MGT-415 Executive Summary

Authors: Sascha Frey, Kuber Malhorta

Using Networks to find influencers amongst customers

Introduction

The aim of this project is to identify key individuals in the customer base that can be targeted to quickly and efficiently spread a positive PR message. These key individuals, which we will call “Influencers” should be connected to many individuals in different communities.

The data used to identify these individuals shows the SMS connections between customers, it will be interpreted as a network. A number of analyses will be applied to identify the most promising influencers.

Preliminary Data Analysis

The data consists of 4’039 customers and 88’234 connections between them. Creating an adjacency matrix and visualizing it (Figure 1) already enables us to make a number of hypotheses.

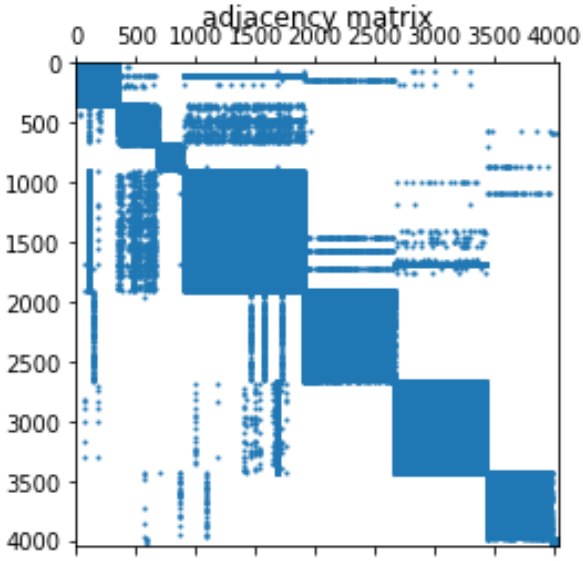

Figure 1: Adjacency Matrix. A coloured block implies a connection between two individuals.

Firstly, one can intuitively identify a number of “communities” in which all individuals are connected to each other. This is visible in the solid blocks along the leading diagonal. We further identify a number of individuals who seem to also have connections outside of their communities, these individuals are potentially the influencers we are looking for. We check if the network is connected, meaning that any two individuals (or nodes) are connected through one or more edges. Whilst intuitively it does not seem to be due to the isolated communities, we find the network to be connected. This is likely due to the cross-community individuals connecting different communities.

Connectivity Analysis

We use a series of measures to look at the importance of each node in terms of its "connectedness". A basic approach is to look at how many direct connections a node has, this is called the degree of the node. Looking at the distribution of the degree of the nodes shows that an overwhelming majority has only a single degree. The distribution is approximately exponential, with a median of 25. There are a few nodes with a very high degree, notably 8 nodes with a degree greater than 250. These are individuals who are extremely well connected and could as such be interesting targets for an advertisement campaign.

Another measure of the importance of a node is it's betweenness measure, this normalized measure looks at how many times a node appears on the shortest path between two nodes. High betweenness implies that a node is an important in connecting the network. We notice that almost all nodes have betweenness zero or very close to zero. To be more explicit, the maximum betweenness is 0.48 and Table 1 shows the number of nodes with betweenness values greater than a given constant. These nodes are very important for the connectivity of the network and should be targeted as influencers.

Table 1: Number of nodes with a betweenness value greater than a given constant.

Betweenness Greater Than:	# of Nodes
0.1	7
0.01	24
0.001	96

An interesting insight is obtained when looking at a further measure, the closeness. This indicator shows how close a node is to the remaining nodes in the network, a shorter average path length induces a larger closeness. The closeness is far more evenly distributed, as is visible in Figure 2, implying that all nodes are equally close but are perhaps connected by only a few individuals. An adequate analogy would be bridges connecting communities either side of a river.

In order to explicitly choose influencers, we choose the 10 nodes with the highest measure for each of the three discussed measures. As expected, there is some overlap and we identify 21 "Influencers" whose customer IDs are given in Annex A.

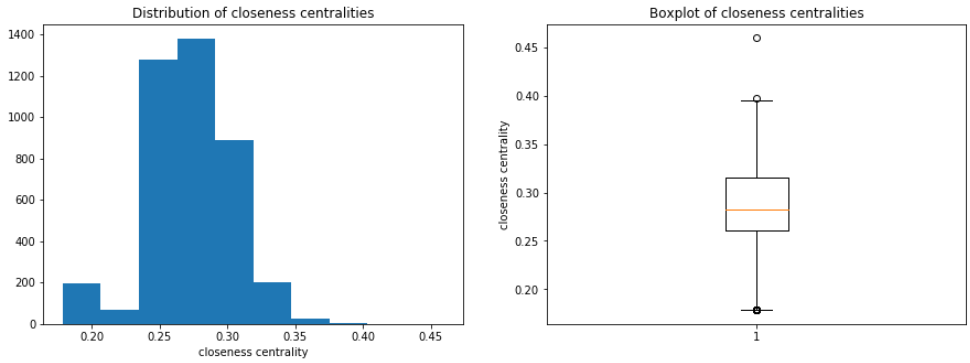

Figure 2: Closeness centrality distribution and boxplot. We see an even distribution which is possibly gaussian.

Analysing the shortest path length justifies the idea of using these influencers, as on average only 3.5 nodes are required to connect any two individual customers. A large number of customers are therefore already reached if only the "Influencers" send the message to their contacts.

Clustering

Another method to identify "influencers" would be to cluster the points according to the three measures (Degree, Betweenness and Closeness). We use a k-means clustering algorithm and apply different penalties for using a higher number of clusters. We find the optimal number to be K=6, as seen in Figure 3.

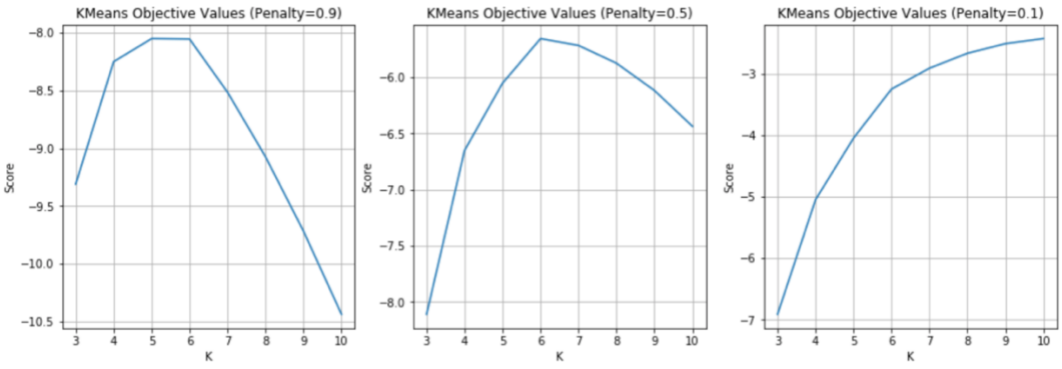

Figure 3 Objective values for different penalties. We see that the optimal value for K is 6.

We notice that one of the clusters created has very high values for all measures, these "important" nodes have a customer ID of 107, 1684, 1912 and 3437. This is a subset of the 21 "Influencers" mentioned above.

Finally, using the fast-greedy algorithm, we identify a number of communities (13 to be exact) and plot them with different colours, further identifying the 21 (resp. 4) influencers in their communities. This visual solution is given as an annex.

We recommend that our company target the 21 or 4 individuals identified, depending on the resources available for the advertising campaign.

Annex A-Visual representation of communities and influencers

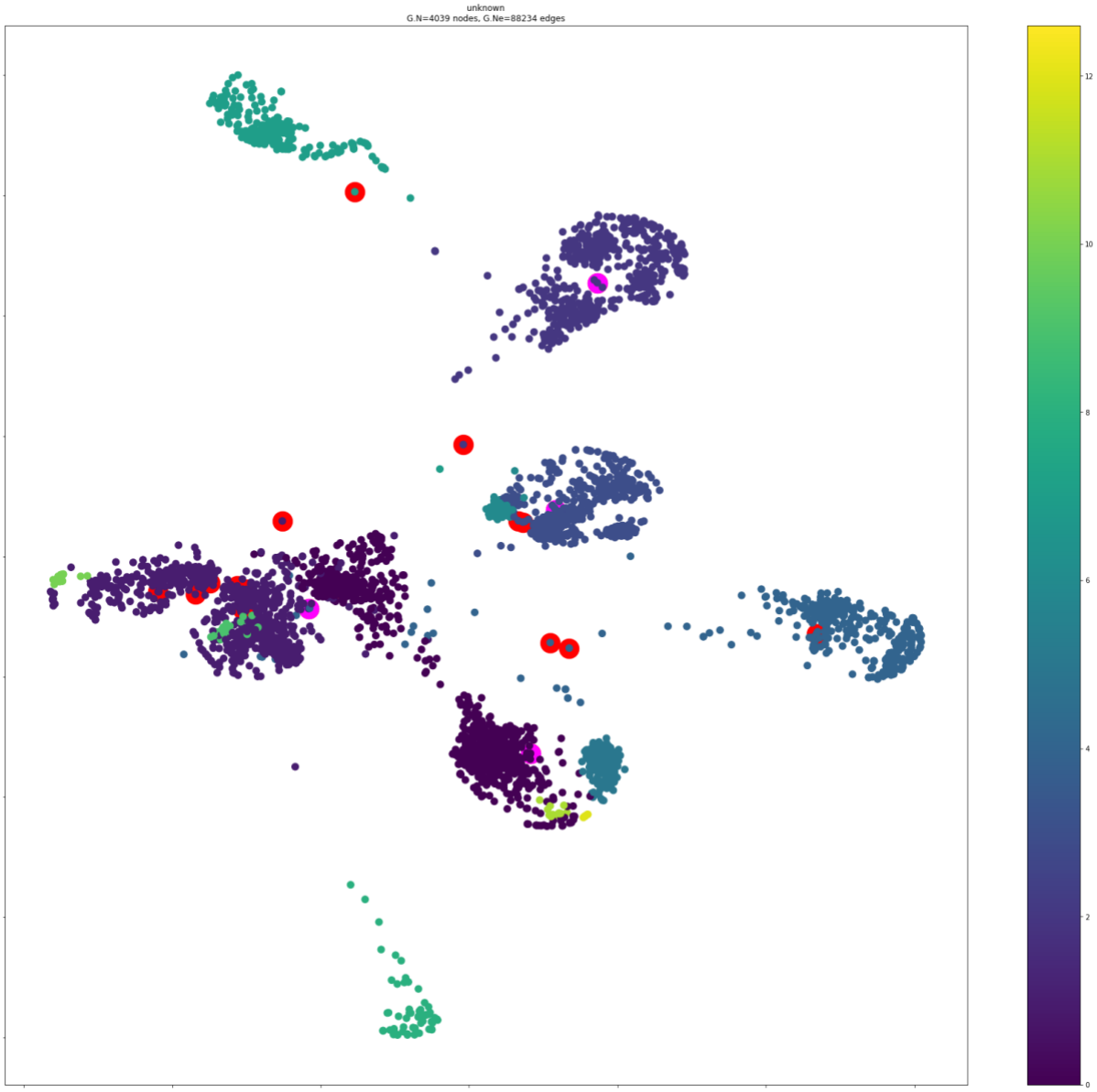

Figure 4: Result of Fast-Greedy algorithm communities. 21 influencers given in red, 4 influencers from clustering given in pink.