Network Analysis of Page Likes from Facebook User Profiles

Kyle Brauner[1]; Anton Kocheturov, PhD[1]; Michal Kosinski, PhD[2]; Panos M. Pardalos, PhD[1]
[1]Center for Applied Optimization, University of Florida, Gainesville, FL; [2]University of Cambridge, Cambridge, UK

Abstract

Since the turn of the 21st century, it has been increasingly shown that online social networks such as Facebook and Twitter display similar characteristics to that of real life networks (e.g. the World Wide Web and metabolic networks). Our research aims to show that when the pages that Facebook users prefer, or "like", on the website are represented as a network of nodes connected by the people who like them, the subsequent graph that is constructed displays many of the key attributes of these existing real world networks. We show that the degree distribution of the resulting graph follows a power law, meaning that the graph is representative of a scale-free network, a major property of many small world networks found in real life. Other calculated properties of our network of Facebook pages, including values such as the average path length, diameter, and clustering coefficient, returned consistent results as well. We believe that these properties of the network topology may be useful in further experiments, such as attempting to improve the prediction accuracy of age and gender of Facebook users.

Introduction

Networks are seen all around us. Some classic examples include:

- Protein-protein networks
- Electric power grids
- World Wide Web

Online social networks such as Facebook have become an increasingly popular network type. We take a rather unique approach to analyzing the pages that Facebook users prefer by analyzing a network in which the pages themselves are represented as nodes of a graph that are connected by the people who like them. We provide evidence to support the notion that this network has very similar traits to that of other real life networks and can even be characterized as a small world network. Properties of small world networks include:

- High clustering coefficient
- Low average path length
- Degree distribution that follows a power law

Dataset Description

We obtained our dataset from the myPersonality project, a popular Facebook application that launched in 2007 and allowed users to take a psychometric exam and. With the users' consent, the application also recorded details of their public Facebook profile in the background, including the pages they liked (1). Our massive dataset contained the following:

- Roughly 46.5 million unique pairings of users and pages
- 221,830 distinct users
- 5,556,502 different pages

The graph that we constructed consisted of only the pages that were liked by at least 100 people. This resulted in a graph consisting of 46,137 nodes representing the pages. Then, for every pair of Facebook pages i and j, we determined 3 separate values:

n_i - number of users who liked item i
n_j - number of users who liked item j
$n_{i,j}$ - number of users who like *both* item i and j

We then drew a directed arc in the graph from node i to node j if $\mathbf{n_{i,j}/n_i \geq r}$, where r is a predefined threshold. We obtained results at incremental threshold levels between 0.2 and 0.5.

Data Characteristics

The calculations of a few important network properties at various threshold levels are shown in Table 1 below. These properties are described as follows:

- **Clustering coefficient** - Provides an overall indication of clustering in the network. If you consider a network of friends, where the people are represented by nodes connected via friendship, a high clustering coefficient implies that the friends of any given person are likely to be friends with one another as well.
- **Diameter** - The longest of all the shortest paths in the network. In other words, it is the shortest distance between the two most distant nodes.
- **Average path length** - Defined as the average number of steps it takes to get from one node of the graph to any other node. It is calculated by finding the mean of the shortest path between all pairs of nodes.

The graph analyzed at threshold r = 0.2 most resembles a small world network. Additionally, the degree distribution of this directed graph follows a power law (Figure 1), indicating that our network is also a scale-free network, another characteristic shared by many real networks.

Threshold, r	Size (number of nodes)	Clustering coefficient	Diameter	Average path length
0.2	46111	0.775720342	5	1.989185589
0.25	45948	0.754603673	6	2.028625738
0.3	45092	0.706791341	9	2.126386169
0.35	42979	0.628105279	12	2.317733431
0.4	39209	0.528405363	14	2.613994462
0.45	33394	0.426103506	16	3.007930237
0.5	25271	0.311839965	19	3.480004691

Table 1. Properties of the largest connected component at various threshold levels.

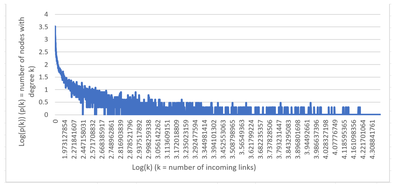

Figure 1. In-degree distribution at threshold r = 0.2.

Concluding Remarks

Small world networks are described as networks that are highly clustered, like regular lattices, have small average path lengths, and are characteristic of a scale-free network (2). As supported by Table 1 and Figure 1, the graph at threshold r = 0.2 most resembles a small world network. The caveat, however, is that the lower the threshold, the weaker the connections are, even though more nodes are involved. As we increase the threshold, less nodes become involved, but there is a greater probability that the *same people* are liking two connected nodes, and so the connections are stronger and thus more valuable. Hence, there is a tradeoff between the number of nodes involved in the network and the strength of the connections.

Future Directions

One major goal of this project is to use properties of this network to predict Facebook user traits such as age, gender, and number of Facebook friends. One approach could involve the following series of steps:

1. Implement clustering analysis to separate the graph into groups, or cliques.
2. For each clique, we could determine the most common traits among the users who liked the pages in the clique.
3. Assign each user to a clique based on the pages they liked on Facebook.
4. Use the most common traits of the clique for prediction.

To do this, we will likely have to find the optimal threshold value to use. Ideally, we would incorporate as many pages as possible into the graph without sacrificing the strength of the connections between them.

Contact Information

Kyle Brauner
Senior, Industrial and Systems Engineering
Email: kbrauner654@ufl.edu
Phone: (518) 281-9896

References

1. Kosinski M, Stillwell D, Graepel Y (2013) Private traits and attributes are predictable from digital records of human behavior. Proceedings of the National Academy of Sciences (PNAS).
2. Watts DJ. Strogatz SH. Collective dynamics of 'small-world' networks. Nature. 1998;393:440–442.
3. Netoff, T.I., Clewley, R., Arno, S., Keck, T., White, J.A. Epilepsy in small-world networks. J Neurosci. 2004;24:8075–8083.