**Nate Bachmeier**

Week 2: Explore Data Analysis

March 17th, 2019

TIM-8101 Principals of Computer Science

Northcentral University

Explore Data Analysis

My former manager used to say that “data and telemetry are the life blood of a service.” Without having customer data to route through our algorithms, hardware is simply expensive piles of silicon. As the bits of data flows through our system data analysis is required to convert them into business intelligence and actionable insights.

An example of this can be found with the *Social Circles: Facebook* data set. The researchers collected the social media networks from 10 people and their 4029 associated friends (McAuley & Leskovec, 2012). Raw text files were provided to describe these networks, but only through data exploration methods can one understand: content, short comings, and future applications.

# Explain How Social Media Data Is Created

There are 2.32 billion monthly users on Facebook with 1.52 billion of them connecting daily (Facebook, 2018). To use Facebook, a person, group or corporation registers an account and then specifies some basic profile information. Afterwards the user can post status updates, upload videos and pictures, or “check-in” by posting their location.

In addition to their basic profile information, the user will also specify their social circles. Examples of social circles might be people that worked at Contoso, live in Seattle, or are friends with Bob. Facebook uses these declared social circles to discover relevant content and then presents a dynamic personalized view to the user (Bosworth & Cox, 2006).

The selected content to show is based on the affinity of relationship with the friend. To measure the affinity of this relationship, the number of actions involving both users is counted, where a higher count signifies a higher affinity (Cox & Bosworth, 2006). Actions could include mentions in a post, liking the friend’s content, or tagging a friend in a check-in. In many scenarios the user will proactively declare the relationship to their friends in terms of being family members or husband/wife.

A common scenario where this can be seen is with a user that is part of a social network consisting of three people: a close friend, their mom and a random stranger. The affinity to their mother and close friend should be higher than the stranger, resulting in the dynamic view optimized away from the stranger.

It is in the self-interest of the platform to identify the undeclared relationships so that appropriate third-party content can be shown. In many cases products of interest to a user’s friend will also be of interest to the user themselves. Unfortunately, without accurate context, incorrect content will be shown and degrade the usefulness of the platform.

Consider a cohort of users that make frequent posts about “Java.” A subset of these users will be interested in *O’Reilly books* versus the others want *Starbucks promotional material*. The platform needs to infer which subset is relevant by inspecting the occupations of the user’s friend list (Facebook, 2008).

# Determine Weaknesses and Improvements to the Data Set

The data set provided by Leskovec is an extract of Facebook ego networks, where each of the friend relationships has been labeled. This data was anonymized such that it is possible to know that user 7 and 14 are members of political party 2—but it is not possible to identify the mapping between “political party 2” and the Democratic or Republican party (McAuley & Leskovec, 2012). The researchers then used an unsupervised machine learning algorithm to infer relationship labels between users.

## Challenge: Phantom Transient Property

The research attempts to infer the relationships of two or more users by examining a combination of profile information and their ego graph structure (McAuley & Leskovec, 2012). For instance: if a friend has the same age, home town and high school—then they must be class mates, correct? Well not necessarily.

Perhaps it is the case that Alice was friends with Bob who associated with Charlie. Alice might have never interacted with Charlie. However, Facebook is continuously looking for new relationships and recommended them. Perhaps Alice bulk accepted a group of friend request that happened to include Charlie.

Assuming there are 2.35 billion monthly users averaging 200 connections each; then and a 1% identification error results in 4.7 billion incorrect inferences. Anecdotally, reviewing my personal profile has identified several instances of this transient relationship.

## Challenge: Account Legitimacy

Another challenge with similar data sets is that it assumes all members of the friend list are honest and legitimate entities. There is minimal authentication of credentials or accuracy of published content.

David might have lied about attending a local community college, having a more prestigious title, or volunteering at a charity. These lies aid David in gaining social credits, as he is publicly more desirable. In the context of this research it results in inaccurate inferences between the entities as the underlying data is wrong. To mitigate these challenges, additional features needs to be added to the data set to determine the validity of the claims.

Then consider Eric, an automated script, attempting to gain access to other user’s private information. To gain this access, Eric’s fake profile would likely use the victim’s public information as part of their own profile—such age, home town, and high school. By repeating these same claims, it is more likely the victim will blindly accept the friend request.

## Challenge: Relationship Weightings

The provided data set does not expose weights for the affinity between the user and the friend. This limits the data set to only making general qualitative inferences. That limits the usefulness of the label as we do not know the strength of that relationship.

To partially mitigate this scenario, the researchers attempt to boost the predicted label through a hierarchical labeling model. The more social circle users overlap the more features in common and thus the stronger the relationship. However, this is another variant of the Charlie dilemma.

One method for addressing this problem is count the number of interactions between the two users score the affinity (Cox & Bosworth, 2006). Alice frequently participates in bi-directional conversations with Bob. During the same period Charlie has only liked one image posted by Alice. If this information was included in the dataset then more actionable labels –friend from school vs member of same school -- can be inferred.

Expanding on the efforts of Cox and Bosworth, consider the communications between Frank and identical twins Gavin and Henry. Frank has posted to Gavin’s home page one short reply to ten different posts. Over the same period Frank has posted two active discussions on five different posts. The naïve counting model would suggest that Frank has an equal affinity to either twin. However, a more complex model would argue that the Frank and Henry have a stronger affinity due to the additional effort involved in the conversations.

# Identify what Research is Based on this Data

Social networking information is used in numerous research scenarios ranging from advertising (Berry, 2015) to fraud detection (Tsikerdekis, 2018). One of the more interesting use cases has been to gain insights into protests.

## Insights into Protests

Historically studying social movements, has been challenging as acquiring reliable data is often complex. This is “primarily due to sampling bias, recruitment, response rate and phrasing of questions (Ogan, Giglou, & d'Haenens, 2017).” These challenges are reduced when the social network is consulted instead of one-on-one direct inquiries.

During the Arab Spring of 2012, it was widely reported that Twitter was responsible for enabling the broad communication among the protestors. This was due to unfiltered and anonymous posting, which removed the fear of governmental retaliation (Kassim, 2012) (Wolman, 2013). That caused a significant uptick in response rates which helped journalist and researchers gain the full story. Further the challenges due to the phrasing questions are reduced, as comments originate from those involved in their own words.

# Predict Future Research for this Data

## Credibility Scoring

It is well publicized that the 2016 presidential election was to some extent manipulated by fake social media. The discussion tends to center around foreign state actors publishing fictional stories to discredit or persuade the public. The notion of propaganda in a campaign is not new, however social media enabled the stories to cheaply and efficiently spread across the internet.

The spreading of inaccurate information is caused by a lack of a “credibility score” that measures how trustworthiness of a publisher and their content. Due to the decentralized nature of the social graph, content would likely receive many credibility scores from many entities. A mechanism would then be required to build a weighted net score for the consumer.

In a recent example, the Morning Show reported that that “Apple headphones cause cancer (Yahoo Newsroom, 2019).” If a credibility scores were available by a respected medical journal and a dozen tin foil hat wearing conspiracy theorist-- then a scoring mechanism needs to appropriately skew the score toward the respected journal.

## Social Manipulation

As annoying and dishonest as spreading fake news through social media is, it often creates profit and opportunity. That leads to an incentive for opposition research to continue the cat-and-mouse game.

For example, if a broad credibility scoring system was deployed across the social media network then researches would look to skew the results in their favor. Perhaps they leverage a botnet to down vote the respected journal while upvoting the conspiracy theorist. This allows the researchers to control the narrative and what is largely deemed factual.

# Summarize How the data is Structured and Stored

The value of a social network is unlocked through the ability to traverse the entity relationships. Most modern systems represent that structure as a series of connected graphs versus flat relational tables.

If the network was stored as a collection of flat relational tables, then each relationship traversal would require joining across a global table index. As the network increases in size so would the index, and therefore the scan time to find associated entities.

In contrast a graph approach is natively implemented with each node maintaining a local array of edges that directly point at the neighboring node. By replacing the join operation with a simple pointer dereference; the foreign entity is retrieved significantly faster. This locality of relationships also enables NoSQL solutions like graph databases to support significantly larger entity counts than traditional relational systems.

# Include a Brief Sample Research Area

One of the challenges with exploring highly connected data sets is missing the forest for the trees. This is partially due to the locality of relationships, which forces us to start at a specific point and walk forward. To address these challenges data visualization can be leveraged to gain a broader understanding.

## Environment Configuration

A visualization layer was not provided with Stanford’s Snap.py repository, which led to the need for an open source solution. The industry has generally standardized on Apache TinkerPop for graph storage and Gremlin as the query engine. An ecosystem of tools like Gephi have been created around these open tools.

To get Leskovec’s data set into TinkerPop a short script was created to generate the relevant Gremlin commands. Next Gephi was connected to TinkerPop for visual filtering and inspection of the graph. Running many of the Gephi’s models required significant hardware resources which was addressed through Amazon EC2 Spot Fleet.

## Observation: The Overall Structure

An initial expectation was there to be multiple disconnected Ego networks, as the sampling was intended to be random. However, from the visualization it was immediately apparent that transient relationships exist to connect all networks in the cohort. This would suggest a limited diversity of user data to the original study. If there are not enough examples for the unsupervised algorithm, then it can run into challenges when applied to other markets.

## Observation: Connectivity of Network 0 and 107

The next experiment was to generate the subgraph for the Ego network 0, with a max depth of 2 neighbors. This view showed that there was a strong affinity with network 107. The shortest path between these two networks ended up being an immediate edge of locale\_127. Making similar observations was relatively easy with Gephi’s visual filtering options.

# Figures

All figures were generated with Gephi and the Facebook data set provided by Leskovec (Bastian M., 2009) (McAuley & Leskovec, 2012).

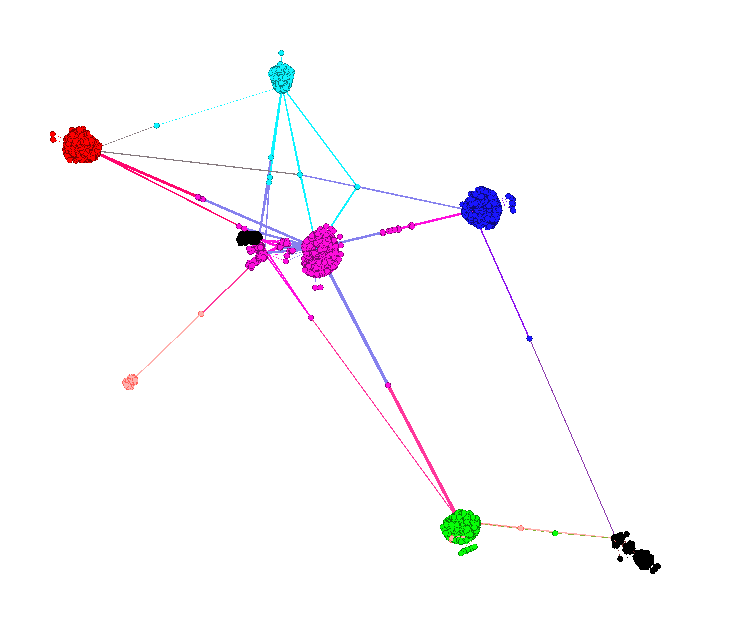


Figure 1: Facebook Network

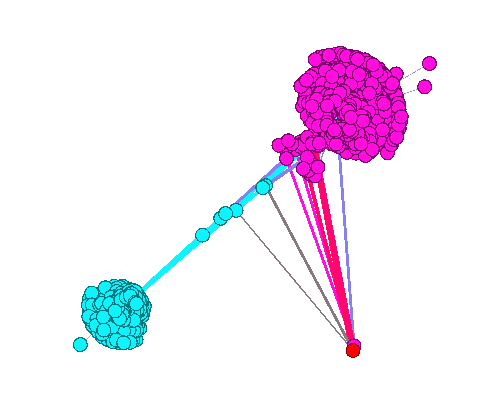


Figure : Network 0 and 107

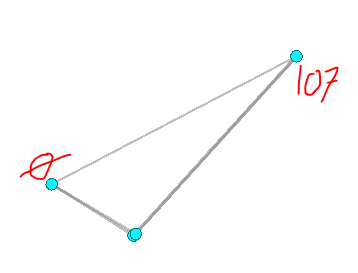


Figure : Shortest Path Between Nodes

# References

Bastian M., H. S. (2009). Gephi: an open source software for exploring and manipulating networks. *International AAAI Conference on Weblogs and Social Media*. AAAI.

Berry, M. (2015, January 21). *TEDx: The future of social media.* Retrieved from YouTube: https://www.youtube.com/watch?v=sm-ZVhC\_yfM

Bosworth, A., & Cox, C. (2006). Systems and methods for generating dynamic relationship-based content personalized for members of a web-based social network. *Patent Number 200800403070*. Retrieved from Patent .

Cox, C., & Bosworth, A. (2006). Systems and methods for measuring user affinity in a social network environment. *Patent Number 20080040475*.

Facebook. (2008). Social Advertisments and Other Information Messages on a Social Networking Website, and Advertising Model for Same. *Patent Number 20090119167*.

Facebook. (2018, December 31). *Company Info.* Retrieved from FB: https://newsroom.fb.com/company-info/

Kassim, S. (2012, July 3). *Twitter Revolution: How the Arab Spring Was Helped by Social Media.* Retrieved from Mic: https://mic.com/articles/10642/twitter-revolution-how-the-arab-spring-was-helped-by-social-media#.nlVQqhvez

McAuley, J., & Leskovec, J. (2012). *Social circles: Faebook.* Retrieved from Stanford Network Analysis Project: http://snap.stanford.edu/data/ego-Facebook.html

Ogan, C., Giglou, R., & d'Haenens, L. (2017). Challenges of conducting survey research related to a social protest movement: Lessons learned from a study of Gezi protests involving the Turkish diaspora in three European countries. *Information Society. Vol. 33 Issue 1*, 1-12.

Rodriguez, M. (2015). The Gremlin Graph Traversal Machine and Language. *ACM Proceedings of the 15th Symposium on Database Programming Languages*, 1-10.

Treadaway, C. &. (2012). *Facebook marketing: An hour a day. .* Indianapolis, Indiana: John Wiley & Sons.

Tsikerdekis, M. (2018). Real-Time Identity-Deception Detection Techniques for Social Media: Optimizations and Challenges. *IEEE Internet Computing IEEE Internet Comput. Internet Computing, IEEE. 22(5)*, 35-45.

Wolman, D. (2013, April 16). *Facebook, Twitter Help the Arab Spring Blossom.* Retrieved from Wired: https://www.wired.com/2013/04/arabspring/

Yahoo Newsroom. (2019, March 16). *Wireless headphones like Apple Airpods 'coud pose cancer risk'.* Retrieved from Yahoo News: https://au.news.yahoo.com/wireless-headphones-like-apple-airpods-pose-cancer-risk-052504674.html