CSCI 5352-001 Project Proposal

Team Members: Irene Beckmen, Santhanankrishnan Ramani, Ruhi Saraf

Background:

In May 2016, 11.5 million documents were leaked from the database of Mossack Fonseca, the fourth largest offshore law firm in the world.These documents, dubbed the Panama Papers, reveal the details of how many global leaders, past and present politicians, and wealthy well connected influencers circumvent taxes via offshore banking.

Problem Statement:

We intend to classify officer nodes as politicians vs non-politicians

Data:

1. There is some exploration done on this dataset which we initially intend to replicate http://www.degeneratestate.org/posts/2016/Jun/30/exploring-the-panama-papers-network/.
2. Dataset we will be using: Panama Papers, found at https://panamapapers.icij.org/

Information About the network:

1. Network Model:
   a. Nodes are of 4 types: Officers, Entities, Intermediaries and Addresses.
   b. Edges are of the type: Shareholder of, Officer of etc.
   c. There are about 800,000 nodes and 1.1 million edges
2. Connected Components: Most nodes are present in one subgraph of 708807 nodes, while there are other smaller ones of around 100-200 nodes.

Bottlenecks and Assumptions:

1. Although the edges in the data are supposed to be directional, for simplicity we intend to start off by looking at them as undirected edges.
2. There are a lot of names / addresses referring to the same person / entity but with small modifications (like adding a title / suffix), at the same time, some names are very common therefore it is debatable if we should merge nodes with similar names or not.

Additional Questions we intend to explore:

1. Centrality: Are a certain type of node (officer, intermediate, address or entity) more central than others? A way to determine this would be to find the average degree of each node type. Or the top 10-20 nodes with the highest degree.
2. We should also look at the length of paths - from one intermediate to another and see if there is a pattern.
3. Is it a small world network?
4. Identifying red flags: Do addresses with high degree indicate potential criminal behavior? Multiple entities with same set of shareholders?