**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 a 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.

**What features could be added?**

## How could those features be included?

# Identify what Research is Based on this Data

## Advertising

Social media is valuable to advertisers as it leads to more targeted ads.

## Fraud

Looking at networks and other graph based data can allow us to easily discover fraudulent relationships.

## Increasing Social Circles

Social media helps grow social groups, such as dating and expanding hobbyist groups.

# Predict Future Uses of Social Media Data

## Use Case 1

## Use Case 2

## Use Case 3

# Summarize how Data is Stored and Structured

## Reverse Indexes

Elastic search / kusto models

## Graph Databases

Neo4J / Neptune or something like that