Visualizing Big Data: Social Network Analysis
By Michael Lieberman
INTRODUCTION AND SCOPE
We live in an interconnected world. What began as a dream of academic researchers has entered the realm of commerce and politics. What began as an experiment run by computer scientists in the IBM labs has morphed into human-computer interactions and social network analysis, and is driven by social media and information visualization.

Social researchers became entrepreneurs, and the phenomenal success of social media platforms such as Twitter, Facebook, Flickr, YouTube, and Wikipedia perceived their output as networks. That is, interconnected actors related to each other. Sociologists introduced the language of social communities, and in academic research laboratories researchers studied motivation, influence, and social roles on the internet.

Many research groups and startup companies have developed tools for social media network analysis and visualization. Today, network metrics has become commonplace and functionality on layout and clustering algorithms has expanded dramatically, producing breakthroughs that have raised the quality of what it is possible to visualize.

Historical Perspective
According to Wikipedia, social network analysis is “the analysis of social networks. Social network analysis views social relationships in terms of network theory, consisting of nodes (representing individual actors within the network) and ties. These networks are often depicted in a social network diagram, where nodes are represented as points and ties are represented as lines.”

Social network analysis has its theoretical roots in the work of early sociologists such as Georg Simmel and Émile Durkheim, who wrote about the importance of studying patterns of relationships that connect social actors. Social scientists have used the concept of "social networks" since early in the 20th century to denote complex sets of relationships between members of social systems at all scales, from interpersonal to international.

In 1954, J. A. Barnes started using the term to describe patterns of ties, encompassing both lay and scientific concepts: bounded groups (e.g., tribes, families) and social categories (e.g., gender, ethnicity). Scholars such as Jacob Moreno have employed the term social network to describe all sorts of social interaction. This usage has expanded in the last 30 years by the use of systematic social network analysis. Social network analysis has found applications in various academic disciplines, as well as important practical applications such as countering money laundering and terrorism. Even corporate organizational structures and classic literature have been analyzed with these tools.
Social Networks

- History: from the dawn of time!
- Theory and method: 1934 ->
- Jacob L. Moreno

![Figure 3.1. Sociogram of a football team](image)

Jacob Moreno’s early social network diagram of positive and negative relationships among members of a football team.


Introduction of Open Source Big Data Synthesize - NodeXL

Marc Smith is a sociologist who specializes in the social organization of online communities and computer-mediated interaction; he and I met at a recent Predictive Analytics conference. Smith leads the Connected Action consulting group in Silicon Valley, California and co-founded the Social Media Research Foundation. Marc has pioneered a revolutionary graphics program, NodeXL, which synthesizes and clusters social network data.

Instead of a complicated listening platform, NodeXL is able to synthesize, for example, Twitter feeds, and produce a relevant graphic and report. NodeXL creates maps that make sense of social media, and that is just the beginning.

Thanks to Marc’s mentoring, I have functionally mastered NodeXL and learned how to make its output relevant to the marketing-research industry.

This paper will present the fundamentals of social network analysis. I will provide sample Twitter and Facebook maps and show how they may be used for understanding brand conversations on the socialmediaosphere. In addition, we will explore marketing applications for hyperlink networks, where a massive potential lies for this technology.
Social Network Analysis
Social network analysis views social relationships in terms of network theory, consisting of nodes (representing individual actors within the network) and ties (which represent relationships between the individuals, such as Facebook friendships, email correspondence, hyperlinks, or Twitter responses). These networks are often depicted in a social network diagram, where nodes are represented as points and ties are represented as lines.

Visualization of social networks is now coming online to make sense of network data and convey the results of analyses. NodeXL is one of the leading, open-source analytic software modules for network visualization. It is free, an Excel add-on, and fairly simple to use (though there is a learning curve). NodeXL deciphers the deluge. A network visualization, shown below, summarizes the software and its user manual.

Network Map of NodeXL Functionality

- Social Network Map of Software Textbook
- Can be applied to Social Network Data
- Can analyze organizational structure/literature
- Can be applied to Marketing Research Visuals

A Marketing Research Example
Supermarket transactions don't jump up and yell: “social network.” So it might come as a surprise to learn how most interactions, be they email-, Facebook-, or even food-purchase-related, can be placed into a network diagram and analyzed.
Network visualization of transactions tells us an enormous amount about how consumers purchase, and can give our client invaluable marketing information on what, for example, to offer on special in order to boost sales of other products as well.

**Social Network of Supermarket Transactions**

Data-mining project results pack the most punch when they are delivered visually. Above is an example of a data-mining output for a major supermarket chain. They want to know which items they should put on special.

When examining the supermarket’s data-mining visual, there are a few things to keep in mind:

- Items are clustered by transactions. Items with the same shade tend to be purchased together.
- The thickness of the red line indicates the number of transactions between its linked items. Heavy red lines, big seller. Dotted connection, not as large.
- The size of the oval represents the ‘between-item connectivity,’ not the transaction rates. This is what we are looking for—product influencers. We’ll explain below.

Examining the graph, we see that there are four ‘clusters.’ Peppers are purchased with shrimp; onions with carrots; broccoli with chicken; and roast beef with mustard.
The managers at the supermarket, however, would like to know which items they need to promote to increase overall sales. Best-selling items are not necessarily the correct answer. Hotdogs, for example, are a popular item. But they are connected mainly to catsup and mustard, and not to other item clusters. The map gives us the valuable—and otherwise unknown—information that lettuce and turkey breast are items which spur growth across all food clusters.

**Sources of Information**

Today’s Big Data mining field is prodigious. More never-ending streams of data are being created every day then were produced for the first four thousand years of human existence.

However, most sources being utilized for the marketing research industry come from a small number of well-known sources.

This is true of Social Network Analysis. In order to filter the brand conversation, the main sources of data are Twitter, Facebook, YouTube, and others shown in the chart above. Each contains one or more social networks.

Social media network maps are a great way to get an overview of your own social media topics and the related ones that matter to you. Maps offer a quick way to assess the type of social structure that has emerged around a topic. These maps make it easy to compare the ways groups in your networks use URLs, hashtags, words, @usernames, and phrases.
Deciphering a Network Graph

A network graph can be summarized by a number of metrics. NodeXL reports these measures for each network. These values can be a useful way to compare networks. Marc Smith, the creator of NodeXL, offers this summary of key metrics.

**Vertices**: the count of unique entities in the network (this is the count of users or accounts)

**Unique Edges**: the count of connected pairs of vertices

**Edges With Duplicates**: the count of connections that are repeated

**Total Edges**: the count of all connections

**Self-Loops**: the count of connections that start and end in the same person (these represent tweets that have no other person’s name in them)

**Reciprocated Vertex Pair Ratio**: the proportion of vertices or users who have a connection returned to them

**Reciprocated Edge Ratio**: the fraction of edges that are matched by an edge in the opposite direction

**Connected Components**: the number of distinct, isolated, sets of connected vertices

**Single-Vertex Connected Components**: the count of isolate vertices, users who do not mention anyone

**Maximum Vertices in a Connected Component**: the count of vertices within the largest set of connected vertices

**Maximum Edges in a Connected Component**: the count of connections within the largest set of connected vertices.

There is quite a bit of information to digest. However, the marketing researcher, the corporate client, and the C-Suite executive need only remember a few of these measures. These are summarized below.
Social network maps are made up of four major factors: Sub-groups. Bridges. Island, and clusters.
Social Network Analysis Using Twitter

Social network analysis is today most commonly used to mine Twitter feeds. Twitter maps have become the most popular and versatile social network graphs for marketing snapshots. These are generally designed to measure a conversation about a hashtag, Twitter account, brand, public policy, or more or less anything that is discussed on Twitter (which, these days, is everything).

There are six basic types of Twitter Social Network Analysis maps. Please Note: In illustration below, "Bazaar" section should read "many medium-sized groups, some isolates"

**6 kinds of Twitter social media networks**

<table>
<thead>
<tr>
<th>Polarized:</th>
<th>In-group:</th>
<th>Brand/Public Topic:</th>
</tr>
</thead>
<tbody>
<tr>
<td>two dense clusters with little interconnection</td>
<td>few disconnected isolates, many connections</td>
<td>many disconnected isolates, some small groups</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Bazaar:</th>
<th>Broadcast:</th>
<th>Support:</th>
</tr>
</thead>
<tbody>
<tr>
<td>many medium sized groups, some isolates</td>
<td>a hub which is retweeted by many disconnected users</td>
<td>a hub which replies to many disconnected users</td>
</tr>
</tbody>
</table>

**Polarized Network:** Most often seen in politics or political issues, this pattern emerges when two groups are split in their opinion on an issue: Two dense clusters with little interconnection. The map shown below was produced during the Arab Spring. It was a map with a commentary on the War on Women. Though the Arab societies might have thrown off their dictatorships in 2011, their deep-seated cultural attitudes towards women have not changed. However, women have begun to speak up, particularly over social media.
The hashtag #monaeltahawy comes from Mona Eltahawy. Mona is a writer and public speaker on feminism and Arab and Muslim issues. The map lets us know in no uncertain terms that not all Arab tweeters agreed with her views of feminism in the Arab world.

In-Group Network: Seen at conferences and among tightly knit groups of people, this type of network rarely ventures outside of its membership. Here’s the In-Group Twitter map for the CASRO National Conference held in October 7-10, 2013 in Westlake Village, California. That is, a map of conference attendees who were tweeting in real time about the conference using the hashtag #CASRO.

From a marketing standpoint, in-group maps are most useful for technical products where users converse over Twitter. In addition, many new platforms that are emerging in the marketing research industry can create hashtags for their products, and then follow comments about those platforms using these twitter maps. These kinds of maps are a must for specialty brands, highly technical products, or specific pharmaceutical products.
The Twitter Brand Map

Below is an example of a Twitter NodeXL Social Network graph. Instead of using listening tools, these Twitter maps organize and visualize content in a way that makes them very easy to interpret.

- We can see who is talking about the brand.
- We can determine the major influencers or connectors and what they are saying. (A connector is someone whose tweets are heard by a large number of people, oftentimes in different ‘clouds’ within the graphic.) If we look at Graph 1, we see an oval that shows several connectors.
- These are people that the brand might want to contact directly through Twitter.
Given that BMW is a major brand, there are a lot of people who are talking about it who are not ‘connected’ to other tweeters. These nodes are referred to as “islands.” Most brands will have a large cluster of islands. On Graph 1, G1 is the large group of dots to the left. These are people not directly connected to the main BMW Twitter account but are discussing topics related to BMW. These are people with whom BMW might want to talk in the future.

The clusters in the BMW map, the multi-colored clouds of names, are people who are in contact with each other either by retweeting or responding. If we look to see which Twitter hashtags cluster in each group we get a pretty good idea of what they are saying about BMW.

What’s exciting about this is that NodeXL allows companies to understand not only what is being discussed in the social networkosphere, but also to identify their most efficient messenger. This allows firms to ferret out prospective customers and identify influencers, and allows brands to test social media campaigns by monitoring NodeXL Twitter maps over specified periods of time.

It would be valuable for a brand to map its Twitter feed for a month. That would give a company a good idea of its brand conversation on Twitter.
**Bazaar Networks:** Bazaar literally means ‘marketplace’. These networks were named for their diverse Twitter clusters. They are most often seen with medium-sized companies or political issues with various levels of community involvement. Currently, a gubernatorial race is brewing in Texas. Of course, Texas is a Red State, and the Democrat candidate, Wendy Davis, has become a lightning rod. This map was produced on January 22, 2014, the day after Davis announced some ‘inconsistencies’ in her bio. As the map shows, there is a bazaar of opinions about the viability of the Davis candidacy.

A table of the NodeXL output, which clusters hashtags and word pairs, shows differing views of Wendy Davis.
Here is the take-home: Using Twitter maps, the clustering of the hashtags often yields the brand conversation, whether it is a bazaar network for a controversial political figure, such as Wendy Davis, a famous brand, like the RitzCarlton, or controversial figures such as Pussy Riot in Russia. While sentiment analysis might ‘dig-down’ into thousands of tweets about a particular subject, Social Network Analysis maps can summarize the Twitter conversation using its clustering capabilities. Moreover, once the map is set, it can be run daily for a month. The reports over that time period reveal the evolving internet conversation.

Broadcast Network: This is another very useful Twitter map, called a broadcast map. In a broadcast map, an individual account dominates the map. Examples might be Lady Gaga, the GOP, a news channel, or the Chicago Bulls.

Forbes publishes a list of the Top 50 Social Influencers. These people many spend a lot of time on Social Networks—they are not celebrities—building the industry of Social Business. Among the Top 50 Forbes Influencers in the Marketing field is Ann Handley, the owner of @marketingprofs. Below is her Twitter Broadcast map.
A look at the different groups reveals that, hashtag clusters differently depending on the social conversation. Below are a few highlights.

- SocialMedia
- SocialSkim
- DigitalMarketing
- Contentmarketing
- Entrepreneurs
- Annuitas
- Business
- MarketingSmarts
- SocialSelling
- SocialListening

Good candidates for Broadcast maps are sites such as @CNNMoney, @YahooNews, or @2morrowknight (Sean Gardner, the #1 Forbes social media influencer). An official football
team’s Twitter feed will show the message the team is sending, and fan reaction. Again, run for a month straight, the broadcast map will provide a good sense of the state of client broadcasts.

Twitter Broadcast maps are also a good measure of a celebrity ‘Brand Equity’. A celebrity that might have once been hot may have cooled off. One can tell easily by scanning up to 18,000 tweets and noting the number of times that celebrity is mentioned. Past celebrities, for example, Sarah Palin or Paris Hilton, made the news when they tweeted. At the time of this writing, however, I would wager that their coverage in the mainstream media would be far less extensive than, say, two years ago. By identifying the number of people tweeting about a subject out of the maximum search, NodeXL can measure the popularity of that subject.

**Support Network:** Think customer support. These types of networks are known to be good at customer service. Organizations such as the UN Global fund provide support for various companies such as Dell Support. Below is an example support network for SAP cloud software. These maps are the least common found in social network analysis.
Social Network Analysis with Facebook

In addition to Twitter, NodeXL analyzes email networks, hyperlinks, FLICKR, and Facebook friends, ‘Likes’ pages, and groups. Below is an example of a Facebook Social Network Analysis map.

**Social Network Analysis and Facebook**

*Michael Lieberman’s Facebook Network*

- G1 are High School Classmates
- G2 Current Friends – Jerusalem
- G3 Family
- G9 Worked at Ristorante Alfredo 1989

For Facebook accounts passwords are required.

Unlike Twitter, Facebook networks are not publically available—a password is required to collect data from a user. Therefore, as an example, I used my own Facebook network. I do not use Facebook for Multivariate Solutions, so my account reflects life clusters. It is instructive to see how NodeXL clusters my Facebook friends.

NodeXL surveys all my friends, and clusters those in my network who are connected to one another. It then maps these people around these shared connections.

Examining my network, we see three dominant clusters. One is my high school class—the reason I originally got on Facebook. The second is my current social circle. The third is family. I
have several other small clusters of friends, but to show how robust NodeXL is, I put a small blue circle around a tiny cluster at the bottom, right-most corner of the map. These are a few people who I worked with at Ristorante Alfredo in New Brunswick, New Jersey in 1986 while I was an undergraduate at Rutgers University.

**Mining Facebook ‘Likes’ Pages.**
Our Social Network Analysis software has the ability to mine Facebook ‘Likes’ pages without a password. For major brands, such as Starbucks, thousands of likes and comments are recorded each day.

### Facebook Like’s Pages - Starbucks

**Comments Made on the Starbucks ‘Likes’**

1/21/2014 – 1/23/2014

<table>
<thead>
<tr>
<th>Graph Metric</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Graph Type</td>
<td>Undirected</td>
</tr>
<tr>
<td>Vertices</td>
<td>11777</td>
</tr>
<tr>
<td>Unique Edges</td>
<td>365085</td>
</tr>
<tr>
<td>Edges With Duplicates</td>
<td>0</td>
</tr>
<tr>
<td>Total Edges</td>
<td>365085</td>
</tr>
</tbody>
</table>

In just two days, more than 360,000 likes and comments were registered on the main Starbucks ‘Likes’ pages. By examining the above graph, we see that nearly 12,000 people visited the site.

This amount of data is best synthesized by platforms that perform Sentiment Analysis; NodeXL would lose the nuance of the comments.

Social Network Analysis visualizations using NodeXL would work wonders, however, with local business Facebook pages. Identification and targeting of local influencers, for instance, a local microbrewery, restaurant, or boutique florist could be performed very effectively. Such a client would simply surrender his or her Facebook password to be analyzed. The amount of data would not overwhelm NodeXL, and local influencers could be easily identified.
Social Network Analysis and the World Wide Web--Hyperlink Networks

The World Wide Web functions as the world's largest network graph wherein web pages link to one another. The tens of billions of pages on the World Wide Web can each be seen as a vertex in a graph whose edges are URL hyperlinks connecting one page to another. There is no larger source of interlinked information.

Professor Robert Ackland of the Australian National University highlights the promise of the web, “These hyperlinks may point to other pages on the same web site or to sites of other organizations, companies, institutions, and nations. The result is literally a web of connections that is often mentioned but rarely seen. The network of related web sites, although based on basic “Web 1.0” technologies, represents important social, economic, and institutional relationships that can reveal significant insights for researchers, policy makers, and corporate strategists.”

Web pages are the untapped gold mine of Big Data information. An entire industry, Search Engine Marketing, has grown up around the late 1990s internet explosion. But who your website is connected to, and who connects to your website, is really the mother lode of network takeaways. With new technology, these links have now come online using free, open-source software.

Collecting and analyzing hyperlink network data allows you to answer questions like: How does your organizational web presence compare with your “offline” brand presence? How does your organizational web presence compare with that of your competitors? Where is the cross-branding taking place?

Hyperlink data can also help you identify new competitors. If you are entering a new market, these data can give you a sense of the current players and their inter-relationships. If your organization is public affairs, what other organizations are linked to it? If you run a non-profit, who links to you? In other words, what is your best fundraising source?

The VOSON Data Provider

The VOSON Data Provider is a NodeXL data import plug-in for the collection and analysis of hyperlink network data. It provides a front-end to a data collection and processing service that is part of the Virtual Observatory for the Study of Online Networks (VOSON). This data collection and processing service consists of a web crawler that crawls “seed sites” nominated by the user, extracting hyperlinks to other sites. The service also finds inbound hyperlinks to a given seed sites.

One would need to set up an account with VOSON, but a free version is available for up to ten databases. Of course, web crawling is nothing new. Spammers have been doing it since the dawn of the web. The beauty of VOSON is that it is a specialized hyperlink search service that works
within the framework of NodeXL. The results, which are delivered in a format much like, for example, Twitter links, makes it easy for non-programmers with a bit of Excel savvy to edit, shape, and ultimately visualize and analyze the data.

Below is a hyperlink social map of our industry. Among the seed sites,

• www.Casro.Org
• www.Kantar.Com
• www.Marketingpower.Org
• www.Marketingresearch.Org
• Yougov.Co.Uk
• www.mvsoluton.Com
• www.Quirks.Com

**Hyperlink Space**

**Marketing Research Industry**

VOSON crawls the web looking for both inbound and outbound links to the seed websites. It is possible to control the number of links one sees. While there are space limitations, once the sites are crawled, the data can be downloaded into Excel in a format that NodeXL can analyze without a great deal of data preparation.
The graph above was filtered, and then various links (among the 4000 or so which emerged from VOSON) were labeled. The results give us a picture of the ‘industry space’ of the limited marketing research websites I asked to crawl.

Next, I decided to try a hyperlink network for an industry that I do a lot of work for, casual dining restaurants. The seed sites were several of the most popular restaurants, such as the Olive Garden, the Cheesecake Factory, Buffalo Wild Wings, and Applebee’s, among a few others.

I have circled some of the seed sites, and labeled some of the more interesting links. What our graph shows us are lists for employment on Craig’s List, local media reviews of restaurants, food blogs, college campuses where specials might have been made for a particular restaurant, and a link between the NHL Detroit Redwings, who co-branded with Applebee’s under the banner of “Applebee's, Your Home for Hockey - Detroit Red Wings - Fan Zone”.

The main advantage of hyperlink network analysis is that it is not as fluid as, say, Twitter. One can run a web crawl for an industry space and be fairly sure that the structure will be intact for a
period of a month or longer. Moreover, hyperlink crawls can be filtered to show only blogs, media, and “Web 2.0” applications.

**Research Challenges for Social Network Analysis**

Personal Twitter networks can only be mined to a limited extent. Currently, the Twitter API (application programming interface) imposes severe limits on this activity. Let's say, for instance, that you are interested in who is following a particular account rather than in what is being tweeted. Twitter imposes strict limits on how many links you may download in a 15-minute time span. In practice, that would mean that downloading a meaningful Twitter account network would take *days*—and often the API will cut off your access in the middle. Efforts to streamline the Twitter API are underway. However, at this time, the limits are still in place.

Additionally a low-tech challenge, attention span, must also be considered. Corporate brand managers and C-suite executives have been slow to embrace this new technology. They glance at a Twitter map and say "That’s nice, what does it mean?" The explanation, though, can take a few minutes.

Researchers are racing to design easily deployable deliverables. When we can explain in one sentence the meaning of a ‘brand conversation’ or an ‘internet industry space,’ then large corporations will adopt the new techniques and add them to their existing arsenal of marketing research applications.

**Conclusion**

We are living in an increasingly saturated world. Facebook, Twitter, Google, smartphones—there is so much marketing noise that even big TV networks and news giants no longer have the ability to guarantee overnight product success.

The result is a growing effort by smart organizations to augment their campaigns through proven media channels with social network analysis. That is, these companies have learned to harness the power of thought leaders, experts, and influencers to promote their products. Social network analysis plays a starring role in the forthcoming drama.
REFERENCES


