Diving Deep into the Panama Papers

Vincent Micheli **Franck Dessimoz** **Christophe Huang**
franck.dessimoz@epfl.ch
vincent.micheli@epfl.ch xingjian.huang@epfl.ch

Abstract

The Panama Papers are arguably the largest leak of confidential data to date. They provide a wealth of knowledge by exposing previously hidden ties between corporations and their ecosystems. This paper presents a social network analysis of the companies and individuals involved in this leak. We identified the countries in which the key players were operating, how we could predict the involvement of a country and what were the underlying relationships between the parties involved. We were able to shed light on the central actors of the network and provide further investigations on their connections.

1 Introduction

The Panama Papers are leaked documents that detail the internal operations of one of the world's biggest firms in incorporation of offshore entities, Mossack Fonseca. This Panamanian law firm and corporate service provider had financial and attorney-client information for more than 210 000 offshore entities revealed to the public. While offshore business entities are legal, reporters found that some of the Mossack Fonseca shell corporations were used for illegal purposes, namely fraud, tax evasion, and evading international sanctions. Therefore one may argue that it is of public and state interest to explore the Panama Papers in order to extract insights about the key players involved. We intend to discover, analyse and explain the underlying relationships inside the Panama Papers. [1]

The primary goals of our project were three-fold:

1. Identify the countries and jurisdictions with the most offshore entities, intermediaries and officers. Test whether the number of entities is linked to the number of intermediaries and officers or not.
2. Predict the involvement of a country based on financial and risk indicators. That is, find the correlates of evasion and discuss the relevance of commonly used metrics by regulators.
3. Quantify the notion of importance in the papers and identify the key actors. Describe the communities embedded inside the network and investigate their constituents.

2 Related Work

One could argue that the most difficult part when investigating leaked data is to preprocess it and make it understandable to the human eye. To this end the ICIJ did a wonderful job. Indeed the leaks have been made available as a database powered by a graph query engine called Neo4j (Miller, 2013). Moreover the ICIJ already performed a great deal of investigative journalism which we will reproduce and complement in the first part of our work (ICIJ, 2018).

3 Data Collection and Description

As stated before we will resort to the Panama Papers section of the ICIJ Offshore Leaks Database available online. In the second part of our work we used a multitude of data sources in order to extract economic and financial indicators about the countries present in the dataset.

The database is distributed into 5 csv files:

- Entity: A company, trust or fund created by an agent. 213634 entries
- Officer: A person or company who plays a role in an offshore entity. 238402 entries

[1] Due to space constraints this paper will not cover our work in its entirety. Analysis available at https://github.com/franckdess/ADA-Project.

- Intermediary: A go-between for someone seeking an offshore corporation and an offshore service provider. 14110 entries
- Address: Contact postal address as it appears in the original databases obtained by the ICIJ. 93454 entries
- Edges: Relationship between two nodes and the nature of the link. 674102 entries

4 Methodology, Models and Methods

In order to obtain key figures about the evolution and situation of Mossack Fonseca's clients we used the python library Pandas (McKinney, 2011). Most of it has been dataframes manipulations.
Moreover we established the strength and direction of monotonic relationships between entities, intermediaries and officers with the Spearman's rank-order correlation statistic. It is the Pearson correlation coefficient between the ranked variables.

$$\begin{aligned} r(s) &= p(ranked_x, ranked_y) \\ &= \frac{cov(ranked_x, ranked_y)}{std(ranked_x)std(ranked_y)} \end{aligned}$$

The correlates of evasion were identified by performing linear regressions for each financial indicator while controlling for Gross Domestic Product and Population. We applied logarithm to the number of links, GDP and population, and we standardized all indicators except for INCSR Primary Concern and FATF greylist, which hold binary values. Each regression had the form:

$log(Number\ \ of\ \ entities) = a \cdot standardized\ Indicator\ of\ Interest + b \cdot log(Population) + c \cdot log(GDP) + E$
with $E \sim N(0, \sigma^2)$

Computations were carried in the python library *Statsmodels* (Seabold and Perktold, 2010).
We then moved on to Social Network Analysis with the python library *Networkx* (Hagberg et al., 2008). Our objective was to zoom in on communities and actors inside the network.
We quantified node importance based on centrality measures:
The Pagerank algorithm performs a random ergodic walk on the graph and outputs the state probability for each node (Page et al., 1999). That is the probability of being on a specific node at each iteration of a random surfer.
Betweenness centrality is based on shortest paths (CS224W, 2018).

$$bet(v) = \sum \frac{\sigma st(v)}{\sigma st}$$

for every pair of nodes s, t.
That is the number of shortest paths between s and t going through v divided by the number of shortest paths between s and t. The higher the betweenness centrality, the more a node acts as a bridge between nodes.
Degree is based on the number of adjacent edges for each node. Meaning that the more neighbours a node has, the more important it may be in the network.
Prior to community detection we assessed the global structure of the graph:
The clustering coefficient is a measure of the degree to which nodes in a graph tend to cluster together. The global clustering coefficient is

$$G = \frac{number\ of\ closed\ triplets}{number\ of\ open\ or\ closed\ triplets}$$

The clustering coefficient can also be computed for each node individually.
The density is a measure of how connected a network is.

$$d = \frac{2m}{n(n-1)}$$

where m is the number of edges and n the number of nodes. That is the ratio of the observed sum of degrees divided by the maximum sum of degrees.

Finally we employed community detection algorithms for specific countries. A community refers to the occurrence of groups of nodes in a network that are more densely connected internally than with the rest of the network (CS224W, 2018). The Louvain method performs communities extraction (Blondel et al., 2008). Initially every node is considered as a community. The communities are traversed, and for each community it is tested whether we can obtain a better clustering by joining it to a neighboring community. The objective function to maximize is the modularity.

$$Q = \sum(1 - \frac{deg(i)deg(j)}{2m})\delta(c_i, c_j)$$

for every pair of nodes i, j. The delta function is 1 if i and j belong to the same community.
The detection and visualisation of communities were carried in *Gephi*, an open graph visualisation platform (Bastian et al., 2009). Node size was a function of centrality measures while the color code indicated communities. Besides their constituents were investigated and visualised using Neo4j.

4.1 Results and findings

The following bar plots present the jurisdictions and countries with the most offshore entities, intermediaries and officers.

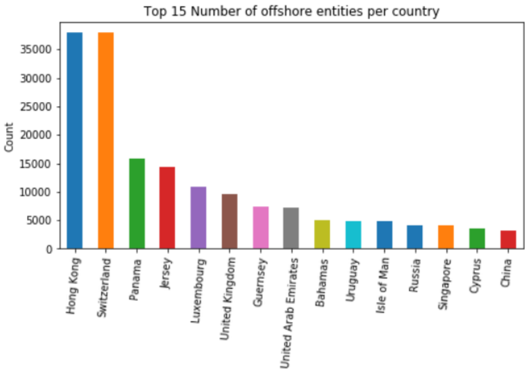

Figure 1: Entities per country

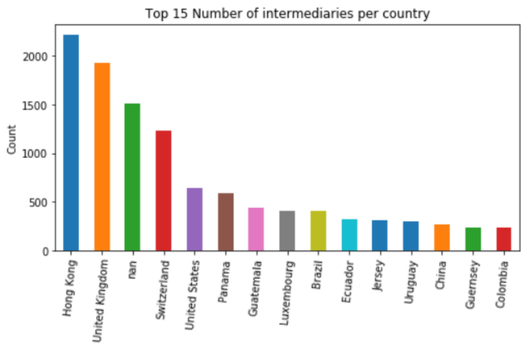

Figure 2: Intermediaries per country[2]

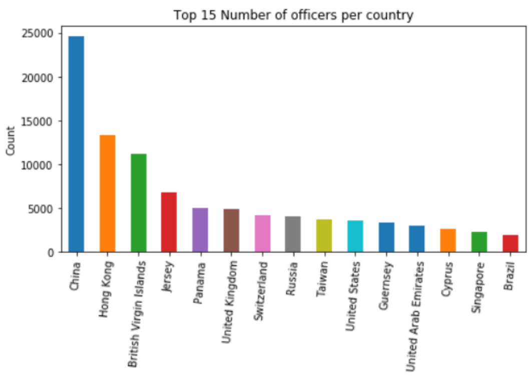

Figure 3: Officers per country

[2]NaN indicates that we could not retrieve the country for the intermediary.

The most represented countries and jurisdictions are tax heavens themselves. They act as intermediary countries for tax evasion.

Entity/Intermediary Correlation	0.89
Entity/Officer Correlation	0.82
Officer/Intermediary Correlation	0.79

Table 1: Spearman's rank order statistics

We confirm the existence of a strong positive monotonic relationship between the number of entities and the number of intermediaries and officers.[3]
Now that we had understood which were the key countries we explained their involvement based on economic and financial indicators. Scores used by various governmental agencies to assess the risk of a country regarding financial secrecy and money laundering were also included. [4]

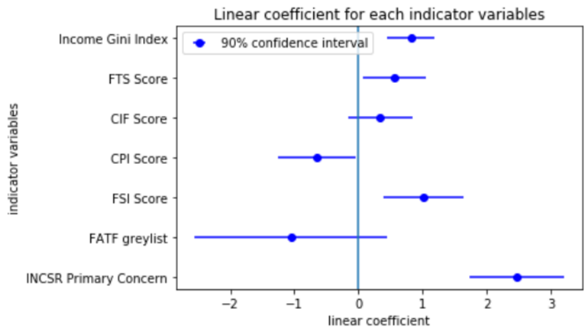

Figure 4: Correlates of evasion

One of the best predictors for the links between a country and Mossack Fonseca is the Financial Secrecy Index (FSI). From the observations above, an increase of 1% of a country in the FSI should results in an increase of roughly 1% in its log number of apparitions in the Panama Papers. There also seems to be an undeniable relationship between Income Inequality (Gini Index) and the number of dealings with Mossack Fonseca, though it remains unclear whether this relationship is causal or not. One explication would be that higher inequality leads to higher number of high-income clients looking to avoid the equally high taxes by transferring their money to a tax heaven.
Mossack Fonseca is not only suspected of helping

[3]The data was inspected for outliers and further testing still suggested a strong positive monotonic relationship. P values are smaller than 10^{-30}.

[4]The datasets used are listed in the repository linked earlier.

tax evasion schemes, but there is also concern that the firm might be involved in money laundering activities. In order to dig into this problem, we decided to take a look at the Financial Action Task Force (FATF) greylist of countries with low anti-money laundering (AML) regulations and practices. The results show that Mossack Fonseca avoids to deal with such countries. On the other hand, the firm is working with a lot of clients from the countries listed in the US State Department's "Primary concern" for money laundering (ML), which are countries currently suspected of taking part in ML activities.

The commonly referred to corruption (Corruption Perception Index, CPI) and illicit financial flow (Cumulative Illicit Flows, CIF) scores used by some regulators have little implication on whether the countries appear in the Panama Papers.

We extracted the most important actors based on centrality measures: [5]

name	type	degree	pager
ACCELONIC LTD.	entity	1007	0.000580
AKARA BLDG.; 24 DE CASTRO STREET; WICKHAMS CAY...	address	1007	0.000538
VELA GAS INVESTMENTS LTD.	entity	493	0.000456
Akara Building; 24 de Castro Street; Wickhams ...	address	813	0.000431
Akara Building; 24 De Castro Street; Wickhams ...	address	686	0.000354

name	type	degree
ORION HOUSE SERVICES (HK) LIMITED	intermediary	7016
MOSSACK FONSECA & CO.	intermediary	4364
PRIME CORPORATE SOLUTIONS SARL	intermediary	4117
OFFSHORE BUSINESS CONSULTANT (INT'L) LIMITED	intermediary	4094
MOSSACK FONSECA & CO. (SINGAPORE) PTE LTD.	intermediary	3888

Figure 5: Key actors based on pagerank and degree

We observed that the degree distribution was a pseudo power law characterized by its heavy tail. Intermediaries being the cause of this distribution as suggested by a small median to mean ratio.

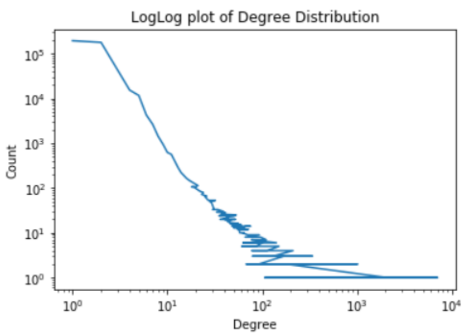

Figure 6: Degree distribution

Moreover we concluded that the network was not a small world because of its zero clustering coefficient. The idea of a highly dispersed network was supported by its density measure $d = 5.37 \cdot 10^{-6}$. This left us with communities detection and investigation. Every country we worked on displayed a community structure based on independent clusters of nodes as indicated by a large modularity metric. We present the results we got on Russias ego graph whose modularity is 0.923. We extracted 6 major actors:

We started our investigation with *FABIAN BONNELAME* and *LEONA GROUP LIMITED*. As shown in the Appendix, *FABIAN BONNELAME* is the shareholder and the beneficiary of *MOWBRAY PACIFIC LIMITED*, *LEGAL CONSULTING SERVICES LIMITED* and the intermediary of *MOWBRAY PACIFIC LIMITED*. We found out that *FABIAN BONNELAME* was registered at many addresses and under different names.

Since we did not find a connection between *FABIAN BONNELAME* and *LEONA GROUP LIMITED* yet, we investigated on *LEONA GROUP LIMITED*. Unsurprisingly, we found out that indeed *FABIAN BONNELAME* was a shareholder of *LEONA GROUP LIMITED*. We also noticed that *UNITRUST CORPORATE SERVICES LTD* was the intermediary of *LEONA GROUP LIMITED* (See Appendix). *UNITRUST CORPORATE SERVICES LTD* is an intermediary company connected to 680 entities and linked to Russia and Canada.

Finally, *EUROGLEN S.A.* and *GLOBAL WEALTH MANAGEMENT CENTER LIMITED* have no direct connections. *EUROGLEN S.A.* is an intermediary company connected to 211 entities, while *GLOBAL WEALTH MANAGEMENT CENTER LIMITED* is also an intermediary company connected to 279 entities.

5 Conclusion

In this paper we identified the countries in which the key players were operating, most often tax heavens themselves such as the British Virgin Islands or Switzerland. We saw that actors were distributed interdependently and that this distribution could be inferred from financial secrecy indexes. Moreover, we shed light on intermediaries as the main drivers of the network. Indeed we quantified node importance based on centrality measures. Finally we observed that the network was mostly dispersed and organized in independent communities which we provided further investigations on their structure.

[5] For the rest of the analysis we focused on the giant component, that is the largest connected component in the network.

References

[Bastian et al.2009] Mathieu Bastian, Sebastien Heymann, Mathieu Jacomy, et al. 2009. Gephi: an open source software for exploring and manipulating networks. *Icwsm*, 8(2009):361–362.

[Blondel et al.2008] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008.

[CS224W2018] CS224W. 2018. Stanford cs224w analysis of networks. `http://web.stanford.edu/class/cs224w/`. Accessed: 2018-12-12.

[Hagberg et al.2008] Aric Hagberg, Pieter Swart, and Daniel S Chult. 2008. Exploring network structure, dynamics, and function using networkx. Technical report, Los Alamos National Lab.(LANL), Los Alamos, NM (United States).

[ICIJ2018] ICIJ. 2018. Icij offshore leaks database. `https://offshoreleaks.icij.org/`. Accessed: 2018-12-12.

[McKinney2011] Wes McKinney. 2011. pandas: a foundational python library for data analysis and statistics. *Python for High Performance and Scientific Computing*, pages 1–9.

[Miller2013] Justin J Miller. 2013. Graph database applications and concepts with neo4j. In *Proceedings of the Southern Association for Information Systems Conference, Atlanta, GA, USA*, volume 2324, page 36.

[Page et al.1999] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab.

[Seabold and Perktold2010] Skipper Seabold and Josef Perktold. 2010. Statsmodels: Econometric and statistical modeling with python. In *Proceedings of the 9th Python in Science Conference*, volume 57, page 61. SciPy society Austin.

6 Appendix: Russia key actors investigated in Gephi and Neo4j

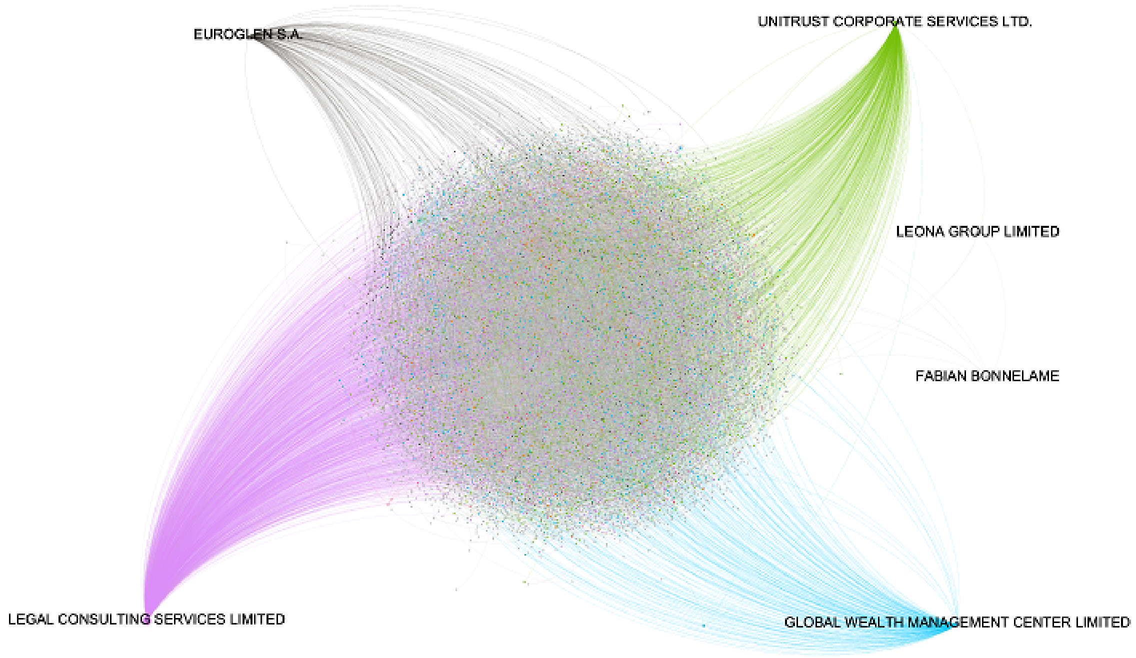

Figure 7: Russia ego graph

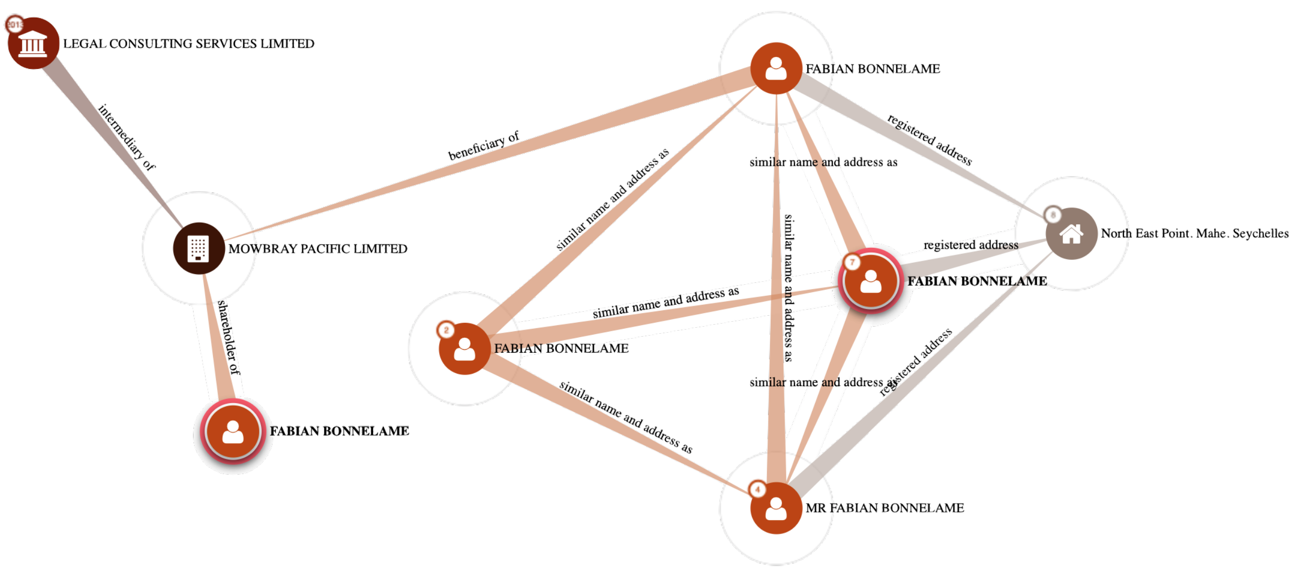

Figure 8: Russia key actors 1

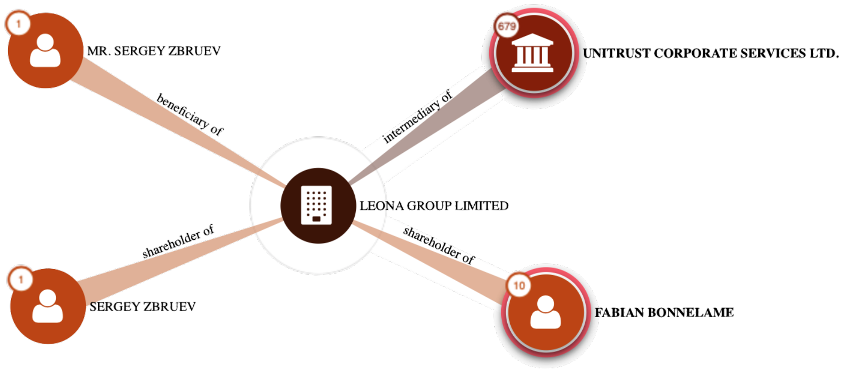

Figure 9: Russia key actors 2