that Barack Obama is the only political figure in my top 100, so I gather that my network tends to skew liberal from a political standpoint.

Now that I have my top 100 users according to my 1st and 2nd degree network, it's time to pare down the data and see what this network looks like in network visualization software to see if some patterns don't emerge.

Bringing data into NetDraw

For my network visualization, I'll be bringing data into NetDraw, which is free software that works in conjunction with UCINET and can be found here: <http://www.analytictech.com/Netdraw/netdraw.htm>. Because I'm not going to get too fancy with my analysis, I'm going to skip UCINET and just go right into NetDraw and use some of the analysis features built into that program.

I want to look at the structure of my top Twitter users, so I'm going to look at two sets of data. The first is the network of my top 100 users (removing noisy users like @MrTweet, @WeFollow, @BarackObama, etc), the users that follow them, and any links in between. That gets us down to 335 nodes and 8,846 relationships (and 8,846 lines of data to import).


Using formulas, I first find any links to the top 100 users, bring those links into a separate sheet, then create a list of all users who link to the top 100 users, plus those top 100 users. From there, I find any

links between that subset of users to complete the dataset. I add the necessary header for NetDraw⁵, where n= the total number of nodes, and it looks like this:

dl	
n = 335	
labels embedded	
format = edgelist	
data:	
_genealogy	footnote
_genealogy	geneabloggers
_genealogy	lookbackmaps
_genealogy	rjseaver
_genealogy	texhisthunter
_genealogy	worldhistory
1968_project	amhistorymuseur
1968_project	amycsc

Figure 8. Sample of dataset ready for import into NetDraw

This sheet then needs to be saved as a tab-delineated text file. In NetDraw, I go to File>Open>UCINET DL Text File>Network (1 Mode), keep the defaults, and point to the file I just saved.

Without going into too much detail on the NetDraw program, a few modifications to this map makes the visualization really useful. First of all, I want to arrange the nodes in a node repulsion and equal edge length bias layout () , which puts the strongest nodes in the center and arranges weaker ones in concentric circles.

Next, by going to Analysis>Reciprocal ties, I can make reciprocal connections red and non-reciprocal ties blue, which makes it easier to see tight-knit, two-way connections.

⁵ You can find detailed instructions about formatting data for import in the NetDraw manual: <http://www.analytictech.com/Netdraw/NetdrawGuide.doc>.

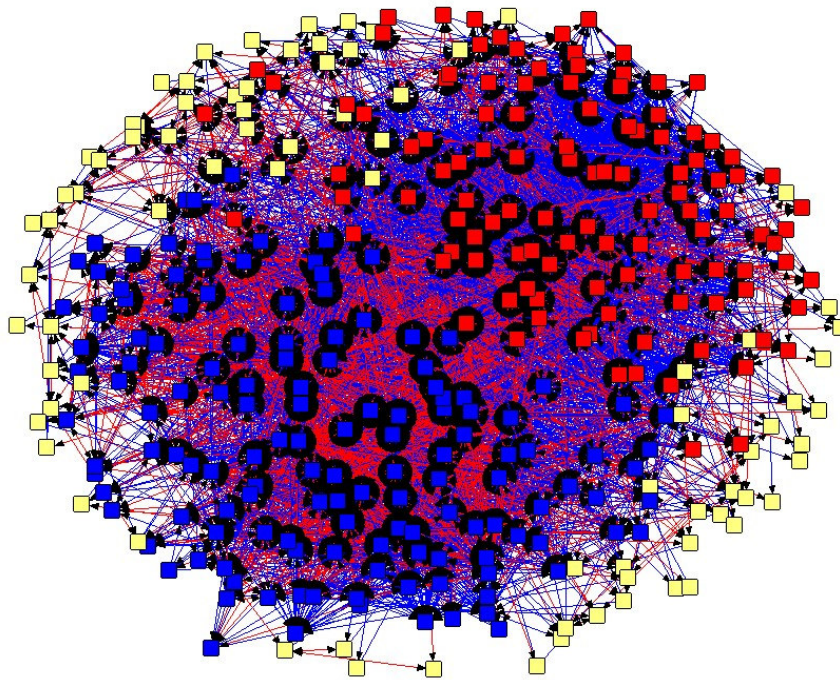



Figure 9. Network of top 96 Friends and their Followers

Finally, I want to break my network into 3 groups or neighborhoods, technically called “factions” in NetDraw. It can take some experimenting to figure out what a good number of factions is, but for my data, I found that there were two essential groups and a third that were sort of periphery. To do this in NetDraw, go to Analysis>Subgroups>Factions and choose the number of factions you want to use.

Your output now looks like figure 10, below, though it’s helpful to turn off the labels (using this button to the top right of the NetDraw program ) , to see the overall patterns more clearly, as can be seen in figure 9.

Of course, there is a lot more you can do with NetDraw to analyze and present your data in a variety of different ways. Hopefully, this is enough to get you started.

About jumpSLIDE networks

Jon Voss started jumpSLIDE networks in 2003 and has been managing IT projects for over a decade. We have worked with a number of non-profits to provide data and network analysis, and training in these systems to perform their own ongoing analysis and comparisons.



Figure 10. Top 96 Friends and their Followers with labels