Quick Analyses Over A Social Network

Yekta Demirci
Electrical and Computer Engineering
University of Waterloo
Waterloo, ON , Canada
ydemirci@uwaterloo.ca

***Abstract*—This report demonstrates the outcomes of a small project where a social network was analyzed by stating two different hypotheses. Firstly, number of friendship distribution was investigated to claim "Number of friendships in a social network do not vary much between the individuals". In order to support the claim, basic statistical tools were used. Then, "People whom have high number of friends tend to connect different communities" was questioned. A modified version of BFS algorithm was implemented to find betweenness of edges to test this claim.**

***Index Terms*—social network, vertex degree, BFS, betweenness**

I. Introduction

Engineering can be considered as the art of modelling real world problems in an abstract way and solving them by using the tools provided by the models adopted. Networks are very typical example of such modellings where individual components are represented by nodes(vertices) and the relations between the individuals are shown by the links(edges). By observing the properties of actual systems and the characteristics of models, networks are divided into four loose categories: social networks, information networks, technological networks, and biological networks[1].

In this report, a social network dataset [2] provided by Jure Leskovec is used to test the hypotheses given in the abstract. The network has 4039 nodes which model facebook accounts(people) and 88234 undirected links which represent the friendship between the individuals. This data is originally used for the mathematical models explained in the paper, "Discovering Social Circles in Ego Networks". The aim was to automatically organize people's personal social networks.

II. Hypothesis I

A. Introduction

In this section, the first claim: "Number of friendships in a social network do not vary much between the individuals" is investigated. At first glance, this problem may seem like it mostly concerns social sciences. However, such claims play crucial roles also in technical works. For instances, in the literature, there are several approaches to estimate size of networks over time[3]. Based on some assumptions, several algorithms were developed to simulate establishing and growing phases of networks. One of the premises in such works[4] was "There is an upper limit on the number of friendships an individual can maintain". It may seem like a basic assumption, yet complex mathematical models were created upon this and a few more premises. Premises are clearly important since if the premises do not reflect the reality in the first place, then the future work will be pointless no matter what.

B. Methodology

In our approach, number of friends of each individual is found. This is quite trivial, degree of each vertex actually tells the number of friends.

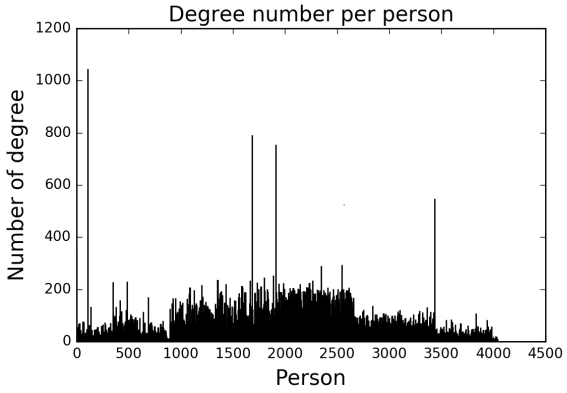

Fig. 1. Number of friends per each individual

As it can be seen in figure 1, there are 4 spikes. A person with 1045 friends, others with 792, 755 and 542(top 4). The rest of the people have less than 347 friends. However this figure does not tell much. Mean and variance distributions should be investigated.

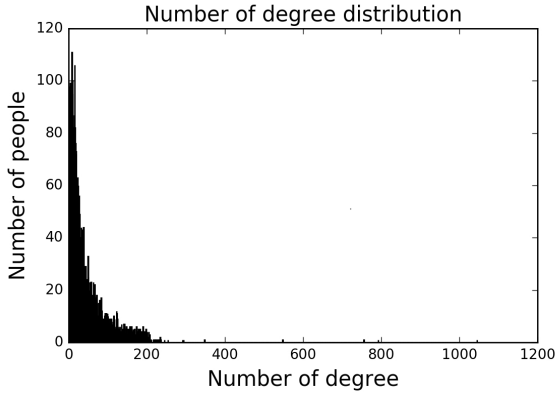

Fig. 2. Distribution of degrees

In figure 2, number of degrees are given in the horizontal axis and the number of people with the regarding degrees are given in the vertical axis. The distribution is skewed right and it reminds Poisson distribution. In average, individuals have 43.7 friends, and the standard deviation is 52.4.

C. Results

Coefficient of variation is 1,2 which is slightly more than 1. Therefore, the variance can be considered as medium-high. The *Hypothesis I* does not seem like a strong claim, since the variance of friendship is not low. Yet, it is wise to not generalize the outcome just by checking a single data-set.

III. Hypothesis II

A. Introduction

The second hypothesis, "People whom have high number of friends tend to connect different communities" tries to find whether *importance* of a person's relations are related with his/her number of friends or not. It is good to explain what is a "community" and *importance*. Community structures are groups which have higher density of edges between them. [5]. The *importance* in this context reflect the status of connecting different communities and not having much alternatives. In a real life example, if there is only a single pair of people that supplies a connection between community "A" and community "B" and if there is not such any other pairs, then these two people have important relations.

B. Methodology

Finding communities in the network may seem like the first step to investigate the given hypothesis. However, in our approach, individual communities are not found. Because, even there has been several published algorithms to partition a network into communities, it is a computationally demanding job.[6]. Instead, a concept called *shortest-path betweenness* is used. It is a divisive clustering algorithm. Basically, this algorithm favors the edges between communities and assign higher values to them and it disvalues the edges within communities.[7] For instance, assume we have two triangles: the first one has vertices "A", "B", "C" and the second one has vertices "X", "Y", "Z". If the two triangles are only connected with a single edge call $< A, X >$ then this edge will be assigned a higher value compared to other edges within the triangles. If there would be two edges that link the triangles, then the betweenness values of these edges will be lower than the single linker, $< A, X >$ case. In other words, the betweenness value gets higher if the edge does not have much alternatives to cross between the communities.

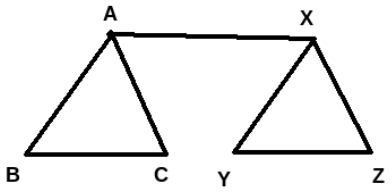

Fig. 3. A basic example to explain betweenness

Firstly, betweenness values of each edges are found. Then the vertices are assigned sum of their edge betweenness. Finally, each vertex betweenness is divided by its degree number to find average value. Otherwise, obviously, the people with highest friends would have higher scores, however we are looking forward to find values in terms of quality rather than quantity.

C. Results

In figure 4 the betweenness values are given per person. Clearly the person whom has the most friends(person with index 107, whom has 1045 friends in figure 1) has the highest betweenness value.

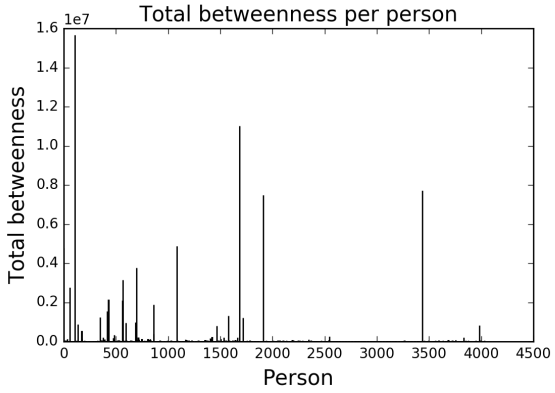

Fig. 4. Betweenness values of each person

When the betweenness values are divided by the degrees, the results are quite interesting. Top three most *important* people are 860, 58 and 594 index wise(Fig. 5),each having 2, 12, 8 friends respectively (Fig. 1). In brief, the person with only 2 friends had the highest *importance* score.The people with the most friends are ranked 12th, 13th, 17th having 1045, 792, 755 friends respectively. As a result, having many friends does not reflect high *importance*.

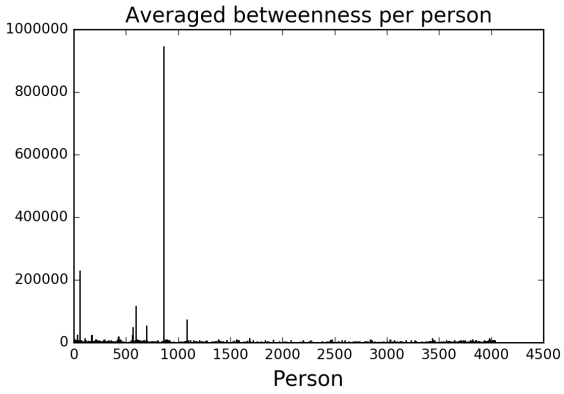

Fig. 5. Average betweenness values of each person

IV. Summary

To sum up, a social network was investigated by claiming two different hypothesis. Among the other data-sets, a social network is chosen since the outcomes could be interpreted better. A quick literature search was done to understand the overall concept. Most of the read papers are authored by Newman or Leskovec.

A link to the codes written to check the hypothesis are given in *Appendix*. For the first one, the code is trivial. However for the second one, the algorithm is implemented according to the "Shortest-path betweenness" description given by Newman. Although the concept is not that complex, implementation contains a few tricks. Several smaller graphs were tested to check the algorithm. It is basically two BFS searches from the source vertex to leaves, and from leaves to the source. "BFS" function given in the implementation is $O(|E|)$ where $|E|$ is the total edge number. Then this algorithm runs for each vertex to find overall 'betweenness'. As a result, total time complexity is $O(|E|*|V|)$ where $|V|$ is the total vertex number. The algorithm took 11 hours on my personal computer to run. The time could be shortened by using multi-threading since each vertex is used as source, independently.

Some other hypotheses could be questioned using community clustering, density of communities, graph growing , etc however it takes quite time to implement, test and run the algorithms over the large network.

Beside, I did not apply the methods used in my previous research like K-means clustering. K-means clustering takes multi-dimensional input. However in networks, there are nodes and links only. I did not take the risk of mapping nodes-links to a multi-dimensional data, since the outcome of clustering heavily depends on the mapping. Also, using K-means clustering was found improper since it ignores the graph structure[8]

V. Appendix

The written codes can be accessed by the given link below.
https://github.com/YektaDemirci/smallProject

The citated papers can be accessed by the giben link below.
https://drive.google.com/open?id=1zOhxYsGUrAKdRc7WNeD3Mk4ghcqmXvsR

References

[1] M. E. J. Newman, "The Structure and Function of Complex Networks" Society for Industrial and Applied Mathematics, 2003, vol. 45, p. 174.
[2] J. Leskovec, "Social circles: Facebook", Accessed on: Dec. 23 2019. [Online]. Available: http://snap.stanford.edu/data/ego-Facebook.html
[3] J. Leskovec, J. Kleinberg, C. Faloutsos, "Graphs over Time: Densification Laws, Shrinking Diameters and Possible Explanations", Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining, Aug. 21 2005.
[4] E. M. Jin, M. Girvan, M. E. J. Newman,"Structure of growing social networks", Physical Review E, Sep. 26 2001, vol. 64.
[5] A. Clauset, M. E. J. Newman, C. Moore, "Finding community structure in very large networks", Physical review E, 2004, vol. 70.
[6] M. E. J. Newman, "Fast algorithm for detecting community structure in networks", Physical review E, 2004, vol. 69.
[7] M. E. J. Newman, M. Girvan, "Finding and evaluating community structure in networks", Physical review E, 2004, vol. 69.
[8] J. McAuley, J. Leskovec, "Discovering Social Circles in Ego Networks", ACM Transactions on Knowledge Discovery from Data (TKDD), 2014 vol. 8 p. 16.