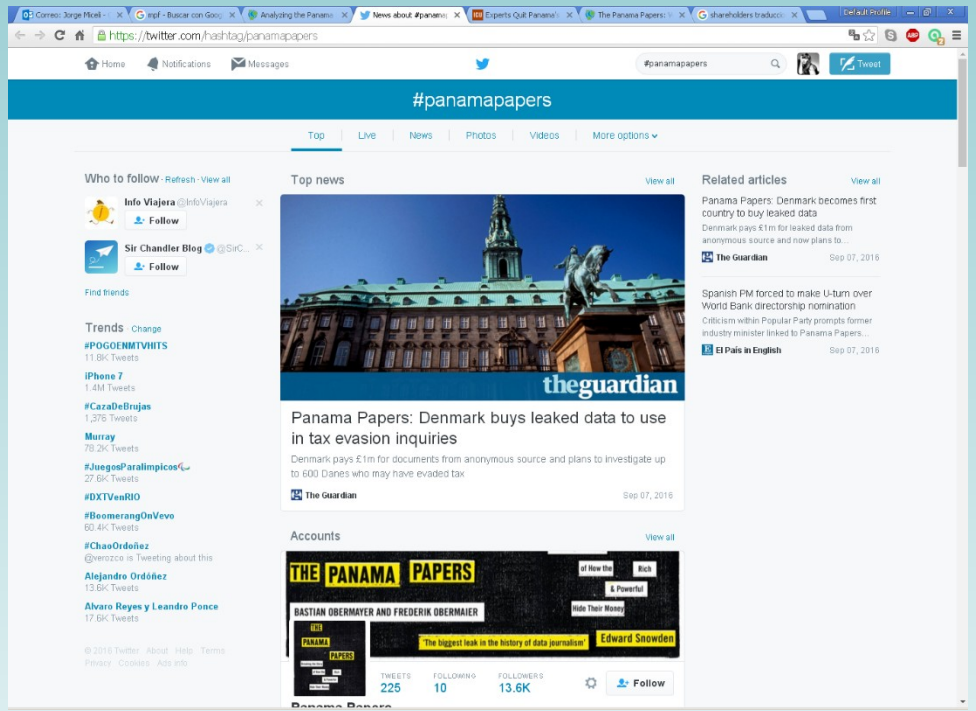
PANAMA PAPERS: redes, mentiras y offshore



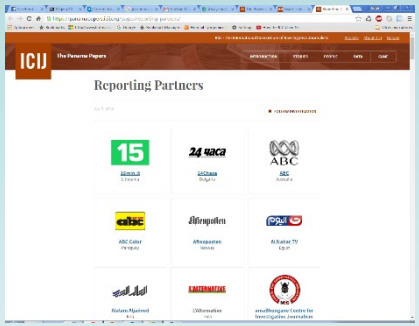
La trama principal

En “Panama papers” se exponen operaciones internas de una de las firmas legales mas importantes en la incorporación de empresas Offshore

MOSSACK FONSECA

2.6 terabytes de datos

Filtración inicial anónima



Medio filtrador

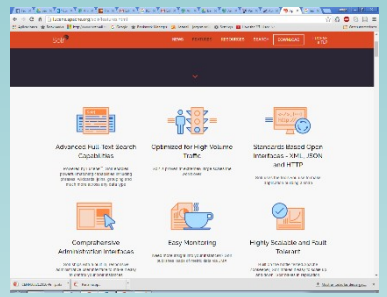
ICIJ (investigación y difusión)

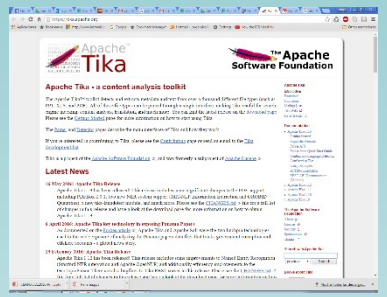
Medios asociados (investigación y difusión)

Procesamiento de datos

"The initial challenge for ICIJ and the Suddeutsche Zeitung was to make these massive amounts of data exploitable by their network of investigative journalists around the world. This implied turning 11.5 unstructured million documents into something meaningful for journalist investigator"

Extracción de metadata (SOLR + TIKA)





Creación de grafos y depuración de datos

Representación como grafos

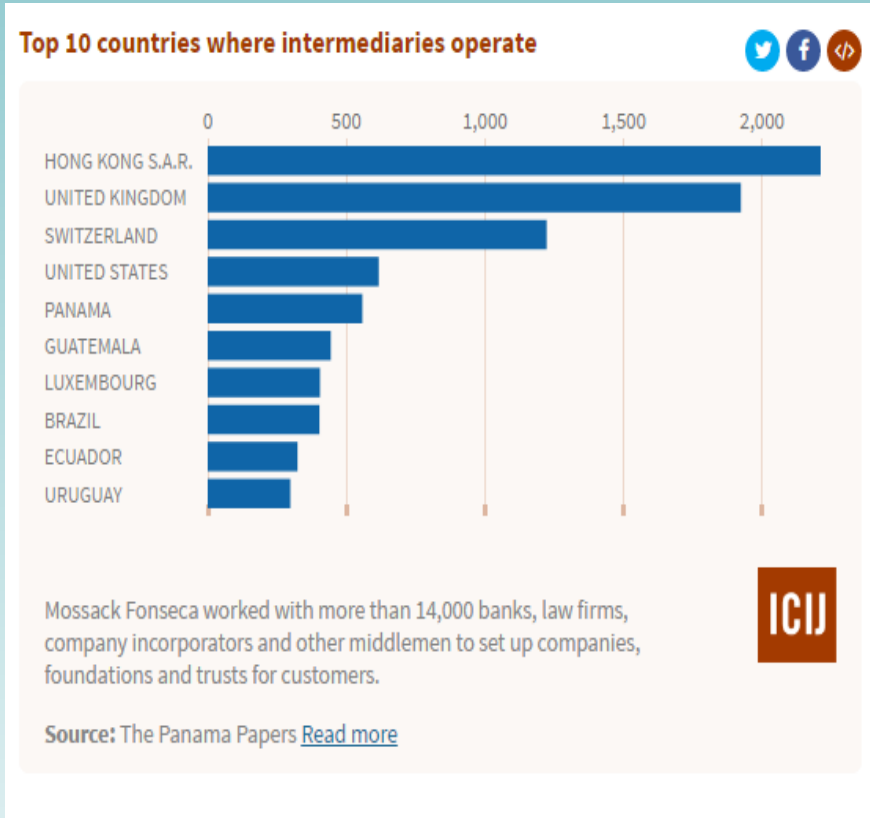
Neo4j

LINKURIOUS
VISUALIZE GRAPH DATA EASILY

Marco temporal y alcance

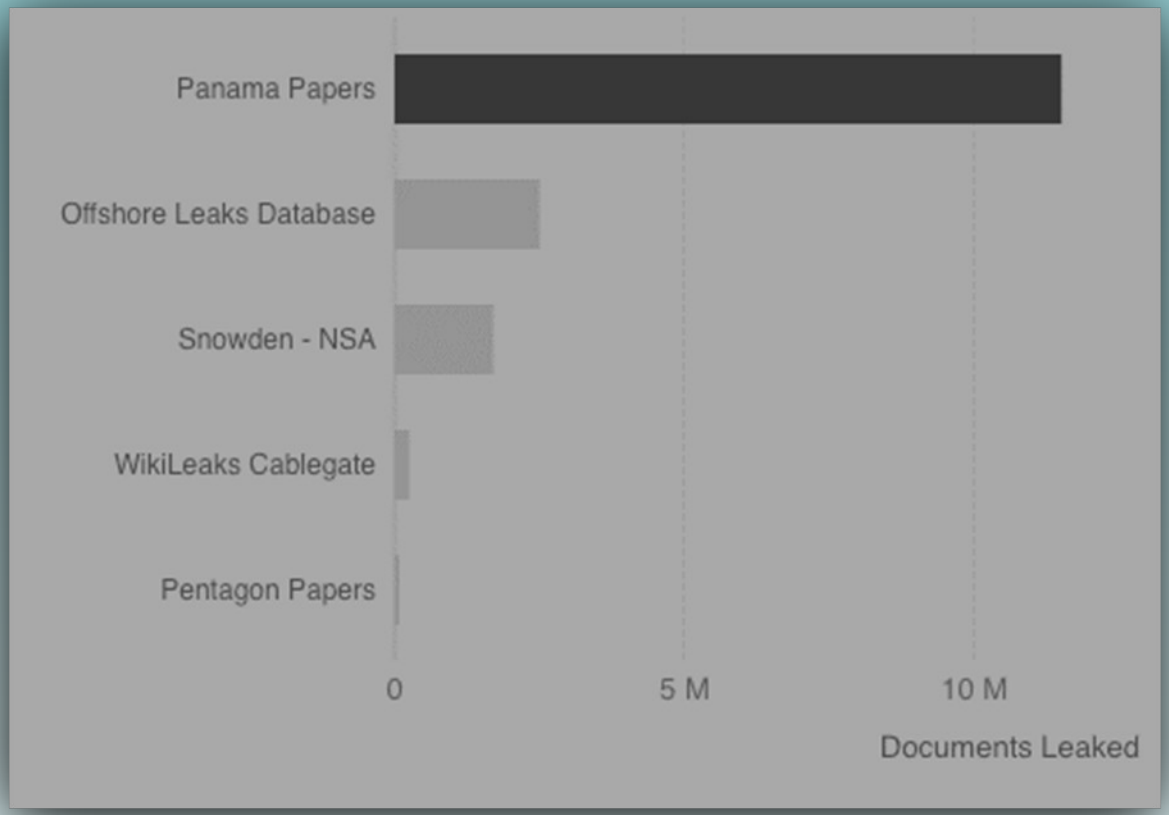
40 años de datos – 210000 compañías – 21 jurisdicciones offshore

14000 bancos, estudios jurídicos y compañías



Volumen absoluto de los datos involucrados

2.6 Terabytes de 11.5 millones de documentos



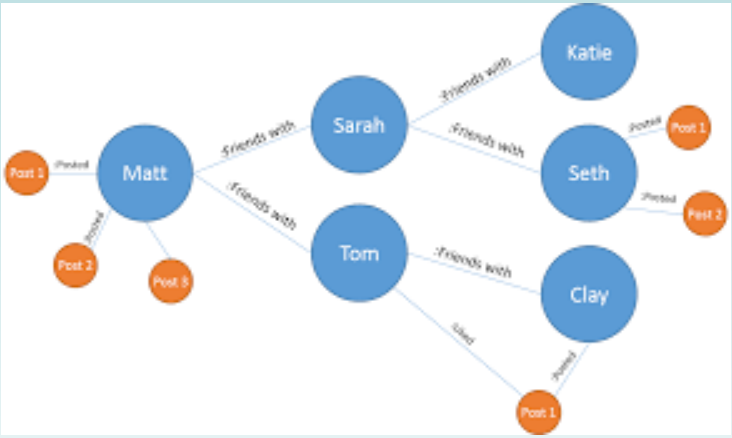
Punto de inflexión: grafos

La tecnología de grafos, aporte fundamental para Panama Papers.

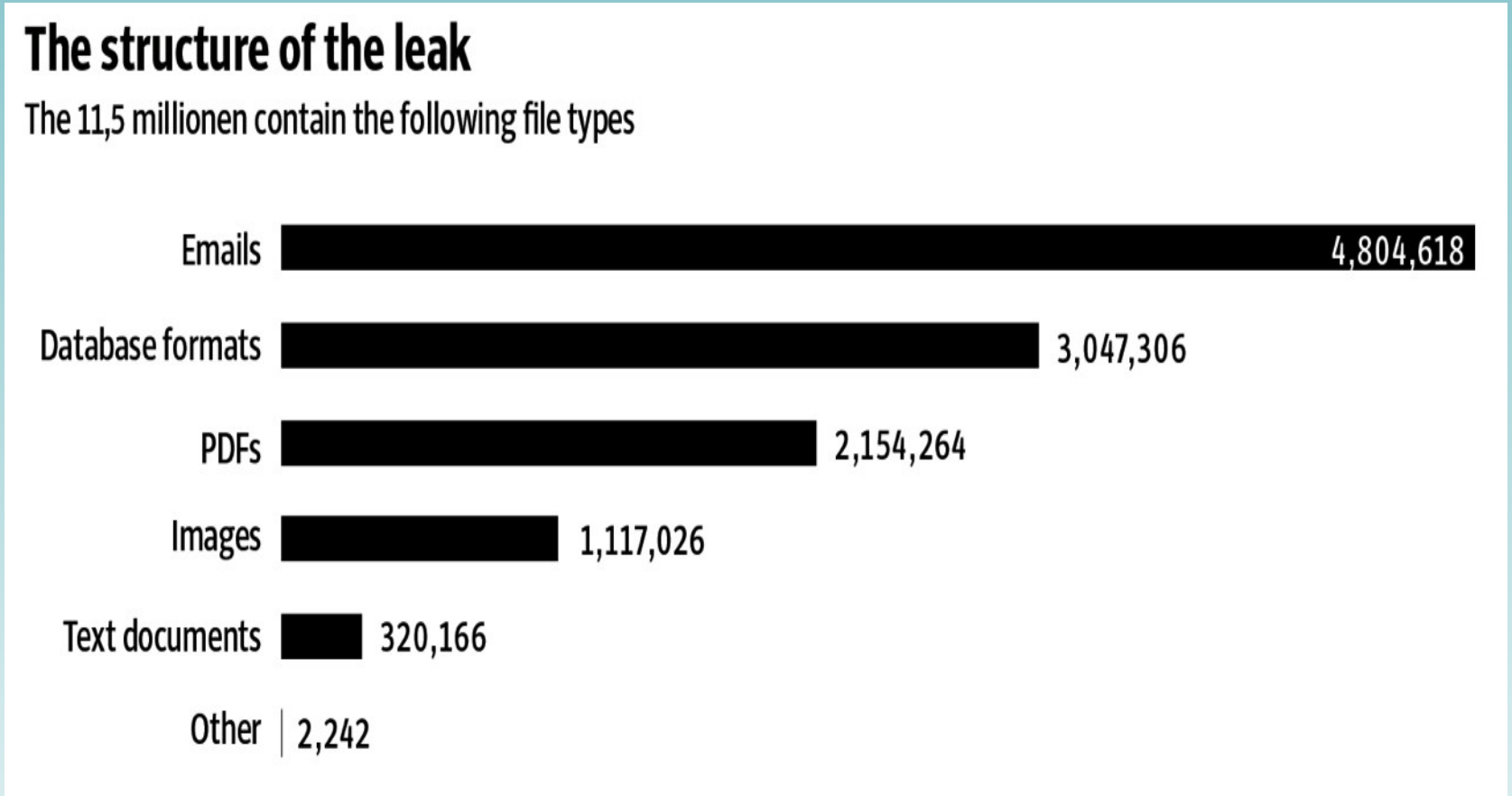
Esta investigación no fue la primera en la que el ICIJ (International Consortium of Investigative Journalists) desarrolló una investigación de este rango.

En 2015 Swiss Leaks. Actividad fraudulenta de 100000 de clients del Banco HSBC en Suiza.

Lo que une a ambas investigaciones es el uso de la llamada “Tecnología de grafos”.



Composición de los datos



Fases de trabajo

The Steps Involved in the Document Analysis

1) Acquire documents
2) Classify documents
    Scan / OCR
    Extract document metadata
3) Whiteboard domain
    Determine entities and their relationships
    Determine potential entity and relationship properties
    Determine sources for those entities and their properties
4) Work out analyzers, rules, parsers and named entity recognition for documents
5) Parse and store document metadata and document and entity relationships
    Parse by author, named entities, dates, sources and classification
6) Infer entity relationships
7) Compute similarities, transitive cover and triangles
8) Analyze data using graph queries and visualizations

Inferencia de relaciones (1): ¿Cómo se infiere una relación en general?

Tipos de relaciones:

1)Algunas relaciones pueden ser inferidas de los documentos:

(:Officer)-[:IS_OFFICER_OF]->(:Company)

2) Otras relaciones pueden ser inferidas analizando intercambio de mails. Si están registrados muchos emails entre una persona y una compañía, podemos inferir que esa persona es cliente de esa compañía.

(:Client)-[:IS_CLIENT_OF]->(:Company)

3) Similar lógica puede aplicarse para inferir relaciones entre entidades que:

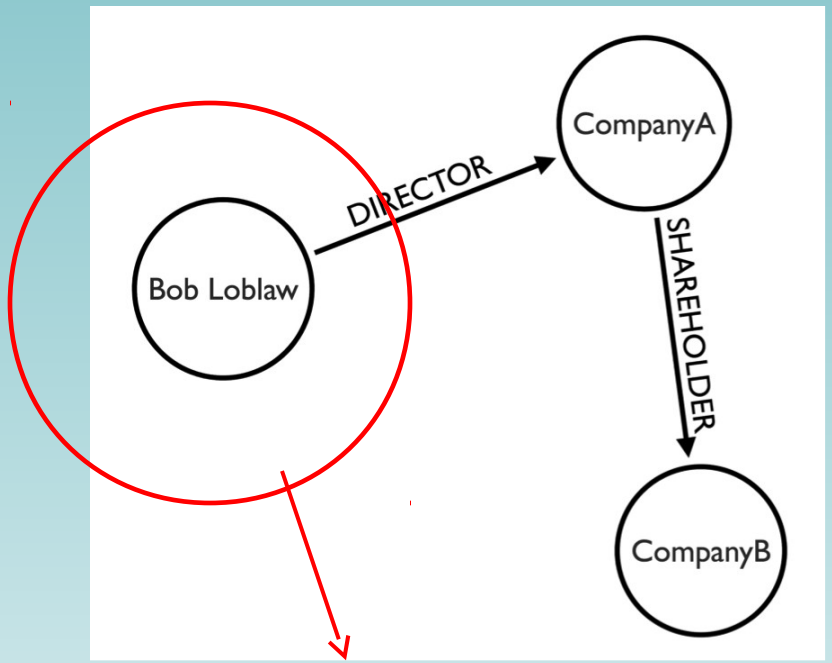
-Comparten el mismo domicilio

-Tienen lazos familiares

-Mantienen relaciones comerciales

-Se comunican regularmente

Inferencia de relaciones (2): ¿Cómo se infiere una relación a través de tríadas?



La localización de tríadas en un grafo sirve para inferir relaciones. En este caso Bob tiene una conexión inferida de la compañía “CompanyB” a través de la compañía “CompanyA”.

Inferencia de relaciones (3): Tipos de metadatos para inferir relaciones en 2 pasos

Paso 1:

Direct metadata -> entities -> relationships to documents

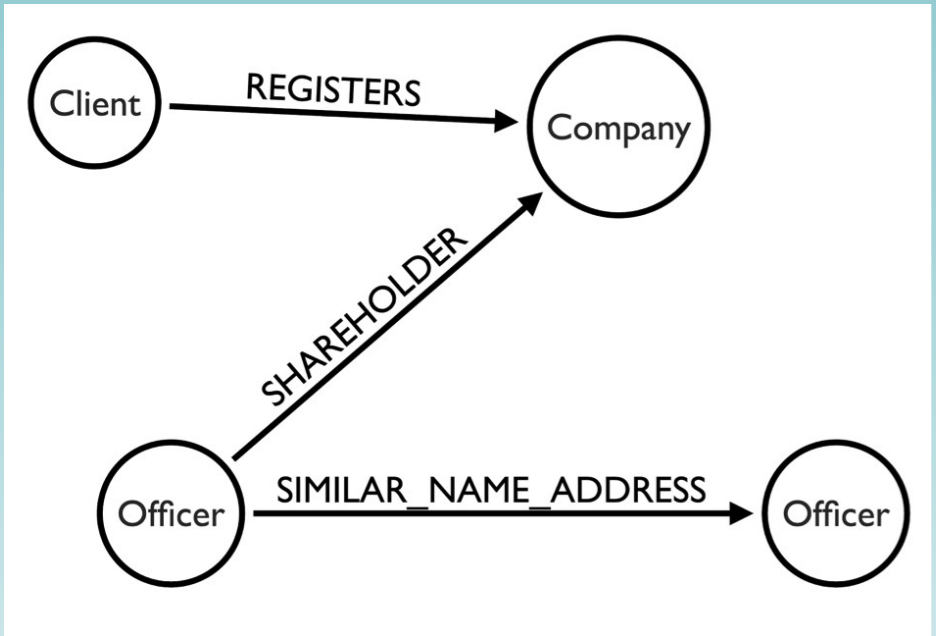
- author, receivers, account-holder, attached to, mentioned, co-located
- Turn plain entities / names into full records using registries and profile documents

Paso 2:

Inferred metadata and information from other sources -> Relationships between entities

- Related to people or organizations from the direct metadata
- Same addresses / organizations
- Find peer groups / rings within fraudulent activities
- Family ties, business relationships
- Part of the communication chain

Modelo de grafos básico usado por ICIJ



Nodos:

- Clients
- Companies
- Addresses
- Officers (both natural people and companies)
  - protector
  - beneficiary, shareholder, director
  - beneficiary
  - shareholder

Relaciones:

- **(:Officer)-[:is officer of]->(:Company)**
- **(:Officer)-[:registered address]->(:Address)**
- **(:Client)-[:registered]->(:Company)**
- **(:Officer)-[:has similar name and address]->(:Address)**

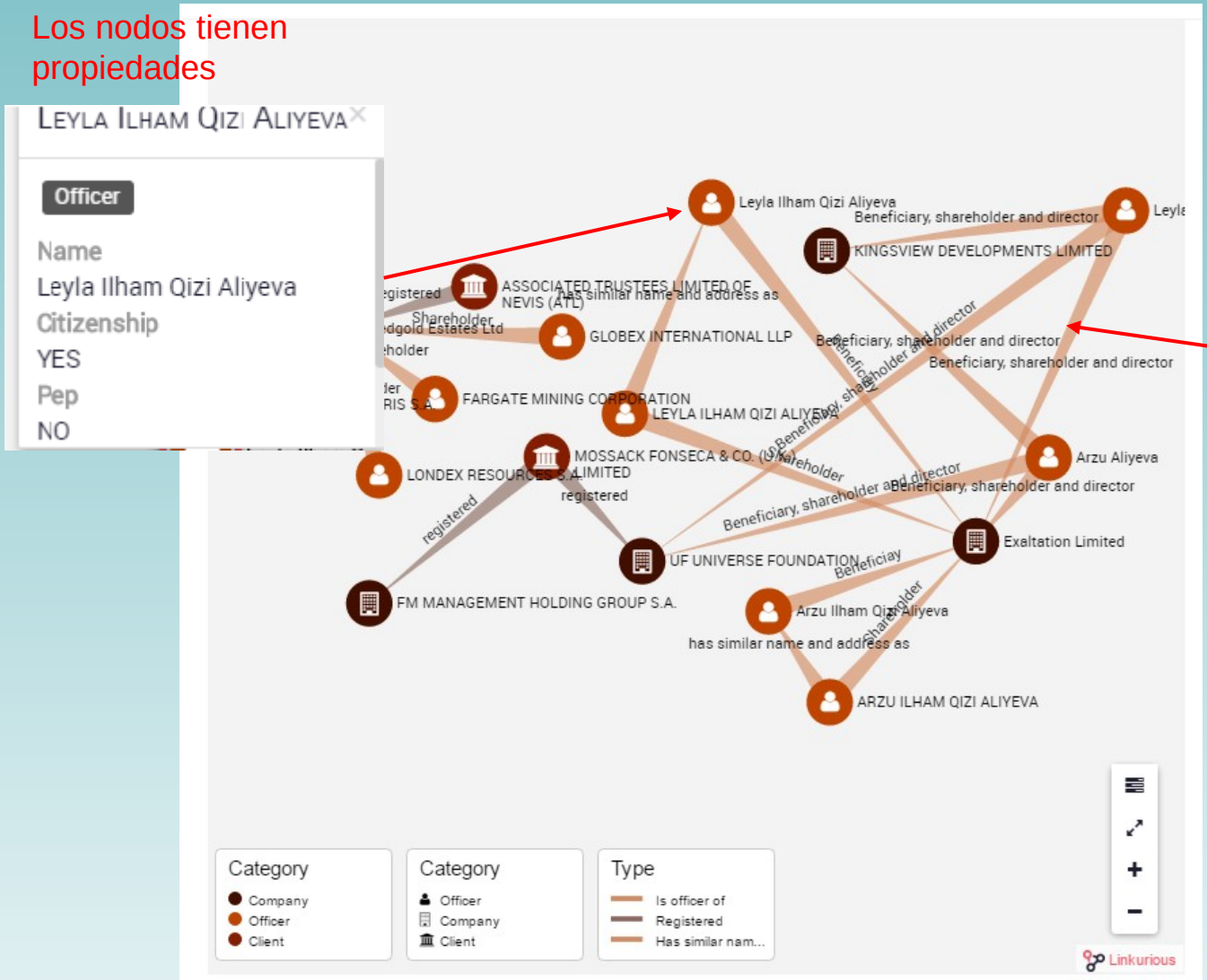
Relaciones que permiten inferior relaciones

Posibles extensiones del modelo de grafos original

- New Entities:
  - Documents: E-Mail, PDF, Contract, DB-Record, …
  - Money Flow: Accounts / Banks / Intermediaries
- New Relationships
  - Family / business ties
  - Conversations
  - Peer Groups / Rings
  - Similar Roles
  - Mentions / Topic-Of
  - Money Flow

Visualización de red total

Los nodos tienen propiedades



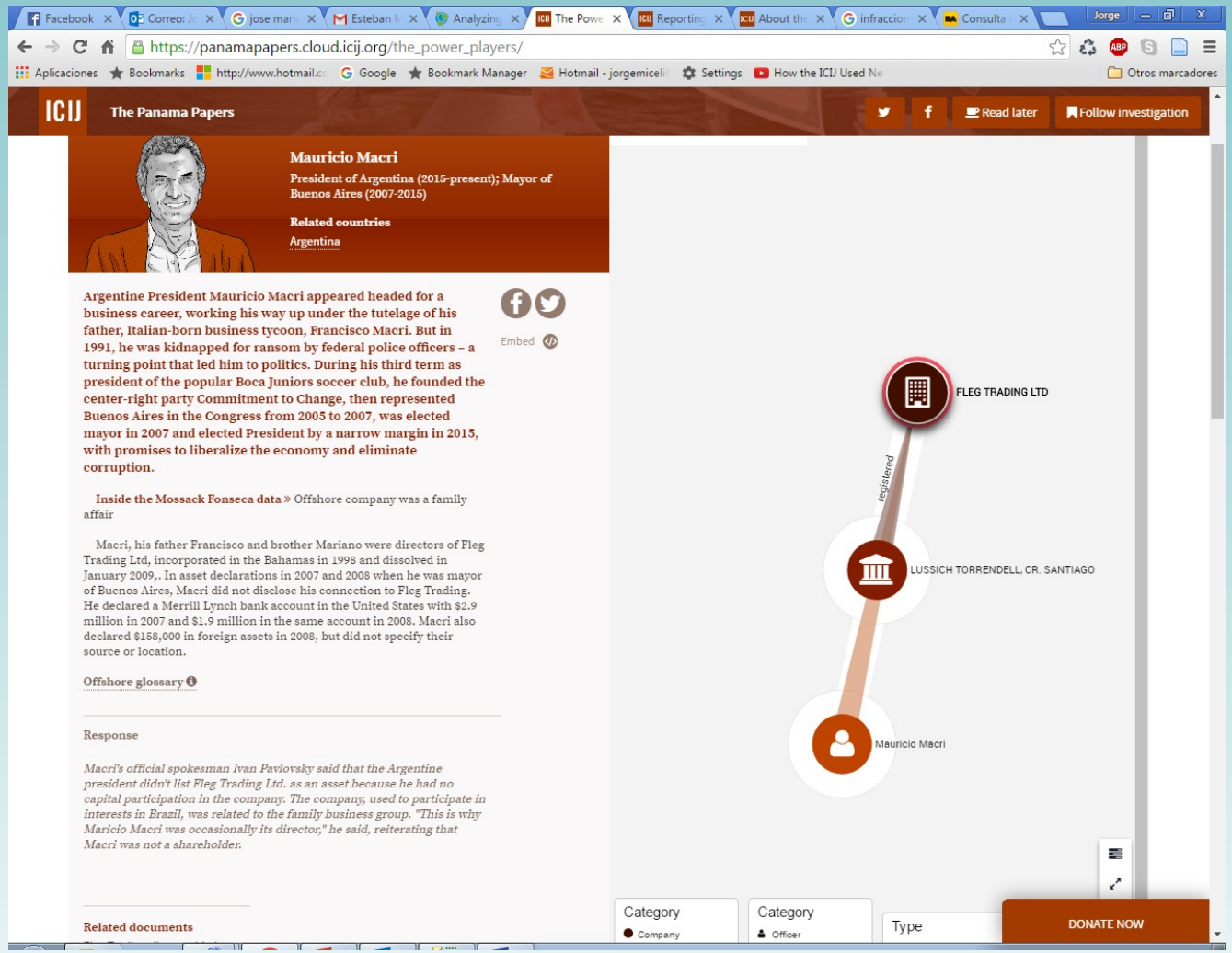
Los lazos tienen propiedades

Name
Beneficiary, shareholder and director
Date_start
02/01/2015
Date_end

https://neo4j.com/blog/analyzing-panama-papers-neo4j/

Visualización de red egocéntrica

https://panamapapers.icij.org/the_power_players/



Aspectos técnicos del lenguaje Cypher (Neo4j)

```
CREATE
(leyla: Officer {name:"Leyla Aliyeva"})-[:IOO_BSD]->(ufu:Company {name:"UF Universe Foundation"}),
(mehriban: Officer {name:"Mehriban Aliyeva"})-[:IOO_PROTECTOR]->(ufu),
(arzu: Officer {name:"Arzu Aliyeva"})-[:IOO_BSD]->(ufu),
(mossack_uk: Client {name:"Mossack Fonseca & Co (UK)"})-[:REGISTERED]->(ufu),
                                        ıy {name:"FM Management Holding Group S.A."}),
```

Representación intuitiva de nodos y relaciones

Family Ties via Last Name:

```
MATCH (o:Officer)
WHERE toLower(o.name) CONTAINS "aliyev"
RETURN o
```

Family Involvements:

```
MATCH (o:Officer) WHERE toLower(o.name) CONTAINS "aliyev"
MATCH (o)-[r]-(c:Company)
RETURN o,r,c
```

Lenguaje de consulta símil SQL

Resolve Duplicate Entities

```
MATCH (o:Officer)
RETURN toLower(split(o.name," ")[0]), collect(o.name) as names, count(*) as count
```

Resolución de nodos duplicados

Shortest Path between Two People

```
MATCH (a:Officer {name: "Mehriban Aliyeva"})
MATCH (b:Officer {name: "Arzu Aliyeva"})
MATCH p=shortestPath((a)-[*]-(b))
RETURN p
```

Camino más corto entre 2 personas

Aspectos relevantes del caso:

Dimensión teórico-metodológica

-Uso del ARS como modelo de representación y metodología de análisis de grandes volúmenes de información.
-Desarrollo de una metodología escalable de procesamiento de información heterogénea con pasos finitos, replicables y perfectibles.

Dimensión técnica

-Posibilidad de procesamiento de grandes volúmenes de datos con elevada velocidad y confiabilidad.
-Pertinencia específica de las bases orientadas a grafos (BDOG) para el tratamiento de esta clase de información.
-Aplicabilidad combinada de distintos paquetes de software (SOLR, TIKA, NEO4j, LINKURIOUS etc.) capaces de operar sobre un conjunto desestructurado de datos y formatearlos, representarlos y analizarlos de manera reticularmente consistente.

Dimensión política

-Importancia crítica de la circulación de información clasificada de incumbencia pública.
-Selectividad de los investigados (ausencia de acusados estadounidenses por ejemplo).

Enrique **Dans**

“El caso “Panama papers” es todo un signo de los tiempos: cambian los negocios, porque lo que no sea razonablemente transparente será cada vez más obligado a serlo, cambia la investigación periodística, y terminará cambiando hasta los marcos legislativos que la regulan. Pero sobre todo, es un esfuerzo brutal y un proyecto de sistemas y tecnologías de información que debería estudiarse en las universidades.”

https://www.enriquedans.com/2016/04/analizando-los-panama-papers-como-proyecto-tecnologico.html

Fuentes:

https://panamapapers.icij.org/
https://neo4j.com/blog/panama-papers/
https://neo4j.com/blog/analyzing-panama-papers-neo4j/
https://panamapapers.icij.org/the_power_players/
https://panamapapers.icij.org/graphs/
https://neo4j.com/blog/tuning-cypher-queries/
https://www.enriquedans.com/2016/04/analizando-los-panama-papers-como-proyecto-tecnologico.html

Muchas gracias