Panama Papers Analysis

*through **Graph Analytics** and **Embeddings***

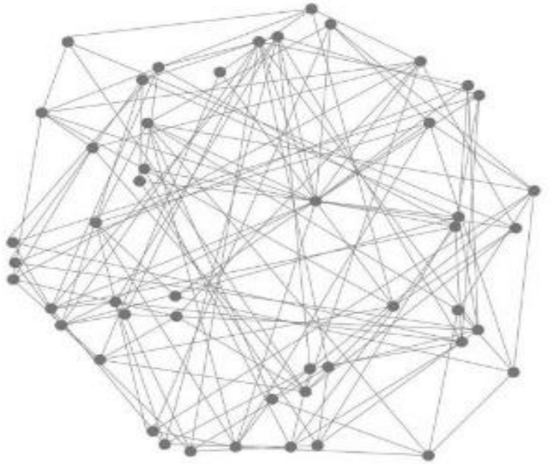
Complex Networks Dynamics

Authors
Manos Chatzakis
(chatzakis@ics.forth.gr)
Eva Chamilaki
(evacham7@gmail.com)

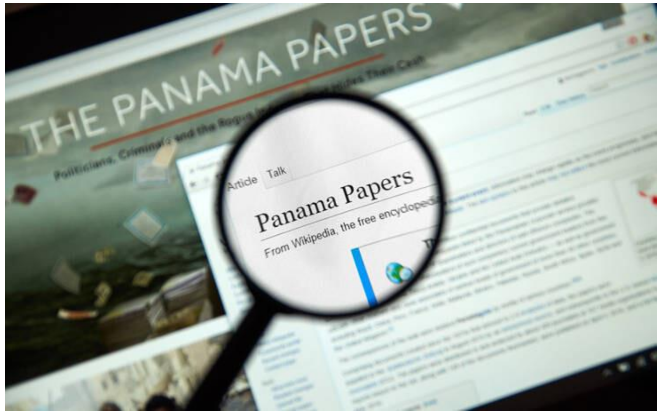
Course Instructor
I.G. Tollis

Contents

- **Introduction** (~3 minutes)
  - Panama Papers
  - Approach
  - Work Outline
- **Graph Analytics** (~3 minutes)
  - Base Dataset Analysis
  - TSP Graph Analytics and Visualization
  - NetworkX Graph Analytics
- **Embeddings** (~3 minutes)
  - Preliminaries
  - Approach
  - RDFsim
  - Use cases
- **Conclusion** (~1 minute)
  - Summary
  - QA



What are the Panama Papers?

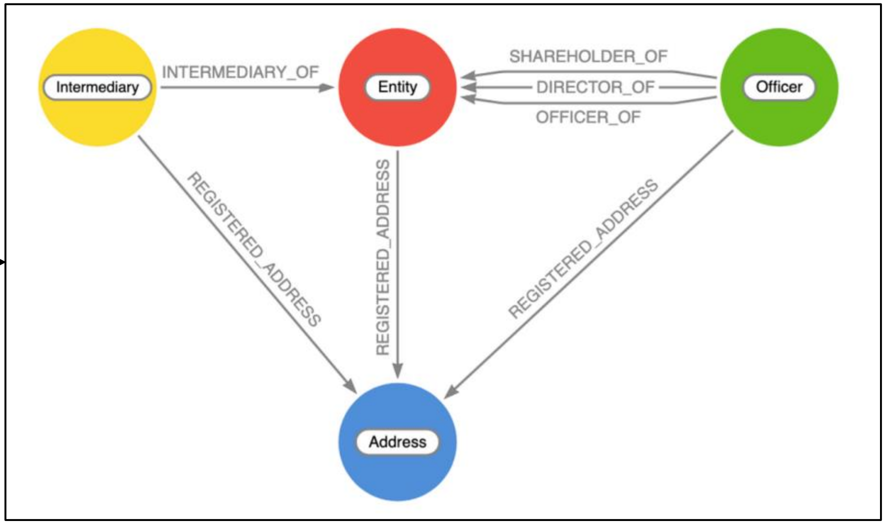


They are documents exposing:

- **Financial activity** of over 200.000 offshores
- **Relations between clients** and other **entities**

They were released to public in **2016**!

- The available panama papers datasets are widely used for **graph analysis**, as they form a specific type of **Fraud Detection Graph**

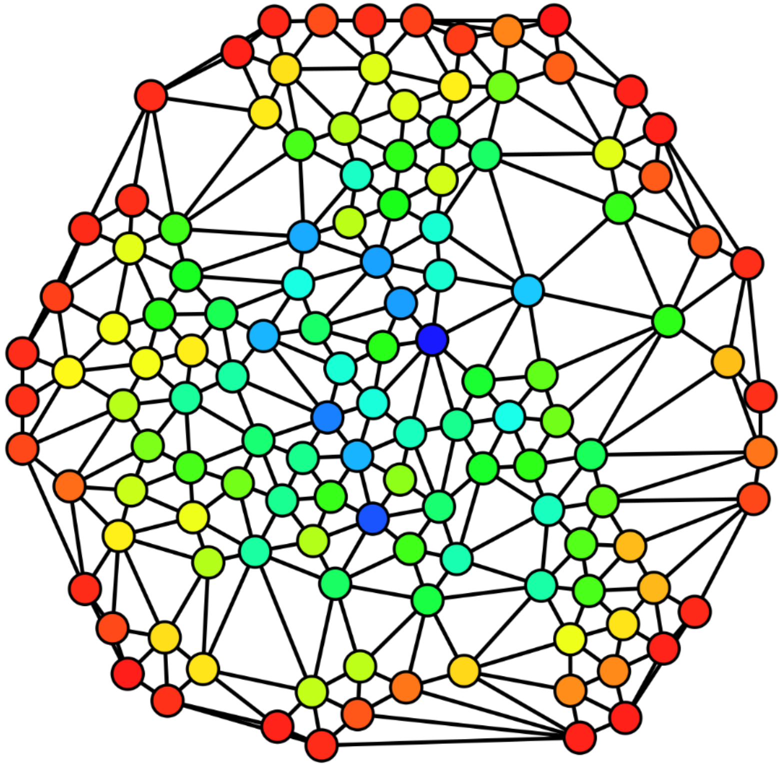


Approach

- It is interesting to use graph analysis algorithms (e.g. ranking, clustering etc.) over a Fraud Detection graph in order to find the **most important nodes.**
- Having information about **which entities of the graph are most "popular"** can lead us to conclusions about interactions of nodes, entities that are most likely to commit frauds in the future and more.

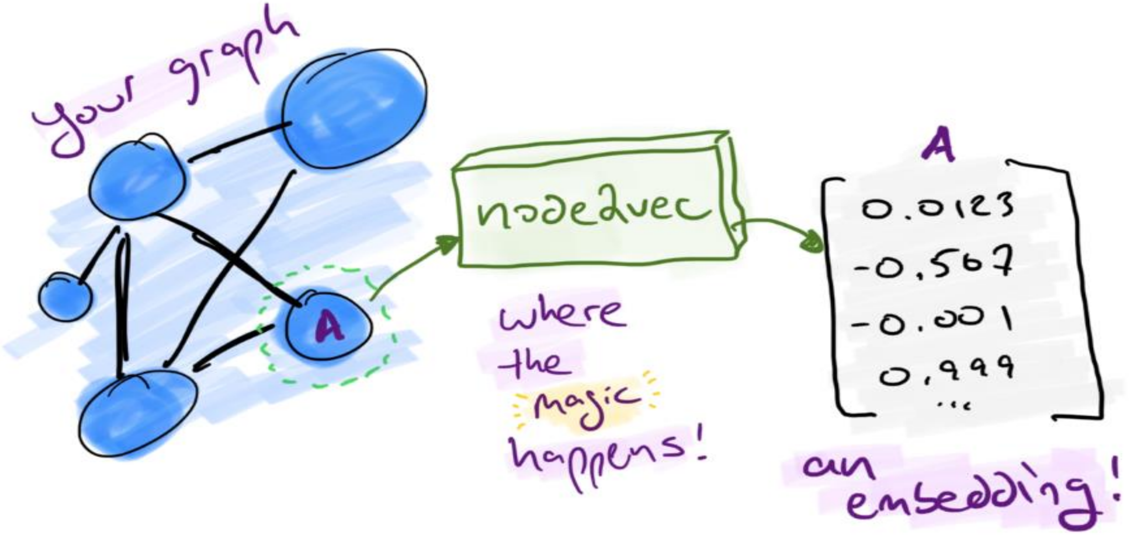
But is just running a bunch of graph analytics algorithms in a network enough, or **could we go one step further?**

How can we **utilize methods from other fields** (e.g. **graph embeddings**) with classic graph analytics methods?

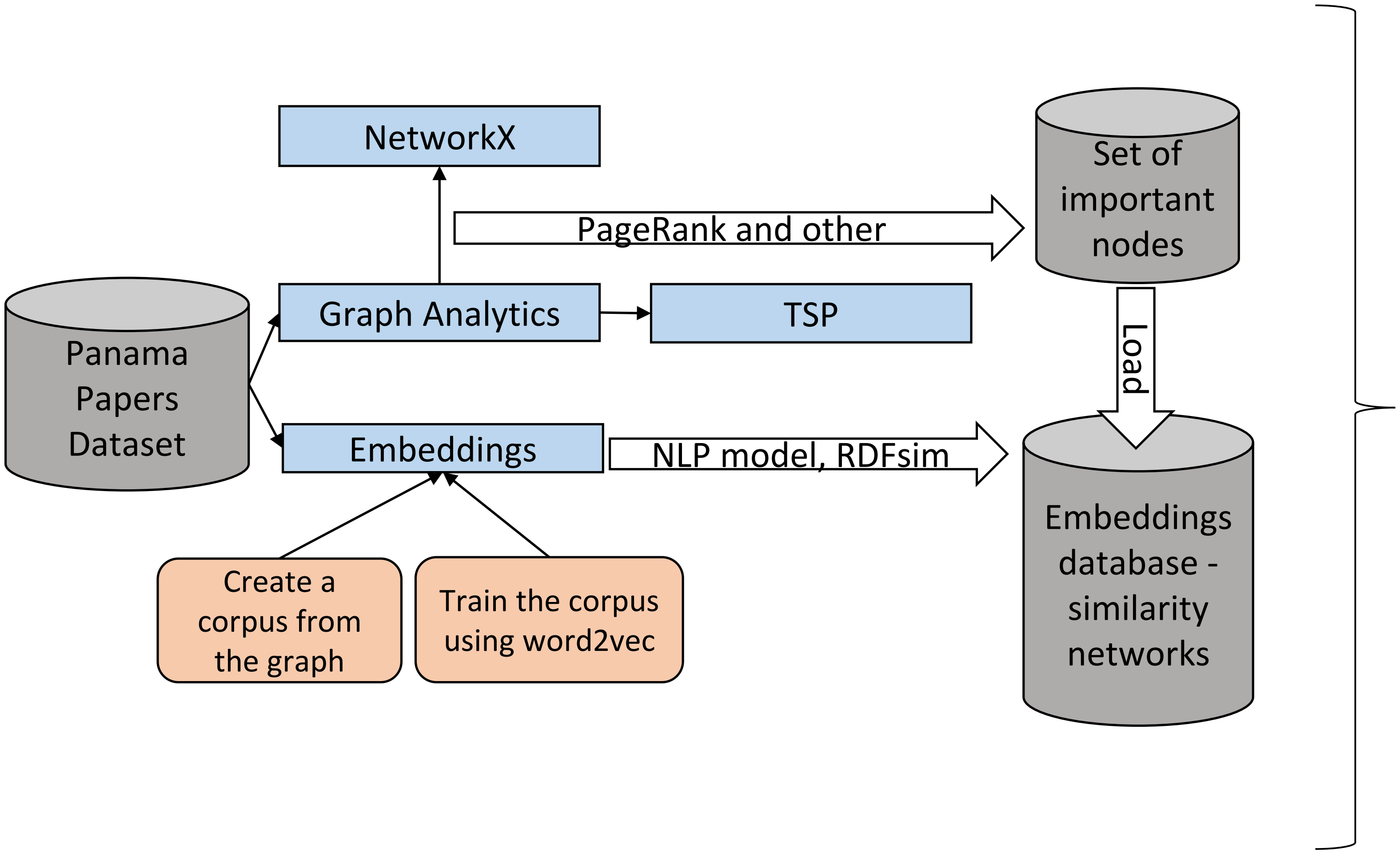


Approach

We propose an approach that **combines** the **results** of **graph analytics** with a **graph embeddings database**, in order to create **similarity networks** of the **most important nodes** of our data, so as to discover relationships between nodes that could not be easily seen before!



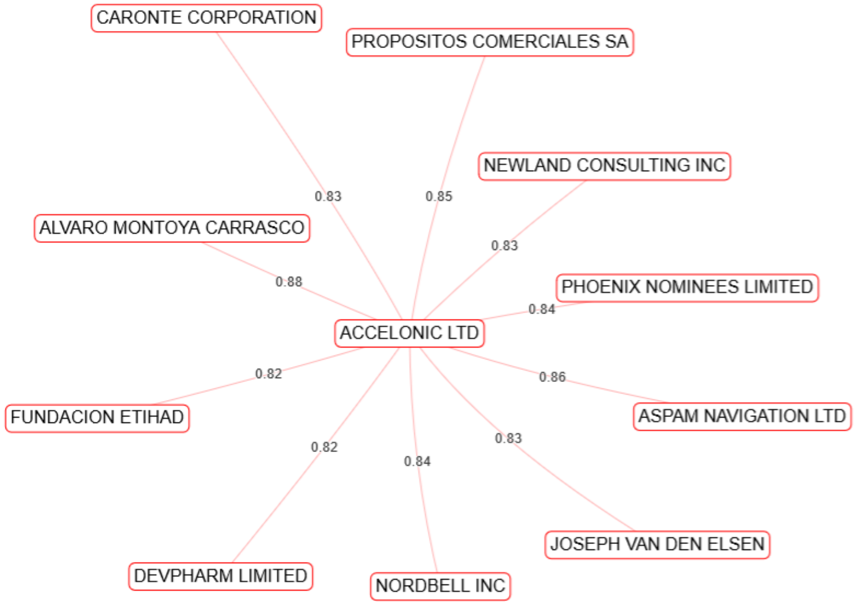
Work Outline

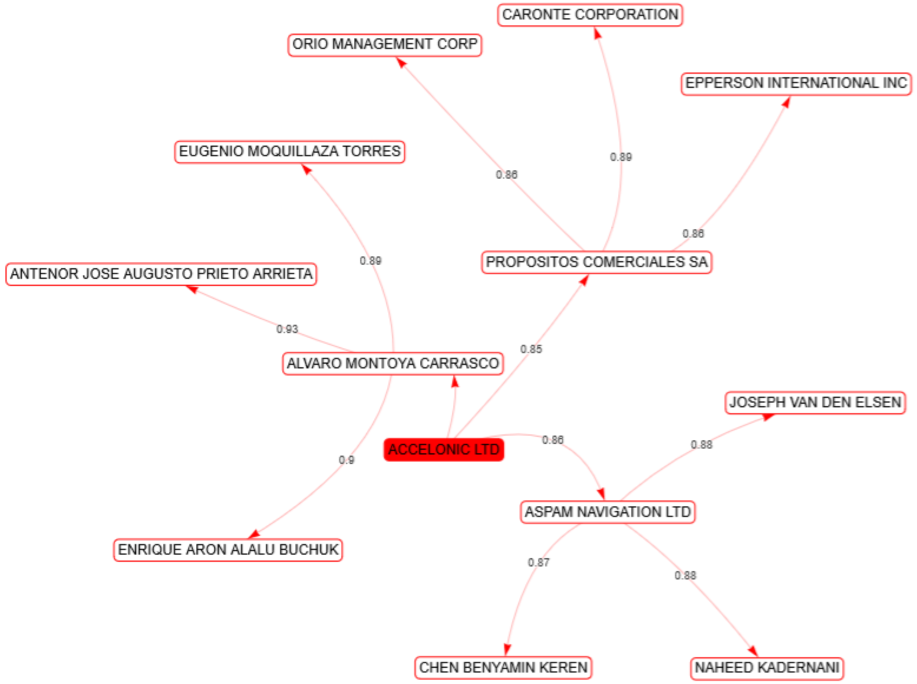


Combine **Graph Analytics** with **Embeddings**!

An early example

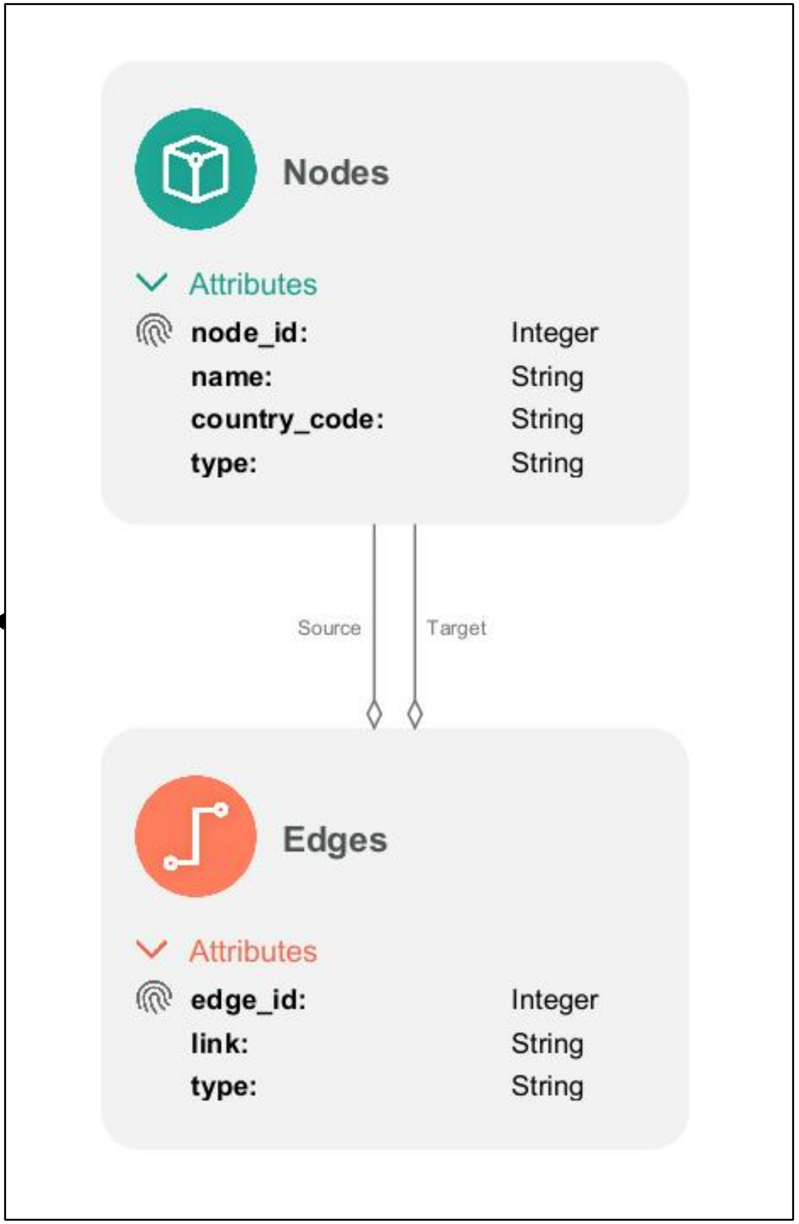
By running **PageRank** and **Eigenvector Centrality**, we discovered that the node with the highest value is **ACCELONIC LTD.** which is an offshore located in Hong-Kong. We could use this entity to locate the **similar nodes** using the graph embeddings database





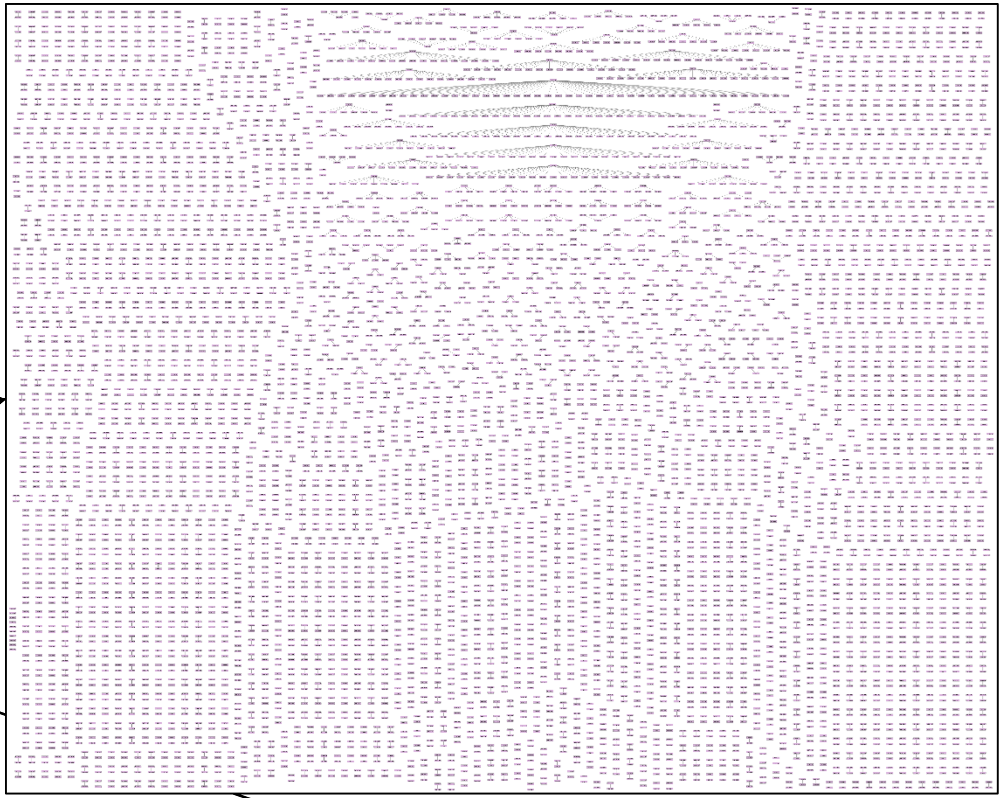
Graph Analytics: TSP

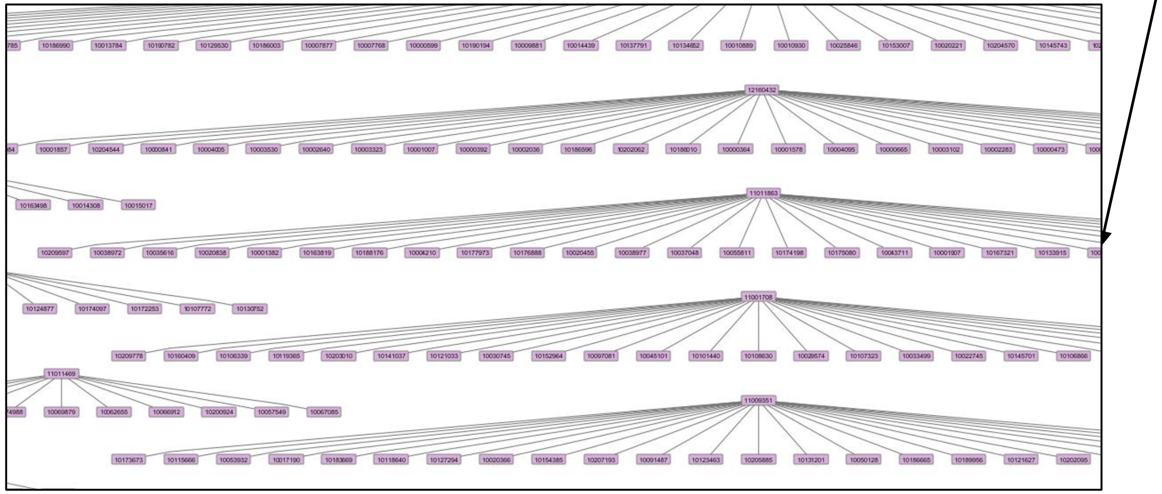
- Given the enormous dataset size, we decided to **explore methods to load subsets of the dataset, without losing information** about the graph:
  - **Simple Subset Chunking**: Selects **N sequential edges** and builds the graph by searching for the corresponding nodes.
  - **Sampling:** Selects **N random edges** and builds the corresponding graph by locating the appropriate nodes.
  - **Reconstruction:** BFS-like approach that begins from a number of source nodes and recreates their direct network by visiting the neighbors.
- **Sampling method** gave the best results, thus we decided to use it to exploit the features of TSP.
- We decided to use a simple schema, **containing only nodes and edges**.

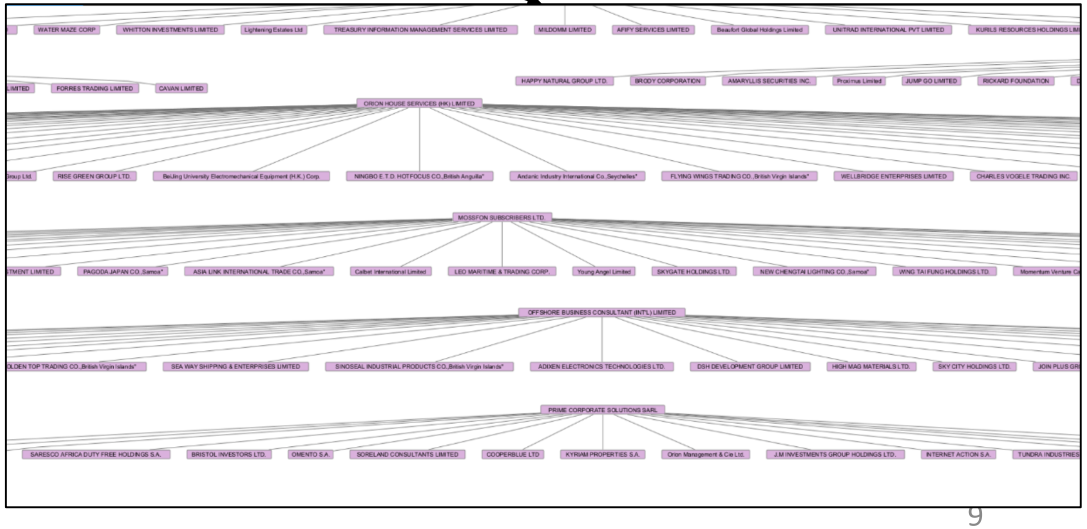


TSP Subgraph Visualization

- We visualized a sampled subset of **4000 edges**.
- By using **reconstruction** and **sampling**, we discovered that the graph consists of several **<u>smaller networks</u>** of nodes and their neighbors.
- We used the TSP previewer and **visualized** the subgraph.
  - We also tried to use the reconstruction method for visualization by selecting a number of random source nodes, but the results were the same.



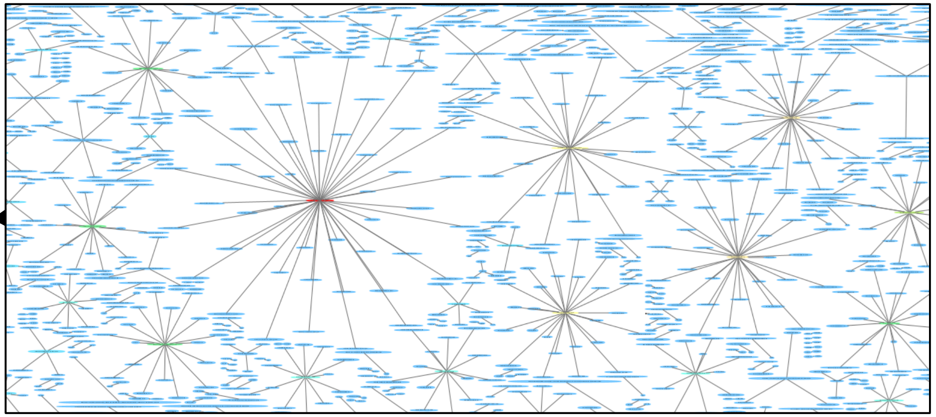


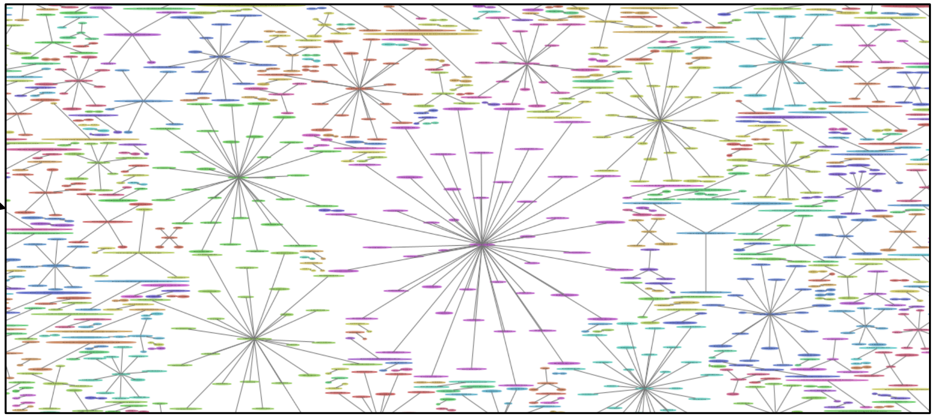


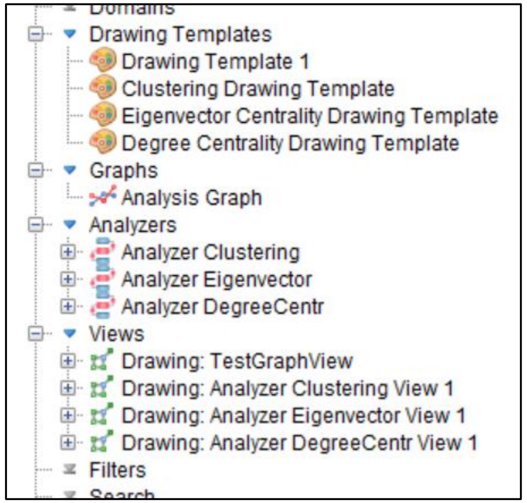
WATER MAZE CORP
WHITTON INVESTMENTS LIMITED
Lightening Estates Ltd
TREASURY INFORMATION MANAGEMENT SERVICES LIMITED
MILDOMM LIMITED
AFIFY SERVICES LIMITED
Beaufort Global Holdings Limited
UNITRAD INTERNATIONAL PVT LIMITED
FORRES TRADING LIMITED
CAVAN LIMITED
HAPPY NATURAL GROUP LTD.
BRODY CORPORATION
AMARYLLIS SECURITIES INC.
Proximus Limited
JUMP GO LIMITED
RICKARD FOUNDATION
ORION HOUSE SERVICES (HK) LIMITED
RISE GREEN GROUP LTD.
BeiJing University Electromechanical Equipment (H.K.) Corp.
NINGBO E.T.D. HOTFOCUS CO.,British Anguilla"
Andanic Industry International Co.,Seychelles"
FLYING WINGS TRADING CO.,British Virgin Islands"
WELLBRIDGE ENTERPRISES LIMITED
CHARLES VOGELE TRADING INC.
MOSSFON SUBSCRIBERS LTD.
PAGODA JAPAN CO.,Samoa"
ASIA LINK INTERNATIONAL TRADE CO.,Samoa"
Calbet International Limited
LEO MARITIME & TRADING CORP.
Young Angel Limited
SKYGATE HOLDINGS LTD.
NEW CHENGTAI LIGHTING CO.,Samoa"
WING TAI FUNG HOLDINGS LTD.
OFFSHORE BUSINESS CONSULTANT (INT'L) LIMITED
SEA WAY SHIPPING & ENTERPRISES LIMITED
SINOSEAL INDUSTRIAL PRODUCTS CO.,British Virgin Islands"
ADIXEN ELECTRONICS TECHNOLOGIES LTD.
DSH DEVELOPMENT GROUP LIMITED
HIGH MAG MATERIALS LTD.
SKY CITY HOLDINGS LTD.
PRIME CORPORATE SOLUTIONS SARL
SARESCO AFRICA DUTY FREE HOLDINGS S.A.
BRISTOL INVESTORS LTD.
OMENTO S.A.
SORELAND CONSULTANTS LIMITED
COOPERBLUE LTD
KYRIAM PROPERTIES S.A.
Orion Management & Cie Ltd.
J.M INVESTMENTS GROUP HOLDINGS LTD.
INTERNET ACTION S.A.

TSP Graph Analytics

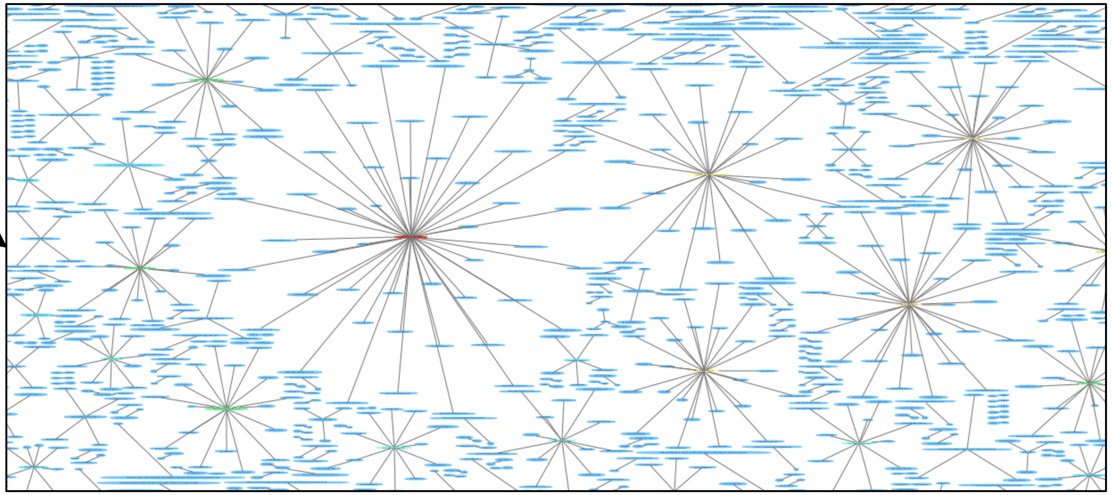
- We used the analyzers of TSP in order to run **Eigenvector centrality**, **Degree centrality** and **Clustering.**
- Eigenvector centrality: We discovered that the most popular node is ***ORION HOUSE SERVICES LIMITED, MOSSACK FONSECA CO.*** and ***OFFSHORE BUSINESS CONSULTANT (INT'L) LIMITED*****.**
- Clustering: We discovered that the most popular node is ***ORION HOUSE SERVICES LIMITED*** and ***MOSSACK FONSECA & CO***.
- Degree Centrality: We discovered that the most popular node is ***ORION HOUSE SERVICES LIMITED*** and ***MOSSFON SUBSCRIBERS LTD*****.**







- Domains
- Drawing Templates
  - Drawing Template 1
  - Clustering Drawing Template
  - Eigenvector Centrality Drawing Template
  - Degree Centrality Drawing Template
- Graphs
  - Analysis Graph
- Analyzers
  - Analyzer Clustering
  - Analyzer Eigenvector
  - Analyzer DegreeCentr
- Views
  - Drawing: TestGraphView
  - Drawing: Analyzer Clustering View 1
  - Drawing: Analyzer Eigenvector View 1
  - Drawing: Analyzer DegreeCentr View 1
- Filters
- Search

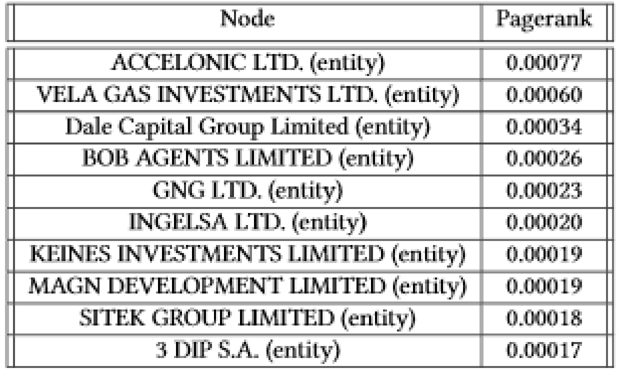


Graph Analytics: NetworkX

- NetworkX is an open source Python API which was able to **load the dataset completely** and run the algorithms we selected.
  - We created a **custom csv dataset parsing script** to load the data and create **mappings** (nodeID, nodeName) using dictionaries.
  - We also implemented some **wrapper functions** for the built-in algorithms of NetworkX to get the results from the simulations.
- We decided to **use the results of NetworkX**, as these results are extracted from the whole dataset.
- The algorithms we used were **PageRank**, **Eigenvector Centrality**, **Degree Centrality** and **Clustering**, as they are the most applicable for locating “popular” nodes in a Fraud Detection graph. After the simulation of the algorithms, we kept only the top results.
- Given that through this analysis we mostly care about the entities, the officers and the intermediaries, we decided to **rule out** the results that were about **nodes representing addresses**.

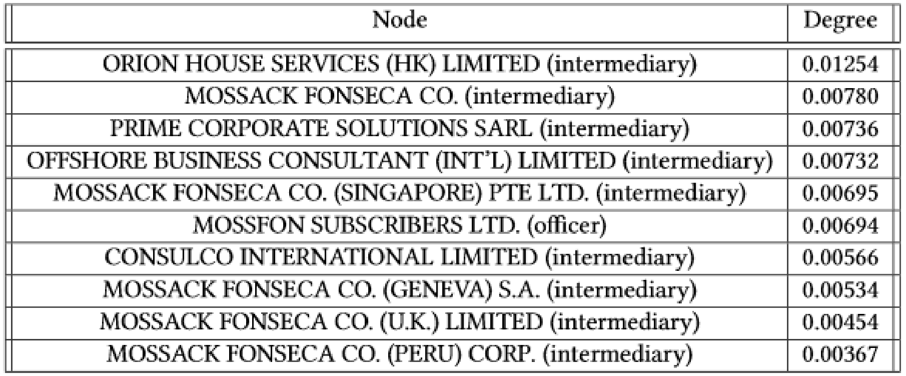
NetworkX Results

- PageRank(*a=0.85,it=100*)



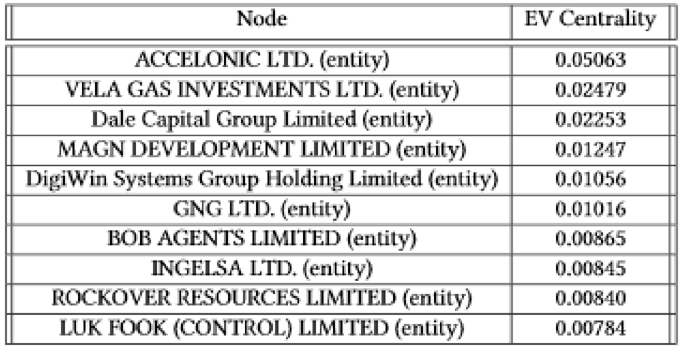
Node	Pagerank
ACCELONIC LTD. (entity)	0.00077
VELA GAS INVESTMENTS LTD. (entity)	0.00060
Dale Capital Group Limited (entity)	0.00034
BOB AGENTS LIMITED (entity)	0.00026
GNG LTD. (entity)	0.00023
INGELSA LTD. (entity)	0.00020
KEINES INVESTMENTS LIMITED (entity)	0.00019
MAGN DEVELOPMENT LIMITED (entity)	0.00019
SITEK GROUP LIMITED (entity)	0.00018
3 DIP S.A. (entity)	0.00017

- Degree Centrality



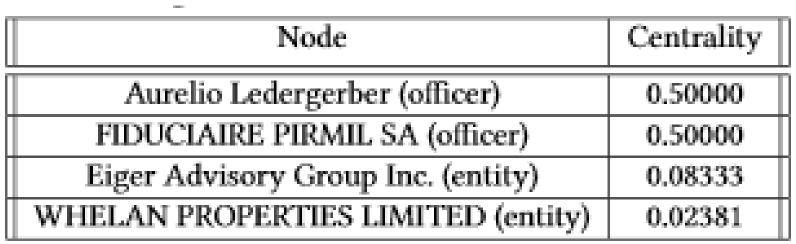
Node	Degree
ORION HOUSE SERVICES (HK) LIMITED (intermediary)	0.01254
MOSSACK FONSECA CO. (intermediary)	0.00780
PRIME CORPORATE SOLUTIONS SARL (intermediary)	0.00736
OFFSHORE BUSINESS CONSULTANT (INT'L) LIMITED (intermediary)	0.00732
MOSSACK FONSECA CO. (SINGAPORE) PTE LTD. (intermediary)	0.00695
MOSSFON SUBSCRIBERS LTD. (officer)	0.00694
CONSULCO INTERNATIONAL LIMITED (intermediary)	0.00566
MOSSACK FONSECA CO. (GENEVA) S.A. (intermediary)	0.00534
MOSSACK FONSECA CO. (U.K.) LIMITED (intermediary)	0.00454
MOSSACK FONSECA CO. (PERU) CORP. (intermediary)	0.00367

- Eigenvector Centrality



Node	EV Centrality
ACCELONIC LTD. (entity)	0.05063
VELA GAS INVESTMENTS LTD. (entity)	0.02479
Dale Capital Group Limited (entity)	0.02253
MAGN DEVELOPMENT LIMITED (entity)	0.01247
DigiWin Systems Group Holding Limited (entity)	0.01056
GNG LTD. (entity)	0.01016
BOB AGENTS LIMITED (entity)	0.00865
INGELSA LTD. (entity)	0.00845
ROCKOVER RESOURCES LIMITED (entity)	0.00840
LUK FOOK (CONTROL) LIMITED (entity)	0.00784

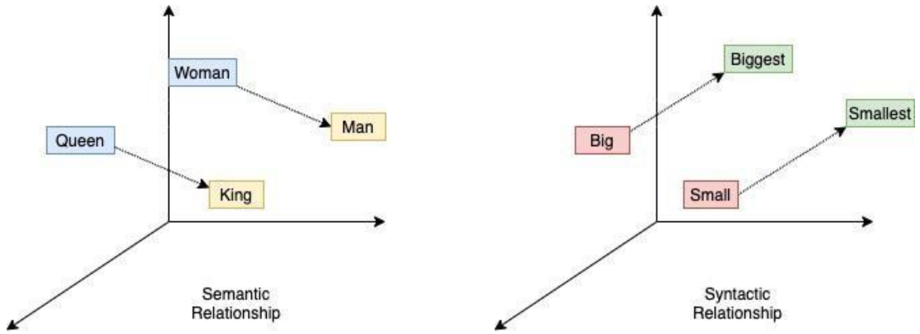
- Clustering

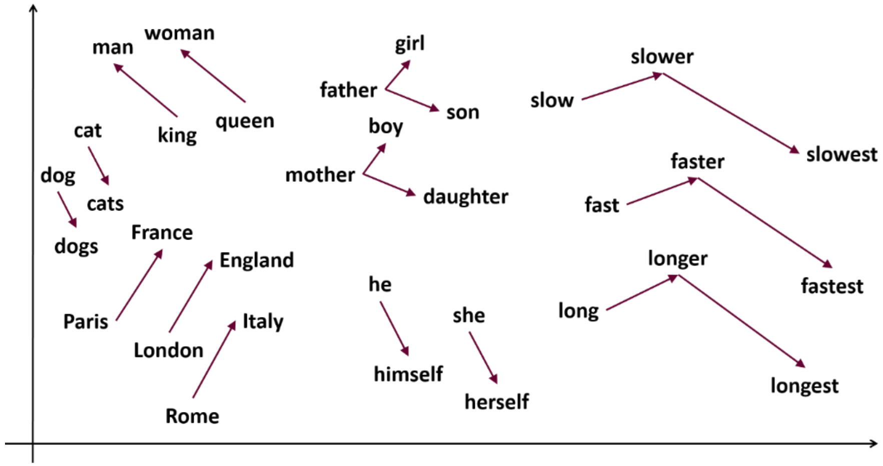


Node	Centrality
Aurelio Ledergerber (officer)	0.50000
FIDUCIAIRE PIRMIL SA (officer)	0.50000
Eiger Advisory Group Inc. (entity)	0.08333
WHELAN PROPERTIES LIMITED (entity)	0.02381

Graph Embeddings (General)

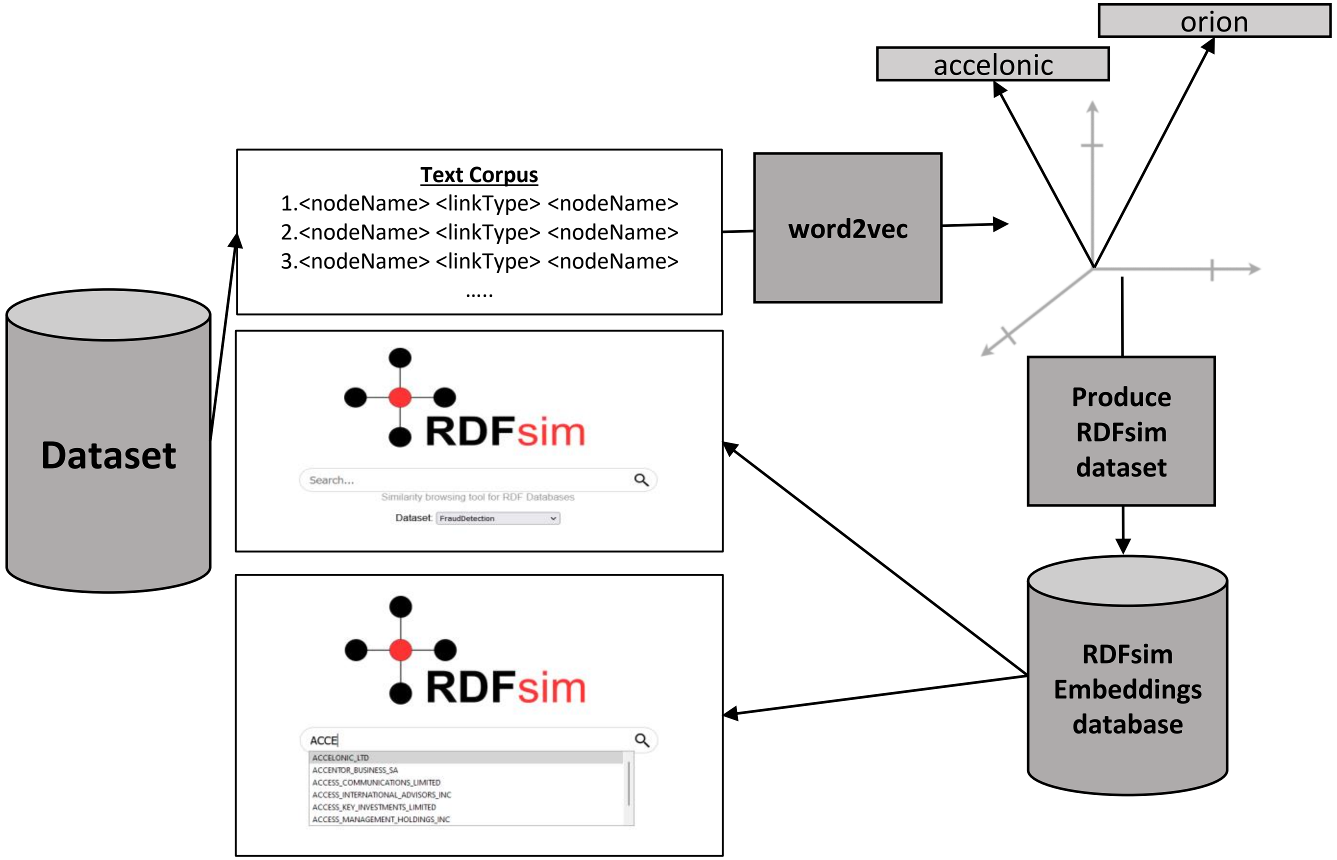
- Embeddings are **vector representations of words**.
- Libraries for embedding creation exploit **pre-trained neural networks** trying to group the words based on their semantic meaning. This means that based on the input, **words that have similar meaning will also have closer vectors**.
- The idea is to **transform the graph into text corpus**, by representing it as triples of the form (“nodeName” “linkName” “nodeName”). This will result in a vector for each node, and **nodes that share many same information will have closer vectors too**.
- This way, we can make operations like **similarity search** i.e. “**give me the first top-K” similar nodes** of the node N.





Graph Embeddings (RDFsim tuned for FD)

- By calculating all similar nodes of every node, we can create a **graph database of embeddings** in order to **develop similarity networks** by exploiting RDFsim.
- **RDFsim** is a **search engine** which is able to **use embedding datasets** and create **similarity networks**. Although it is designed for **knowledge graphs**, we managed to **port the Fraud Detection** data.
- This way we can use the search engine features of RDFsim (e.g. search for a specific node of the graph) and **create networks** of any kind and depth.



Text Corpus

1.<nodeName> <linkType> <nodeName>
2.<nodeName> <linkType> <nodeName>
3.<nodeName> <linkType> <nodeName>
.....

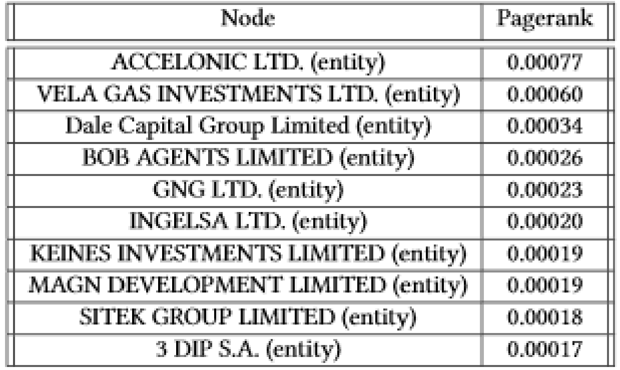
Dataset: FraudDetection

ACCE

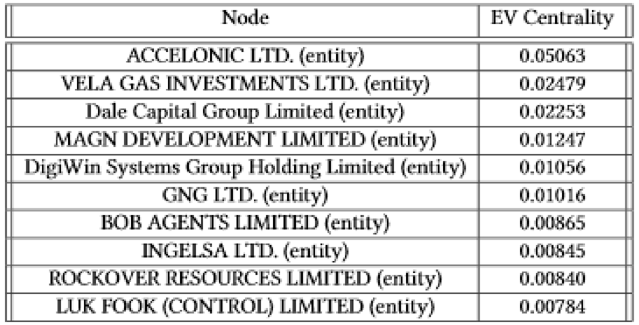
ACCELONIC_LTD
ACCENTOR_BUSINESS_SA
ACCESS_COMMUNICATIONS_LIMITED
ACCESS_INTERNATIONAL_ADVISORS_INC
ACCESS_KEY_INVESTMENTS_LIMITED
ACCESS_MANAGEMENT_HOLDINGS_INC

Graph Embeddings (Examples)

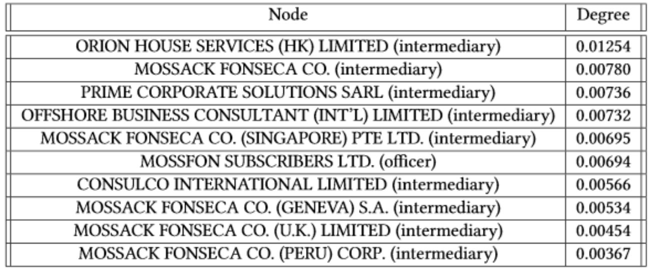
- For this presentation, we show the results about the **important nodes** from graph analytics.
- More about the vocabulary creation, embedding training and RDFsim can be found in the official paper, published in Sep. 2021.
- The nodes that we will explore are:
  - **ACCELONIC LTD**: Important entity with high PageRank and EV score.
  - **VELA GAS INVESTMENTS LTD**: Important entity with high PageRank and EV score.
  - **ORION HOUSE SERVICES (HK) LIMITED**: Important intermediary node with high degree centrality score.
  - Note: Although we selected only three of the important nodes from all the algorithms we simulated, the search engine can create the results **for any node existing in the database**.



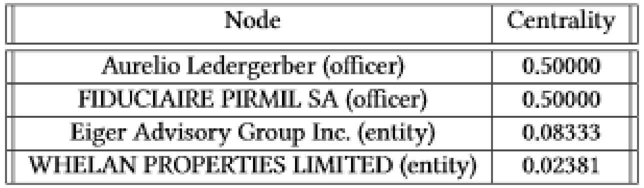
Node	Pagerank
ACCELONIC LTD. (entity)	0.00077
VELA GAS INVESTMENTS LTD. (entity)	0.00060
Dale Capital Group Limited (entity)	0.00034
BOB AGENTS LIMITED (entity)	0.00026
GNG LTD. (entity)	0.00023
INGELSA LTD. (entity)	0.00020
KEINES INVESTMENTS LIMITED (entity)	0.00019
MAGN DEVELOPMENT LIMITED (entity)	0.00019
SITEK GROUP LIMITED (entity)	0.00018
3 DIP S.A. (entity)	0.00017



Node	EV Centrality
ACCELONIC LTD. (entity)	0.05063
VELA GAS INVESTMENTS LTD. (entity)	0.02479
Dale Capital Group Limited (entity)	0.02253
MAGN DEVELOPMENT LIMITED (entity)	0.01247
DigiWin Systems Group Holding Limited (entity)	0.01056
GNG LTD. (entity)	0.01016
BOB AGENTS LIMITED (entity)	0.00865
INGELSA LTD. (entity)	0.00845
ROCKOVER RESOURCES LIMITED (entity)	0.00840
LUK FOOK (CONTROL) LIMITED (entity)	0.00784

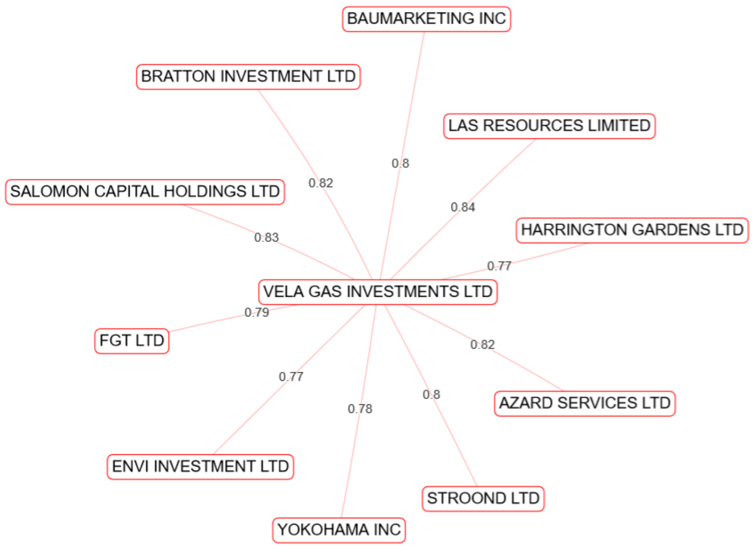


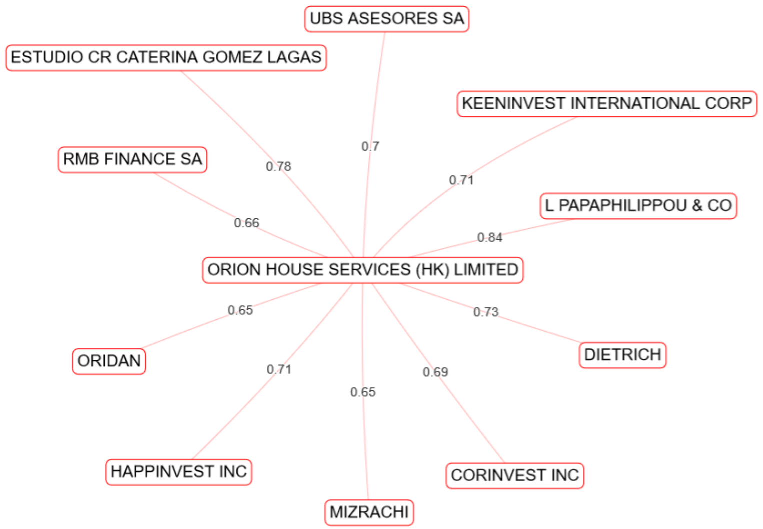
Node	Degree
ORION HOUSE SERVICES (HK) LIMITED (intermediary)	0.01254
MOSSACK FONSECA CO. (intermediary)	0.00780
PRIME CORPORATE SOLUTIONS SARL (intermediary)	0.00736
OFFSHORE BUSINESS CONSULTANT (INT'L) LIMITED (intermediary)	0.00732
MOSSACK FONSECA CO. (SINGAPORE) PTE LTD. (intermediary)	0.00695
MOSSFON SUBSCRIBERS LTD. (officer)	0.00694
CONSULCO INTERNATIONAL LIMITED (intermediary)	0.00566
MOSSACK FONSECA CO. (GENEVA) S.A. (intermediary)	0.00534
MOSSACK FONSECA CO. (U.K.) LIMITED (intermediary)	0.00454
MOSSACK FONSECA CO. (PERU) CORP. (intermediary)	0.00367

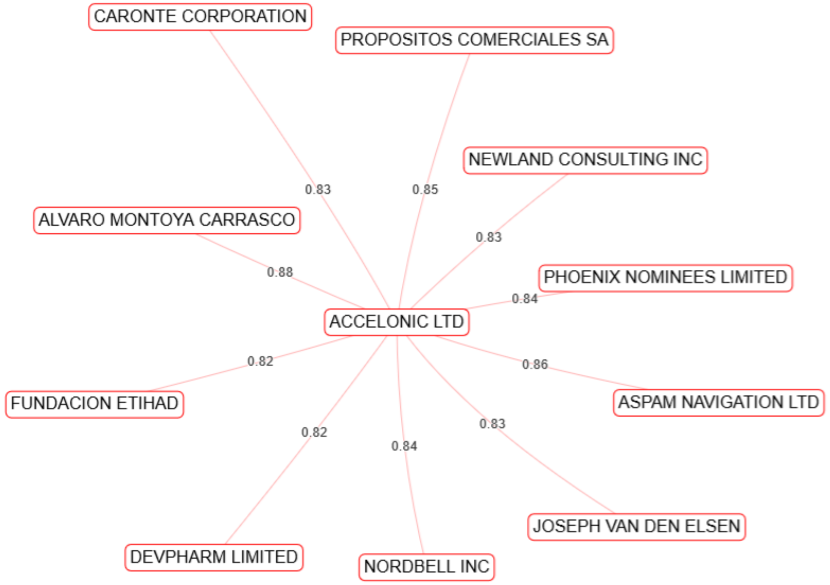


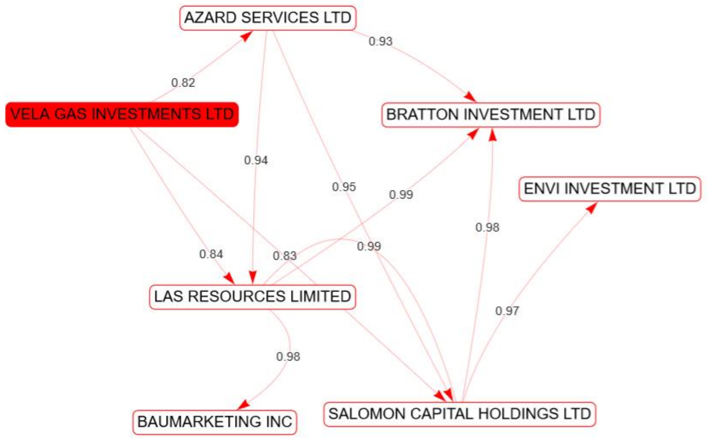
Node	Centrality
Aurelio Ledergerber (officer)	0.50000
FIDUCIAIRE PIRMIL SA (officer)	0.50000
Eiger Advisory Group Inc. (entity)	0.08333
WHELAN PROPERTIES LIMITED (entity)	0.02381

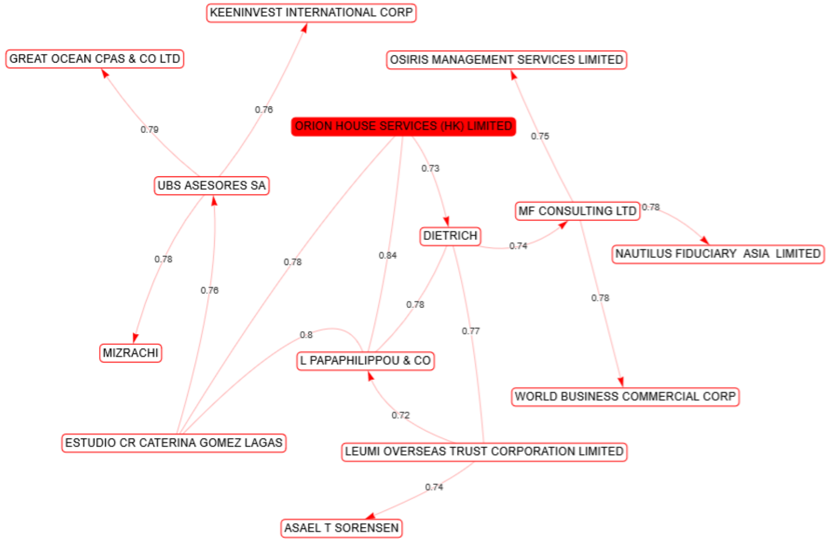
Examples (using RDFsim)

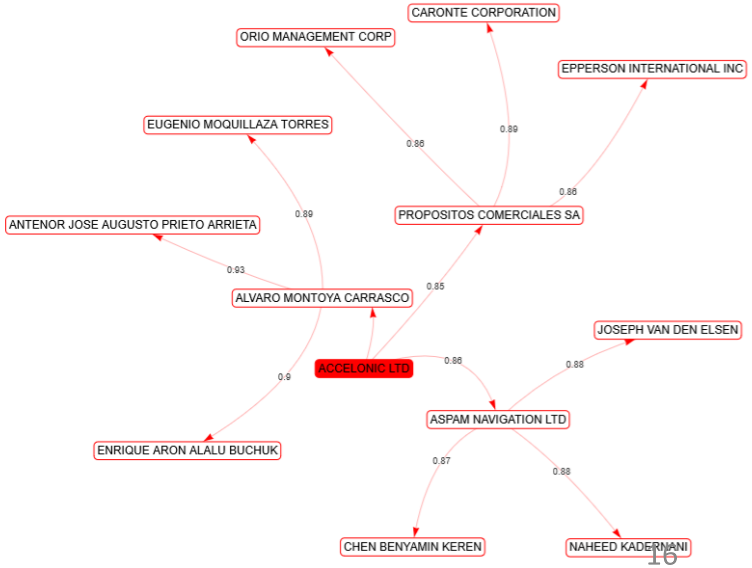









KEENINVEST INTERNATIONAL CORP
GREAT OCEAN CPAS & CO LTD
OSIRIS MANAGEMENT SERVICES LIMITED
ORION HOUSE SERVICES (HK) LIMITED
UBS ASESORES SA
MF CONSULTING LTD
DIETRICH
NAUTILUS FIDUCIARY ASIA LIMITED
MIZRACHI
L PAPAPHILIPPOU & CO
WORLD BUSINESS COMMERCIAL CORP
ESTUDIO CR CATERINA GOMEZ LAGAS
LEUMI OVERSEAS TRUST CORPORATION LIMITED
ASAEL T SORENSEN
0.76
0.79
0.75
0.73
0.78
0.74
0.78
0.78
0.84
0.76
0.78
0.78
0.77
0.8
0.72
0.74



Understanding the results

Lets look at the similarity network of ACCELONIC LTD.

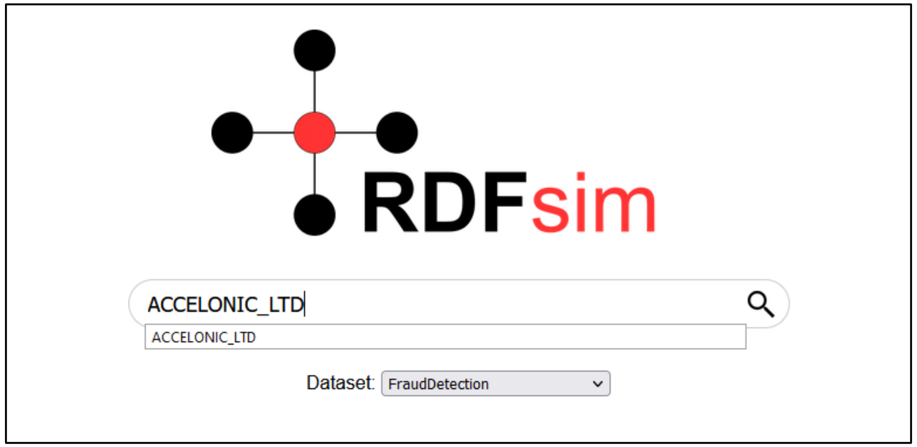
Using this network, **we can discover connections between offshores and other entities that were not clear before**.

For example, we see that our current node has high similarity score with **ASPAM NAVIGATION LTD.** *(the similarity score is based on how close are the corresponding vectors, i.e. it is the **cosine similarity**)* .

This result could mean that **these two nodes may have connections to similar entities, officers, companies and more**.

Given that our base data are dense and might have many connections between nodes, **it could be difficult to discover such connections**, or even worse, such connections **could not even exist in the starting dataset** (as a path or direct edge).

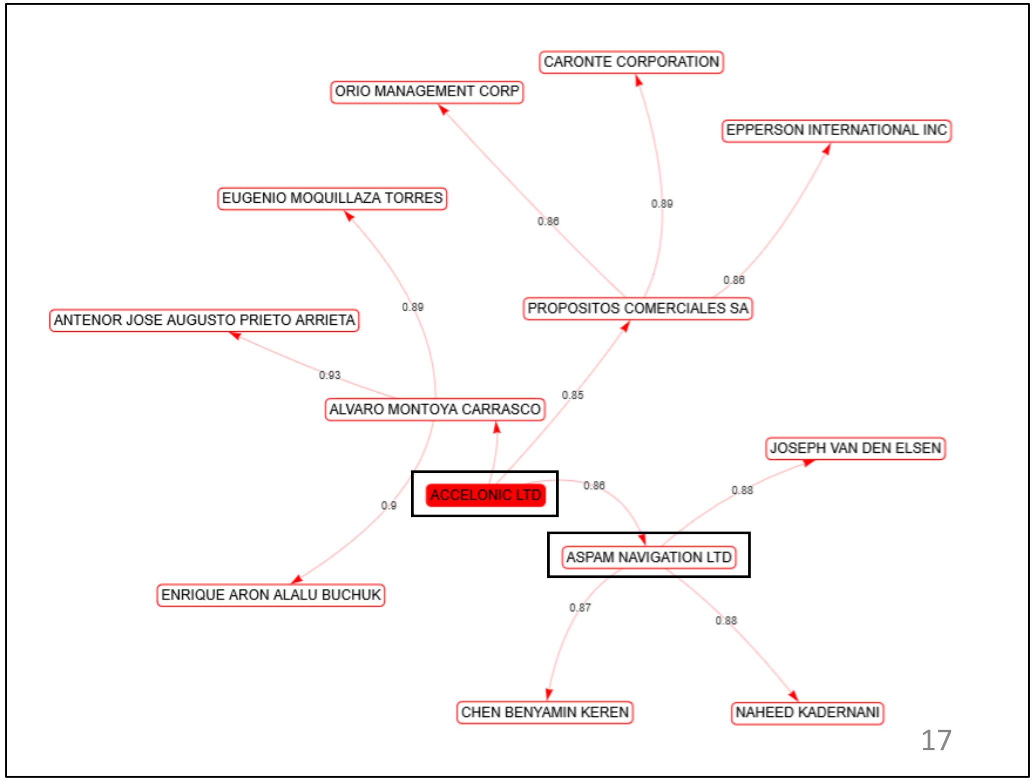
Exploiting graph embeddings could **offer a new way to discover relationships between the nodes of the graph, group offshores with same fraudulent activity etc**.



ACCELONIC_LTD

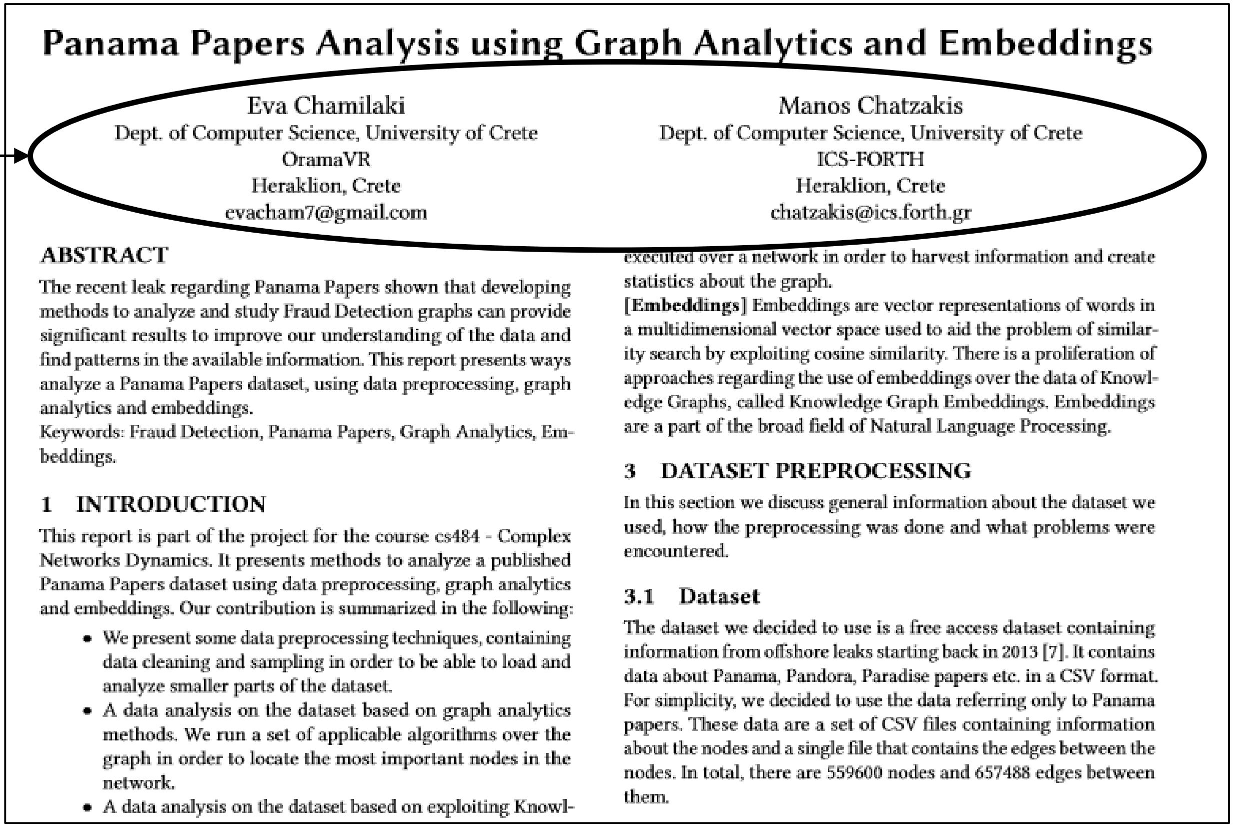
ACCELONIC_LTD

Dataset: FraudDetection



Conclusion

Authors are sorted by their **patience** ☺ ..



Panama Papers Analysis using Graph Analytics and Embeddings

Eva Chamilaki
Dept. of Computer Science, University of Crete
OramaVR
Heraklion, Crete
evacham7@gmail.com

Manos Chatzakis
Dept. of Computer Science, University of Crete
ICS-FORTH
Heraklion, Crete
chatzakis@ics.forth.gr

ABSTRACT

The recent leak regarding Panama Papers shown that developing methods to analyze and study Fraud Detection graphs can provide significant results to improve our understanding of the data and find patterns in the available information. This report presents ways analyze a Panama Papers dataset, using data preprocessing, graph analytics and embeddings.

Keywords: Fraud Detection, Panama Papers, Graph Analytics, Embeddings.

1 INTRODUCTION

This report is part of the project for the course cs484 - Complex Networks Dynamics. It presents methods to analyze a published Panama Papers dataset using data preprocessing, graph analytics and embeddings. Our contribution is summarized in the following:

- We present some data preprocessing techniques, containing data cleaning and sampling in order to be able to load and analyze smaller parts of the dataset.
- A data analysis on the dataset based on graph analytics methods. We run a set of applicable algorithms over the graph in order to locate the most important nodes in the network.
- A data analysis on the dataset based on exploiting Knowl-

executed over a network in order to harvest information and create statistics about the graph.

[Embeddings] Embeddings are vector representations of words in a multidimensional vector space used to aid the problem of similarity search by exploiting cosine similarity. There is a proliferation of approaches regarding the use of embeddings over the data of Knowledge Graphs, called Knowledge Graph Embeddings. Embeddings are a part of the broad field of Natural Language Processing.

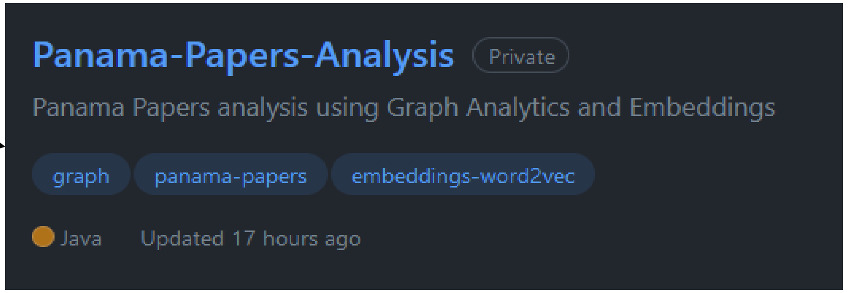
3 DATASET PREPROCESSING

In this section we discuss general information about the dataset we used, how the preprocessing was done and what problems were encountered.

3.1 Dataset

The dataset we decided to use is a free access dataset containing information from offshore leaks starting back in 2013 [7]. It contains data about Panama, Pandora, Paradise papers etc. in a CSV format. For simplicity, we decided to use the data referring only to Panama papers. These data are a set of CSV files containing information about the nodes and a single file that contains the edges between the nodes. In total, there are 559600 nodes and 657488 edges between them.

- We presented an analysis which **combines graph analytics** with **embeddings**.
- The analysis resulted in a **set of important nodes** over the network and the **creation of** a Fraud Detection Graph **Embeddings Database**.
- An extended analysis of our work including more figures, algorithms and explanations is available in the **repository** of our project.



Panama-Papers-Analysis Private

Panama Papers analysis using Graph Analytics and Embeddings

graph panama-papers embeddings-word2vec

Java Updated 17 hours ago

Thank You!

Any Questions?