**Nate Bachmeier**

Week 2: Explore Data Analysis

March 17th, 2019

TIM-8101 Principals of Computer Science

Northcentral University

Explore Data Analysis

A former manager used to frequently state that “data and telemetry are the life blood of a service.” Without the actual bits to route through our algorithms, hardware is simply expensive piles of silicon. However, that is not to say having data will magically provide business intelligence. On the contrary one must leverage data analysis methodologies to transform data into actionable insights.

An example of this can be found with McAuley and Leskovec’s “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

As of December 2018; 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. Afterward 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 of Bob. Using these declared social circles, Facebook then finds content that the user’s friends have published (Bosworth & Cox, 2006). This content is dynamically presented as a personalized view.

When selecting which content to show; the frequency of selection is based on the affinity of the friend’s relationship. 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. Not all relationships need to be calculated; in many scenarios the user will proactively declare relationships 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 circle consisting of three people: good friend, their mom and a random stranger. The affinity between the user and their mother would be high due to the declared relationship and the good friend through the number of shared posts. Therefore, when determining which status update should be shown to the user-- a more posts from the mother and friend are desired than the random stranger.

It is in the self-interest of the platform to identify the undeclared relationships so that third party content can be appropriately provided. 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 inters which subset is relevant by inspecting the occupations of the user’s friend list (Facebook, 2008). Having more relevant advertisements results in higher click-through counts gaining the platform more revenue potential.

# Determine Weakness and Improvements to the Data Set

The data set provided by Leskovec is an extract of ten users and the manually labeled relationships of members of their friend list. This data was then 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 relationships between users.

## Challenge: Phantom Transient Property

The research attempts to infer the relationships of two user by examining a combination of profile information and the ego graph structure (McAuley & Leskovec, 2012). For instance: if a friend has the same age, home town and high school—then they must be former 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 right now has identified several Charlies.

## Challenge: Account Legitimacy

Another challenge with this data set 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 with 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 incorrect inferences between the entities.

Meanwhile Eric is an automated script that is attempting to gain access to restricted areas of the Facebook graph. An example could include content with access control lists scoped to the friend group. To become a member of the friend group a fake profile could be created using the victim’s public information—such age, home town, and high school.

## Challenge: Relationship Weightings

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

To partially mitigate this scenario, the researchers attempt to boost the label by determining how many other relationships exist with the same label for exist for given user. However, this is another variant of the Charlie dilemma.

One method for addressing this problem is count the number of impressions between the two nodes (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 the communications between Frank and identical twins Gavin and Henry could be explored. Frank has posted to Gavin’s 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.

**Mitigations and Enhancements**

To address these challenges the data set needs to be enhanced to include additional context and feature information. The context could include

# Identify what Research is Based on this Data

## Insights into Protests

Historically there have been challenges with studying social movements 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 inquiry.

During the Arab Spring of 2012, it was widely report that Twitter was responsible for enabling the broad communication between the protestors. This was due to the 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 of phrasing questions are removed 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 both the publisher and the material itself are trustworthy. Many agencies will attempt to publish credibility scores based on their own agendas. A credibility chain would then need to exist for the reviewers to determine the amount their individual input impacts the net score.

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 conspiracy theorist with tin foil hats-- then a mechanism should exist to weigh all thirteen responses and appropriately bias toward the journal.

# 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 an index. As the network increases in size so would the index, and therefore the required 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 dereferenced pointer; the foreign entity is retrieved significantly faster. This locality of reference information enables graph-based systems to scale to significantly larger entity sizes 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.

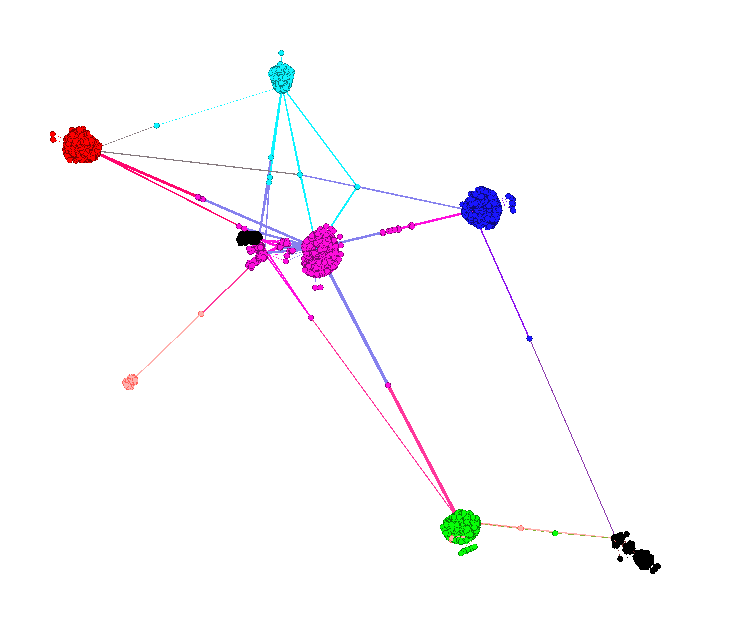
## Environment Configuration

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 broadly standardized on Apache TinkerPop and Gremlin-- as the storage and querying engine. An ecosystem of tools like Gephi have been created around Apache TinkerPop.

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 the use Amazon EC2 Spot Fleet.

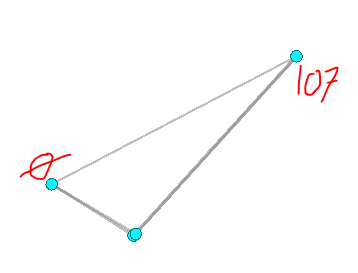
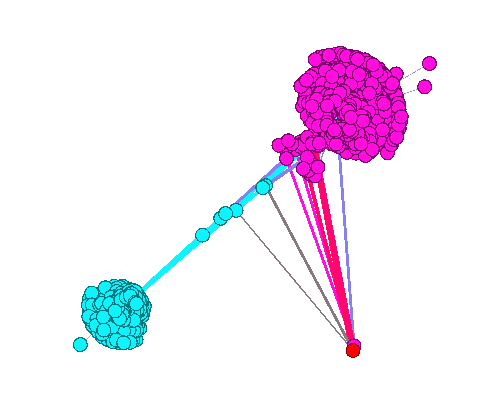
## Observation: The Overall Structure

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 a transient relationship exists 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 due to the visual filtering approach.



## Observation: Confirming Behavior Outside of Visual

The final experiment was to confirm the results of the visualization tool by using the Gremlin query interface. As expected the native queries provide the same data values, which gives confidence in continuing additional research in this direction.

Mar 18, 2019 1:58:45 AM java.util.prefs.FileSystemPreferences$1 run

INFO: Created user preferences directory.

\,,,/

(o o)

-----oOOo-(3)-oOOo-----

plugin activated: tinkerpop.server

plugin activated: tinkerpop.utilities

plugin activated: tinkerpop.tinkergraph

gremlin> graph = TinkerGraph.open();

==>tinkergraph[vertices:0 edges:0]

gremlin> graph.io(graphml()).readGraph('/tmp/Facebook.graphml');

==>null

gremlin> graph

==>tinkergraph[vertices:4031 edges:39368]

gremlin> g = graph.traversal()

==>graphtraversalsource[tinkergraph[vertices:4031 edges:39368], standard]

gremlin> g.V('0').repeat(out()).until(id().is('107').and().simplePath()).path().limit(1)

==>[v[0],v[107]]

gremlin> g.V().has('birthday\_0')

==>v[348]

==>v[414]

gremlin> g.V('348').out().out().hasId('107').path().limit(1)

==>[v[348],v[414],v[107]]