# JURION IPG USE-CASE

## RE-ENGINEERING A COMPLEX RELATIONAL DATABASE APPLICATION

Christian Dirschl
Chief Content Architect
Wolters Kluwer

Kevin Feeney
CEO, DataChemist

Gavin Mendel Gleason
CTO, DataChemist

ALIGNED
Software and Data Engineering

DataChemist

Wolters Kluwer

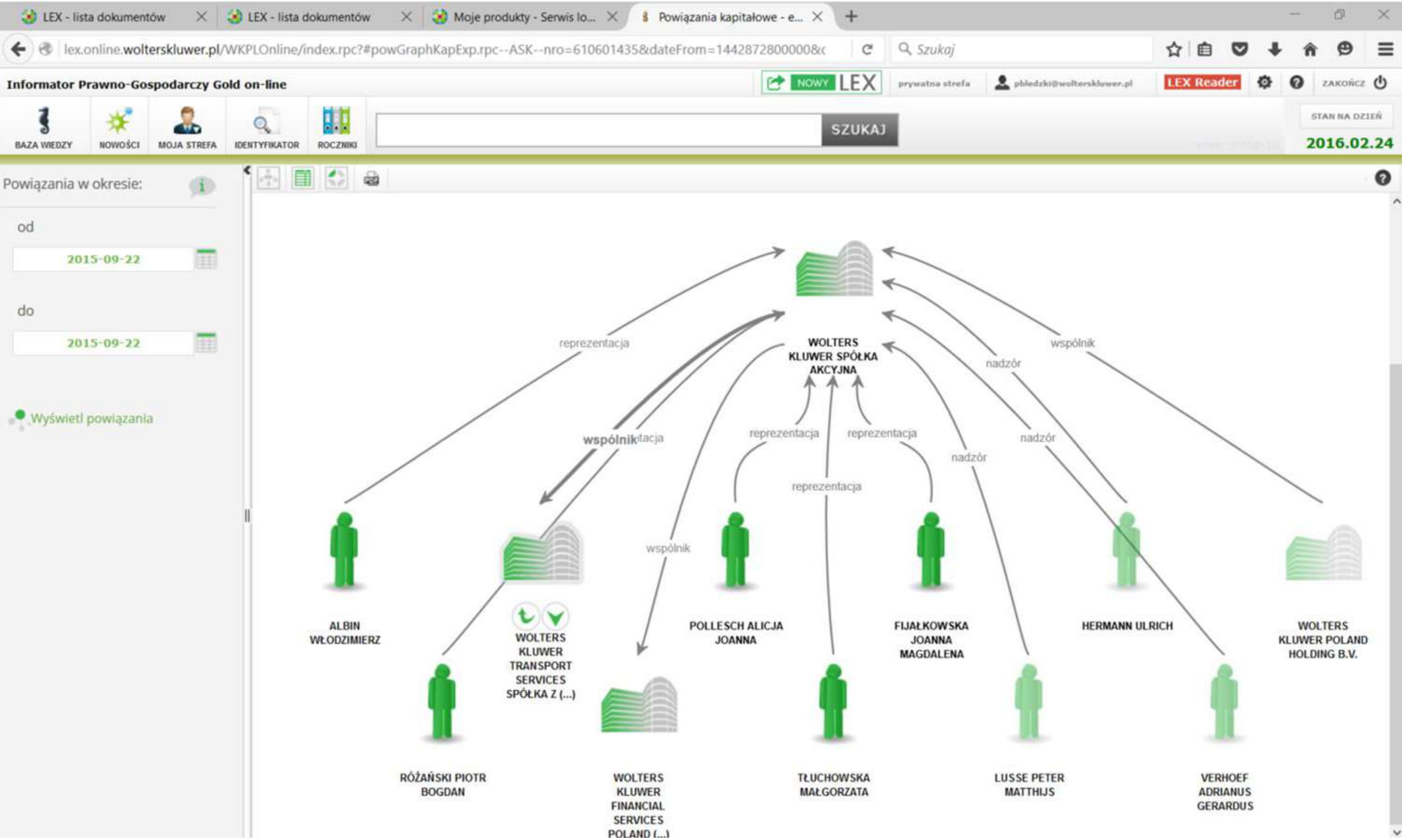
# IPG USE CASE INTRODUCTION

ALIGNED
Software and Data Engineering

DataChemist

Wolters Kluwer

# Legal-Commercial Information System (IPG Gold) product – graph view



lex.online.wolterskluwer.pl/WKPLOnline/index.rpc?#powGraphKapExp.rpc--ASK--nro=610601435&dateFrom=1442872800000&c

Informator Prawno-Gospodarczy Gold on-line

STAN NA DZIEŃ 2016.02.24

Powiązania w okresie:

od 2015-09-22

do 2015-09-22

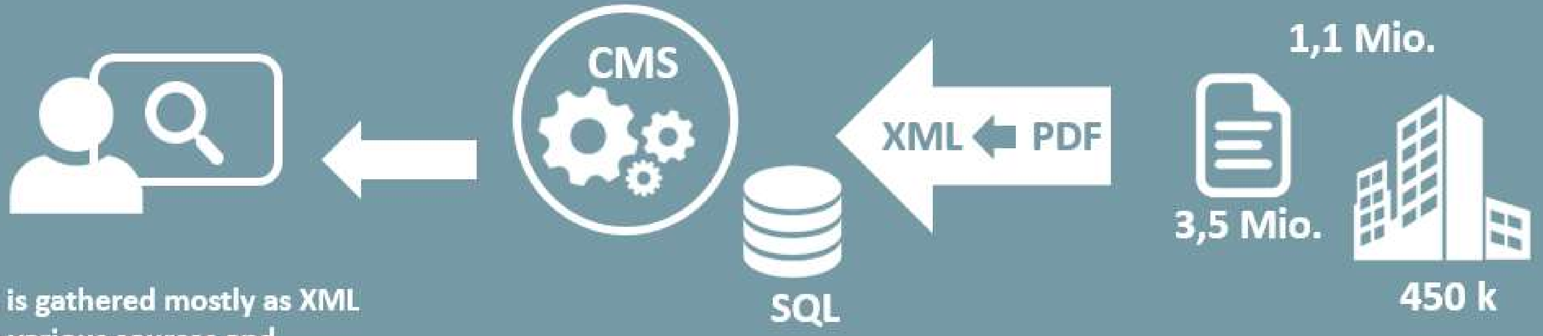
ALIGNED Software and Data Engineering

DataChemist

Wolters Kluwer

# IPG Problem Statement

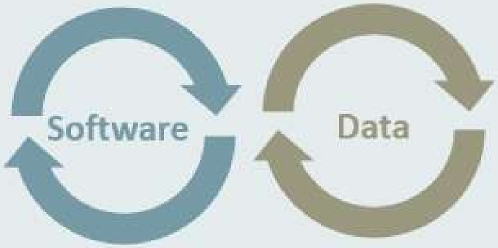
**IPG – a Commercial Intelligence System by Wolters Kluwer Poland with …**



Data is gathered mostly as XML from various sources and processed through a proprietary CMS and a standard SQL database for content enrichment and validation into a final search index which serves data for the end user.

Data is also expected to be enriched with data originating from new sources including publicly available repositories and third party datasets.

A major problem is data quality including missing or incoherent data as well as semantic inconsistencies which could be detected and corrected by using Aligned tools.



The software development lifecycle ist mostly autonomous from the data lifecycle. Both of them not changing very often, but are expected to change in the next few years as major upgrades are planned.

# Data Complexity

**450k companies**

**1,1 Mio people**

**3,5 Mio documents**

**Spatial data and administrative division data (2,5k counties)**

**Legacy DB model in Polish**

+

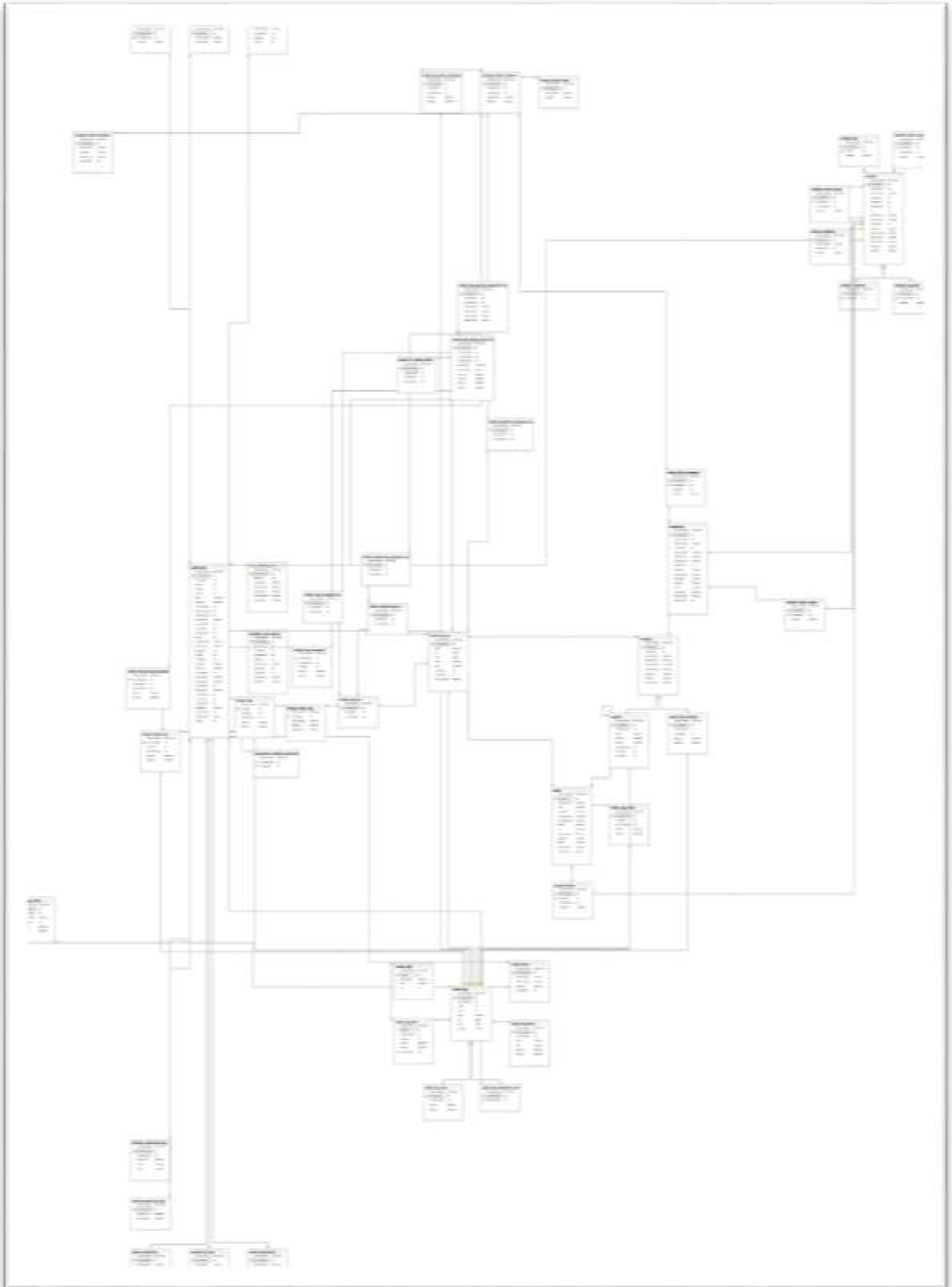
**Complex Schema**

**50** types of Companies

**20** types of relations between companies & people

**70** types of events/ documents related to companies or people

**30** Types of roles



# Data Complexity example

# 32 Unsolvable Scenarios

Basic Datatype Errors

9, 25, 26, 27, 28

Invalid email address

Temporal Constraints

10, 11, 12, 13, 14, 15, 17, 21

Same receiver and trustee

Inconsistent Data

2, 8, 22, 30

Multiple shareholders in sole shareholder company

Missing Mandatory Properties

1, 3, 4, 5, 6, 7, 16, 18, 19, 20, 23, 24

No trustee in bankruptcy

Data Model Complexity

29

Relationship model in main table is incomprehensible

Temporal Queries

31

Find relationship at any time between any 2 entities

Recursive Queries

32

Subsidiarity Loop: Company A owns B, owns C, owns A.

# IPG – DATA CHEMIST SOLUTION

ALIGNED
Software and Data Engineering

DataChemist

Wolters Kluwer

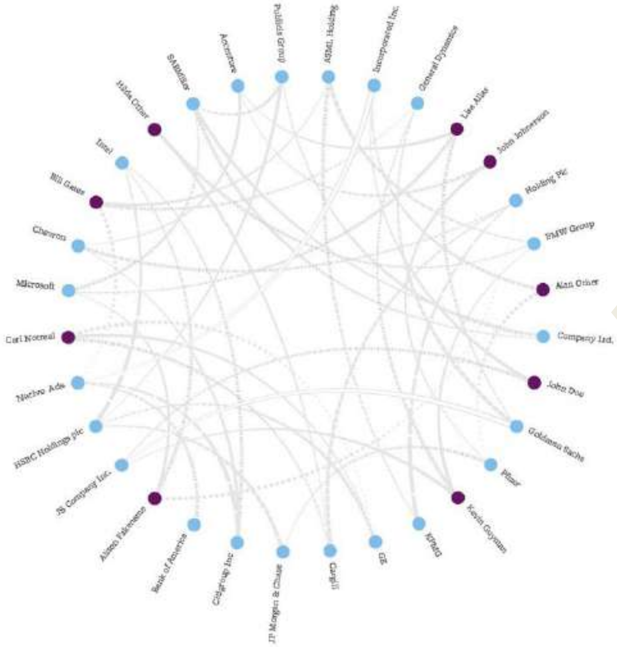
# DataChemist

Closed World OWL Reasoning Engine

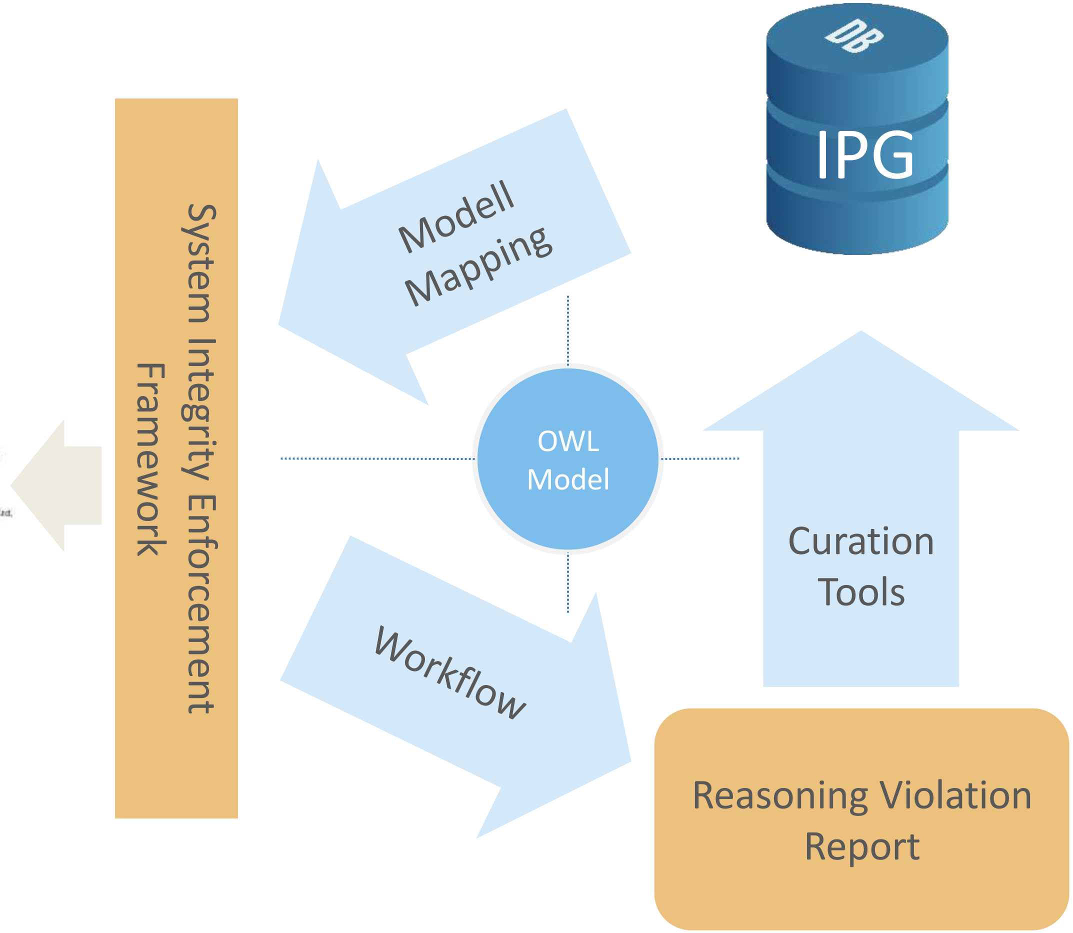
Fast ACID in-memory datastore with integrated logic engine

WOQL – model aware query language

# How it works



High Quality – Ontology Conformant Knowledge Graph



DB
IPG
System Integrity Enforcement Framework
Modell Mapping
OWL Model
Curation Tools
Workflow
Reasoning Violation Report

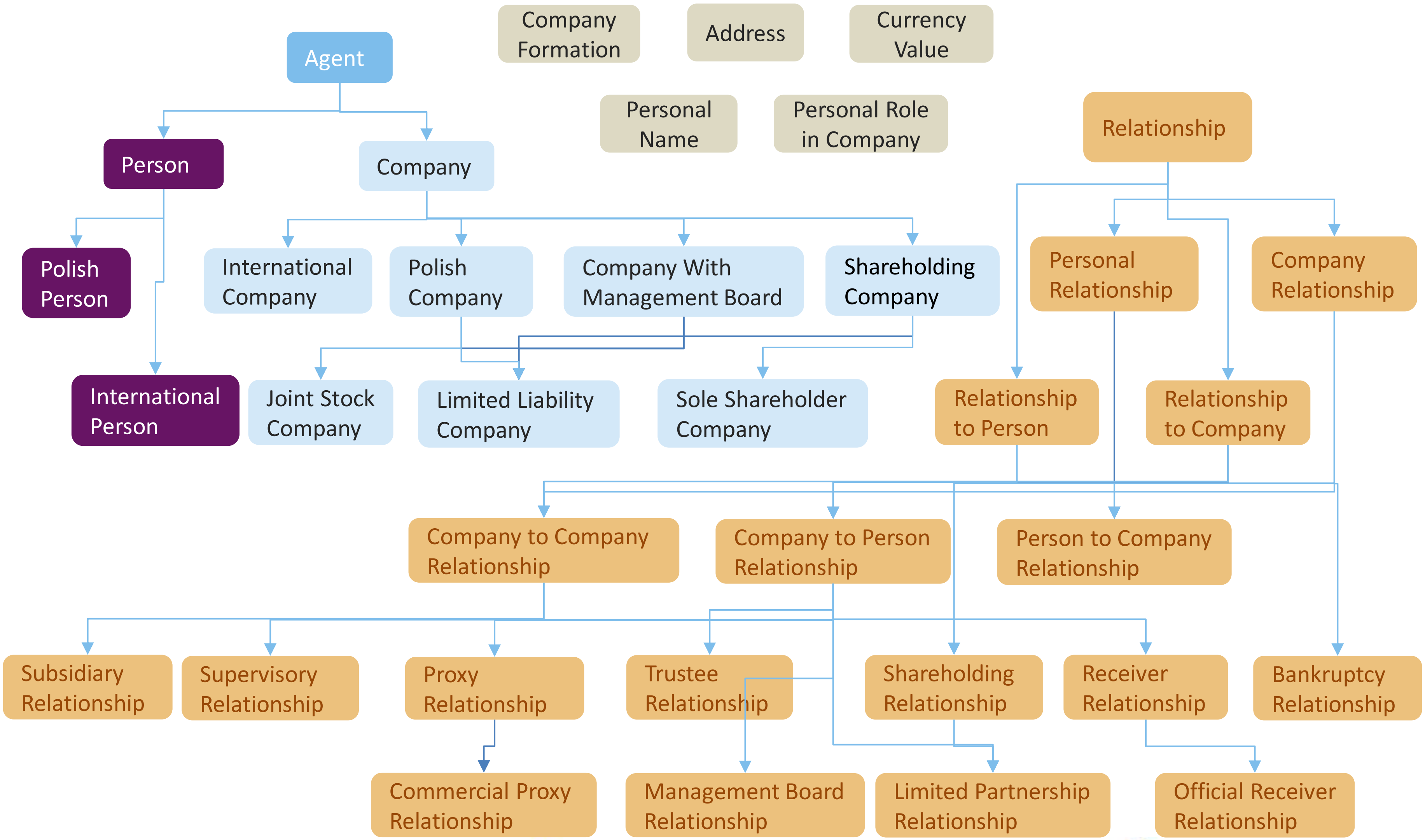
Model driven tools

ALIGNED Software and Data Engineering

DataChemist

Wolters Kluwer

# Semantic Model – 36 classes

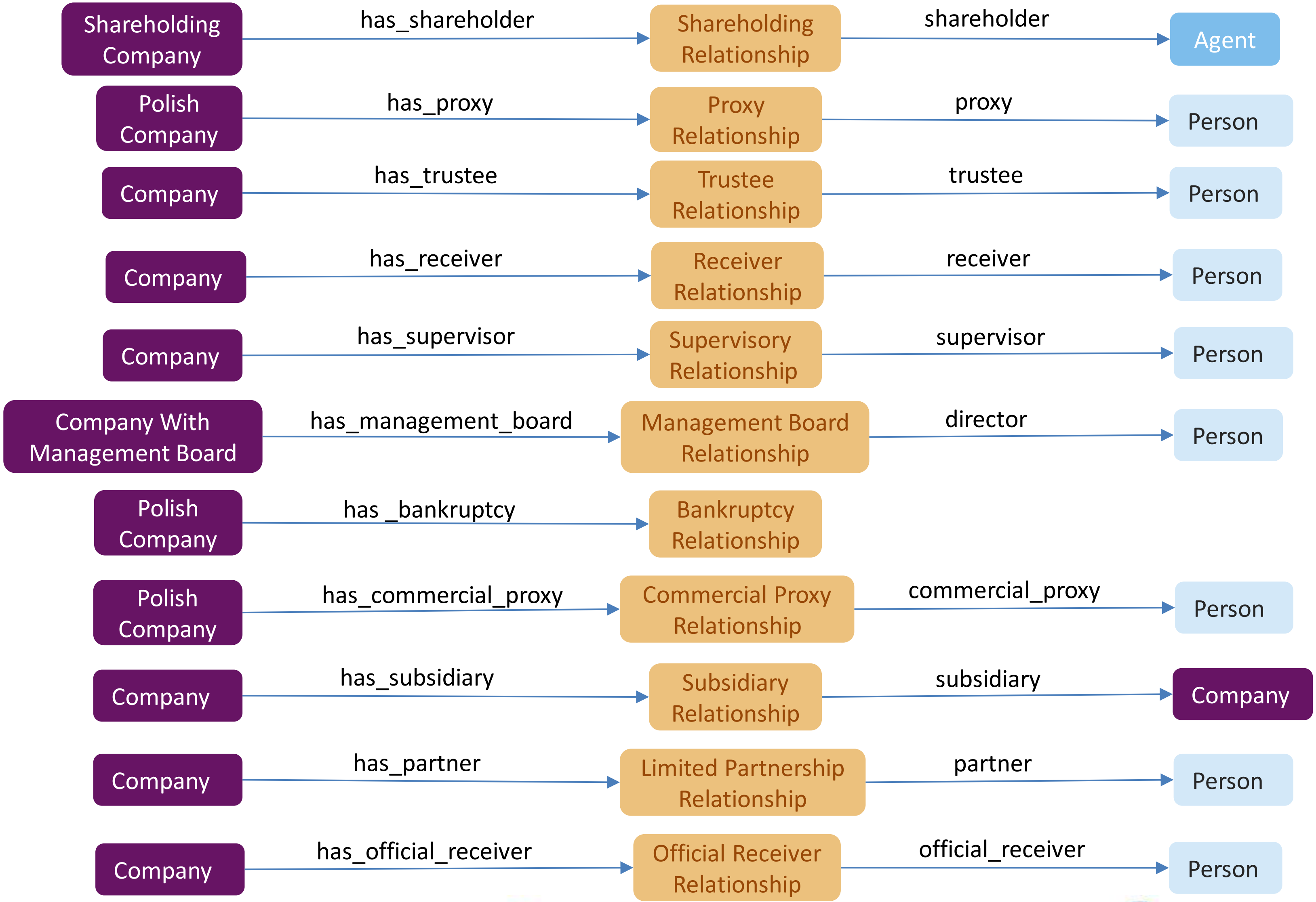


ALIGNED Software and Data Engineering

DataChemist

Wolters Kluwer

# Semantic Model – 21 relationship properties



Shareholding Company → has_shareholder → Shareholding Relationship → shareholder → Agent

Polish Company → has_proxy → Proxy Relationship → proxy → Person

Company → has_trustee → Trustee Relationship → trustee → Person

Company → has_receiver → Receiver Relationship → receiver → Person

Company → has_supervisor → Supervisory Relationship → supervisor → Person

Company With Management Board → has_management_board → Management Board Relationship → director → Person

Polish Company → has _bankruptcy → Bankruptcy Relationship

Polish Company → has_commercial_proxy → Commercial Proxy Relationship → commercial_proxy → Person

Company → has_subsidiary → Subsidiary Relationship → subsidiary → Company

Company → has_partner → Limited Partnership Relationship → partner → Person

Company → has_official_receiver → Official Receiver Relationship → official_receiver → Person

# Semantic Model – 36 simple properties

### Company

| | |
|---|---|
| company_name | String |

### Polish Company

| | |
|---|---|
| annual_report | string |
| formation | Company Formation |
| krs | integer |
| region | integer |
| nip | integer |

### Agent

| | |
|---|---|
| address | Address |

### Person

| | |
|---|---|
| personal_name | Personal Name |
| given_name | Personal Name |

### Polish Person

| | |
|---|---|
| pesel | pesel |

### Company Formation

| | |
|---|---|
| formation_method | string |
| formation_circumstances | string |

### Address

| | |
|---|---|
| email | email |
| website | url |
| postal_street | string |
| postal_number | string |
| postal_locality | string |
| postal_code | string |

### Personal Role in Company

| | |
|---|---|
| personal_role_name | string |

### Currency Value

| | |
|---|---|
| currency_value | float |
| currency_unit | string |

### Shareholding Relationship

| | |
|---|---|
| wholly_owned | string |
| liability | decimal, string |
| number_of_shares | string |
| ShareholdingCompany | Currency Value |

### Trustee Relationship

| | |
|---|---|
| legal_basis | string |
| appointment_date | dateTime |

### Management Board Relationship

| | |
|---|---|
| board_type | string |
| management_board_role | string |
| management_role_suspended | boolean |

### Limited Partnership Relationship

| | |
|---|---|
| liability | decimal |

### Bankruptcy Relationship

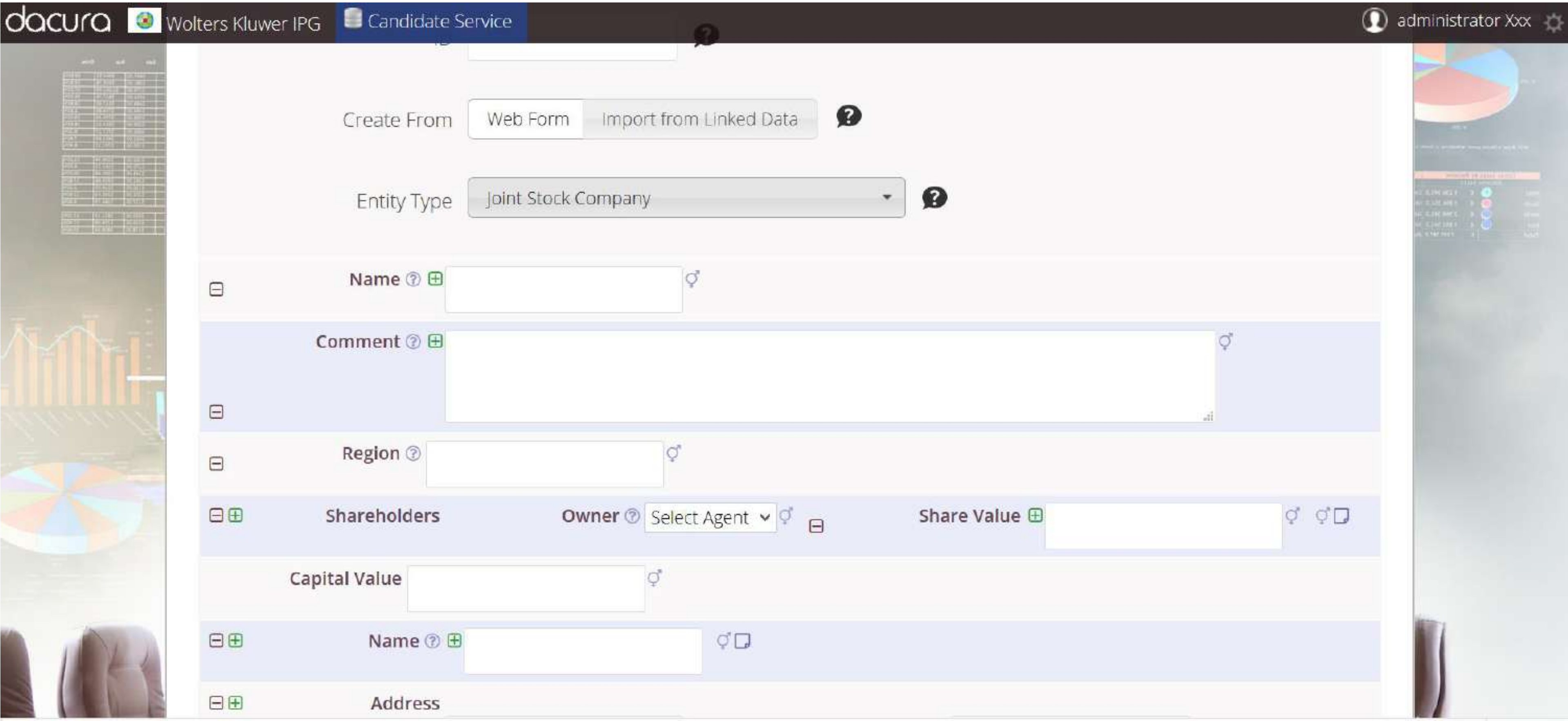
| | |
|---|---|
| announcement | string |
| termination | string |
| method | string |
| repeal | string |

### Proxy Relationship

| | |
|---|---|
| proxy_info | string |
| proxy_type | string |

+ 16 constraints

# Ontology Editing & Visualisation



Create From: Web Form | Import from Linked Data

Entity Type: Joint Stock Company

Name

Comment

Region

Shareholders — Owner: Select Agent — Share Value

Capital Value

Name

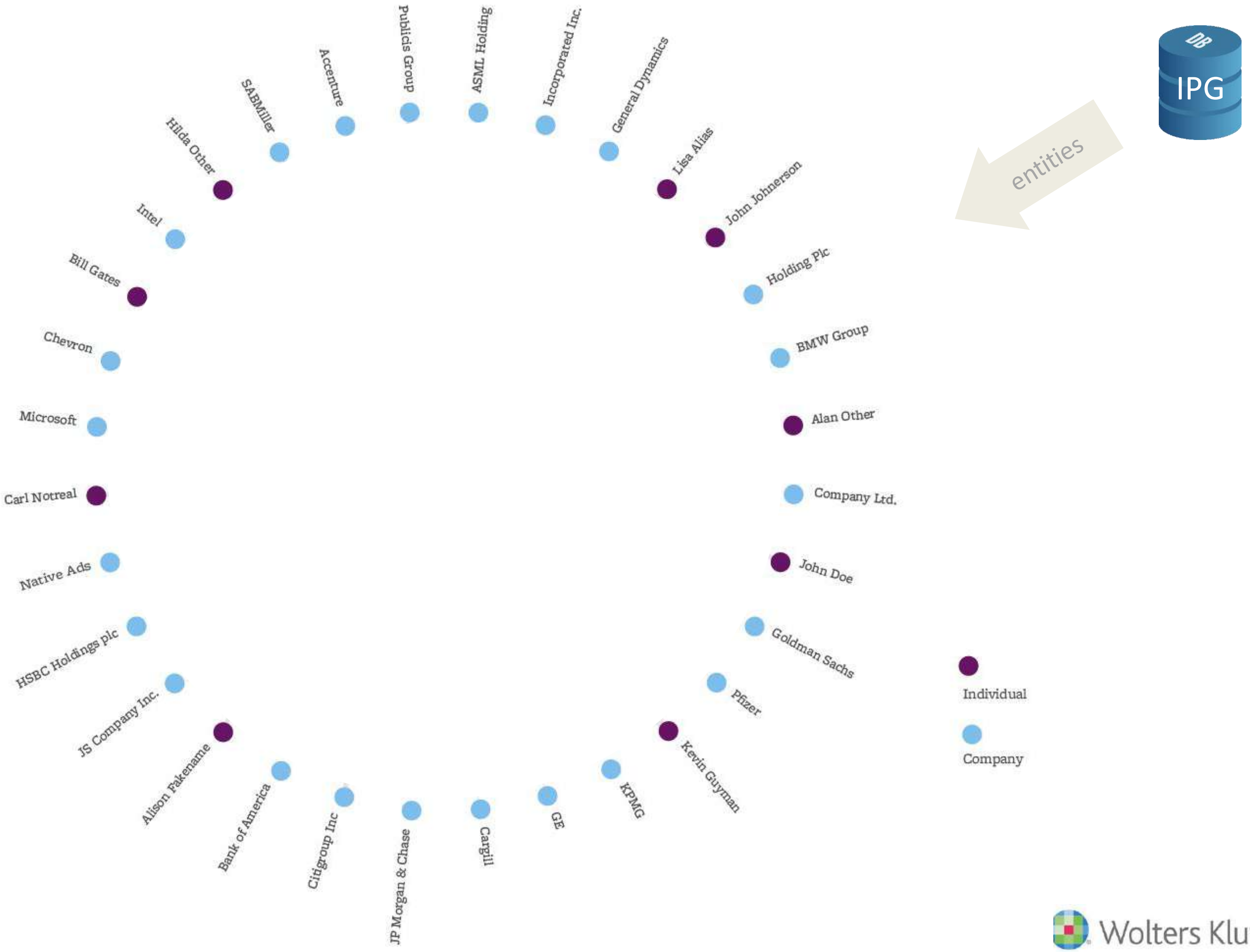
Address

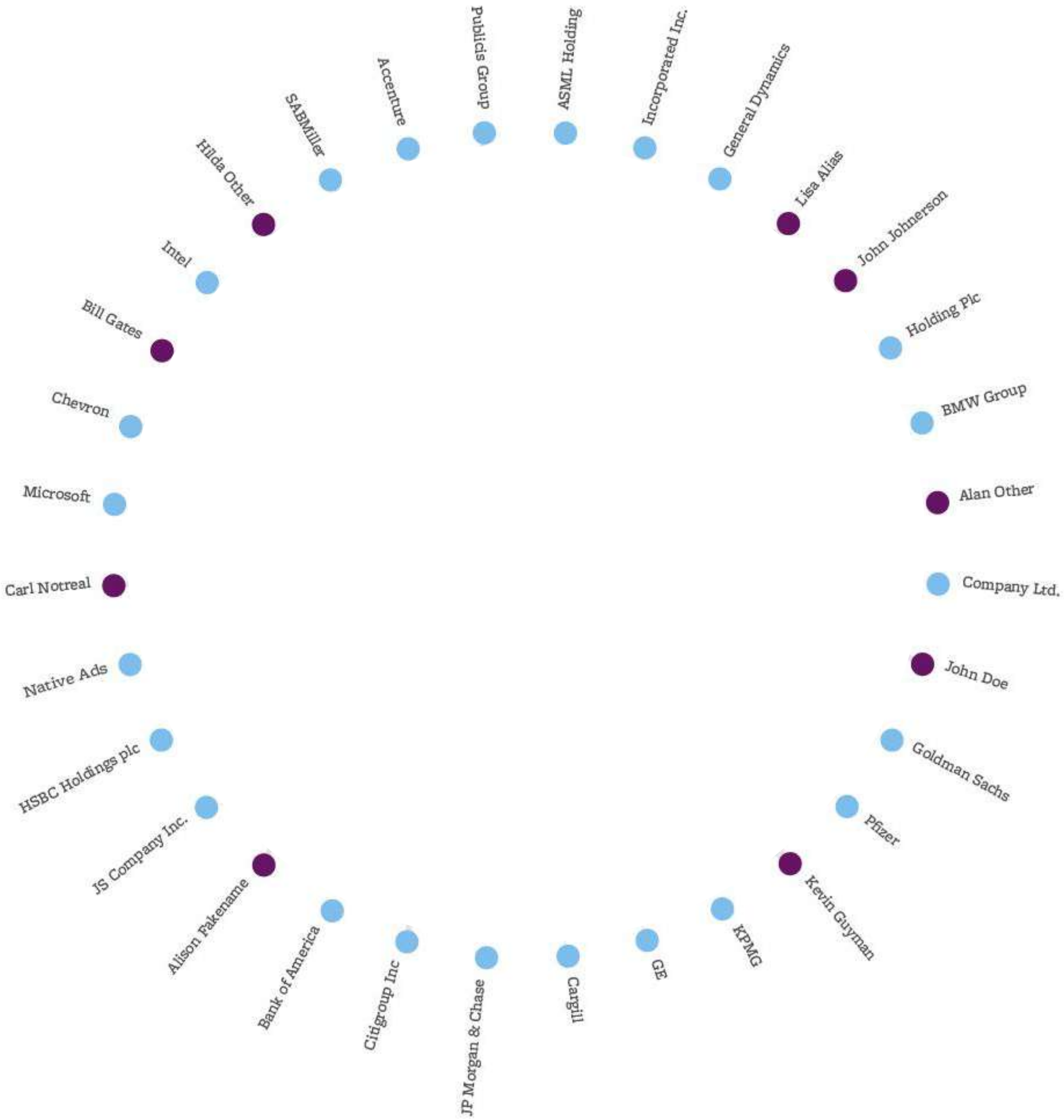
# KNOWLEDGE GRAPH CONSTRUCTION

ALIGNED
Software and Data Engineering

DataChemist

Wolters Kluwer


DB
IPG
entities
Publicis Group
ASML Holding
Incorporated Inc.
General Dynamics
Lisa Alias
John Johnerson
Holding Plc
BMW Group
Alan Other
Company Ltd.
John Doe
Goldman Sachs
Pfizer
Kevin Guyman
KPMG
GE
Cargill
JP Morgan & Chase
Citigroup Inc
Bank of America
Alison Fakename
JS Company Inc.
HSBC Holdings plc
Native Ads
Carl Notreal
Microsoft
Chevron
Bill Gates
Intel
Hilda Other
SABMiller
Accenture
● Individual
● Company



IPG

## Alan Other

PESEL

- 32321212

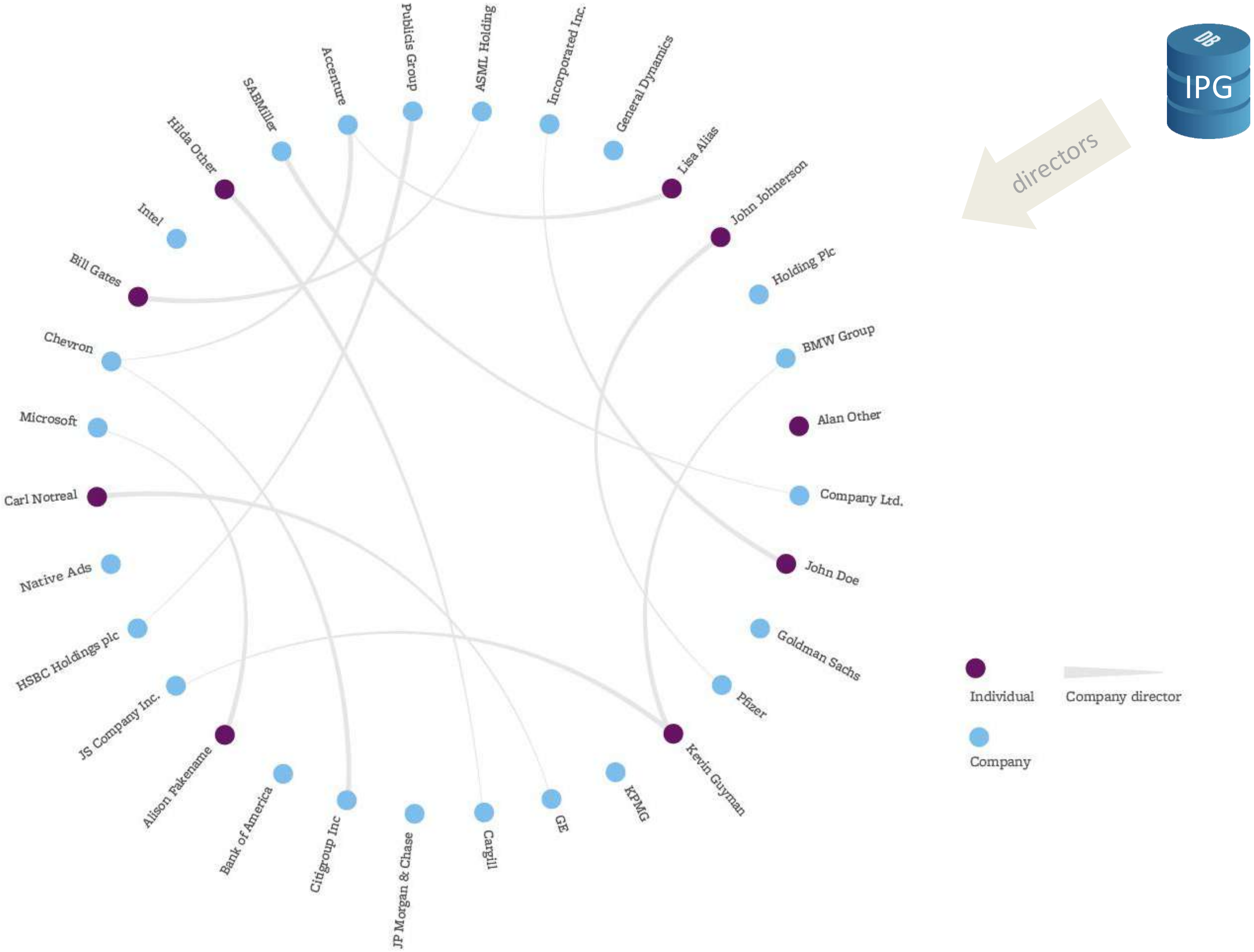
EMAIL

- alan.edgelord@gmail.com

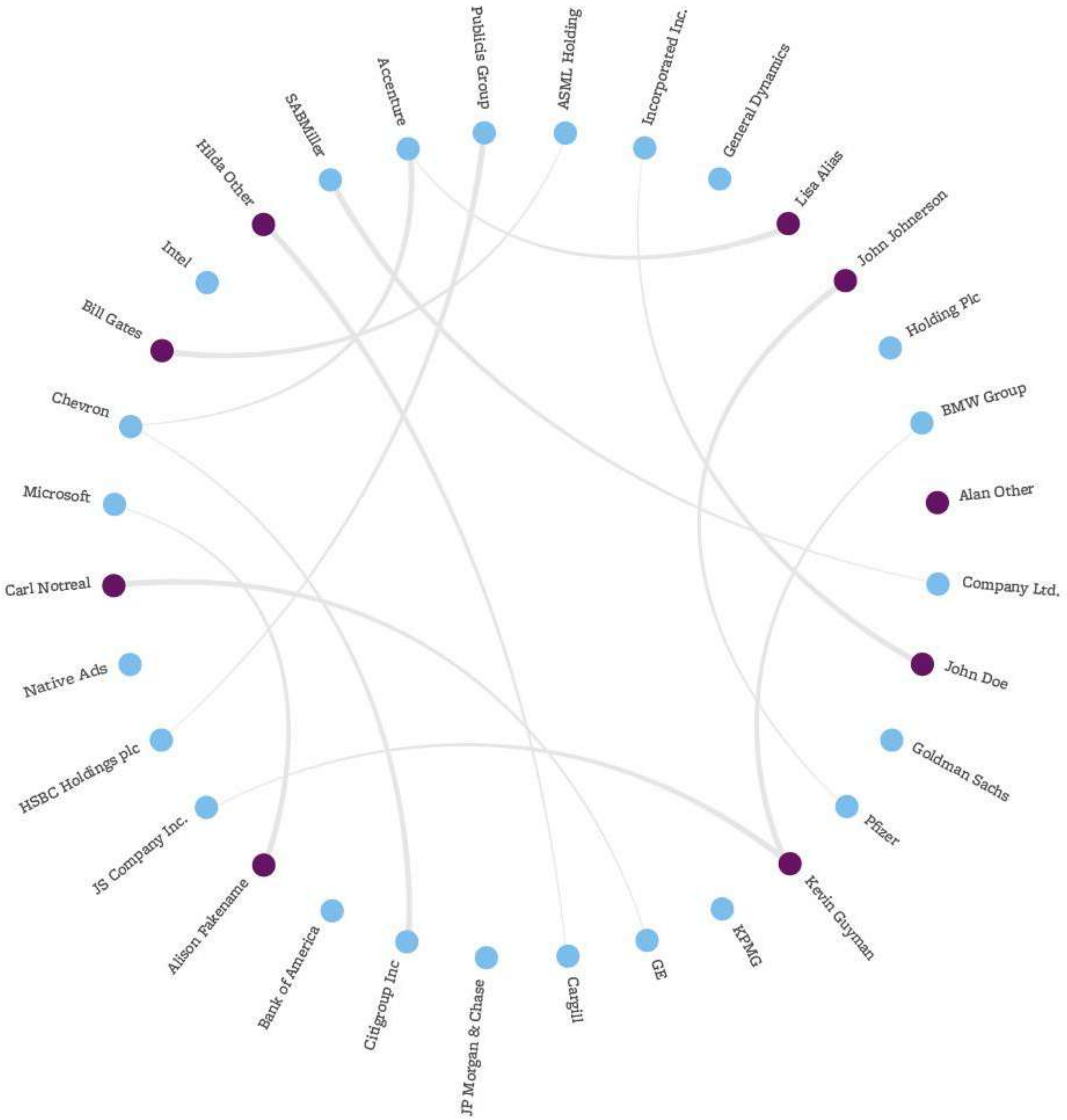
! ALERT: Invalid PESEL

● Individual

● Company



IPG
DB
directors
Publicis Group
ASML Holding
Incorporated Inc.
General Dynamics
Lisa Alias
John Johnerson
Holding Plc
BMW Group
Alan Other
Company Ltd.
John Doe
Goldman Sachs
Pfizer
Kevin Guyman
KPMG
GE
Cargill
JP Morgan & Chase
Citigroup Inc
Bank of America
Alison Fakename
JS Company Inc.
HSBC Holdings plc
Native Ads
Carl Notreal
Microsoft
Chevron
Bill Gates
Intel
Hilda Other
SABMiller
Accenture
Individual
Company director
Company

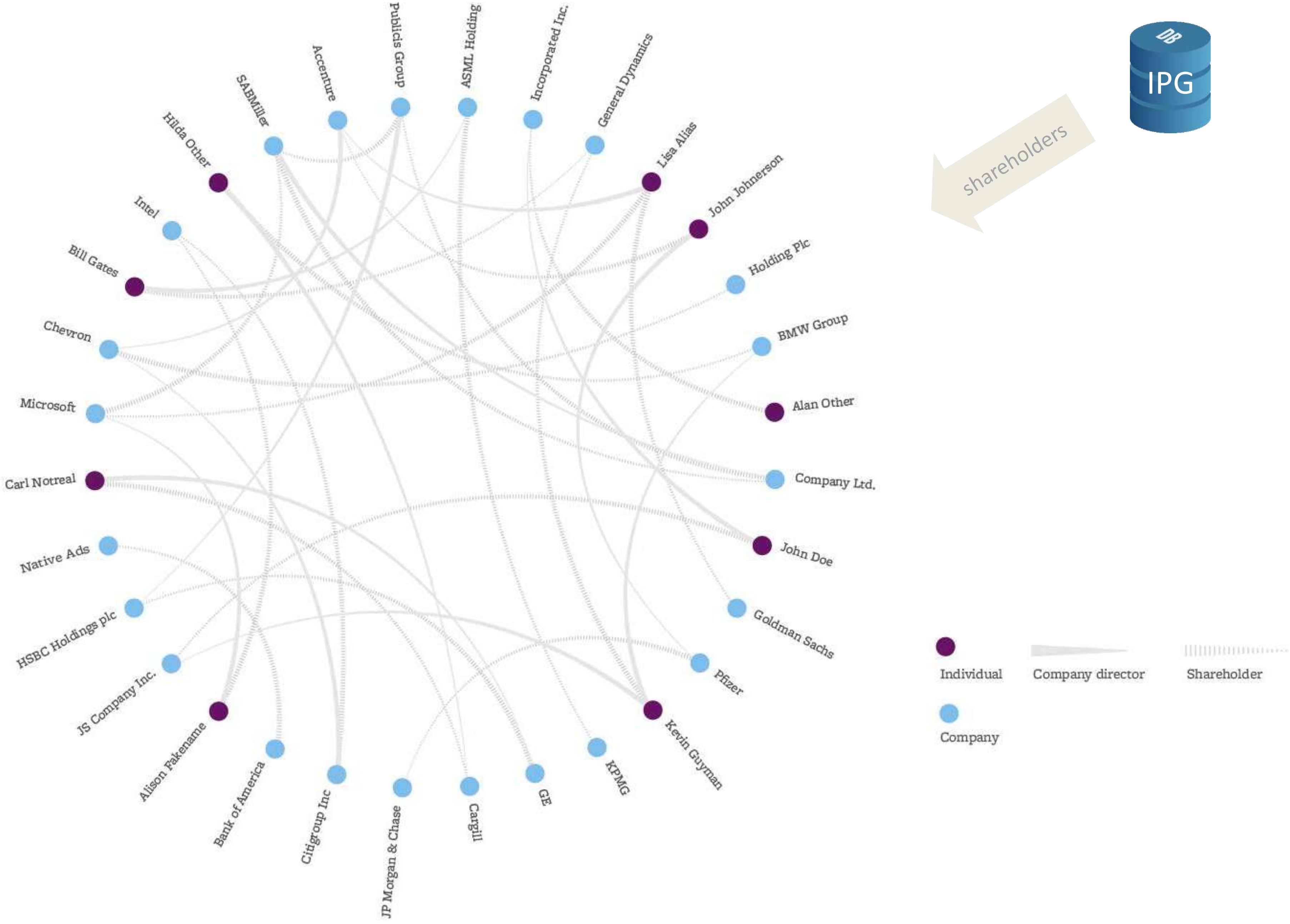


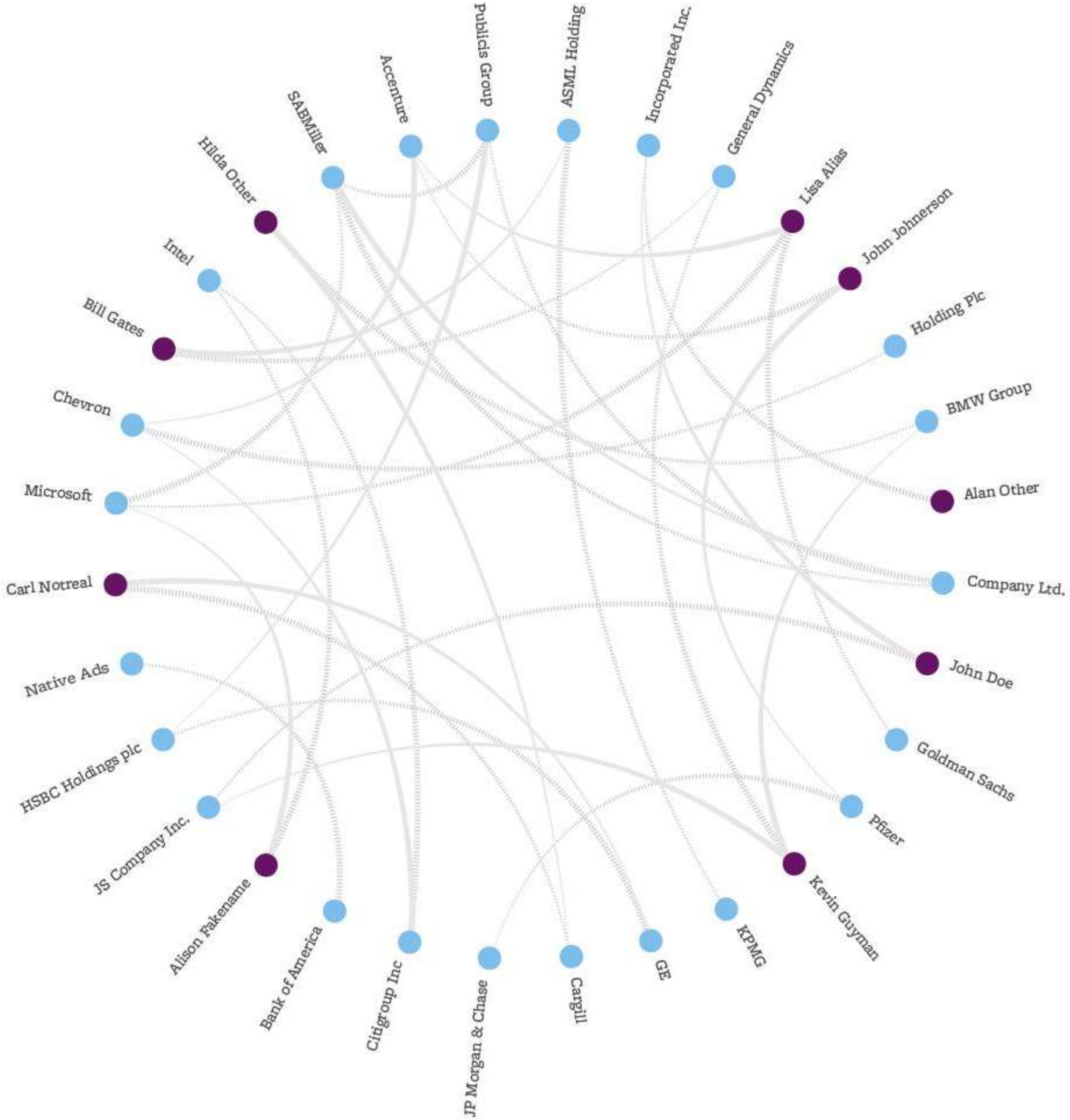
John Doe

DIRECTOR OF
- Incorporated Inc.

BOARD ROLE
- Chairman

! ALERT: Member of board of company that has no board


Publicis Group
ASML Holding
Incorporated Inc.
General Dynamics
Lisa Alias
John Johnerson
Holding Plc
BMW Group
Alan Other
Company Ltd.
John Doe
Goldman Sachs
Pfizer
Kevin Guyman
KPMG
GE
Cargill
JP Morgan & Chase
Citigroup Inc
Bank of America
Alison Fakename
JS Company Inc.
HSBC Holdings plc
Native Ads
Carl Notreal
Microsoft
Chevron
Bill Gates
Intel
Hilda Other
SABMiller
Accenture
DB
IPG
shareholders
Individual
Company director
Shareholder
Company



IPG

## AMSL Holdings

| Shareholders | 2 |
|---|---|

- Bill Gates
- KPMG

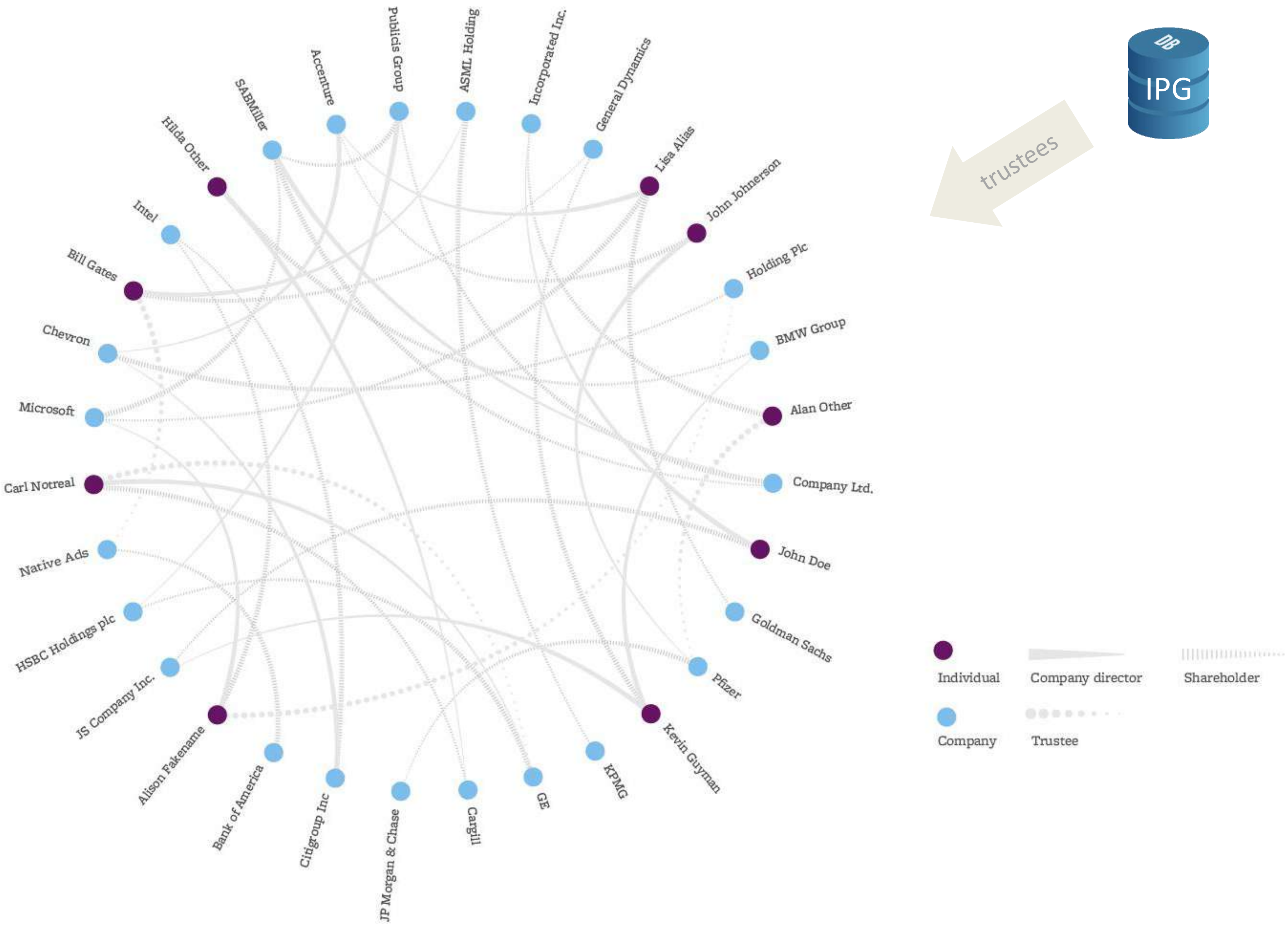
! ALERT: Multiple shareholders of Sole Shareholding Company

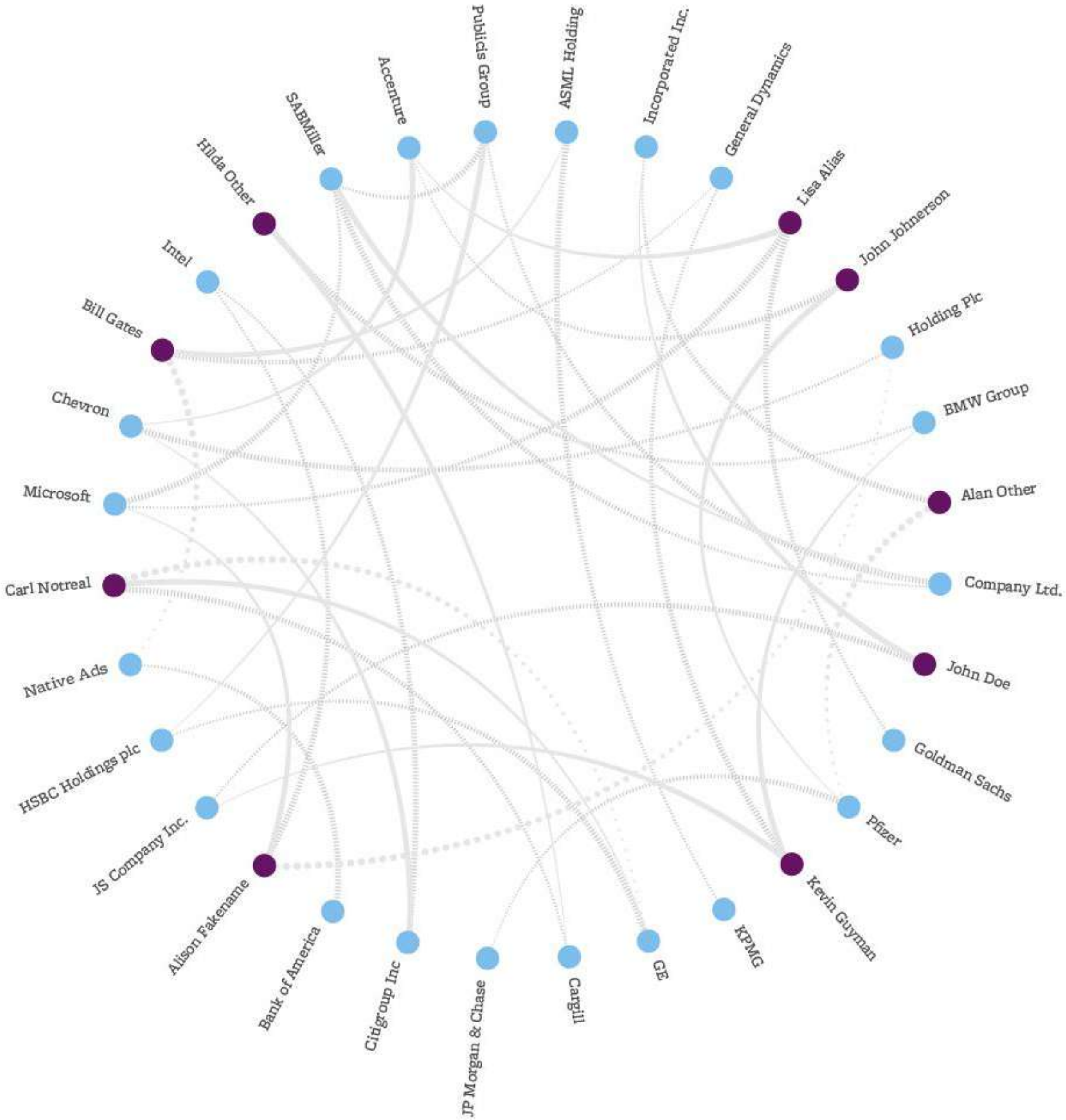
● Individual

Company director

Shareholder

● Company


Publicis Group
ASML Holding
Incorporated Inc.
General Dynamics
Lisa Alias
John Johnerson
Holding Plc
BMW Group
Alan Other
Company Ltd.
John Doe
Goldman Sachs
Pfizer
Kevin Guyman
KPMG
GE
Cargill
JP Morgan & Chase
Citigroup Inc
Bank of America
Alison Fakename
JS Company Inc.
HSBC Holdings plc
Native Ads
Carl Notreal
Microsoft
Chevron
Bill Gates
Intel
Hilda Other
SABMiller
Accenture
DB
IPG
trustees
Individual
Company director
Shareholder
Company
Trustee

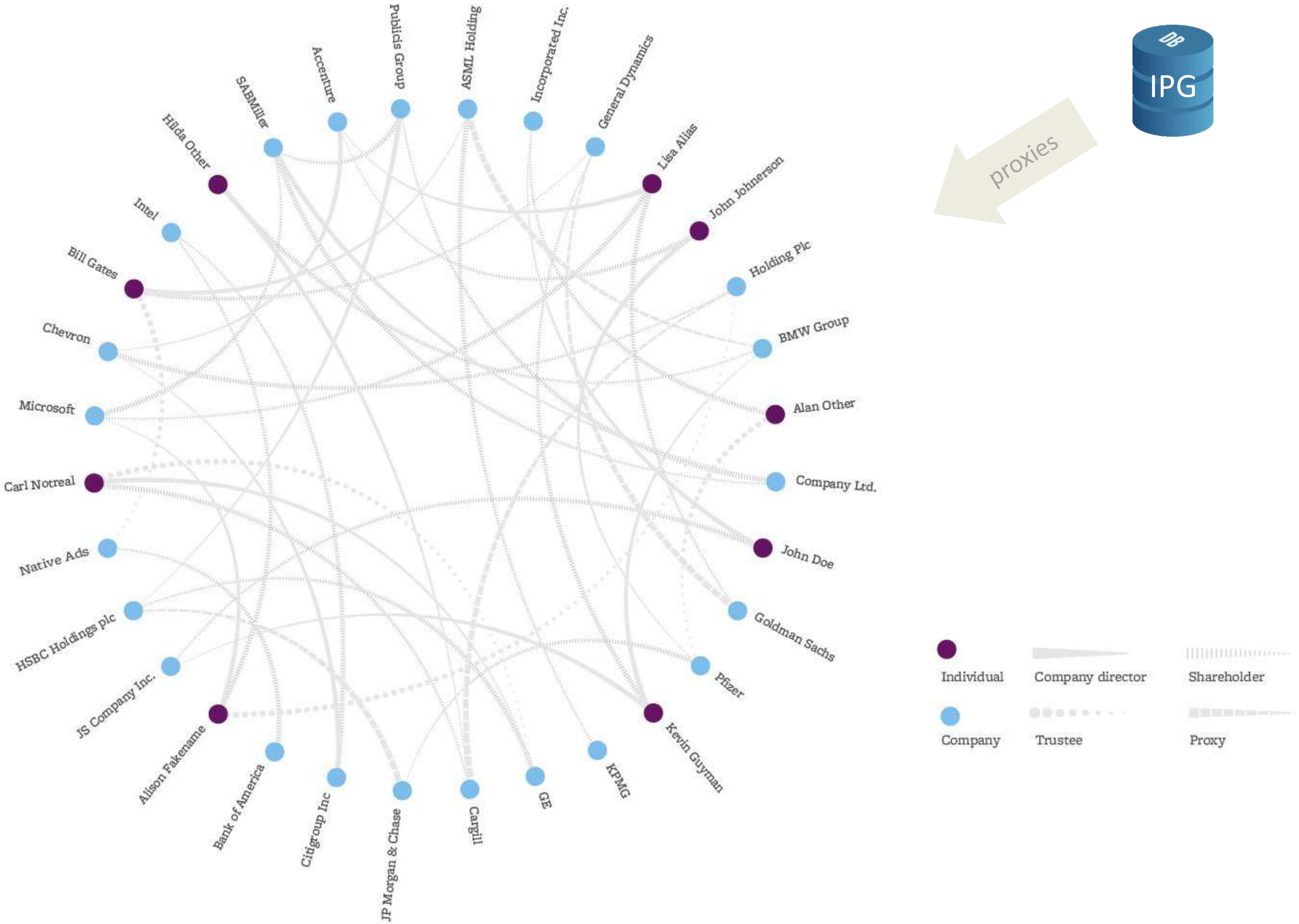


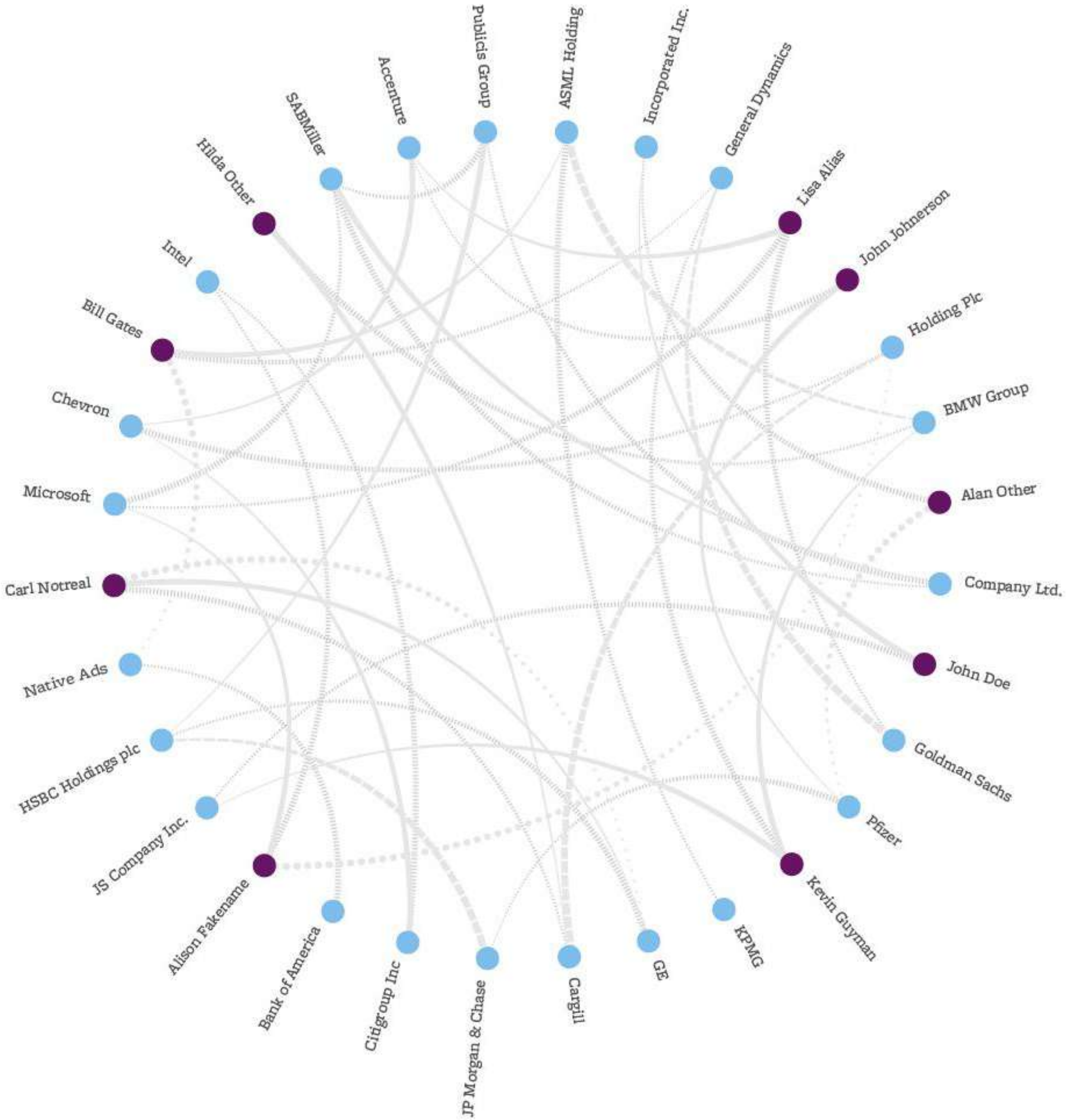
## Accenture

BANKRUPTCY
- From: 01/02/2005
- To: 02/09/2005

! ALERT: Bankruptcy without Trustee.

● Individual
Company director
Shareholder
● Company
Trustee


IPG
DB
proxies
Publicis Group
ASML Holding
Incorporated Inc.
General Dynamics
Lisa Alias
John Johnerson
Holding Plc
BMW Group
Alan Other
Company Ltd.
John Doe
Goldman Sachs
Pfizer
Kevin Guyman
KPMG
GE
Cargill
JP Morgan & Chase
Citigroup Inc
Bank of America
Alison Fakename
JS Company Inc.
HSBC Holdings plc
Native Ads
Carl Notreal
Microsoft
Chevron
Bill Gates
Intel
Hilda Other
SABMiller
Accenture
Individual
Company director
Shareholder
Company
Trustee
Proxy



## Carl Notreal

COMMERCIAL PROXY OF

- GE

SHAREHOLDINGS 1

- GE

❗ ALERT: commercial proxy without proxy type

● Individual

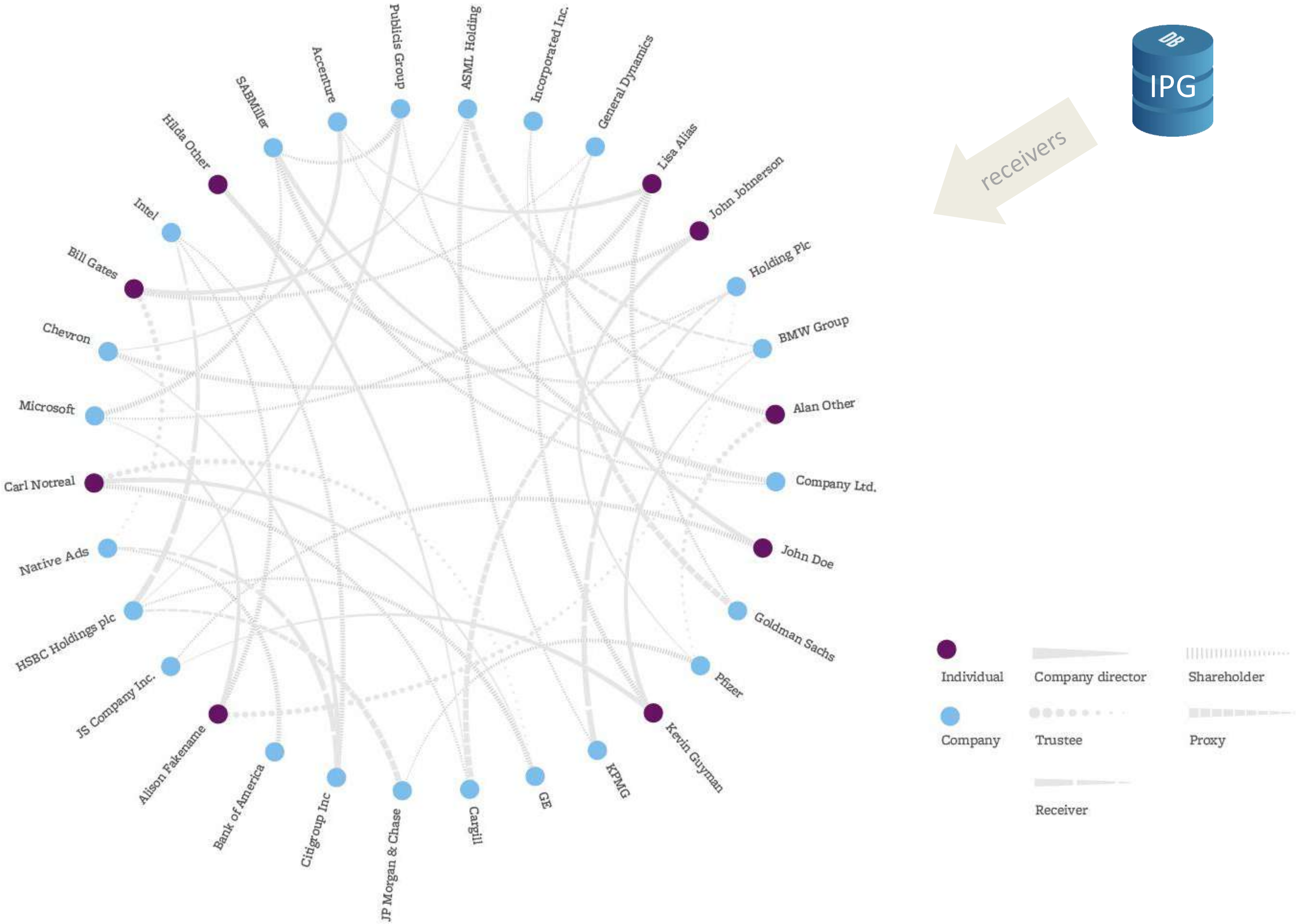
Company director

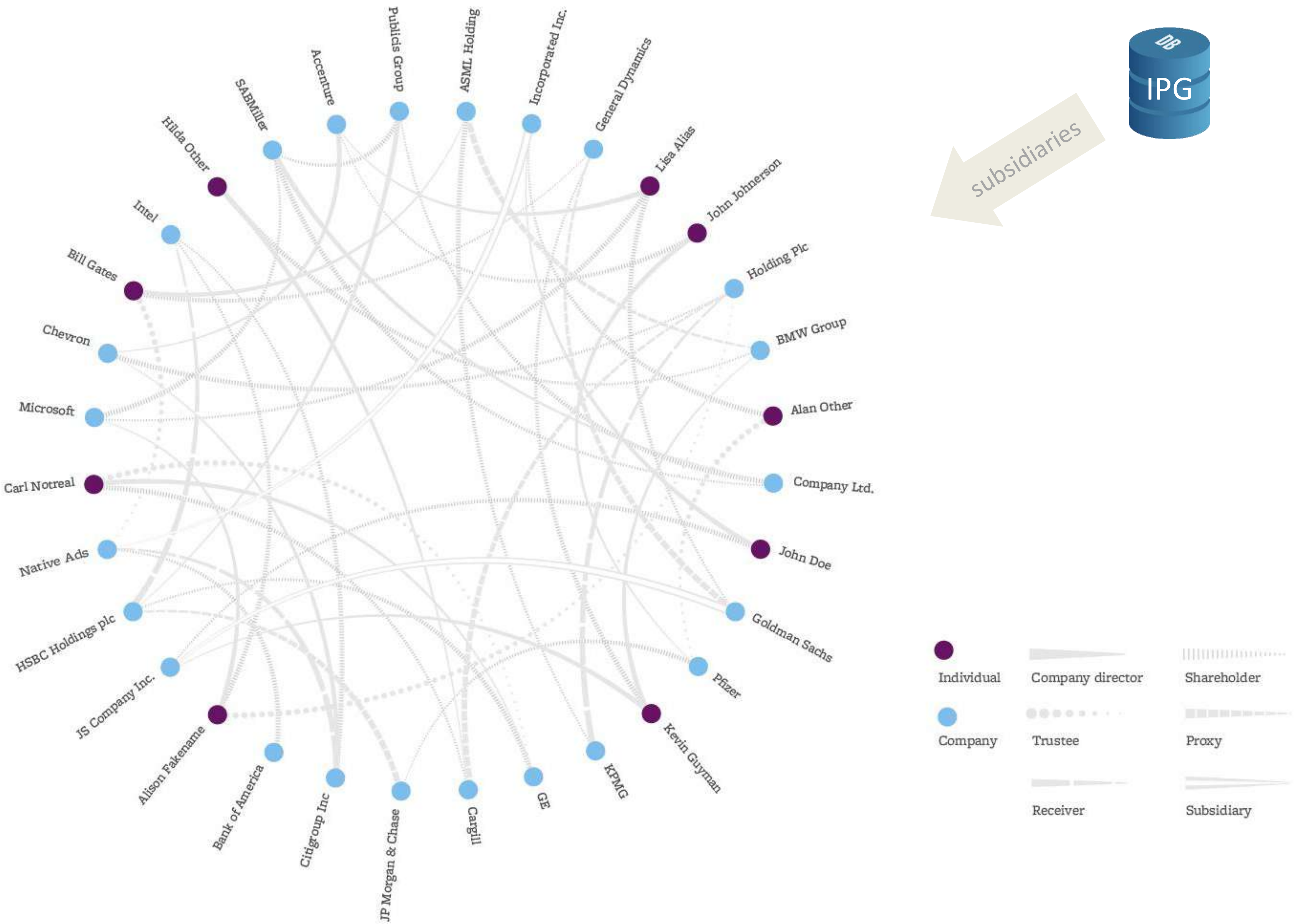
Shareholder

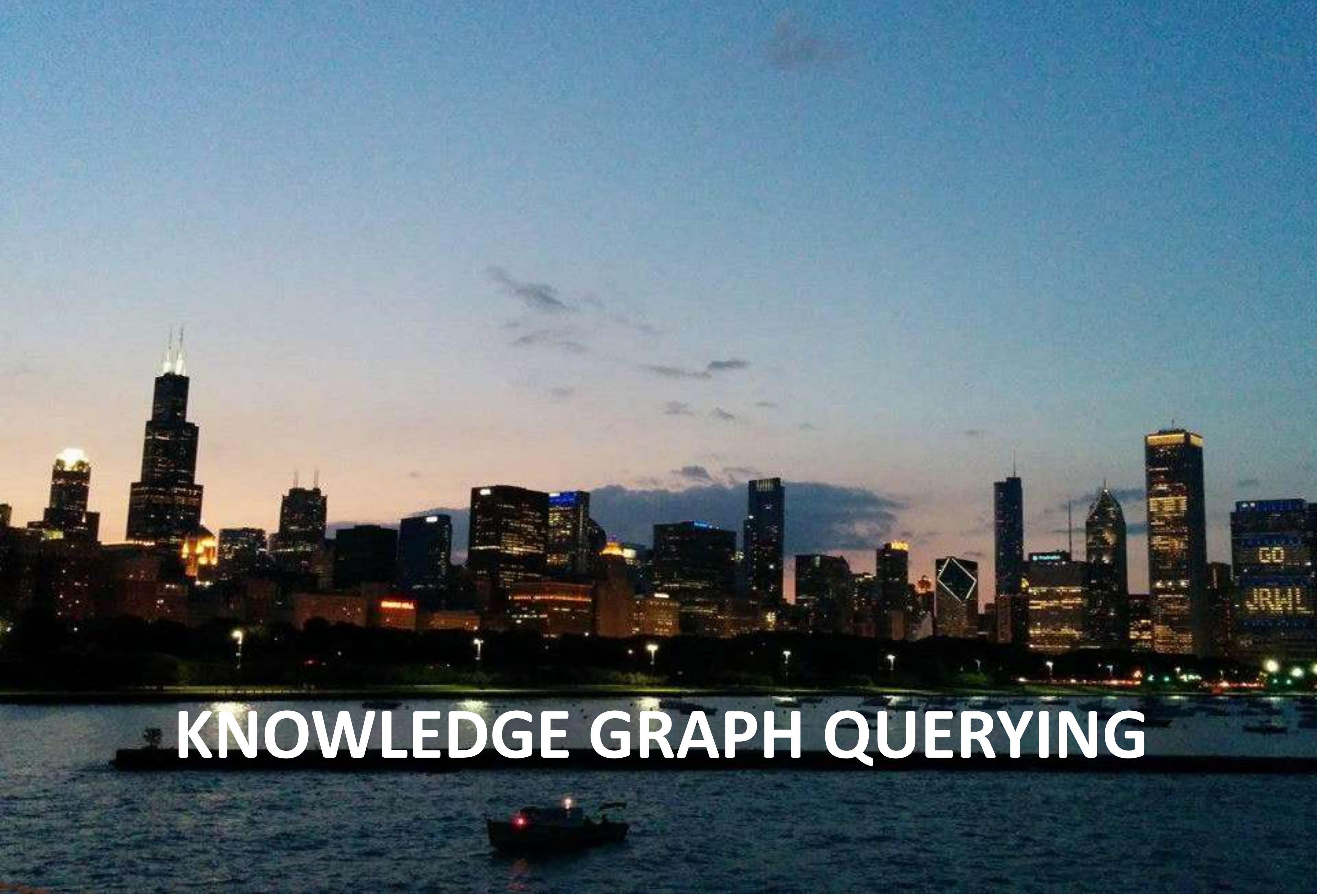
● Company

Trustee

Proxy


IPG
DB
receivers
Publicis Group
ASML Holding
Incorporated Inc.
General Dynamics
Lisa Alias
John Johnerson
Holding Plc
BMW Group
Alan Other
Company Ltd.
John Doe
Goldman Sachs
Pfizer
Kevin Guyman
KPMG
GE
Cargill
JP Morgan & Chase
Citigroup Inc
Bank of America
Alison Fakename
JS Company Inc.
HSBC Holdings plc
Native Ads
Carl Notreal
Microsoft
Chevron
Bill Gates
Intel
Hilda Other
SABMiller
Accenture
Individual
Company
Company director
Trustee
Receiver
Shareholder
Proxy


IPG
subsidiaries
Publicis Group
ASML Holding
Incorporated Inc.
General Dynamics
Lisa Alias
John Johnerson
Holding Plc
BMW Group
Alan Other
Company Ltd.
John Doe
Goldman Sachs
Pfizer
Kevin Guyman
KPMG
GE
Cargill
JP Morgan & Chase
Citigroup Inc
Bank of America
Alison Fakename
JS Company Inc.
HSBC Holdings plc
Native Ads
Carl Notreal
Microsoft
Chevron
Bill Gates
Intel
Hilda Other
SABMiller
Accenture
Individual
Company
Company director
Trustee
Receiver
Shareholder
Proxy
Subsidiary

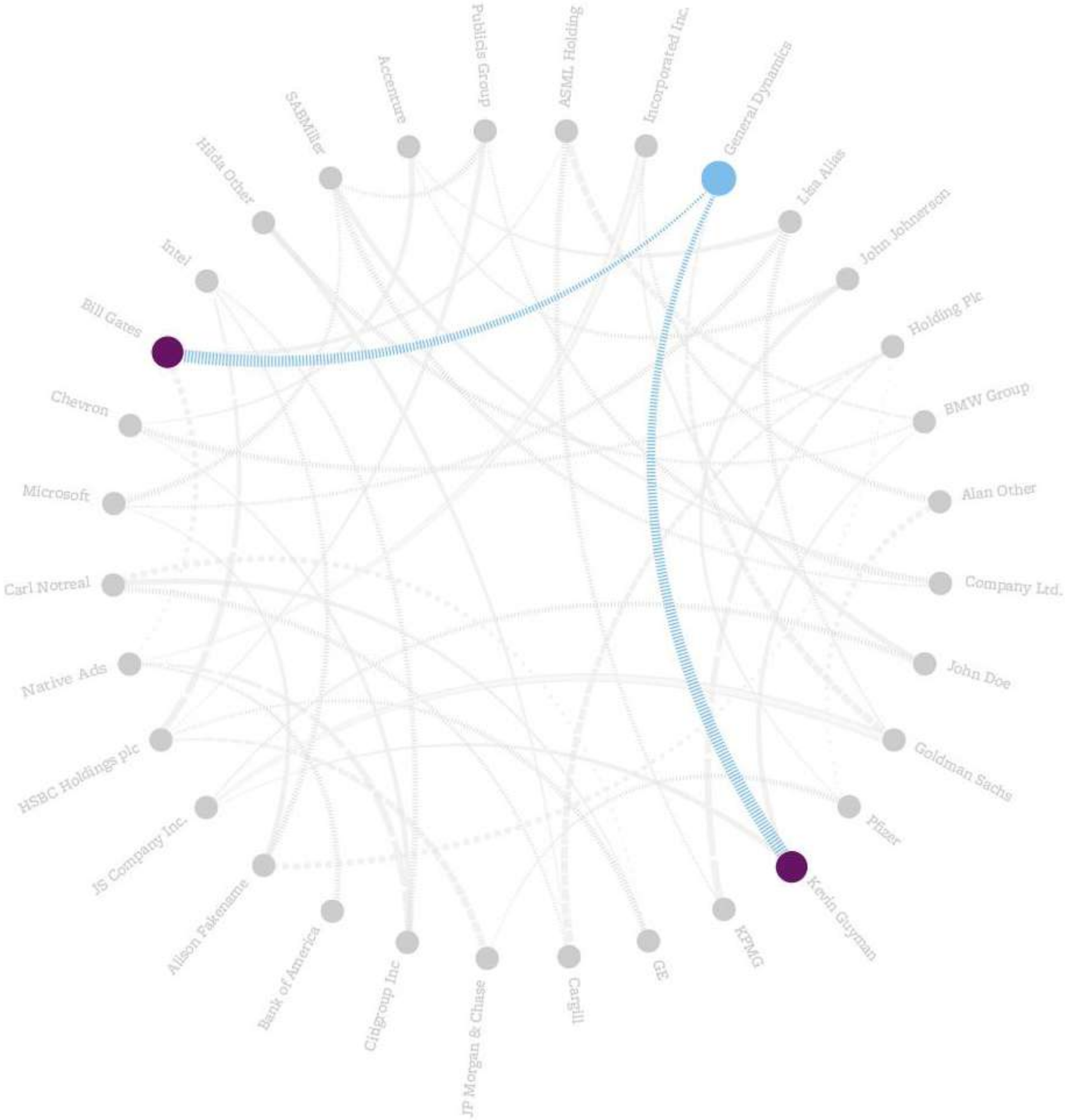


# KNOWLEDGE GRAPH QUERYING

ALIGNED
Software and Data Engineering

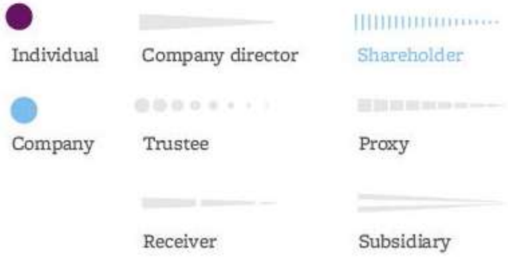
DataChemist

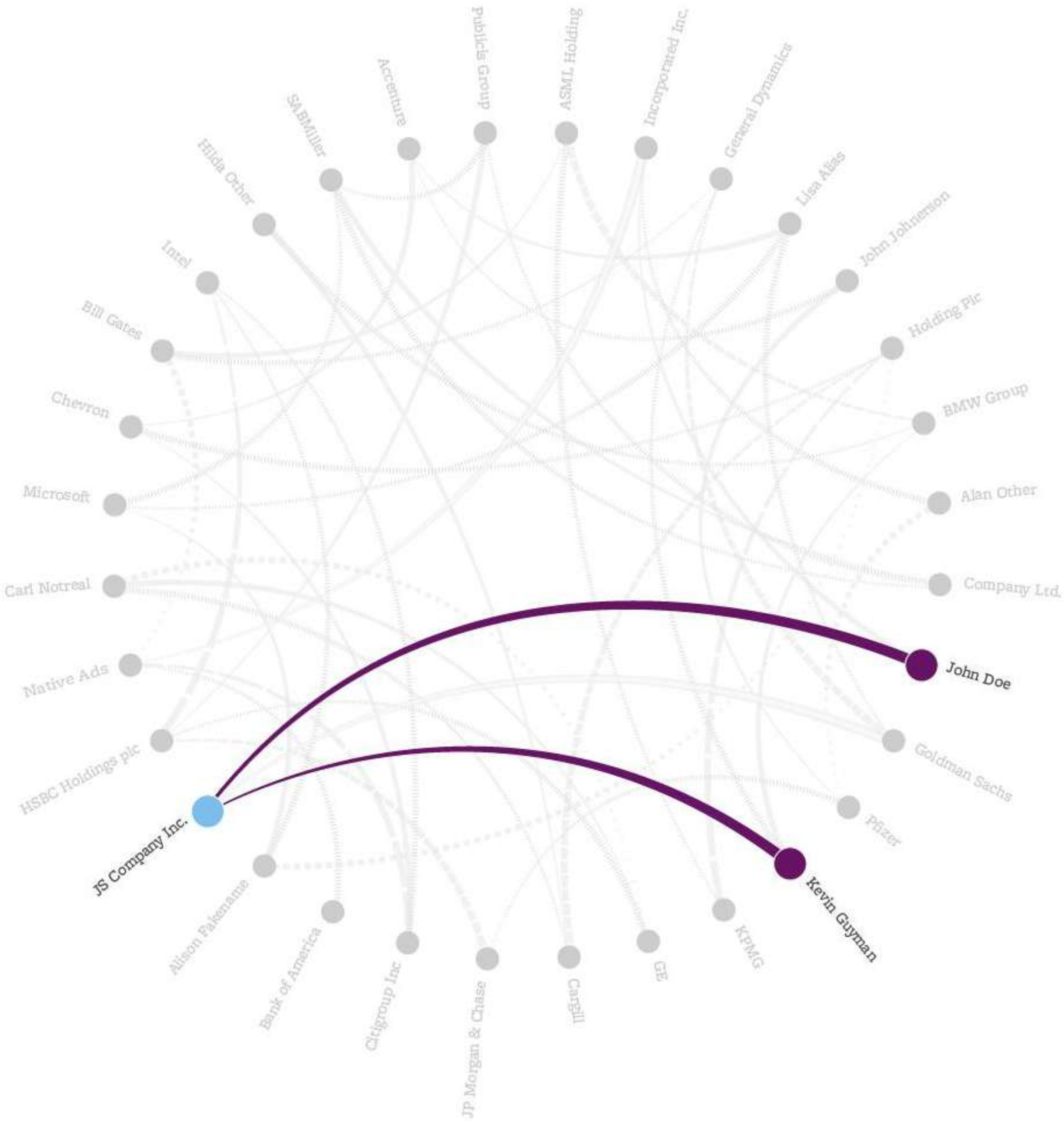
Wolters Kluwer



## General Dynamics Shareholders

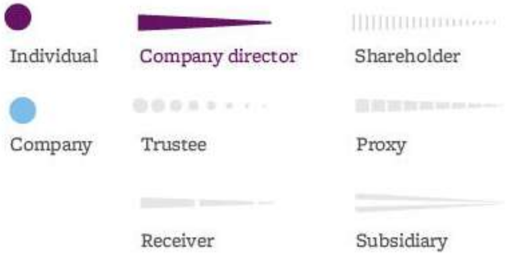
```
>  (x:Company).name ~=  'General dynamics‘ & x.shareholder -> (y:Person)
```

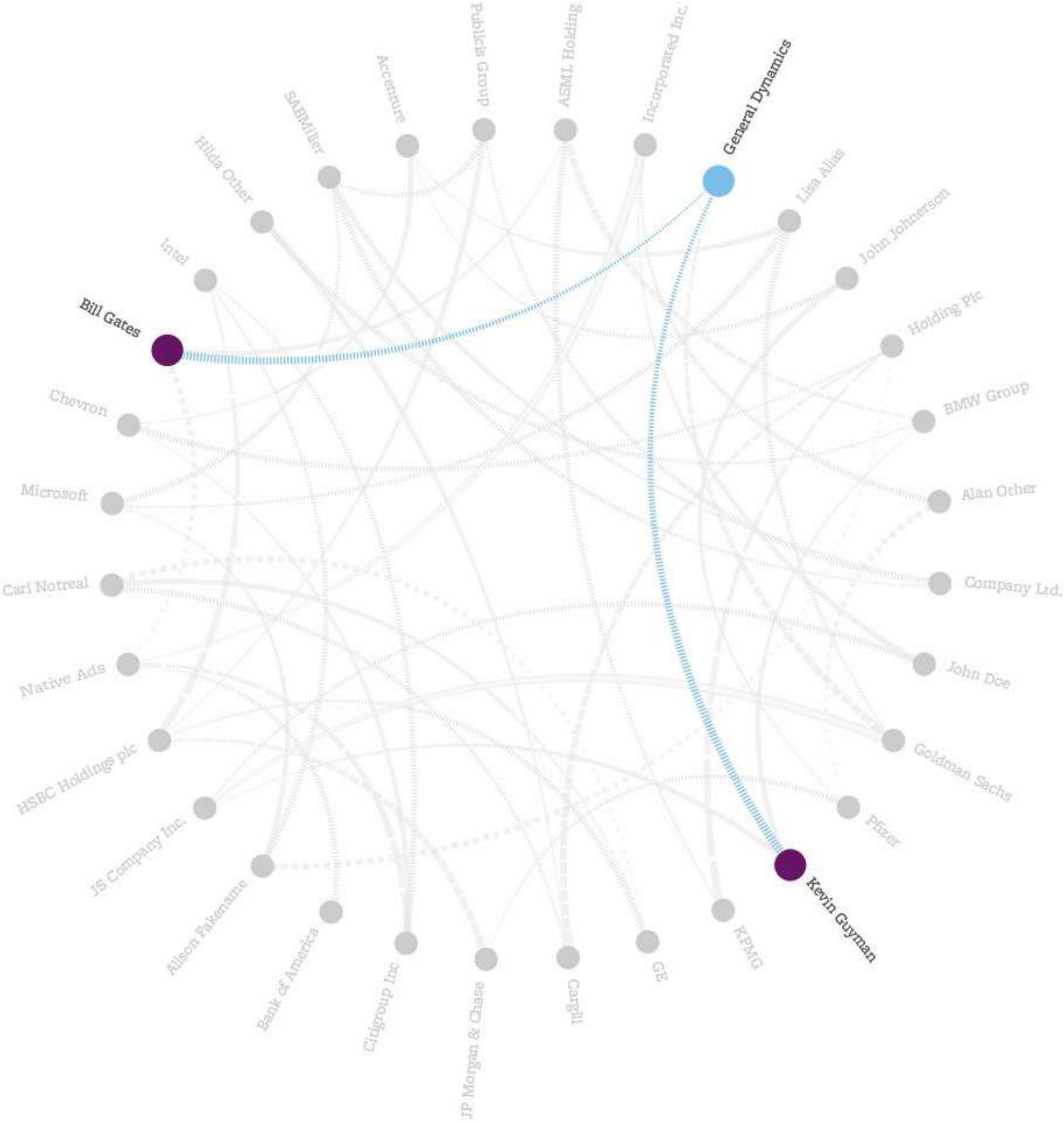




## Directors of JS Company Inc.

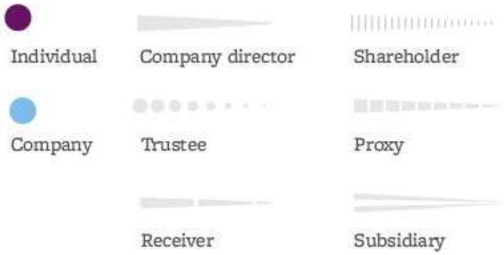
```
>  (x:Company).name ~= 'JS Company'
& x.director -> (y:Person)
```

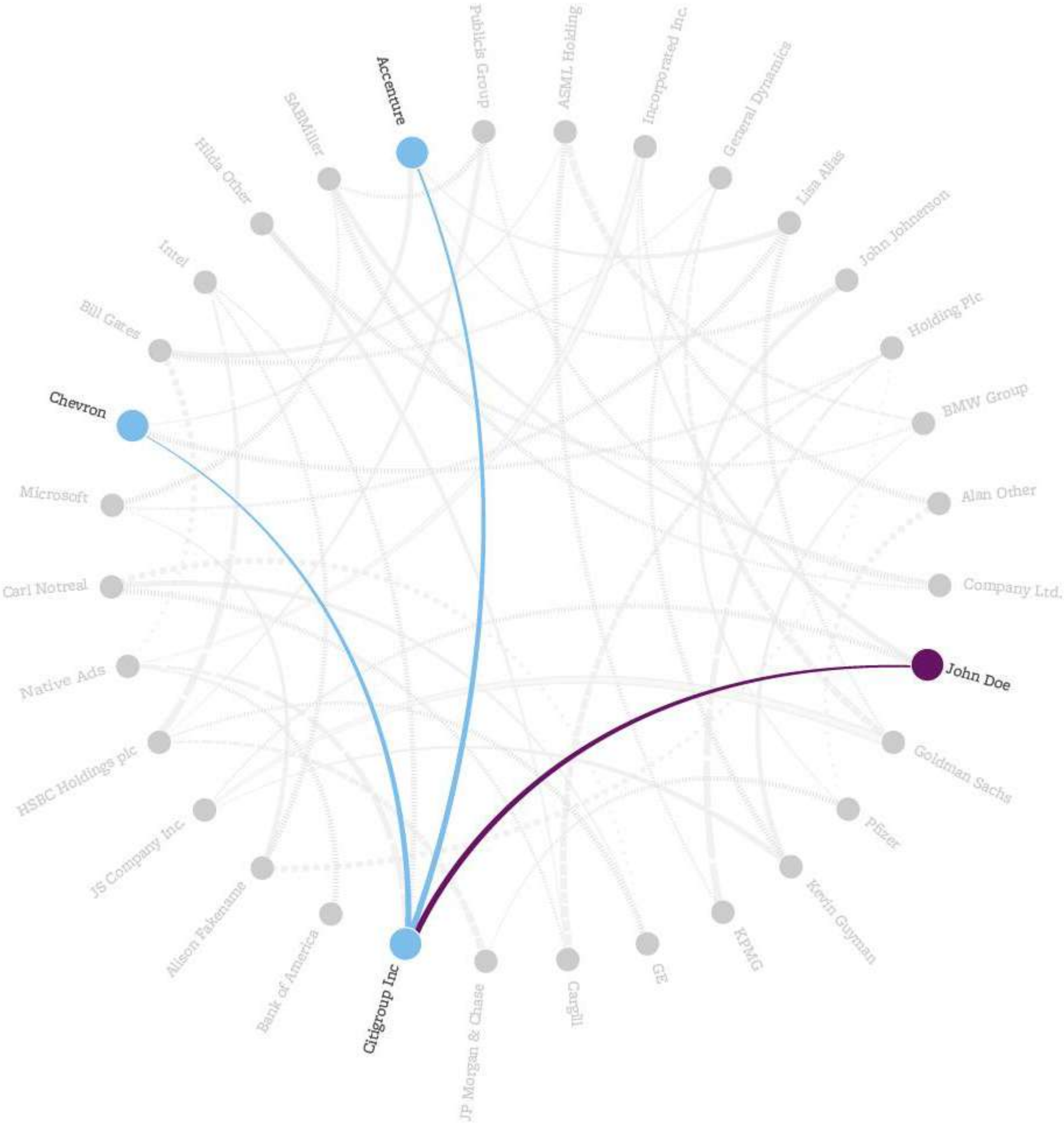




## People Linked to General Dynamics

```
>  (x:Company).name ~= 'General dynamics‘ & x.* -> (y:Person)
```





## All Citigroup Inc. Connections

> (x:Company).name ~= 'City Group' & x.* -> (y: [Person | Company])

● Individual

● Company

Company director

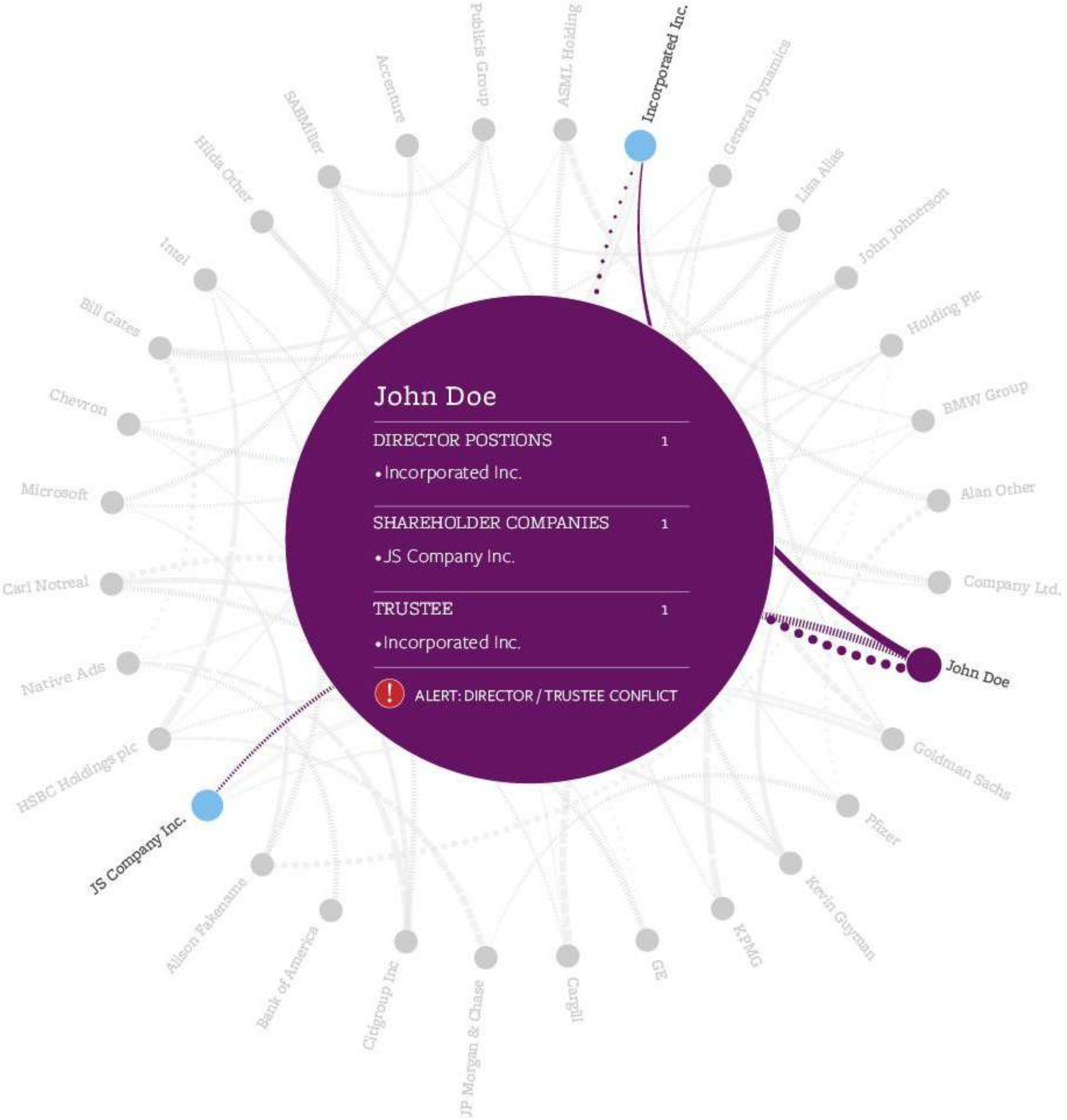
Trustee

Receiver

Shareholder

Proxy

Subsidiary



John Doe

DIRECTOR POSTIONS 1

- Incorporated Inc.

SHAREHOLDER COMPANIES 1

- JS Company Inc.

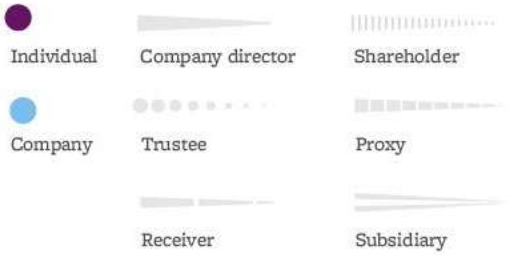
TRUSTEE 1

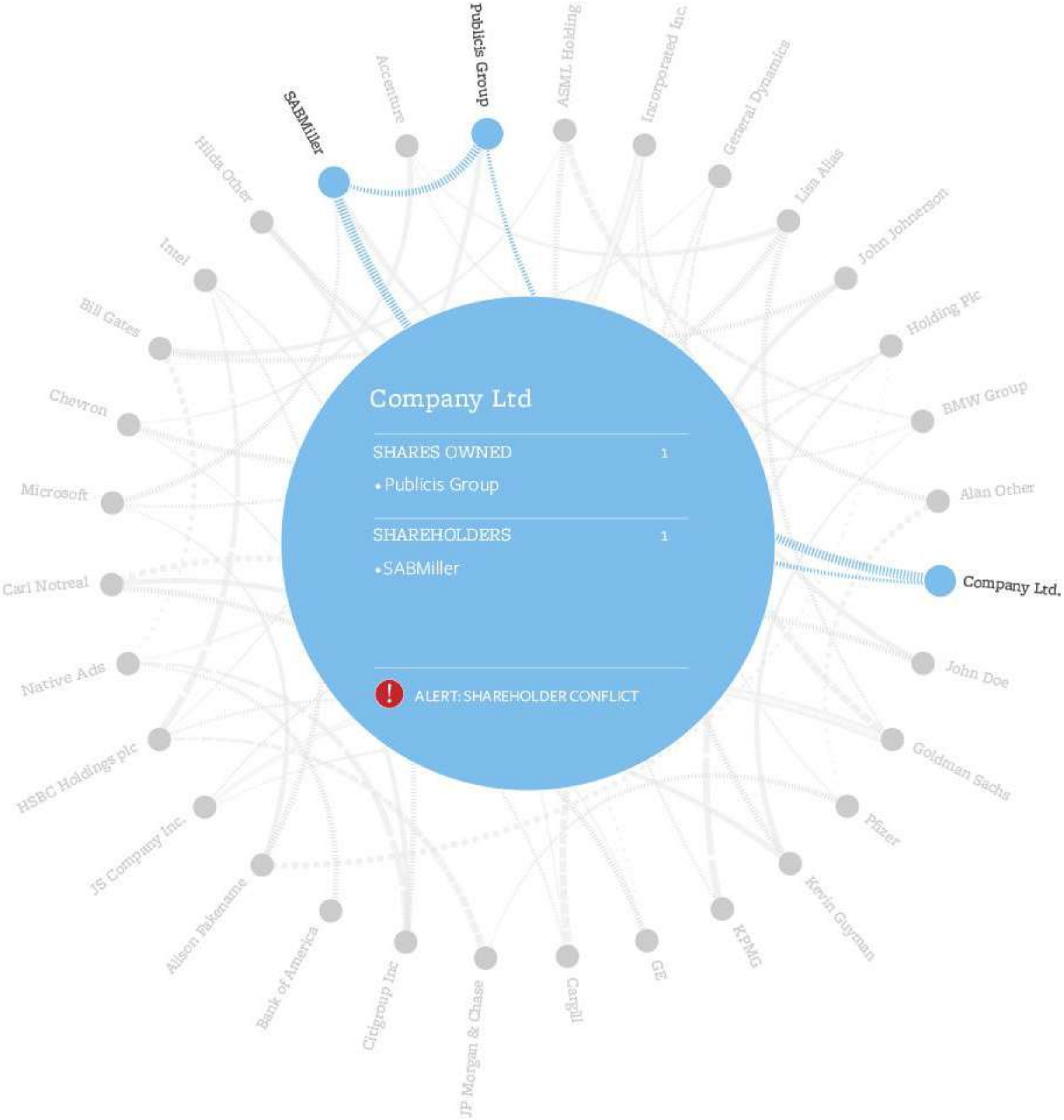
- Incorporated Inc.

! ALERT: DIRECTOR / TRUSTEE CONFLICT

## Temporal Constraints

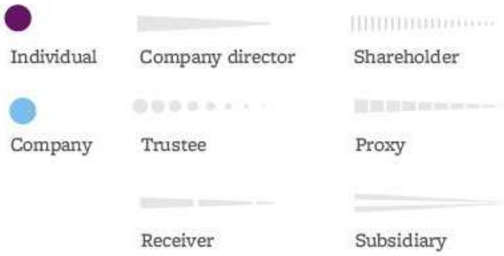
> (x:Person) = (y:Company).director & (x:Person) = y.trustee & (_.drector.lifespan) >< (_.trustee.lifespan)

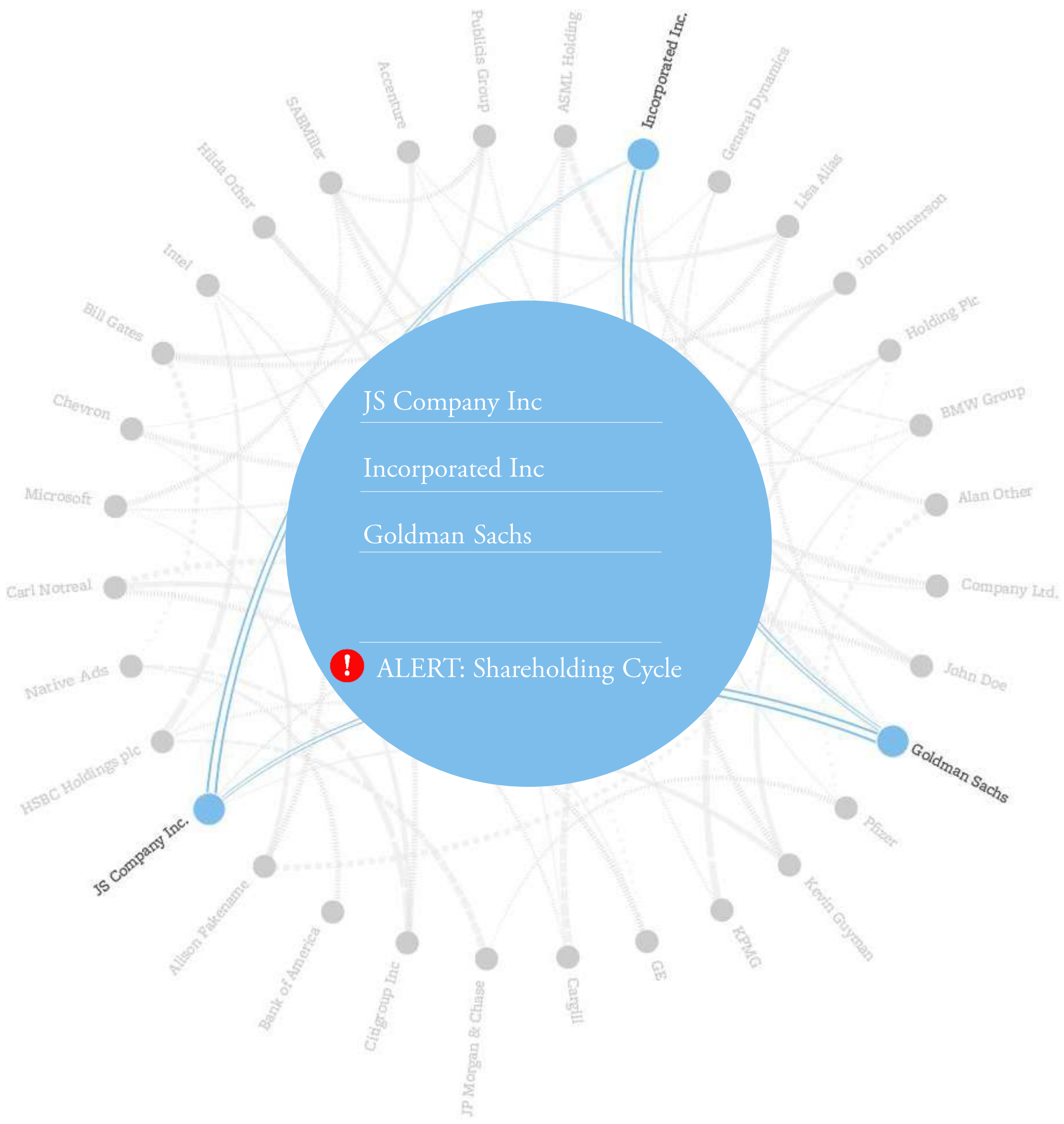




## Cross Shareholding Restrictions

```
>  (x:Company).name ~= 'Publicis Group'
& (y:Company).name ~= 'SABMiller'
& x.shareholder -> (z:Company)
& y.shareholder -> z
```





JS Company Inc

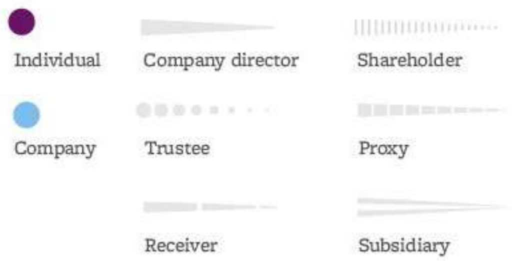
Incorporated Inc

Goldman Sachs

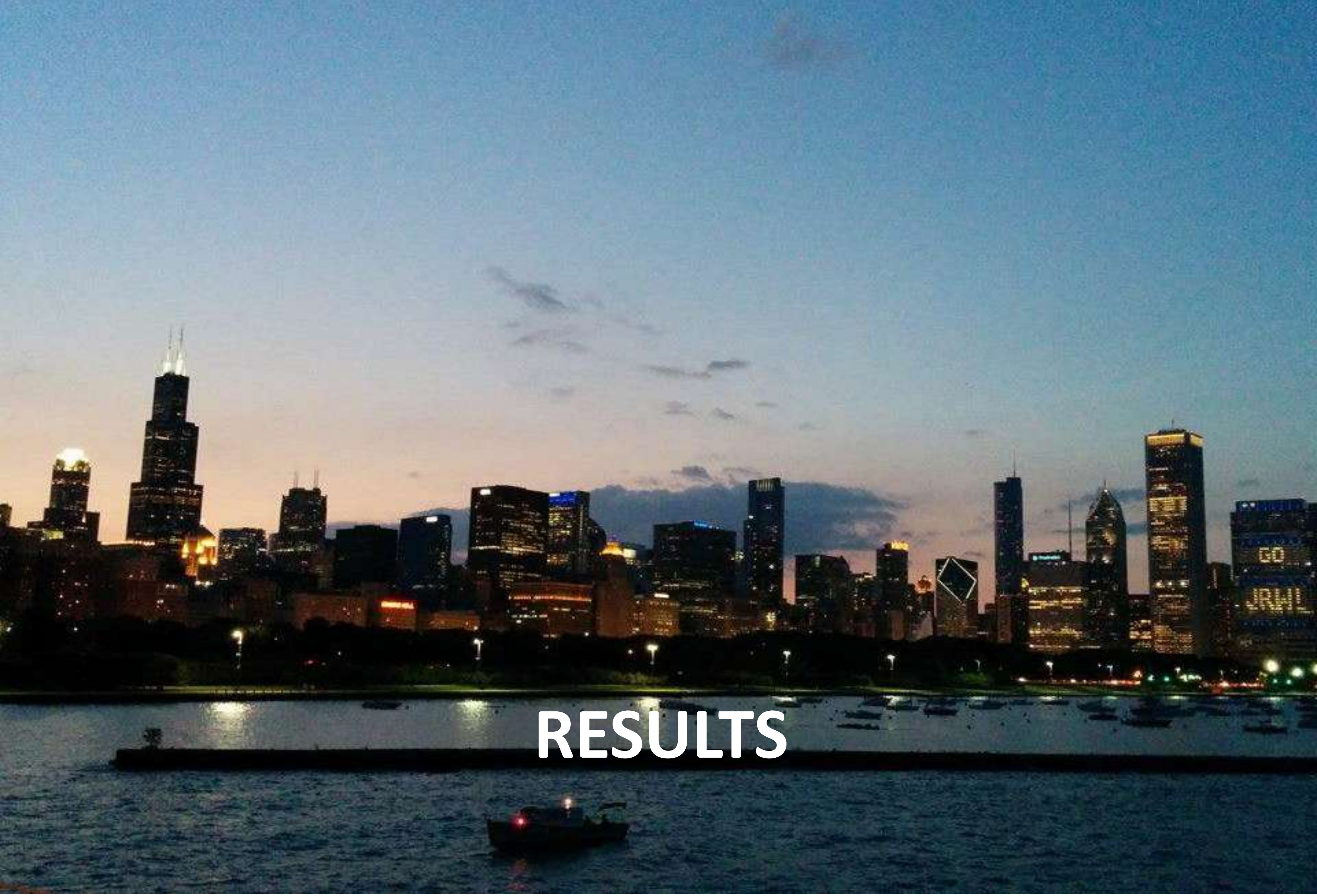
! ALERT: Shareholding Cycle

## Recursive Queries

> x.shareholder -> (_.shareholder)* -> x



# Live Demonstration



# RESULTS

# Solving the Unsolvables

| Error Type | Scenarios | Solved | Partially | Unsolved | Errors Detected |
|---|---|---|---|---|---|
| Basic Datatype | 5 | 5 | 0 | 0 | 8,500 |
| Missing Mandatory Properties | 12 | 10 | 2 | 0 | 10,032 |
| Temporal Constraints | 8 | 4 | 4 | 0 | 12,320 |
| Inconsistent Data | 4 | 2 | 1 | 1 | 1,000 |
| Temporal Queries | 1 | 1 | 0 | 0 | 5,324 |
| Recursive Queries | 1 | 1 | 0 | 0 | 909 |
| Model Complexity | 1 | 0 | 1 | 0 | NA |
| Total | 32 | 23 | 8 | 1 | **~40,000** |

ALIGNED Software and Data Engineering

DataChemist

Wolters Kluwer

# Other Findings

- The IPG Use-Case was added in the second half of the project. The work described here began in July 2017
- IPG is a very large dataset: 100 million SQL rows. This translated into 2 billion triples with provenance information included. We had to handle files > 100GB
  - Dealing with the scale of the data was by far the largest challenge – every piece of our tool-chain had to be rebuilt to deal with the size and speed requirements. Even *ed* breaks at that scale.
- IPG has a schema that has evolved over >15 years in response to immediate business needs.
  - The second largest challenge was deciphering the schema.
- A very large number of errors were found beyond the 32 unsolvables – many referential integrity violations, duplicates, inconsistent dates, typos….
  - The third largest challenge was parsing inconsistent formats used for the same field
- The first complete demonstration of running queries over the entire 2 billion-triple dataset was delivered on 5/3/2018
  - work is ongoing to complete the partial solutions.
- We estimate that our solution is 1-2 orders of magnitude faster and cheaper than existing methods. With the scaling work, we required 10 person months; without, 3 person months.

# Significance to Wolters Kluwer

- Creating domain specific knowledge models that drive new business and applications are at the core of our global WK corporate strategy (LegalTech, FinTech, Health, etc.)
- These applications are all over the place
- We need to semantify our data and we do not have the resources to build everything from scratch again
- This approach addresses several major challenges that we have to solve

**Knowledge graphs as a necessary ingredient in AI applications are now at the core of interest for companies. WKD can tell from its own industry, but also SWC from their customer side.**

# Questions?

Kevin Feeney & Gavin Mendel Gleason (CEO / CTO DataChemist)
kevin@datachemist.com gavin@datachemist.com

Christian Dirschl – Chief Content Architect, Wolters Kluwer Germany Christian.Dirschl@wolterskluwer.com