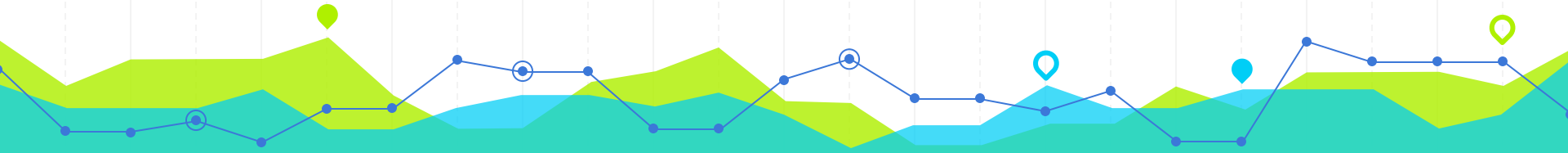


MC536 - Trabalho Final

Prof. André Santanchè
2°S 2019

Dupla LPT:

Gabriel Henrique Rosa Oswaldo - 172185

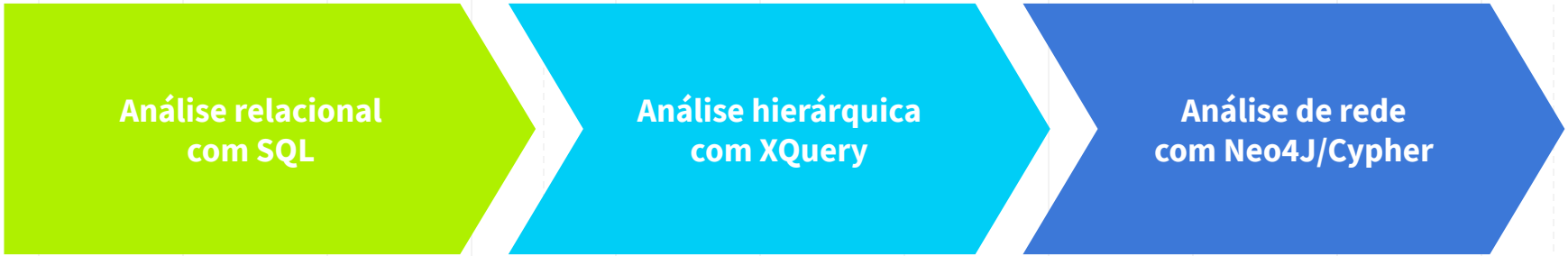
Lucas Silva Lopes Do Carmo - 202110

Dupla JSS:

Gabriel De Alcantara Bomfim Silveira - 197244

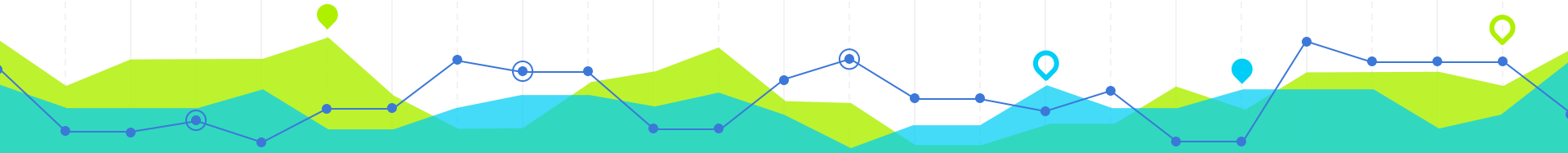
Vitor Coppo Ferreira - 206956

NOSSO PROCESSO



GitHub

https://github.com/gabrieloswaldo/mc536-trabalho



Análise relacional com SQL

1

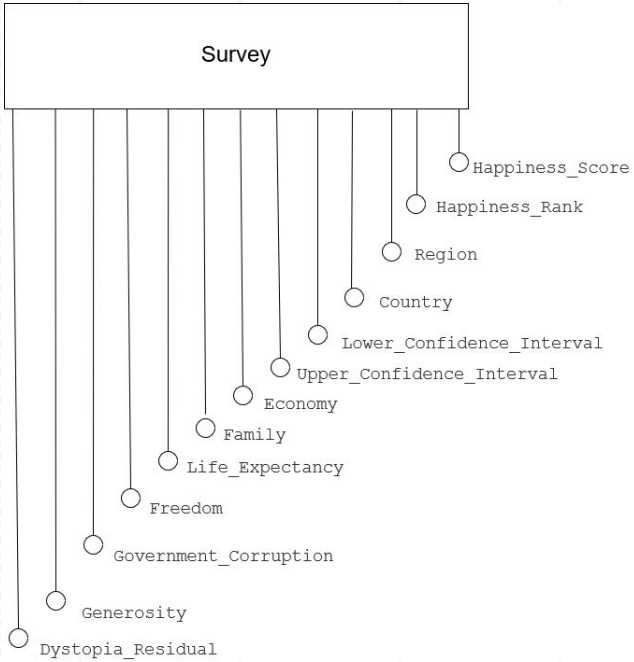
Modelos relacional e lógico, e análises em SQL

Dupla LPT

- Fonte de dados: [World Happiness Report 2016]
- Pontuações e classificações de felicidade nos países
- Um total de 157 instâncias e 13 variáveis, as quais incluem:

  - País
  - Região
  - Rank de felicidade
  - Pontuação de felicidade
  - Intervalos de confiança
  - Distopia Residual
  - **Família**
  - **Expectativa de Vida**
  - **Liberdade**
  - **Confiança no Governo**
  - **Generosidade**
  - **Economia (PIB)**

Modelo relacional:



Modelo lógico:

Survey (Country, Region, Happiness_Rank, Happiness_Score, Lower_Confidence_Interval, Upper_Confidence_Interval, Economy, Family, Life_Expectancy, Freedom, Government_Corruption, Generosity, Dystopia_Residual)

Análises

Agrupamento por região:

Média da pontuação de felicidade dos países, agrupados por região.

index	REGION	AVG_HAP_SCORE	COUNTRIES
0	Australia and New Zealand	7.323499999999999	2
1	North America	7.254	2
2	Western Europe	6.685666666666665	21
3	Latin America and Caribbean	6.10175	24
4	Eastern Asia	5.624166666666667	6
5	Middle East and Northern Africa	5.386052631578948	19
6	Central and Eastern Europe	5.3706896551724155	29
7	Southeastern Asia	5.338888888888889	9
8	Southern Asia	4.563285714285714	7
9	Sub-Saharan Africa	4.136421052631578	38

Análises

Correlações:

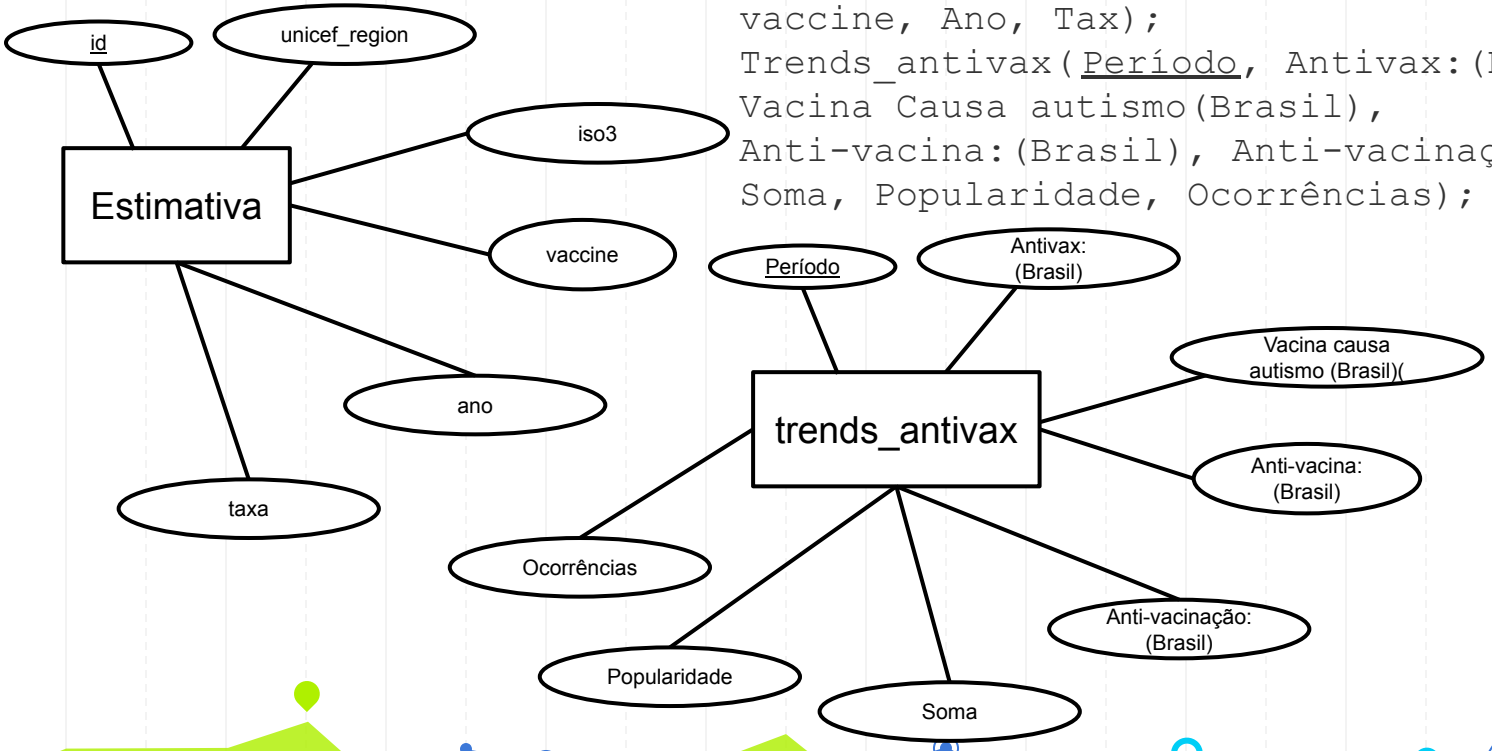
Uso do **coeficiente de Pearson** para avaliar o grau das relações entre a pontuação de felicidade e os seis fatores, determinando quais destes tem maior influência na pontuação

index	RELATION	COEFICIENT
0	economy	0.7903220167261241
1	life_expectancy	0.7653843344336755
2	family	0.7392515774070099
3	freedom	0.5668266730968905
4	gov_corruption	0.4020322451472926
5	generosity	0.15684779640360982

Dupla JSS

- Fontes de Dados: Unicef Child Immunization Report e Google Trends
- Estimativas de Imunização infantil ao longo dos últimos 10 anos em diversos países

Modelo relacional:



Modelo lógico:

```
Immunization_Estimate(id, unicef_region, iso3,
vaccine, Ano, Tax);
Trends_antivax(Período, Antivax:(Brasil),
Vacina Causa autismo(Brasil),
Anti-vacina:(Brasil), Anti-vacinação:(Brasil),
Soma, Popularidade, Ocorrências);
```

Análises

Medias:

Cálculo das médias de taxas de vacinação dos últimos 10 anos no Brasil, por vacina.

index	VACCINE	MEDIA_TAXA
0	DTP1	97
1	HIB3	94
2	DTP3	93
3	HEPB3	94
4	IPV1	91
5	HEPBB	87
6	BCG	96
7	MCV2	70
8	POL3	91
9	ROTAC	85
10	MCV1	95
11	RCV1	95
12	YFV	43
13	PCV3	80

Análises

Numero de vacinas acima/abaixo da média:

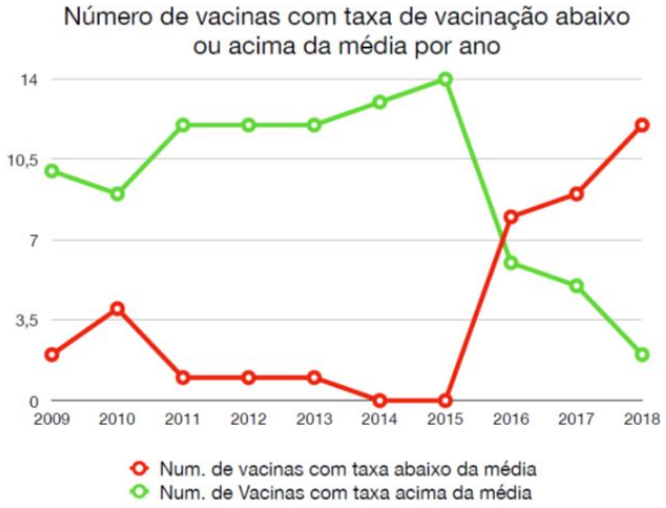
Comparação das taxas de vacinação com a média de cada ano.

Acima

index	ANO	COUNT(*)
1	2018	2
0	2017	5
9	2016	6
8	2015	14
7	2014	13
6	2013	12
5	2012	12
4	2011	12
3	2010	9
2	2009	10

Abaixo

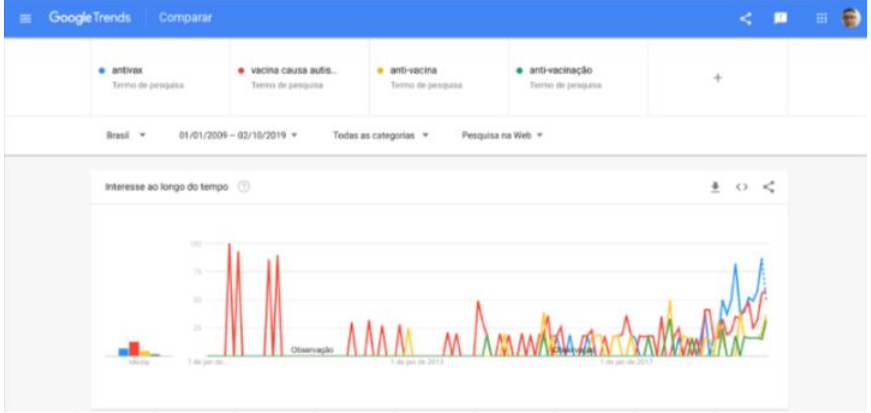
index	ANO	COUNT(*)
1	2018	12
0	2017	9
7	2016	8
6	2013	1
5	2012	1
4	2011	1
3	2010	4
2	2009	2

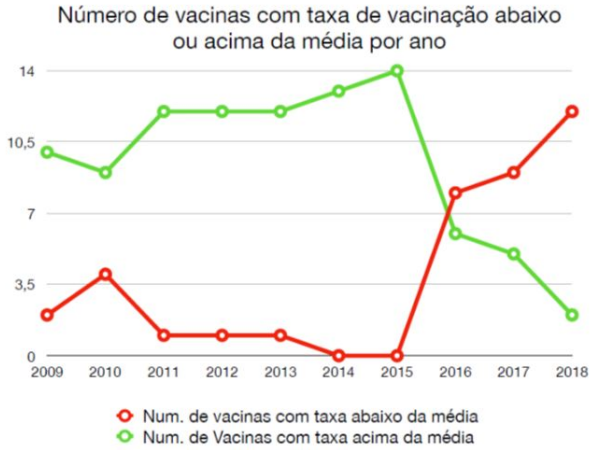


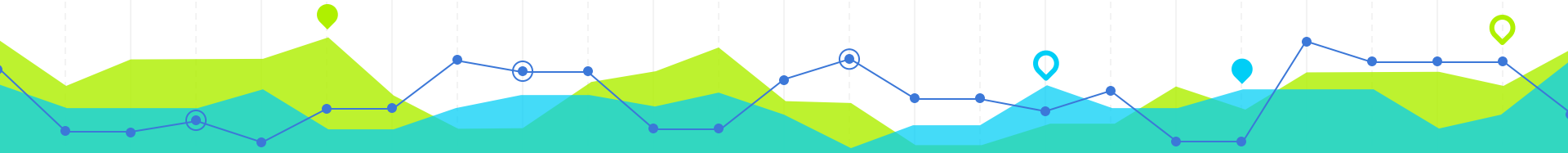
Análises

Trends anti-vacinação:

Visualização das tendências de pesquisa quanto ao movimento anti-vacinação, para comparação com a quantidade de vacinas acima da média







Análise hierárquica com XQuery

2

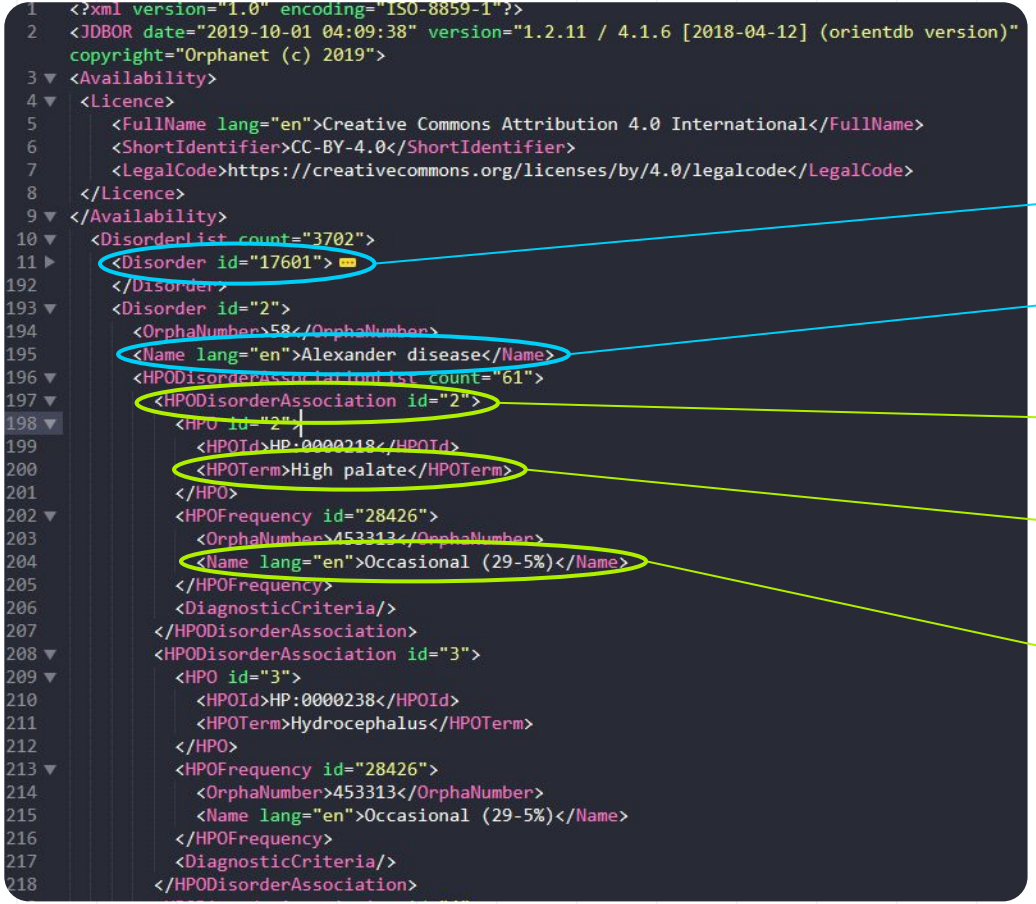
Consultas com abordagens hierárquicas

Orphadata

Rare Disease-Phenotype DB

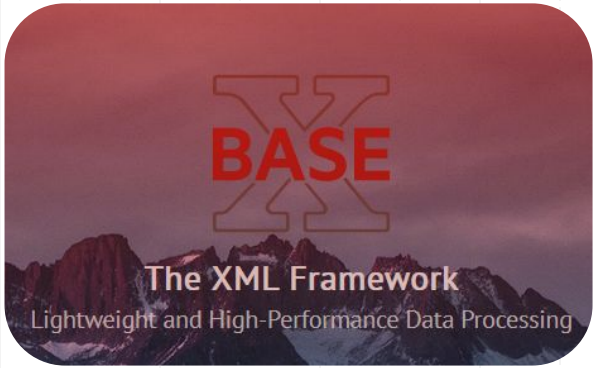
- http://www.orphadata.org/cgi-bin/index.php

- http://www.orphadata.org/data/xml/en_product4_HPO.xml


Doença rara
Nome da doença rara
Distúrbio associado a doença
Nome do distúrbio
Frequência com que o distúrbio se manifesta naquela doença

Preparação dos dados

- try.zorba/io não suportou a quantidade de dados
- Usamos então o BaseX
  - XML framework
  - XQuery 3.1 processor



Análises

XQuery 1: retorna a quantidade de doenças raras cadastradas na base;

XQuery 2: retorna a quantidade de distúrbios associados a doença rara com o id especificado;

XQuery 3: retorna a lista formatada de distúrbios com a frequência acima de “Frequente”, associadas a doença rara com o id especificado;

XQuery 4: retorna uma lista de doenças que possuem o sintoma especificado com frequência acima de “Frequent (79-30%)”;

XQuery 5: retorna uma lista de doenças que possuem ambos os sintomas especificados com uma frequência acima de ocasional.

POR QUÊ O MODELO HIERÁRQUICO?

Padronizacao

Modelo hierárquico e padrão no mundo todo, o que torna dados nesse modelo facilmente entendidos

Transito de dados

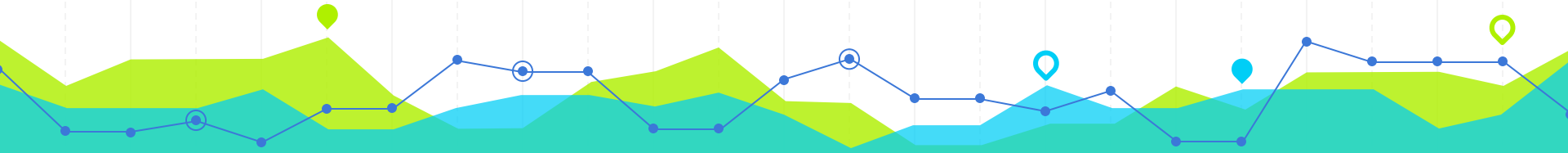
Inserir dados se torna fácil no decorrer do processo. Por exemplo, para colocar um nó, basta ligá-lo ao nó pai

Metadados

guardar ou vincular dados em qualquer formato, graças à liberdade dada ao usuário de definir suas marcações

Facilidade de busca

Nao e preciso de queries avançadas para procura de um dado. Assim basta um caminhamento em árvore para a busca do mesmo



Análise de rede com Neo4J/Cypher

3

Consultas com abordagens de rede

Suicide Rates Overview 1985 to 2016

Compares socio-economic info with suicide rates by year and country

- Fonte de dados: https://www.kaggle.com/russellyates88/suicide-rates-overview-1985-to-2016#master.csv

Preparação dos dados

- Dividimos o csv em duas partes, para compor as análises em rede
- Utilizamos então o Neo4J para o processamento de consultas em Cypher

Análises

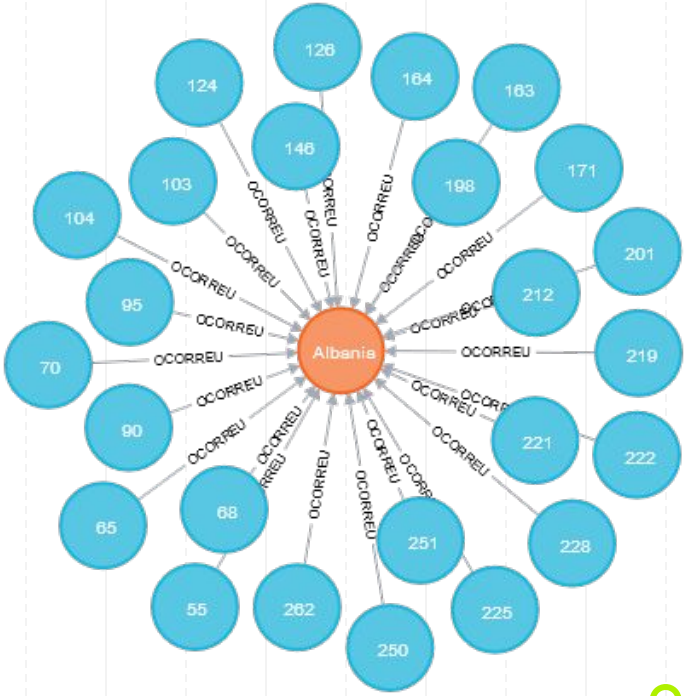
Importação dos dados

```
LOAD CSV WITH HEADERS FROM 'https://raw.githubusercontent.com/gabrieloswaldo/mc536-
trabalho/master/jupyter/data/suicidio-paises.csv' AS line
CREATE (:Pais {id: line.Id_pais, name: line.country})

LOAD CSV WITH HEADERS FROM 'https://raw.githubusercontent.com/gabrieloswaldo/mc536-
trabalho/master/jupyter/data/suicidios-casos.csv' AS line CREATE (:Suicidio {id: line.Id, sex:
line.sex, age: line.age, generation: line.generation})
```

Análises

Cria as relações de casos de suicídio com os países que ocorreram

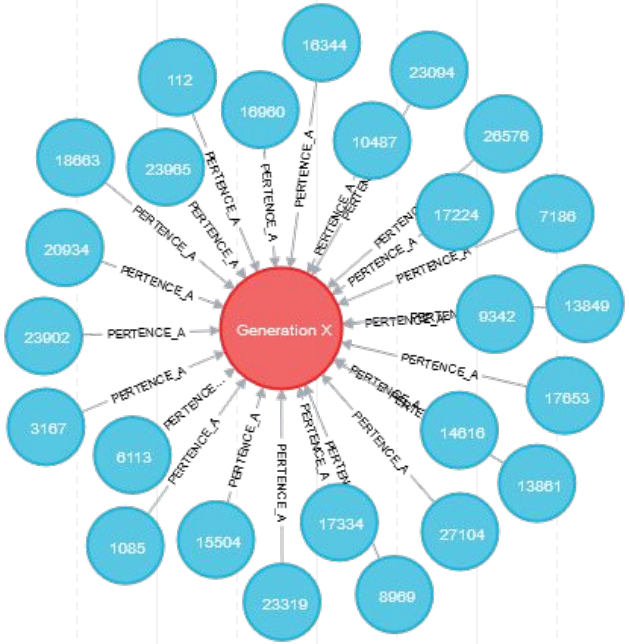
```
CREATE INDEX ON :Suicidio(id)
CREATE INDEX ON :Pais(name)
LOAD CSV WITH HEADERS FROM 'https://raw.githubusercontent.com/gabrieloswaldo/mc536-trabalho/master/jupyter/data/suicidios-relations.csv' AS csvLine
MATCH (p:Pais {name: csvLine.country})
MATCH (c:Suicidio {id: csvLine.Id})
CREATE (c)-[:OCORREU {ano: csvLine.year}]->(p)
```



Análises

Cria um nó para cada geração e cria a relação entre os casos de suicídio e a qual geração a pessoa pertenceu

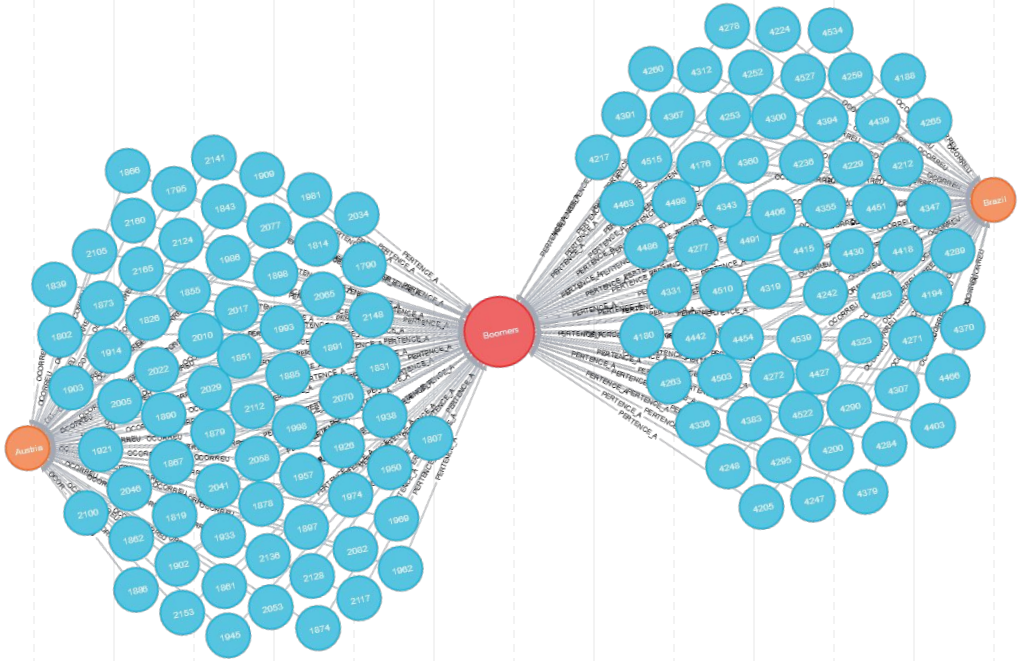
```
LOAD CSV WITH HEADERS FROM 'https://raw.githubusercontent.com/gabrieloswaldo/mc536-trabalho/master/jupyter/data/suicidios-casos.csv' AS line
MERGE (g:Generation {generation: line.generation})

MATCH (g:Generation)
MATCH (s:Suicidio)
WHERE s.generation = g.generation
CREATE (s)-[:PERTENCE_A]->(g)
```



Análises

Retorna a rede de relações entre os suicídios de pessoas e qual geração ela pertenceu, especificada para dois países

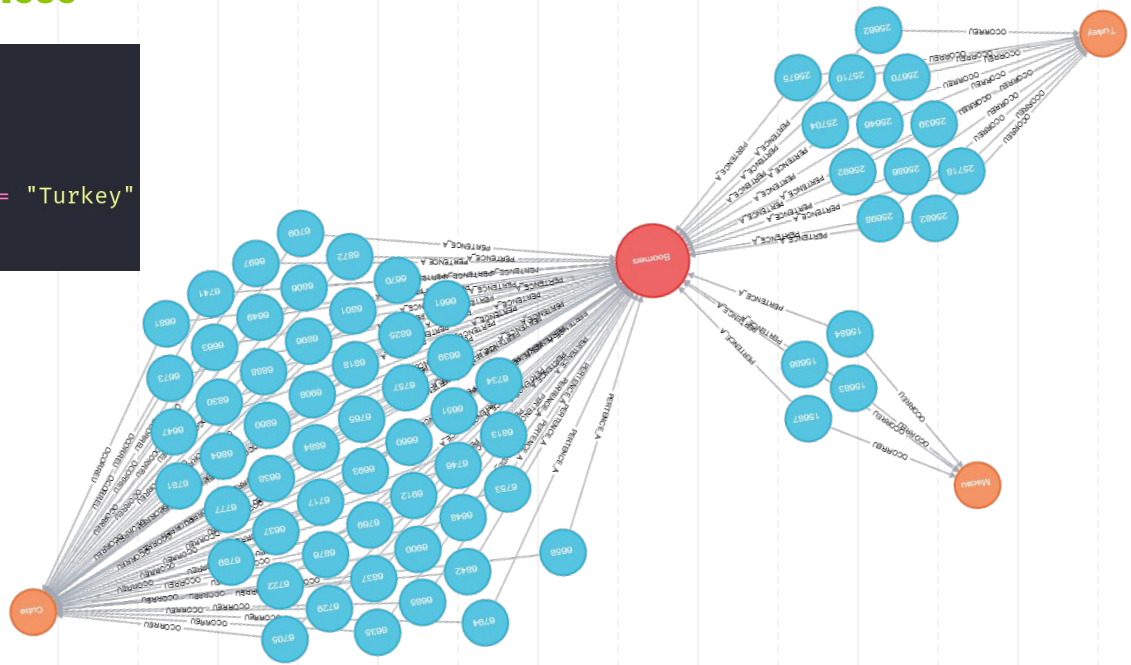
```
MATCH (g:Generation {generation:"Boomers"})
MATCH (s:Suicidio)-[:OCORREU]->(p)
WHERE p.name = "Brazil" OR p.name = "Austria"
RETURN (s)-[:PERTENCE_A]->(g), p
```



Análises

Retorna a rede de relações entre os suicídios de pessoas e qual geração ela pertenceu, especificada para três países

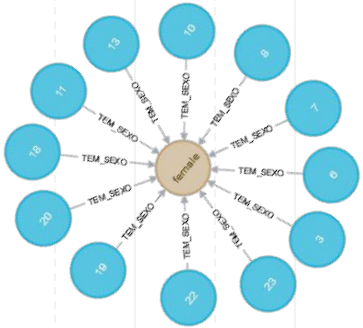
```
MATCH (g:Generation {generation:"Boomers"})
MATCH (s:Suicidio)-[:OCORREU]→(p)
WHERE p.name = "Cuba" OR p.name = "Macau" OR p.name = "Turkey"
RETURN (s)-[:PERTENCE_A]→(g), p
```

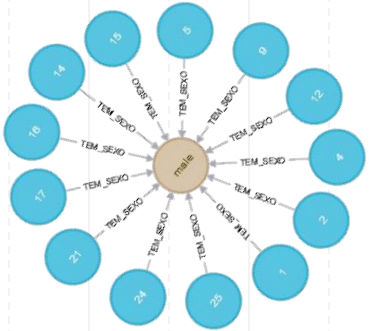


Análises

Cria relações entre suicidas Homens com o nó Male e suicidas mulheres com o nó Female.

```
LOAD CSV WITH HEADERS FROM 'https://raw.githubusercontent.com/gabrieloswaldo/mc536-trabalho/master/jupyter/data/suicidios-casos.csv' AS line
MERGE (s:Sex {sex: line.sex})

MATCH (sex:Sex)
MATCH (sui:Suicidio)
WHERE sui.sex = sex.sex
CREATE (sui)-[:TEM_SEXO]→(sex)
```

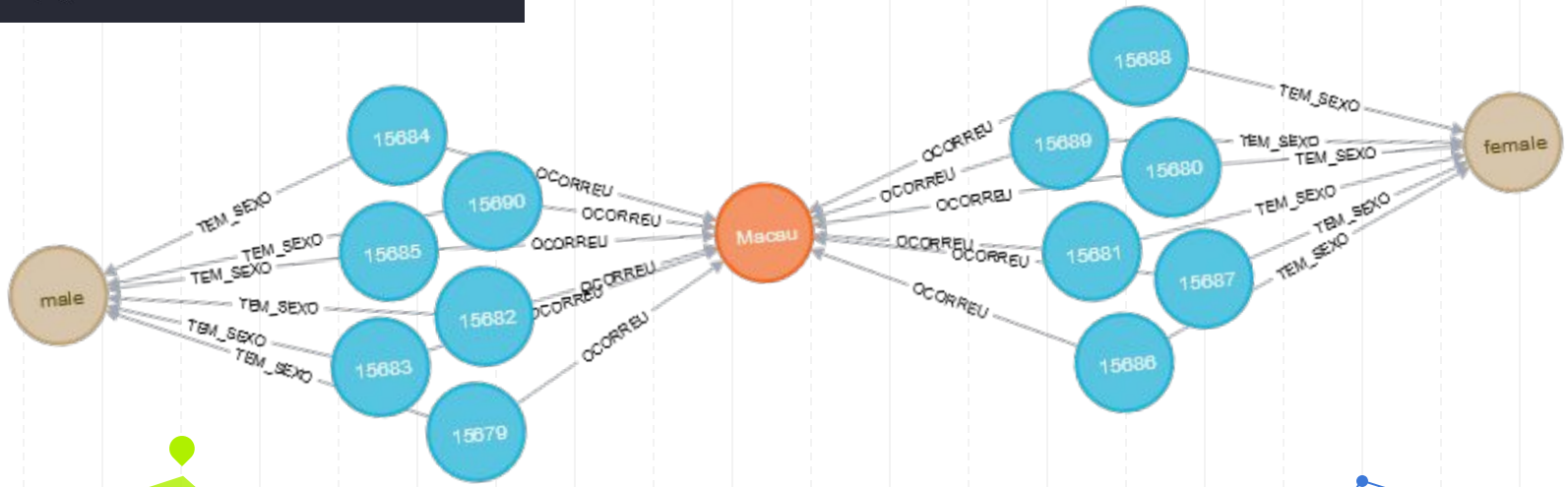




Análises

Criacao de relacionamentos

```
MATCH (s:Suicidio)-[:OCORREU]→(p:Pais)
MATCH (s:Suicidio)-[:TEM_SEXO]→(sex:Sex)
WHERE p.name = "Macau"
RETURN s, sex, p
```



Análises

Pagerank no grafo e retorna o score para os países com mais suicídios

```
CALL algo.pageRank.stream('Page', 'LINKS', {iterations:20, dampingFactor:0.85})
YIELD nodeId, score
RETURN algo.asNode(nodeId).name AS name,score
ORDER BY score DESC
```

name	score
"Austria"	48.85517120361328
"Iceland"	48.85517120361328
"Mauritius"	48.85517120361328
"Netherlands"	48.85517120361328
"Argentina"	47.58015823364258
"Belgium"	47.58015823364258
"Brazil"	47.58015823364258
"Chile"	47.58015823364258
"Colombia"	47.58015823364258
"Ecuador"	47.58015823364258
"Greece"	47.58015823364258
"Israel"	47.58015823364258
"Italy"	47.58015823364258
"Japan"	47.58015823364258
"Luxembourg"	47.58015823364258

POR QUÊ O MODELO DE REDE?

Grafos

Modelo extremamente visual, que facilita interpretação e entendimento por pessoas

.csv

Grafos de rede podem ser montados a partir de formatos comunmente encontrados, como csv e tsv

Ranking Algorithms

Existem algoritmos que permitem ranqueamento, agrupamento por comunidade, entre outros.

OBRIGADO!

Alguma pergunta?