Management Summary For

MGT-415: Problem Set 2

Do Carmo Maria
Fan Xi
Genc Murat
Nyambuu Lkham
Trichilo Giulio

April 7, 2019

EPFL

Introductory Information about the Network

In this report we provide some introductory information as well as basic network measures related to the "Telco" network dataset, which provides a node-to-node relationship of SMS messages sent by customers of the telecom provider.

The network consists of 4039 unique customers, which represent the nodes, among which 88234 SMS messages have been sent at least once (a link is present if user A has sent user B at least one sms). The network is undirected and unweighted, and no external variables were provided. However, a relatively rich analysis can be performed nonetheless.

Indeed, this report will begin by outlining the basic empirical features which characterize the network. We then attempt to detect the most influential nodes in the network (the 'viral' users) through centrality measures. Once identified, we perform community detection on the network in order to gain a deeper understanding of the network topology.

1. Network Statistics: Degree Distribution and Clustering

The network's degree distribution is initially plotted, both on a linear and on a log scale. We notice the distribution looks quasi-linear on a log scale, implying it may be distributed according to a power law: a vast majority of nodes have sent a relatively small amount of sms messages (and are hence not very well connected to the entire customer base, 43.7 connections on average) whereas a very few nodes are connected to a large part of the network, the node with the highest degree has 1045 connections: almost 1/4 of the entire network!

In order to investigate the powerlaw property, indicating the network is scale-free, Exponential, Lognormal, and powerlaw distributions were compared through distribution distance. The best fit for the network would indeed appear to be a powerlaw distribution (and indeed the R test statistic is 0.18 in favor of the powerlaw), however we do notice that true TCDF does diverge from a theoretical powerlaw for a large part of the middle section, before descending again towards the end. This implies the network is not truly scale free, and some constraints in terms of average number of neighbours exist either due to the underlying telecommunications structure, or more broadly due to the general make-up of the customer base. Indeed, having more or less active overall customers would change what the tail CDF looks like and hence how close the network really is to being scale free.

With regards to clustering, we see that the network exhibits a strong high clustering property, and to a lesser extent a small world property. Indeed, the network is sparse: density is only 0.01, however average clustering is rather high at 0.6. The clustering to density ratio is a very high 55.6 times: this implies that this real world network is 55x more dense than you would expect if the network were randomly generated. This underlines the importance of maintaining active connections in a telecommunications network, as higher consumption is intrinsically linked with higher revenue.

The small world property can be seen from the fact that the diameter is only 8 links, and the average shortest path length is less than half: it takes on average only four people so that the network diffuses information across large distances, and the maximum shortest distance it takes to get information from one user to another is 8.

Now that we have discussed properties of the network as a whole, centrality measures will help determine which nodes are those which really influence the network behavior.

2. Centrality Measures

We begin this section by finding out which nodes have the top 10 highest degrees. As mentioned, the largest node is connected to 1045 others, and this tails off relatively quickly with the 10th most connected person having only 235 connections. Unsurprisingly, the top 10 is still much above than the average which lies at 43.7, which is arguably much lower once the influence of these high degree nodes is removed.

At this point, centrality measures are computed on the network in order to verify whether the intuitive measure of importance, a high degree, agrees with what centrality measures consider important. In particular we note that degree centrality, betweenness centrality and closeness centrality, and to a lesser extent pagerank agree, whereas eigenvector and katz centrality do not. Betweenness and closeness centrality are both measures which attribute high importance to nodes that are connected to high degree nodes, and which serve as 'connection hubs' for the other nodes in the network. Pagerank attributes a higher score to those nodes on which there is a higher probability that a random walker with stochastic death will end up on. Eigenvector and katz centrality on the other hand attribute importance to a node when others around it have a high measure of the same centrality.

The fact that the first group plus pagerank agrees while the two "neighbor" centralities don't is a clear signal that being important in this network means having a high degree as well as being connected to other high degree nodes. This sort of 'popularity' mechanism is rather typical in telecommunication networks where the big players can diffuse information through the network at high volume.

3. Community Detection

In this section the Louvain algorithm is used in order to detect communities. The random seed parameter had to be fixed as the algorithm's stochastic component would sometimes generate extra (or a couple extra) communities with a very small number of nodes, which is nonsensical. The resolution parameter was also adjusted as a result of this section's analysis, to ensure that communities were balanced in relative terms. This is not due to wanting to induce some prior belief on the number of communities, but rather that when the algorithm's output resulted in a pair communities with identical composition half the size of all the others there was no need to consider them as individual communities.

The community detection algorithm finds 14 communities with a few hundred members on average. This implies that the telecom provider is faced with a set of relatively homogeneous subgroups in its customer base. For each community, we can calculate correlation. We find that about half the communities are assortative, whereas the remaining half is dissortative. An assortative community means important members are connected to important members more frequently than to less important nodes, whereas dissortative communities imply high degree nodes act as hubs or connection points for many other low degree nodes. This implies that even though the sizes of communities are relatively homogeneous, the telecom company might have

to cater differently to assortative and dissortative communities, given the disparate message structure.

By inspecting the induced community graph we see that there are a few communities which are very strongly interlinked, these may constitute the core of the network, but a majority of these is actually rather disparate and on its own terms. We display the degree centrality score (the top 10 nodes) in each community. Unsurprisingly, we find that the top 10 nodes in the network belong to the most dense communities. This means that the telecom provider will be able to reach the most immediate members relatively easily, however when it comes to "distant" communities perhaps extra effort is required, and it should therefore be evaluated whether the extra effort will be worth the implied cost of maintaining distant customers.

4. Node Selection Algorithm

Based on the analysis and definition above, an algorithm that is able to find out the group of target to whom the telecom company should deliver the message in order to achieve maximum effect of advertising with minimum cost. In such an algorithm, cost can be defined as the sum of square each of the target group member's neighbors, and other alternatives can be the target group members amount, or target group members sum of path, etc. And the effect of advertising can be defined as the ratio of neighboring nodes to the whole set of nodes, while other alternatives can be more complicated using discount factors and path length.

Nodes are selected mainly based on the centrality, which means that the nodes that has top centrality should be selected since they shows broader connection inside the network. Considering the cost, an algorithm should be developed to optimize the combination of cost and effect of advertising, using the top centrality nodes as a base. Nonetheless, reaching every member of the community can be costly especially in assortative networks considering the nodes with a low degree; so different campaign strategies can be developed for assortative and dissortative communities.