Data Challenge 2

Rajan Bhargava (rb3152), Gaurang Sadekar (gss2147)

December 15, 2016

Contents

1 Introduction

The goal of this exercise is to demonstrate a glass ceiling effect against a set of people in a real world data-set. This can be done in multiple ways - depending on the data-set, the type of minority we are looking at, and how we describe the glass ceiling effect.

There are several social network datasets available on the internet, and most contain either directly or indirectly some type of information for us to be able to separate them into different sets of people. For example, some datasets may contain names from which race can be inferred with reasonably high accuracy. People can also be divided in terms of gender, location or by where they work. The glass ceiling effect can also be defined in multiple ways. Some of the ways people have described it in the past are below -

> *The glass ceiling"... is the unseen, yet unbreak- able barrier that keeps minorities and women from rising to the upper rungs of the corporate ladder, regardless of their qualifications or achievements.* [1]

and a more mathematical one is

> [2] defined the glass ceiling effect as *a result of three factors, combined in a mathematical model (biased preferential attachment model) – homophily, minority-majority partition, preferential attachment (rich get richer). Rigorously, the notion of "wealth" in this network is defined as the number of connections each person has, and was applied to a data set of mentors-mentees, where a connection was created between a mentor and a student if they work together.*

In this report, we discuss a specific dataset, the Google Plus Ego Network, and 2 sets of people that this network demonstrates a glass ceiling effect against. We also define our metric of determining the level/score of a user so that we can quantify the effect of this glass ceiling. We then show that the network exhibits this glass ceiling against 2 different sets of people -

1. Women
2. People not working at Google - The more interesting one :)

2 Data-set and Attributes

We are using the Google+ Ego network data from snap.stanford.edu. This dataset has the following statistics:

Individual Ego Networks	132
Total Nodes	107614
Total Edges	13673453

This data is held in the `gplus.tar.gz` file. Each ego network is represented by 6 files in total. All files associated with a single ego network use the same node ID.

- `nodeID.circles`: Each line has a unique circle identifier, and all the node IDs that belong to that circle.
- `nodeID.featnames`: A list of names and indices of all the binary features given for each node.
- `nodeID.egofeat`: The binary feature vector of the ego whose network these files represent.
- `nodeID.feat`: Each line has a node ID that is a node that this ego follows, along with a binary vector of all its features.
- `nodeID.edges`: Each line defines an edge, either between 2 feats, or between one of the feat nodes and some other node. Basically, this file can include vertices that are at most 2 edges away from the ego.

- `nodeID.followers`: This is a list of nodes that follow the ego, ie. there is an edge from each of these nodes to the ego in the graph.

The total graph made using all of the files is almost 1 GB in size. Since using the entire graph for analysis was too expensive and time consuming, we use a subset of the graph picked by sampling. Out of the 132 egos, we randomly select 24 (20% sample), and build a graph using those files. This results in a graph that contains 92091 nodes and 4555522 edges, which satisfies the requirements of the challenge.

3 Majority - Minority Partition

To investigate the presence of a glass ceiling, we need to define an attribute which we can split the nodes of the graph on. We have investigated the glass ceiling effect using 2 majority - minority partitions:

- Gender: Since almost all the papers in the reference material use gender as the partition, we want to see if Google+ does as well. The gender of a node is identified by the first 3 entries in its vector, with [1 0 0] denoting male, [0 1 0] denoting female, and all other combinations being considered as 'Other', which can mean either that the node's gender is not available or that the node doesn't identify as a male or female. Since nodes on Google+ don't necessarily represent people (they can be organizations or communities as well) and the number of nodes having the gender vector [0 0 1] are too few, we group all these together under the gender label 'Other', which denotes simply that the node is neither male nor female.

- There is a very common joke in tech that you only use Google+ if you are an employee of Google. While this is not true in the absolute sense (there are many more non-Google affiliates than Google affiliates), it is interesting to investigate in the relative sense.
  By identifying users as affiliated to Google or not, we can attempt to see if the metric score for Google affiliates is higher than that for non - Google affiliates in the higher ranges of the metric, which would prove that Google 'employees' are indeed power users of Google+, and non-Googlers aren't.

4 Metric for Detecting Inequality

In the data-set we have chosen, we do not have information about the entire network. We only have information about networks centered around a limited number of vertices, and we later combine them to form 1 large network. One thing that jumps our from a social network like Google Plus is that every profile has a set of followers which are represented as in-edges in this data-set. We can count the number of in-edges for every node, and use that as a feature for our metric, which represents the "popularity" of that node. This directly relates to finding a glass ceiling since popularity is one of the main factors used to determine the existence of a glass ceiling effect.

Google plus as a social network contains much more information about connections than just a list of followers. When a person follows another person's profile, they can choose to put them in specific **circles** of their choice. It could be said that the closer or more related two people are, the more things they will have in common and will be put in a larger number of circles than people who are not as close. This represents the strength of the connection. Thus, we decided to use this as another feature in our metric. We decided that there are 2 different things at play here - the number of cirlces that a person is put in, and the sum of fractions of circles they are put in.

We now have 3 features on which we can base our metric

1. Number of followers - Number of in-edges of a node in the network.
2. Total number of circles the user is in - Sum of the number of circles other people have put this user in.
3. Total fraction of circles the user is in - Sum of the fractions of circles of each follower this user is in.

For example, if node A has 2 followers B and C. B has 5 circles and includes A in 3 of them, while C has 8 and includes A in 2. The feature values of A in this case are -

1. Number of followers - 2
2. Total number of circles the user is in - $3 + 2 = 5$
3. Total fraction of circles the user is in - $\frac{3}{5} + \frac{2}{8} = \frac{17}{20}$

We look at each feature independently and check it demonstrates the glass ceiling effect. The total score for a node is a linear combination of these feature values, the weights for which we decide by analyzing the feature value table. This, along with other results is described in the next section.

5 Results

We looked at each feature separately, and noticed that the average score for females for every feature was greater than the average score for males for that feature. We noticed the same for the other division of users, where non-googlers had a lower score. This is a good indication that there might be a glass ceiling effect in the dataset.

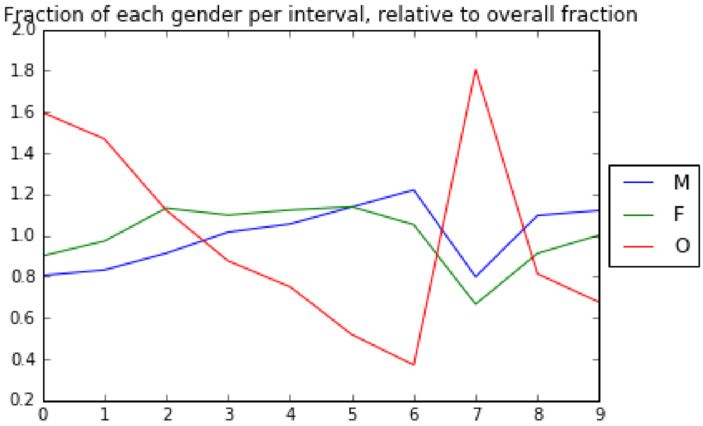

Figure 1: Fraction of each gender relative to original fraction for increasing score percentile intervals

We tuned feature weights to 0.01, 0.39 and 0.6 based on the values that the features took so that we weigh features equally, and such that we can easily observe the glass ceiling effect. We calculated a score for each node (user) in the graph, sort them based on these and divide them by percentiles in differences of 10. We then looked at the fraction of males in each percentile, and divided it by the fraction of males overall. We repeated the same for females, googlers and non-googlers and plot these in Figure 1 and Figure 2. The formula for the Y-axis is -

$$\frac{\left(\frac{\#categorynodesinpercentileset}{\#nodesinpercentileset}\right)}{\left(\frac{\#categorynodesoverall}{\#nodesoverall}\right)}$$

We can see in these figures that as we move to higher scores, the fraction of men in that percentile tends to increase, and the difference is the biggest in the highest bracket, which means that the dataset does exhibit a glass ceiling effect against women with respect to our metric of evaluation.

Another interesting observation, and a much clearer example of the glass ceiling effect is the other sample case in Figure 2. We see here that while the fraction of non-googlers remains more or less constant, the fraction of googlers steadily increases which increase in score percentile. At the highest level, the different is by far the largest, indicating that this effect is clearly present in this network against non-googlers. Thus, by our metric, majority of top scoring users on Google Plus are people at Google, giving a justification to the joke we mentioned earlier.

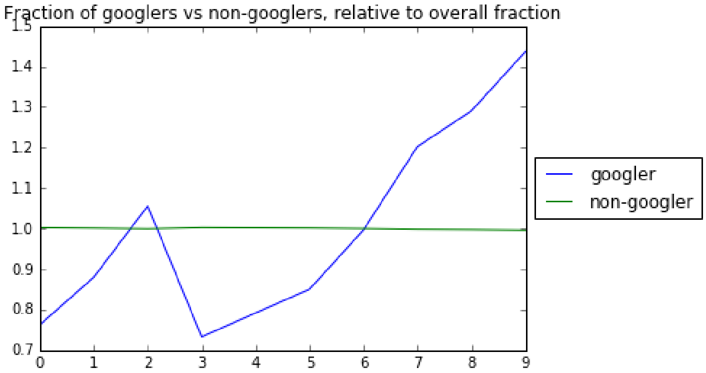

Figure 2: Fraction of each type relative to original fraction for increasing score percentile intervals

6 References

1. Federal Glass Ceiling Commission. "Solid investments: Making full use of the nation's human capital. US Government, Department of Labor. Washington, DC: US Government Printing Office. Retrieved February 2013." (1995).

2. Avin, Chen, et al. "Homophily and the glass ceiling effect in social networks." Proceedings of the 2015 Conference on Innovations in Theoretical Computer Science. ACM, 2015.