

Graph Based Analysis of Panama Papers

Sakshi Srivastava
Department of CSE
MNNIT Allahabad
India

Anil Kumar Singh
Department of CSE
MNNIT Allahabad
India

Abstract—Graphs have become increasingly important in many applications and domains such as querying relationships in social networks or managing rich meta data generated in scientific computing. The Panama Papers(Unprecedented leak of 11.5m data from the database of the worlds fourth biggest offshore law firm, Mossack fonseca) investigation used Neo4j as a key technology to find connections in leaked data.The International consortium of Investigative Journalists is a global network, which is used to analyze massive leaks over the world.This work attempts to analyze a subset of data from panama papers.The data is modeled as graph database using Neo4J for analysis of Panama Papers which results fraud ring over offshore entities.

Index Terms—panama paper, graph database, fraud detection, page rank, fraud detection algorithm, offshore entity, fraud ring, investigative journalist, international consortium

I. INTRODUCTION

Graph are pervasive whose volume and heterogeneity are strongly growing wall to wall to analyze and manage huge enormous amount of graph data which is applicable over globally defined application.The Panama papers allude to a massive leak of information enclosing more than 11 million documents consist of more than 2,50,000 offshore entity originated in Panama city [1] [6]. The Analysis and Management of "Panama Papers" entities and relationship using Neo4j software is propulsive force for fraud detection over leak data set. Innumerable wealthy individual including government officials from various countries, have been suspected. The sudden release of document or suspicious information has led to various lawsuit and resignation [3]. The Lawyer and government agencies of various countries continue to read and trim out the information as at the end of the legal proceeding was not insight. Although having offshore entities is not an illegal activity but such entities have been used to evade income tax, launder moved, hide assets and evade sanction. The show stopper was John Doe who leaked the information to Suddeutsche Zeitung, a German news paper, he leaked the information because of his own strong disfavour for income inequality and because of his own strong disfavour and because of his view regarding same of entities and individuals who are facing injustice. The International consortium of Investigation journalist (ICIJ) aided in the document dissemination. Journalist from 107 media outlet in 80 countries seek out the document and started publishing different articles [2] about them and also released some documents, which triggered a war of nerves among the media.

A. The System

The structure of leaked data is as follows : Mossack Fonseca created a folder for each carapace firm, where each folder consist of emails, contacts, scanned documents and transcript. In some case there are various documents consisting thousand pages. Firstly, the data had to be sequentially indexed through which searching become easier in this sea of information Suddeutsche Zeitung used Nuix, used the same no of program in the end first step that is used by International investigator. ICIJ and Suddeutsche Zeitung uploaded million of different document on high performance Computer and applied (OCR) Optical character recognition that is used to mutate data into machine readable format and which makes searching of information or data more easier by using OCR produce the process convert image such as signed contracts and scanned Id's into searchable text. The step enables journalist to seek out through a wide part of the leak by using simple search must be similar to Google. International criminals, politicians and well known professional athletes, among others are listed by journalist compilation searching the leak for the names which was in the list is possibly done using digital processing. Powerful search algorithm compared the list of documents consisting 11.5 million document set.

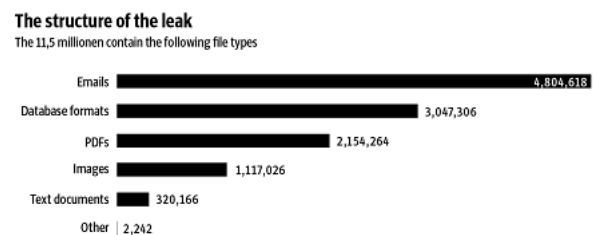


Fig. 1. Leak structure

B. Database

The Panama Papers used Neo4j to design graph database the naively embraces relation between various named vertices, which is used to process, store and query connections efficiently. In a native graph database acquiring nodes and relationship between them is an efficient and a constant time operation and allow traversal of million of connection per second. It is independent on total size of data set ,graph databases at the level of managing complex queries and highly connected data.

C. Neo4j

Neo4J's graph platform is specifically reduced to map, analyze, store and traverse networks of connected data that reveals hidden relationship and invisible facts. Neo4J is the world's leading and trending graph database. It is a high performance graph storage consisting all the properties of a robust and mature database as like a ACID transaction and an efficient query language. Programmer working with this flexible network of nodes and relationship enjoys all the benefits of enterprise quality graph database instead of static table formation. Neo4J propose high degree of magnitude performance benefits as compared relational database results in many application.

D. Cypher Query

Cypher is defined as the property graph query language which is used commercially by different multinational companies and researchers working under data mining [8]. Cypher uses ASCII method to represent different patterns. Nodes(e:entity) and Relationship(r:relationship) in cypher queries works as a key system for different complex queries.

II. MODELING

The raw data is imported and modeled as graph as shown below:



Fig. 2. Process Flow

1) *Raw Data*: The International Investigative Journalist (ICIJ) reveals a searchable database of entity, officers, intermediary and address named nodes in the Panama papers. Additionally in the online search tool, the ICIJ also broadcasted downloadable version of the same information on its site. Downloaded data set is given as follows.

Name	Date Modified	Kind
Addresses.csv	May 13, 2016, 10:14 AM	comma-separated values
all_edges.csv	May 13, 2016, 10:14 AM	comma-separated values
Entities.csv	May 13, 2016, 10:14 AM	comma-separated values
Intermediaries.csv	May 13, 2016, 10:14 AM	comma-separated values
Officers.csv	May 13, 2016, 10:14 AM	comma-separated values

Fig. 3. Dataset

A. Importing Procedure

When we are using data set in the form of CSV files for loading a database each node must have a unique node identifiers for establishing relationship between nodes in the same process. In the following example node identifiers are stored as properties of every node. In Neo4J node identifiers may heed later for cross references to another systems, traceability etc. After a completed import if it don't want the identifier to preserve then it should not specify a name of properties in the field(:ID). When calling Neo4j admin

import, it is possible to import only nodes using the import-tool, but for establishing relationships among imported nodes will have to be created later as the initial tool works only for initial graph population.

B. Graph Database

The Panama papers designed over graph database that inherently embraces relationships, which is able to store process and query connection efficiently [5]. Graph database powered by Neo4J that defines a structure of data as nodes(the node icon seen in the visualization) and the relationships(the links between those node icons).

C. ICIJ Data-structure

ICIJ received a raw data from Mosack Fonseca leaked dataset, which have been passed over OCR(Optical character recognition)tool to make in readable format. The data set observed contains both structured(CSV file) and unstructured data(JSON file). The JSON data is normalized and nodes and edges data is given in different CSVs. Python notebook is used to load and normalize the JSON data and Play around with it.



Fig. 4. Data Structure

III. METHODOLOGY

The process used to analyze the Panama Papers network, are perform in the following steps:

1) *Data Preparation:* The ICIJ (International consortium of investigative journalist) has unveiled a highly connected network of offshore tax structure used all over the world [4]. These structure of Panama Papers were unveiled from leaked financial documents and were analyzed by the journalists [7]. The first step to build a graph model is to extract those named entities from the documents and their meta-data. This includes entity, officer, intermediary and addresses. These entities become nodes in the graph and then extract relationship between nodes.

2) Importing Nodes :-

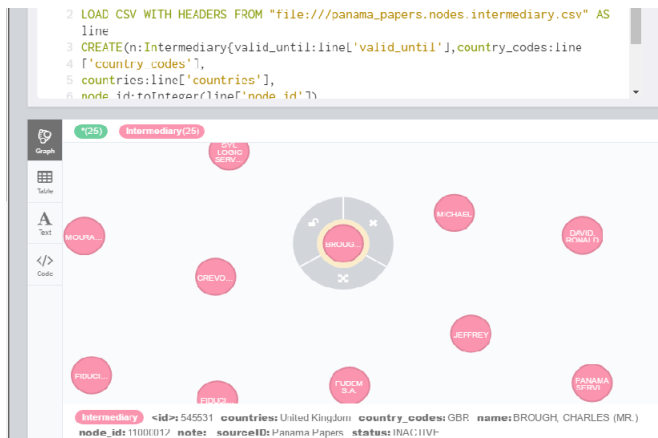


Fig. 5. Intermediary
Similarly other nodes(officer,address and Intermediary) have been imported.

3) Importing Relationship: -

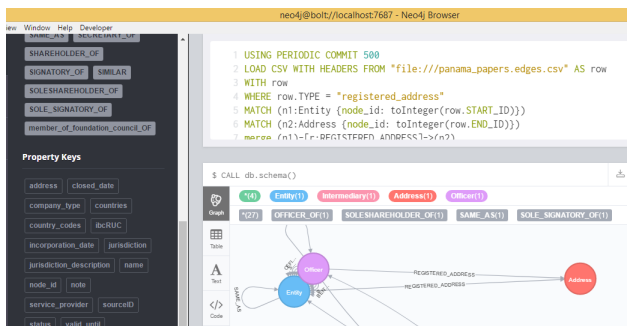


Fig. 6. Relationship between nodes
Similarly other relations have been imported using unique identities between nodes.

4) *Modeling:* -The domain model used by the ICIJ is really basic, just containing four types of entities (Officer, entity, intermediary, Address) and relationships between them.

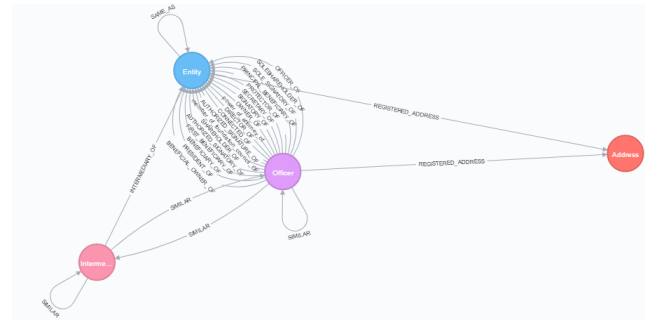


Fig. 7. Modeling

A graph database framework, or simply graph database, is a framework particularly intended for overseeing diagram like information following the fundamental standards of database frameworks. The graph databases are picking up importance in the business because of their utilization in a few areas where charts and system investigation are required. Prevalent chart databases are Neptune [9], Cosmos, Neo4j and Titan .The essential reflection behind a database framework is its database demonstrate. A Database Model ought to characterize three primary parts: an arrangement of information structure types (i.e. the information show), an arrangement of inquiry administrators or deduction rules (i.e. the question dialect), and an arrangement of respectability rules. With regards to diagram databases, a Graph Database Model is where information structures for the pattern or potentially cases are demonstrated as charts (or speculations of them), the information control is communicated by chart arranged activities (i.e. a chart inquiry language), and suitable uprightness limitations are characterized over the diagram structure.

A. Data Extraction

Analysis of data globally releases by International consortium of investigative journalist (ICIJ). The "ICIJ offshore" database defines the web of a relationships between companies and individual people of offshore companies based on tax haven [4]. The analysis of Panama papers leak consist of directed an unweighted network based on commercial registration of all companies involved in tax haven scandal and the relationship between those involved entities are: "director of" refreshing to the person appointed to the company management address through which it was so easy to establish the company origin of company "shareholder of" which holds a stack in an offshore company "intermediary of". If it mediates companies in access to offshore similar of company. If it intercede companies in access to offshore company among other attributes. The first version of database was released in june 2013, having since then more sources and informal are incorporated. The original data set war obtained through information leakage from law firm Mossack Fonseca which was obtained by an anonymous source. The role of secrecy the involved the relationship between companies located in tax haven and other offshore

entities was removed and other offshore intuit was removed and the real owner of the companies and their origin was identified. It has some inconsistency and omissions because of dealing with a database of information leakage which was not supported by official business work.

IV. ALGORITHM OVERVIEW

Graph algorithms are usually ment for computing metrics of graph's node and it's relationships. Neo4j Graph Algorithms is a library that provides coherently implemented, parallel versions of graph algorithms for Neo4j 3.x defined as Ciper methods. These algorithms are used to analyze graph database in Neo4J and to determine the importance of distinct nodes in a network:

A. Page Rank

Pang Rank algorithm individually applied to any network or domain which counts the quality or no of links that determines how important an entity is in a graph.

B. Betweenness Centrality

Betweenness centrality is used to discover nodes that works as a bridge from one part of a graph to another which results as elementary measure of the control that a node apply over the flow throughout graph.

C. Closeness Centrality

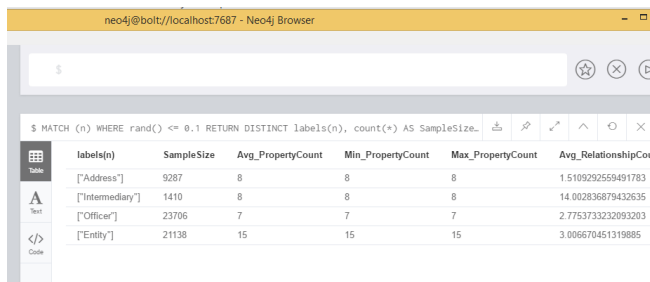
The Closeness centrality measure distance between one graph node to all other node and it's score defines how close a node is present i.e. minimum distance between nodes.

V. ANALYSIS

Initially to build graph model we need to extract those named entities from the documents and their meta-data. This includes entities, officer, intermediary and addresses. These entities become nodes in the graph. This similar logic to create relationships between entities that share the same address, have family ties or business relationships or that regularly communicate.

A. Analysis of Nodes and relationship

1) *No of nodes*: -It states no of nodes or entities in panama papers data set.

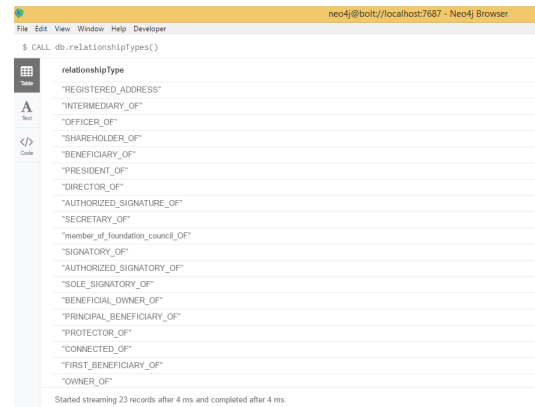


The screenshot shows the Neo4j Browser interface with a Cypher query: `$ MATCH (n) WHERE rand() <= 0.1 RETURN DISTINCT labels(n), count(*) AS SampleSize..` The results are displayed in a table with the following data:

labels(n)	SampleSize	Avg_PropertyCount	Min_PropertyCount	Max_PropertyCount	Avg_RelationshipCou
["Address"]	9287	8	8	8	1.5109292559491783
["Intermediary"]	1410	8	8	8	14.002836879432635
["Officer"]	23706	7	7	7	2.7753733232093203
["Entity"]	21138	15	15	15	3.006670451319885

Fig. 8. Nodes type:

2) *No of Relationship*: - It states total no of relationship established between different nodes having unique identifiers.



The screenshot shows the Neo4j Browser interface displaying a list of relationship types. The list includes: "REGISTERED_ADDRESS", "INTERMEDIARY_OF", "OFFICER_OF", "SHAREHOLDER_OF", "BENEFICIARY_OF", "PRESIDENT_OF", "DIRECTOR_OF", "AUTHORIZED_SIGNATORY_OF", "SECRETARY_OF", "member_of_foundation_council_OF", "SIGNATORY_OF", "AUTHORIZED_SIGNATORY_OF", "SOLE_SIGNATORY_OF", "BENEFICIAL_OWNER_OF", "PRINCIPAL_BENEFICIARY_OF", "PROTECTOR_OF", "CONNECTED_OF", "FIRST_BENEFICIARY_OF", and "OWNER_OF".

Fig. 9. Relationship type:

B. Queries over database

`MATCH (c:Entity)-[r]-(o:Officer) WHERE c.name = "Exaltation Limited" RETURN *`

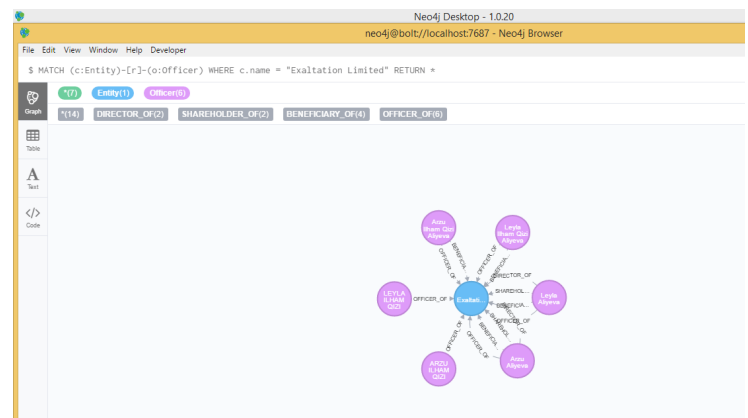
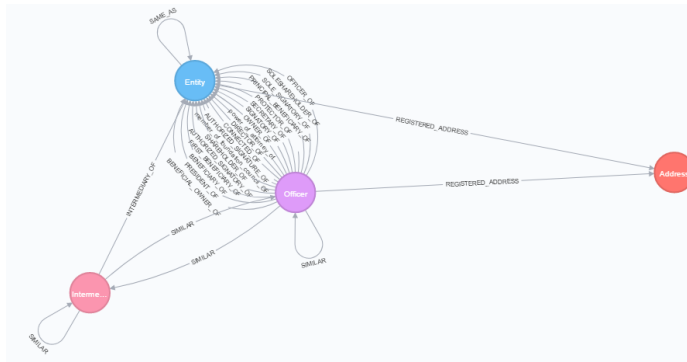


Fig. 10. Query Processing

C. Application of Algorithm

1) *Page-Rank of a node*: -Page Rank algorithm describe how important an entity is over whole database. As a result of which it gives score to no of entities who have been visited maximum no of times with other entities.

Paper graph, which showcase a fraud rings over the database schema that perceive officers(entity) who were working as an intermediary with synthetic identities forming fraud ring between offshore entities.



After Applying Fraud
detection algorithm

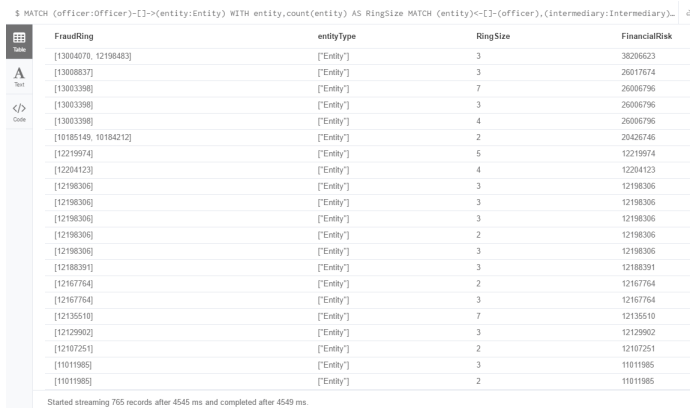


Fig. 15. Fraud ring

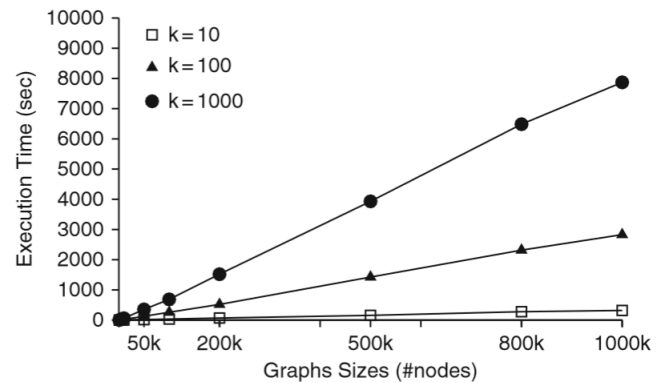
Above flow defines processing of analyzed data with fraud detection query results to be fraud risk over relationship

No_of_Objects	MySQL:S0	Neo4j:S0	MySQL:S1	Neo4j:S1	MySQL:S2	Neo4j:S2
100	19.56	8	33	12.65	111.34	19.57
500	281.38	10	333.96	17	620.56	21

VII. CONCLUSION & FUTURE SCOPE

Graph databases are more adaptable than Relational databases as new connections can be added to graph

databases without the need to rebuild the pattern once more. With such a distinction in the inquiry recovery time of MySQL and Neo4j, Neo4j can be utilized for business purposes like site interface structures and Social network. We have analyzed a Panama Papers using Neo4j and defined a fraud detection technique in search of synthetic identities and to deploy fraud rings efficiently. So, after analyzing one of the greatest Leak data set we can specialize the method by the statement "By using Graph data structure we can perceive the hidden facts over the Big data". The project can further be extended to design an algorithm which lead to greater scrutiny and fundamental redesigns of corporate security structures.



REFERENCES

- [1] Luke Harding. What are the panama papers? a guide to historys biggest data leak. The Guardian, 5, 2016.
- [2] Bastian Obermayer and Frederik Obermaier. The Panama papers: breaking the story of how the rich and powerful hide their money. Oneworld Publications, 2016.
- [3] James ODonovan, Hannes F Wagner, and Stefan Zeume. The value of offshore secretsevidence from the panama papers. 2016.
- [4] <https://www.icij.org/investigations/panama-papers/>
- [5] <https://neo4j.com/news/neo4j-powers-panama-papers-investigation/>
- [6] <https://datafloq.com/read/panama-papers-its-all-about-the-data/2072>
- [7] <https://qz.com/654027/the-five-most-important-graphs-from-these-panama-papers-leaks/>
- [8] Einstein, A., B. Podolsky, and N. Rosen, 1935, Can quantum-mechanical description of physical reality be considered complete?, Phys. Rev. 47, 777-780.
- [9] bookwebber2018programmatic.A programmatic introduction to neo4j, Webber, Jim and Robinson, Ian, Addison-Wesley Professional, 2018