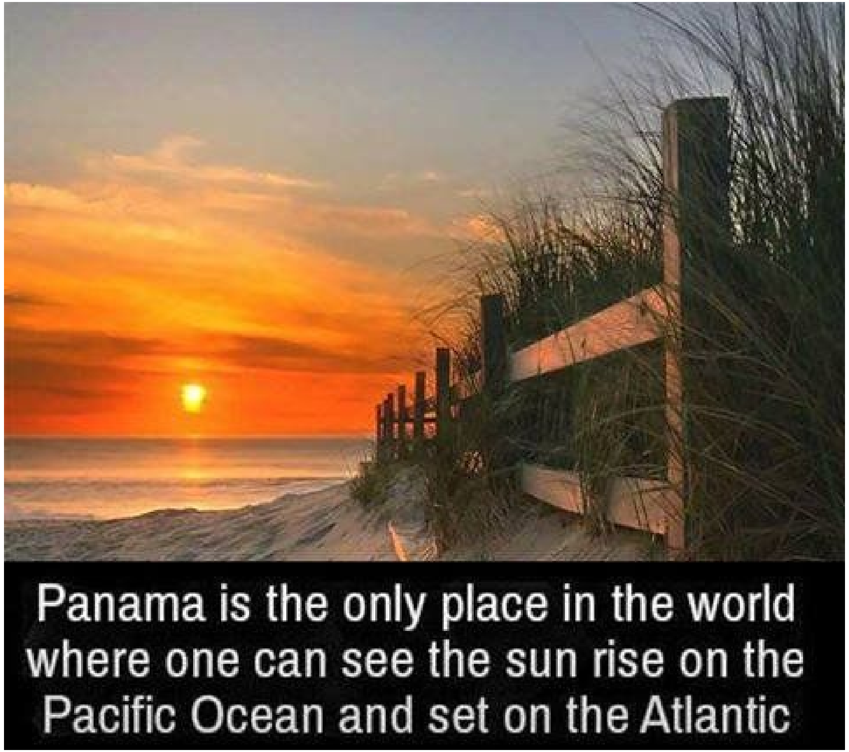
Identifying Politicians in the PANAMA PAPERS

Irene Beckman, Ruhi Saraf, Santhanankrishnan Ramani

Panama Papers

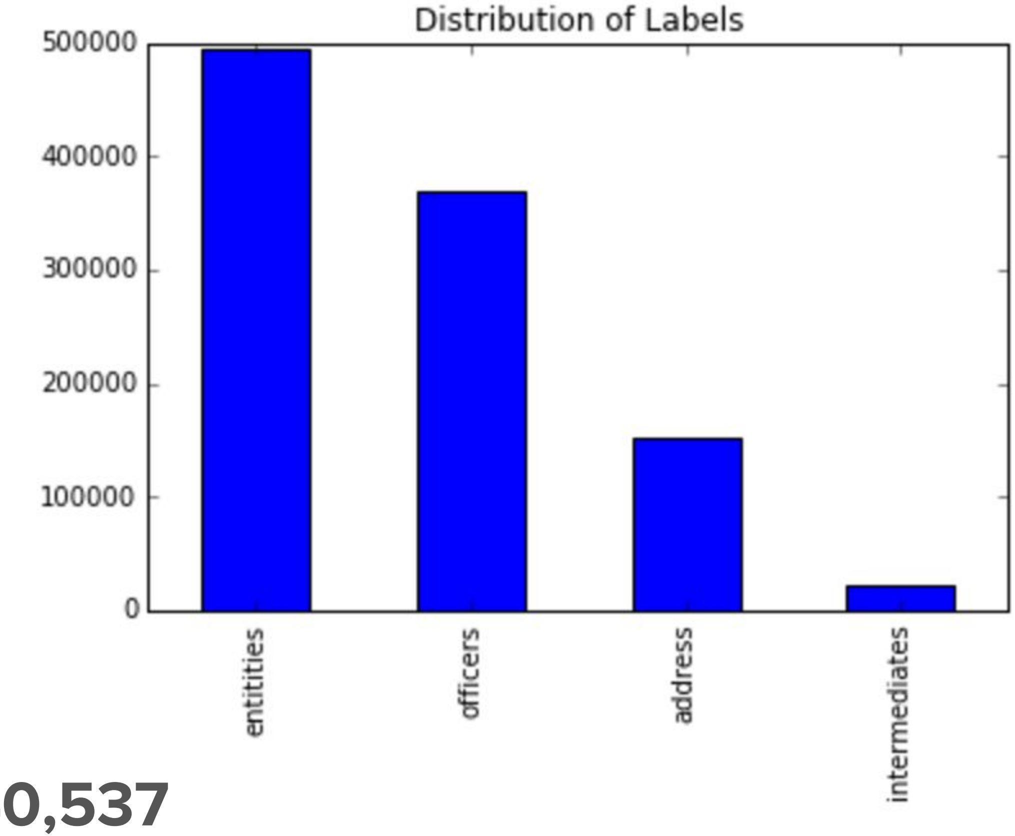
Early 2016 11.5 million emails, power-of-attorney letters, and internal notes from the law firm of Mossack Fonseca were leaked to the press.


Panama is the only place in the world where one can see the sun rise on the Pacific Ocean and set on the Atlantic

- The leak shed light on the financial dealing and details of:
  - World political leaders
  - Fraudsters
  - Drug traffickers
  - Billionaires
  - Celebrities
  - Sports stars and more
- In some cases, brought about new political scandals

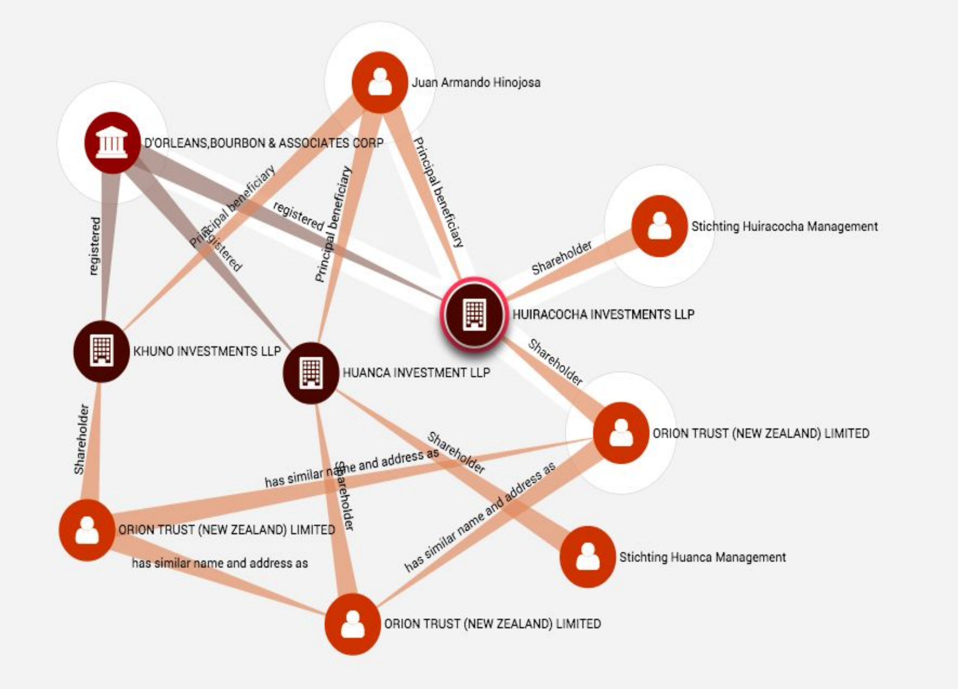
The Dataset: Nodes

- Entities
- Officers
  - People: ≅200,000
  - Politicians: 102
- Intermediaries
- Addresses



Total number of nodes: **1,040,537**

The Dataset: Edges

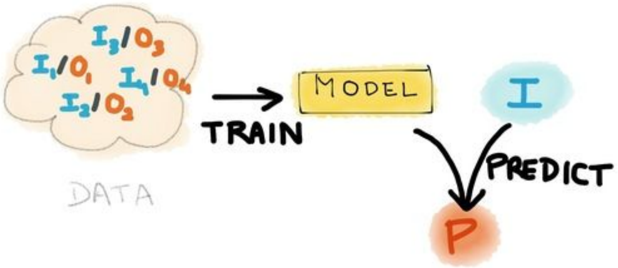

Juan Armando Hinojosa
D'ORLEANS,BOURBON & ASSOCIATES CORP
Stichting Huiracocha Management
HUIRACOCHA INVESTMENTS LLP
KHUNO INVESTMENTS LLP
HUANCA INVESTMENT LLP
ORION TRUST (NEW ZEALAND) LIMITED
ORION TRUST (NEW ZEALAND) LIMITED
Stichting Huanca Management
ORION TRUST (NEW ZEALAND) LIMITED
registered
Principal beneficiary
registered
registered
Principal beneficiary
Principal beneficiary
Shareholder
Shareholder
Shareholder
Shareholder
Shareholder
has similar name and address as
has similar name and address as
has similar name and address as

- Any affiliation between the nodes
  - beneficiary
  - shareholder
  - registered
  - similar name
  - director

Total number of edges: **4,429,023**

Hypothesis

Can we identify people officer nodes with political affiliation by modeling this as a binary classification problem?



Data Parsing and Filtering

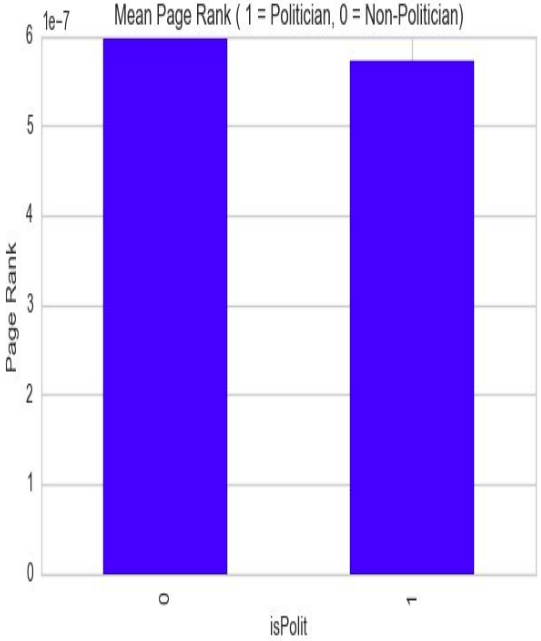
- *Problem*: Large data with distinct nodes types made it hard to pinpoint a subset within officer types
    - *Solution*: Reducing the data set to only include people from officers list
- *Problem*: No label for politicians and affiliates
    - *Solution*: manually labeled all 102 known politicians
- *Problem*: Duplicate names with a variety of spelling for same node
    - *Solution*: Due to the unknown extent of this issue, we could not fix it globally and so left the data as it came

Feature Engineering

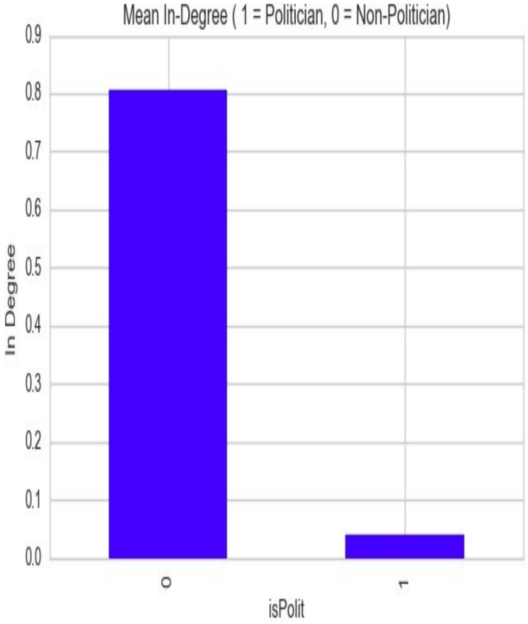
- Vertex Level Measures
- Ego Networks
- Use of directed and undirected representation of graphs
- Motifs

Vertex Level Measures

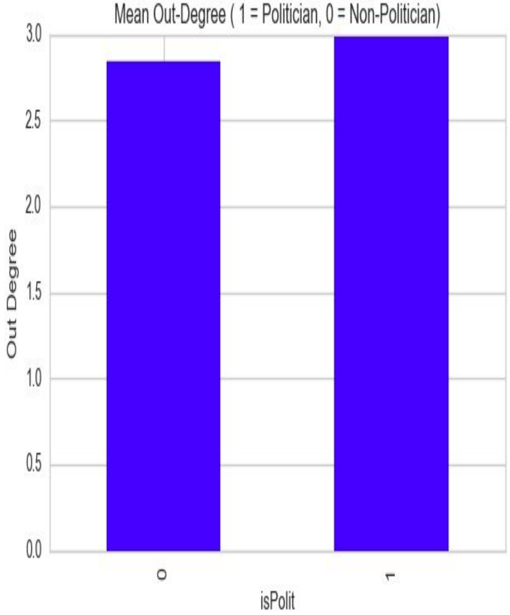
Page Rank



In Degree

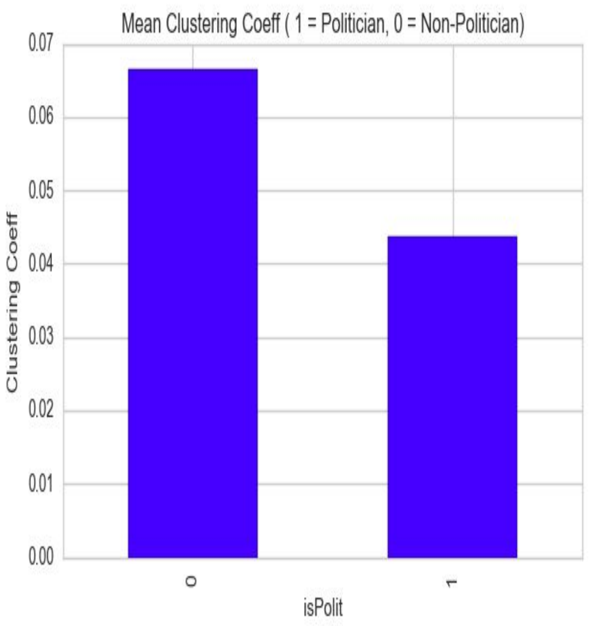


Out Degree

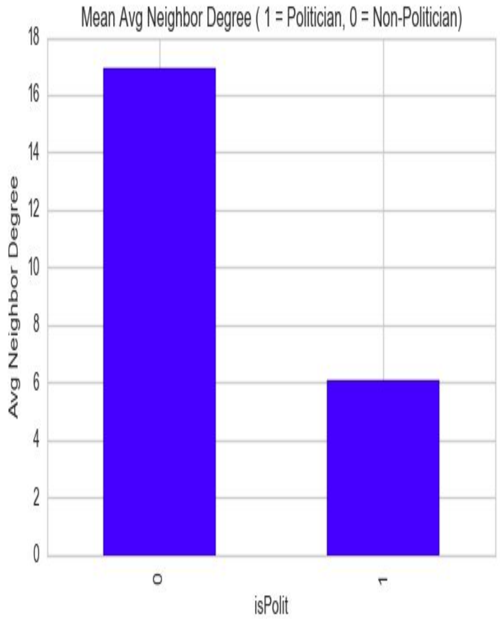


Vertex Level Measures

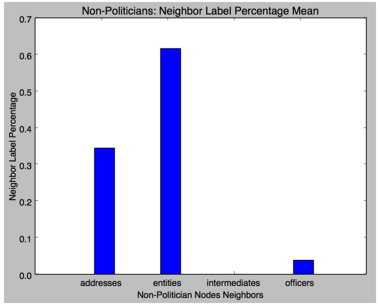
Clustering Coefficient

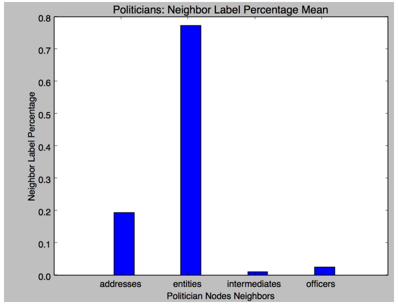


Avg. Neighbor Degree



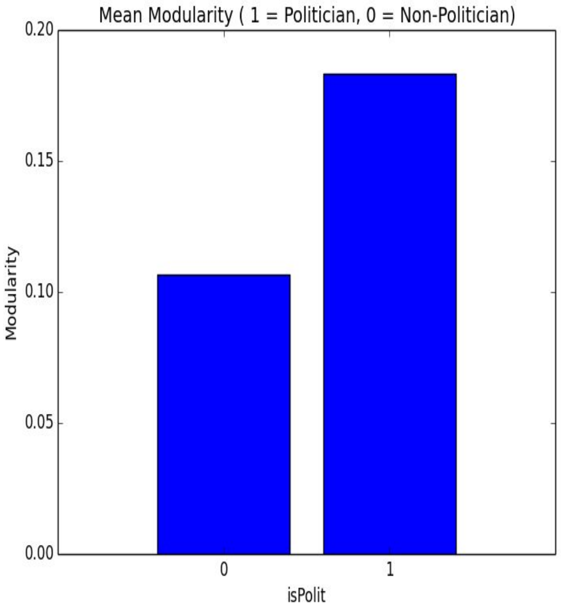
Neighbor Labels



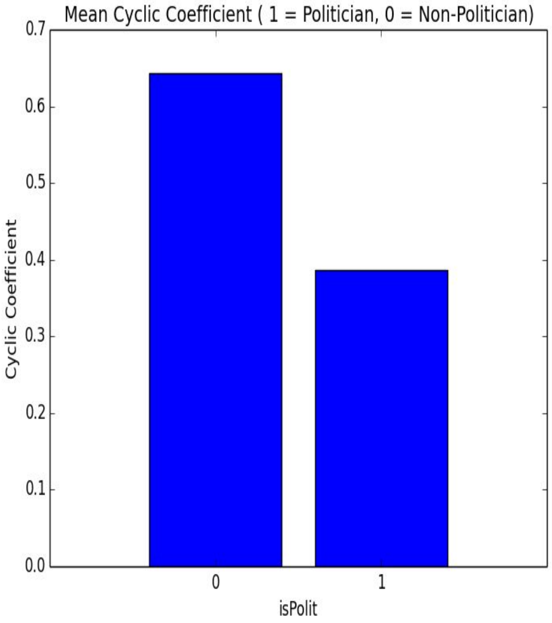


Ego Network Measures

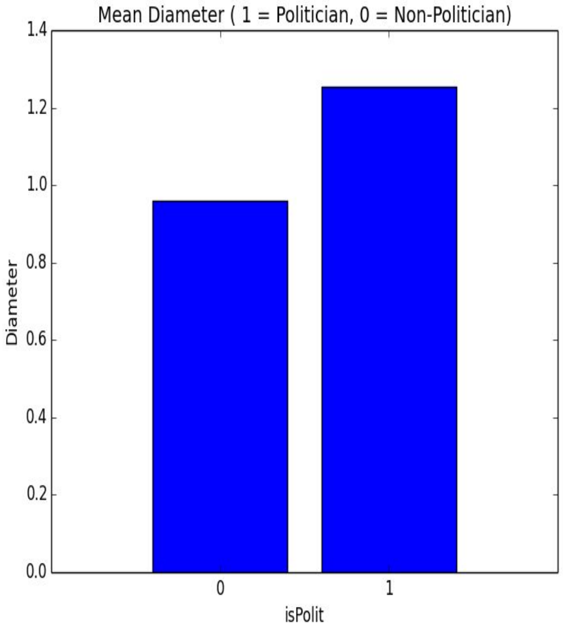
Modularity



Cyclic Coefficient (w/ ego)

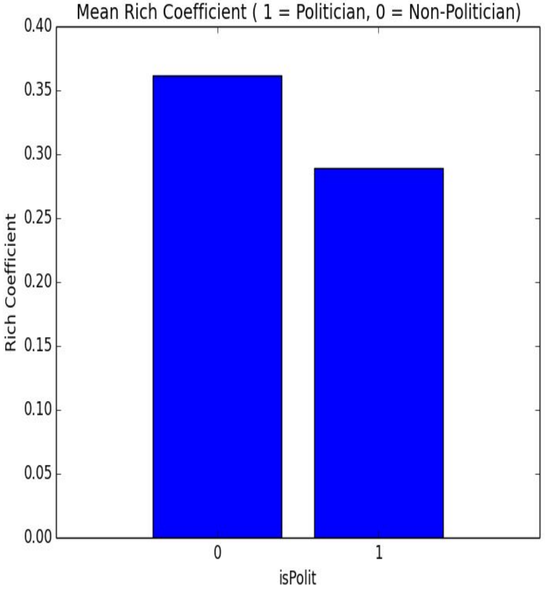


Diameter

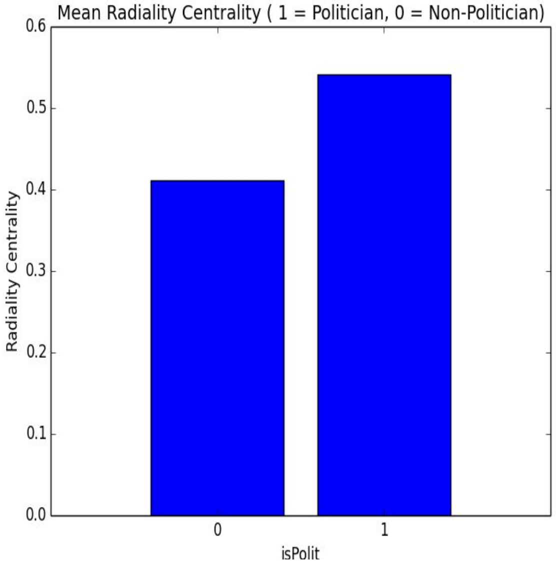


Ego Network Measures

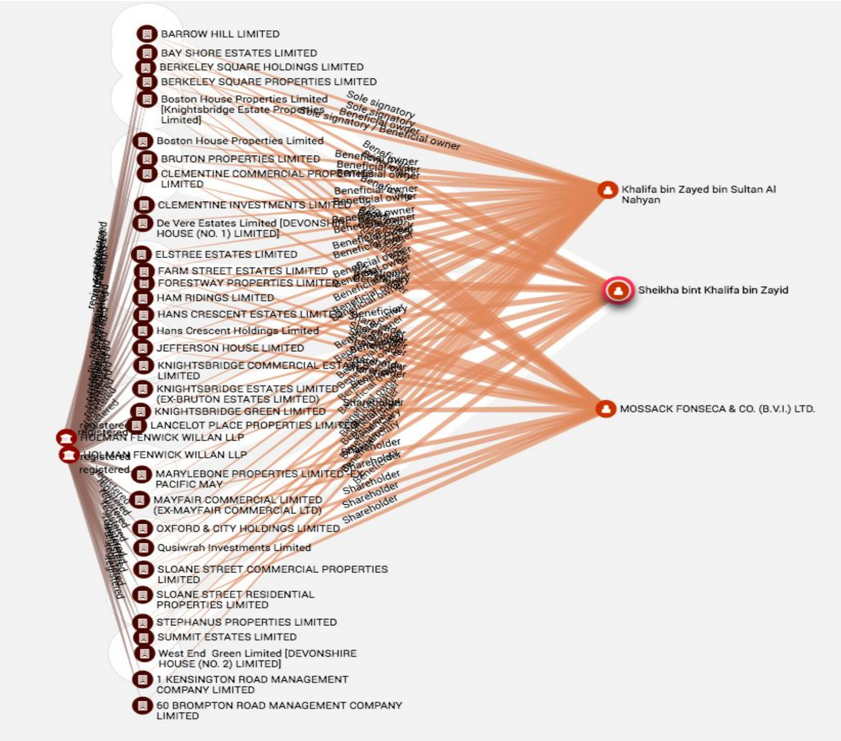
Rich Club Coefficient

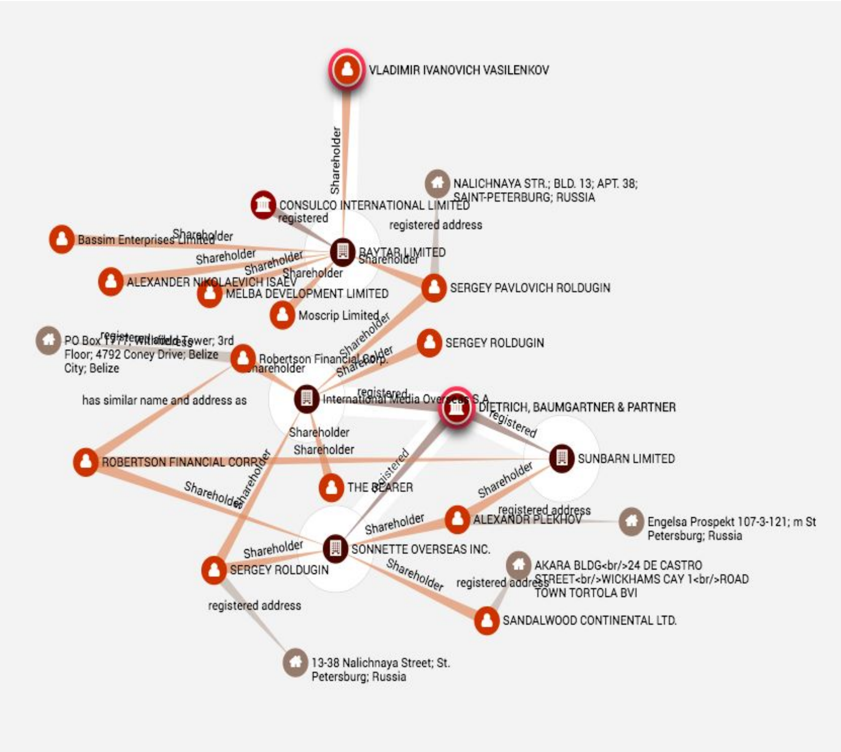


Radiality Centrality (w/ ego)



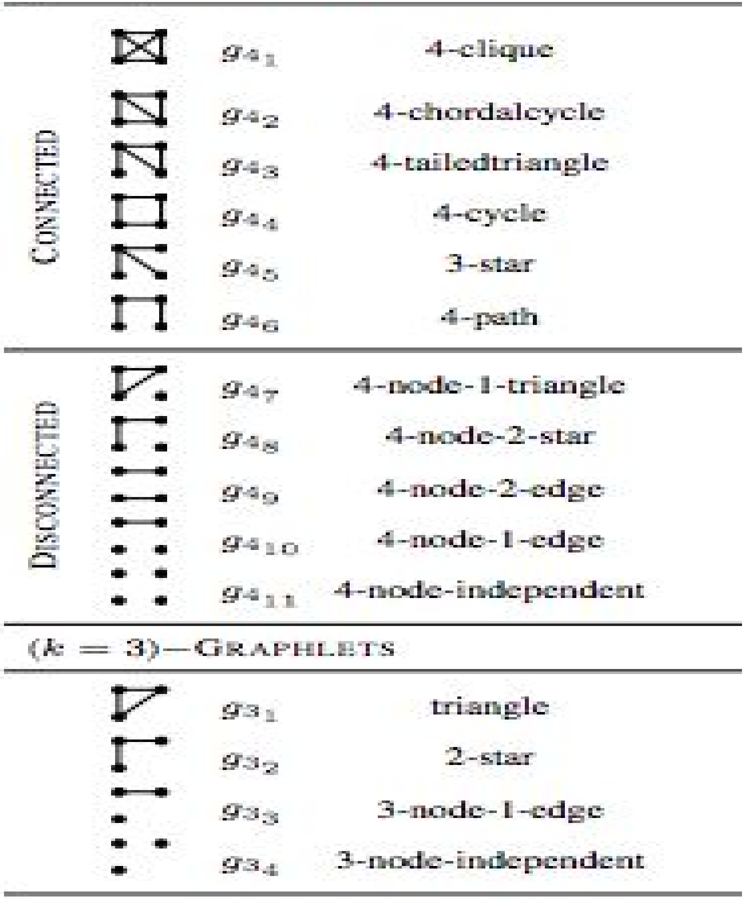
Structure of Network


BARROW HILL LIMITED
BAY SHORE ESTATES LIMITED
BERKELEY SQUARE HOLDINGS LIMITED
BERKELEY SQUARE PROPERTIES LIMITED
Boston House Properties Limited [Knightsbridge Estate Properties Limited]
Boston House Properties Limited
BRUTON PROPERTIES LIMITED
CLEMENTINE COMMERCIAL PROPERTIES LIMITED
CLEMENTINE INVESTMENTS LIMITED
De Vere Estates Limited [DEVONSHIRE HOUSE (NO. 1) LIMITED]
ELSTREE ESTATES LIMITED
FARM STREET ESTATES LIMITED
FORESTWAY PROPERTIES LIMITED
HAM RIDINGS LIMITED
HANS CRESCENT ESTATES LIMITED
Hans Crescent Holdings Limited
JEFFERSON HOUSE LIMITED
KNIGHTSBRIDGE COMMERCIAL ESTATES LIMITED
KNIGHTSBRIDGE ESTATES LIMITED (EX-BRUTON ESTATES LIMITED)
KNIGHTSBRIDGE GREEN LIMITED
LANCELOT PLACE PROPERTIES LIMITED
HOLMAN FENWICK WILLAN LLP
HOLMAN FENWICK WILLAN LLP
MARYLEBONE PROPERTIES LIMITED (EX-PACIFIC MAY
MAYFAIR COMMERCIAL LIMITED (EX-MAYFAIR COMMERCIAL LTD)
OXFORD & CITY HOLDINGS LIMITED
Qusiwrah Investments Limited
SLOANE STREET COMMERCIAL PROPERTIES LIMITED
SLOANE STREET RESIDENTIAL PROPERTIES LIMITED
STEPHANUS PROPERTIES LIMITED
SUMMIT ESTATES LIMITED
West End Green Limited [DEVONSHIRE HOUSE (NO. 2) LIMITED]
1 KENSINGTON ROAD MANAGEMENT COMPANY LIMITED
60 BROMPTON ROAD MANAGEMENT COMPANY LIMITED
Khalifa bin Zayed bin Sultan Al Nahyan
Sheikha bint Khalifa bin Zayid
MOSSACK FONSECA & CO. (B.V.I.) LTD.
Sole signatory
Sole signatory
Sole signatory / Beneficial owner
Beneficial owner
Beneficiary
Shareholder
registered


VLADIMIR IVANOVICH VASILENKOV
Shareholder
NALICHNAYA STR.; BLD. 13; APT. 38; SAINT-PETERBURG; RUSSIA
CONSULCO INTERNATIONAL LIMITED
registered
registered address
Bassim Enterprises Limited
Shareholder
BAYTAR LIMITED
Shareholder
ALEXANDER NIKOLAEVICH ISAEV
MELBA DEVELOPMENT LIMITED
SERGEY PAVLOVICH ROLDUGIN
Moscrip Limited
PO Box 1777; Withfield Tower; 3rd Floor; 4792 Coney Drive; Belize City; Belize
registered address
Robertson Financial Corp.
Shareholder
SERGEY ROLDUGIN
has similar name and address as
International Media Overseas S.A.
registered
DIETRICH, BAUMGARTNER & PARTNER
registered
ROBERTSON FINANCIAL CORP.
Shareholder
SUNBARN LIMITED
THE BEARER
ALEXANDR PLEKHOV
Engelsa Prospekt 107-3-121; m St Petersburg; Russia
SONNETTE OVERSEAS INC.
SERGEY ROLDUGIN
AKARA BLDG<br/>24 DE CASTRO STREET<br/>WICKHAMS CAY 1<br/>ROAD TOWN TORTOLA BVI
registered address
SANDALWOOD CONTINENTAL LTD.
13-38 Nalichnaya Street; St. Petersburg; Russia

Motifs

- Node/Edge Motifs
- Ego Network Motifs
- Neighbors Edge Motifs



CONNECTED	g_{4_1}	4-clique
	g_{4_2}	4-chordalcycle
	g_{4_3}	4-tailedtriangle
	g_{4_4}	4-cycle
	g_{4_5}	3-star
	g_{4_6}	4-path
DISCONNECTED	g_{4_7}	4-node-1-triangle
	g_{4_8}	4-node-2-star
	g_{4_9}	4-node-2-edge
	$g_{4_{10}}$	4-node-1-edge
	$g_{4_{11}}$	4-node-independent
$(k = 3)$—GRAPHLETS		
	g_{3_1}	triangle
	g_{3_2}	2-star
	g_{3_3}	3-node-1-edge
	g_{3_4}	3-node-independent

Sampling and Class Imbalance

Issue: Combating Imbalanced Classes

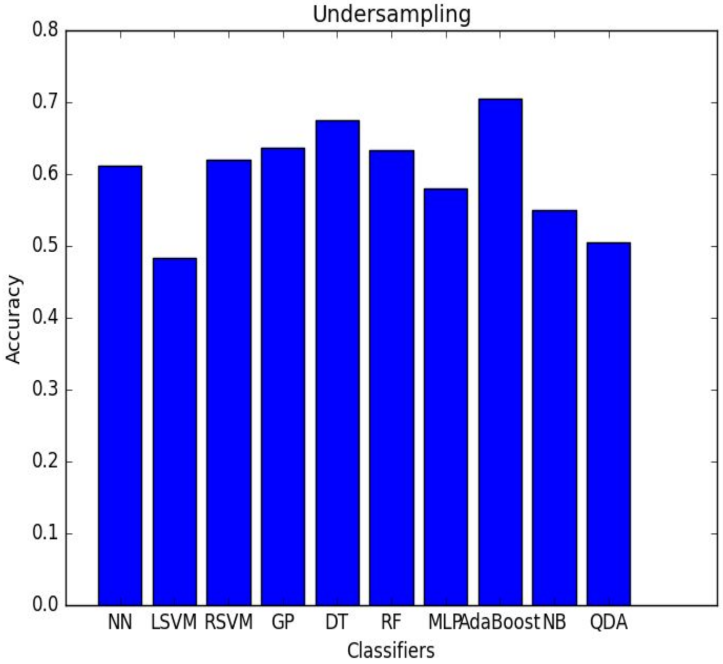
What we tried:

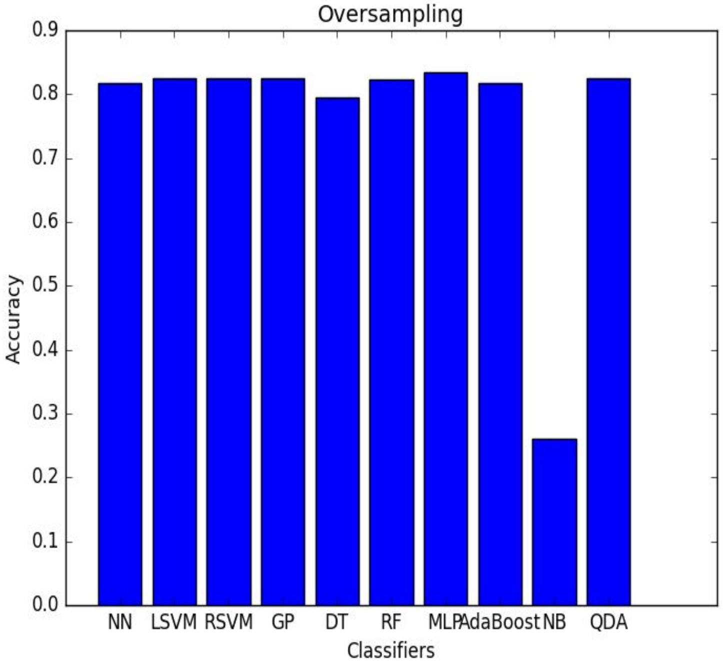
1. Resampling: Undersampling, Oversampling, Unequal Ratios
2. Changed Performance Metric

Method

- Various Classifiers (Random Forest, SVC, Neural Net, AdaBoost)
- Parameter Optimization using Grid Search
- 10-fold testing to compute average accuracy

Results


Undersampling
Accuracy
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
NN
LSVM
RSVM
GP
DT
RF
MLP
AdaBoost
NB
QDA
Classifiers


Oversampling
Accuracy
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
NN
LSVM
RSVM
GP
DT
RF
MLP
AdaBoost
NB
QDA
Classifiers

Future Work:

- Feature Selection
- Try penalized models
- Ensemble Classification using Multiple subsets
- Balanced Cascade
- Incorporate country/region data manually
  (meta data associated with the nodes)

Questions?