Machine learning enabled by network graphs: the power of connecting your data

Clair Sullivan, PhD
Data Science Advocate

@CJLovesData1

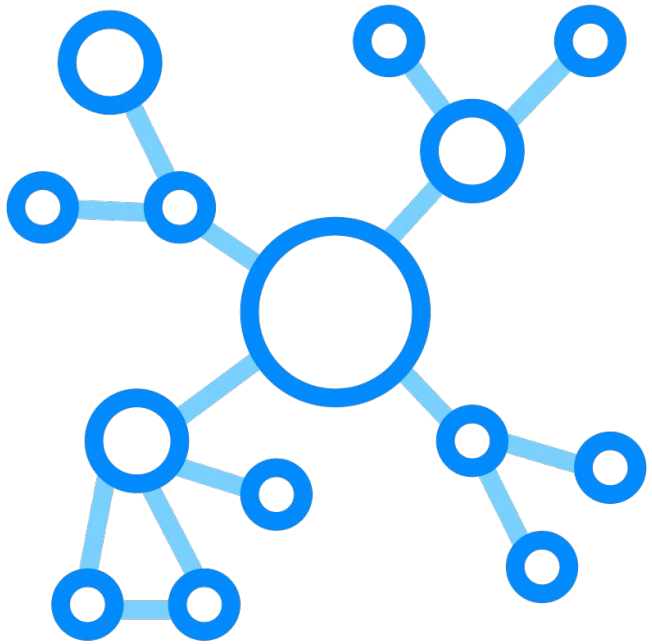
Neo4j, Inc. All rights reserved 2021

Introduction

- What is a graph?
- Why can you (sometimes) get better solutions when augmenting your data with a graph?
- What is a “graph-y” problem?
- How do you create one?
- What can you do with one once you have it?
- Machine learning with graphs

Neo4j, Inc. All rights reserved 2021

What is a graph?

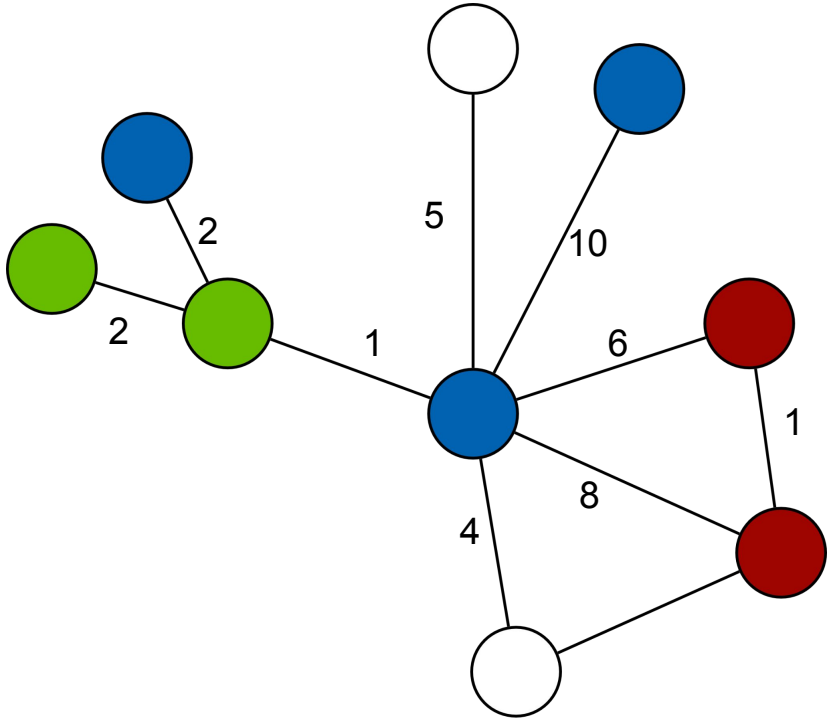


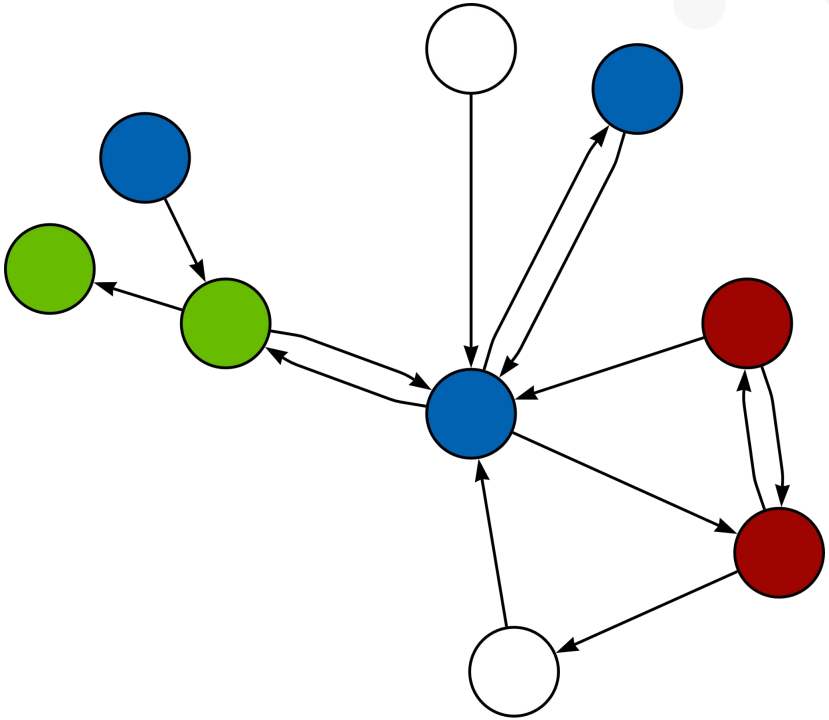
Common examples

- Social media
- Internet routing
- Maps, wayfinding
- Recommender systems
- Search
- Knowledge graphs, question answering

Neo4j, Inc. All rights reserved 2021

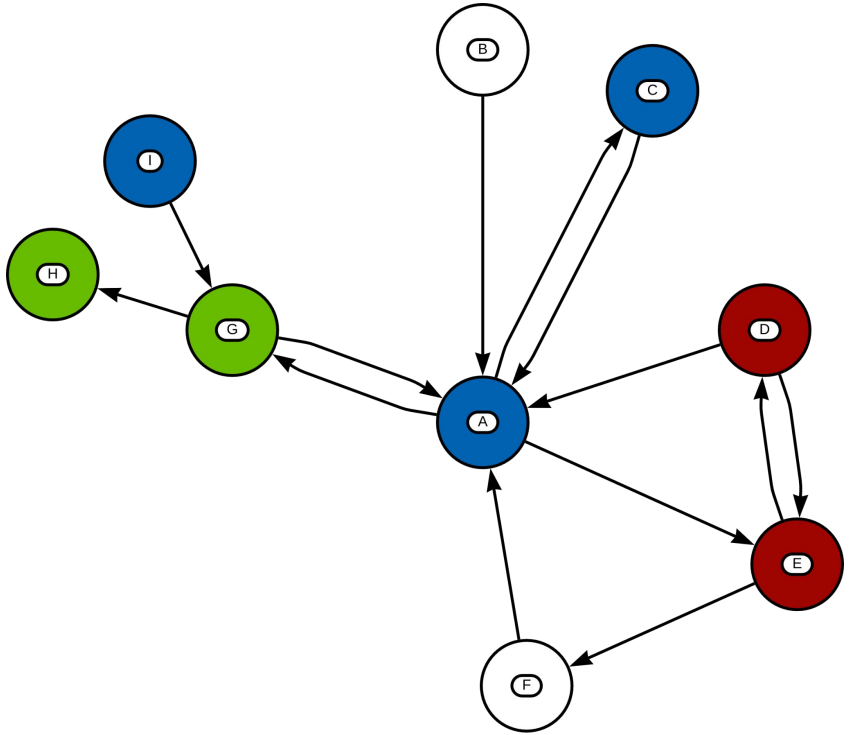
Directed vs. Undirected vs. Weighted





Neo4j, Inc. All rights reserved 2021

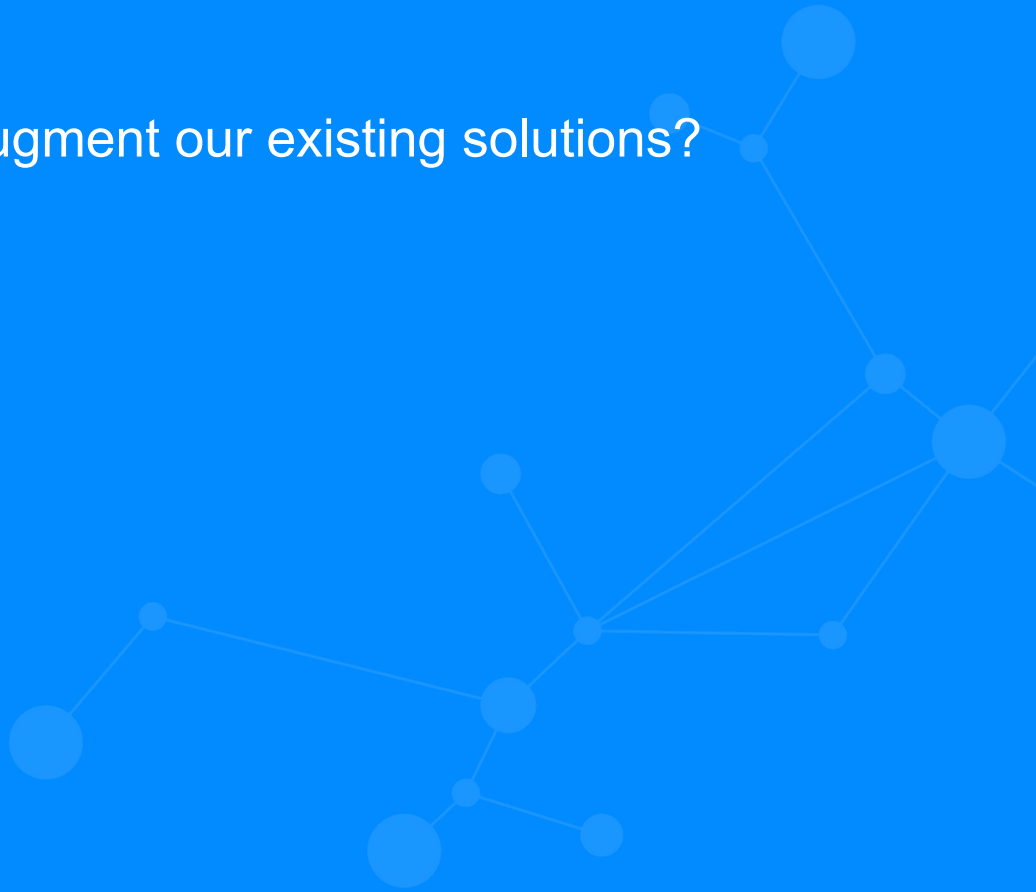
Adjacency matrix



	A	B	C	D	E	F	G	H	I
A	0	0	1	0	1	0	1	0	0
B	1	0	0	0	0	0	0	0	0
C	1	0	0	0	0	0	0	0	0
D	1	0	0	0	1	0	0	0	0
E	0	0	0	1	0	1	0	0	0
F	1	0	0	0	0	0	0	0	0
G	1	0	0	0	0	0	0	1	0
H	0	0	0	0	0	0	0	0	0
I	0	0	0	0	0	0	1	0	0

Neo4j, Inc. All rights reserved 2021

Why does using a graph augment our existing solutions?



Neo4j, Inc. All rights reserved 2021

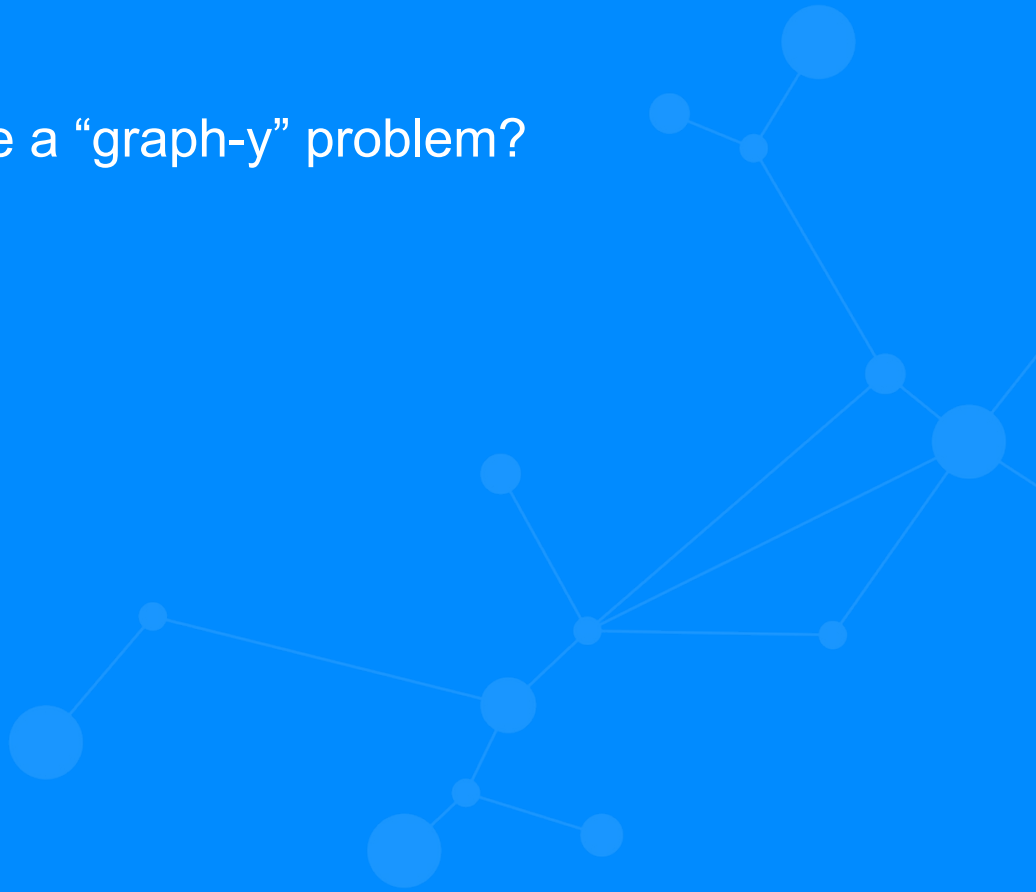
A churn prediction problem

Neo4j, Inc. All rights reserved 2021

A recommendation engine problem

Neo4j, Inc. All rights reserved 2021

How do you know you have a “graph-y” problem?



Neo4j, Inc. All rights reserved 2021

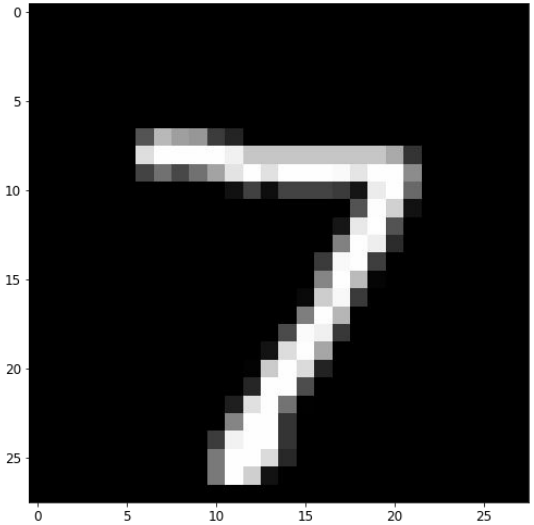
How do you know if you have a “graph-y” problem?

user_id	onbo_completed	active_days	weight_logs	sleep_logs	meal_logs	churn	gender_Female	gender_Male	gender_Undefined	age_class_10s	age_class_20s	age_class_30s	age_class_40s	age_class_50s	age_class_60s and more	age_class_Undefined
1	0	-1.009702	-1.393732	-0.535802	-0.513158	1	1	0	0	1	0	0	0	0	0	0
2	1	1.503925	0.767708	1.178228	-0.513158	0	1	0	0	0	0	1	0	0	0	0
3	1	0.048879	0.077230	-0.535802	-0.513158	1	1	0	0	0	0	1	0	0	0	0
4	1	-1.009702	0.077230	-0.535802	-0.513158	1	1	0	0	1	0	0	0	0	0	0
5	1	0.955027	0.767708	1.178228	-0.513158	1	1	0	0	0	1	0	0	0	0	0
6	1	1.503925	1.718841	1.548042	2.068715	0	1	0	0	1	0	0	0	0	0	0
7	1	0.048879	-1.393732	-1.397882	-0.513158	1	0	0	1	0	0	0	0	1	0	0
8	1	1.369578	0.767708	1.373593	-0.513158	1	1	0	0	0	1	0	0	0	0	0
9	1	-1.009702	0.077230	-0.535802	-0.513158	0	0	1	0	0	1	0	0	0	0	0
10	1	-1.009702	0.077230	1.178228	-0.513158	0	1	0	0	1	0	0	0	0	0	0

https://medium.com/finc-engineering/user-churn-prediction-using-neural-network-with-keras-c48f23ef4e8b

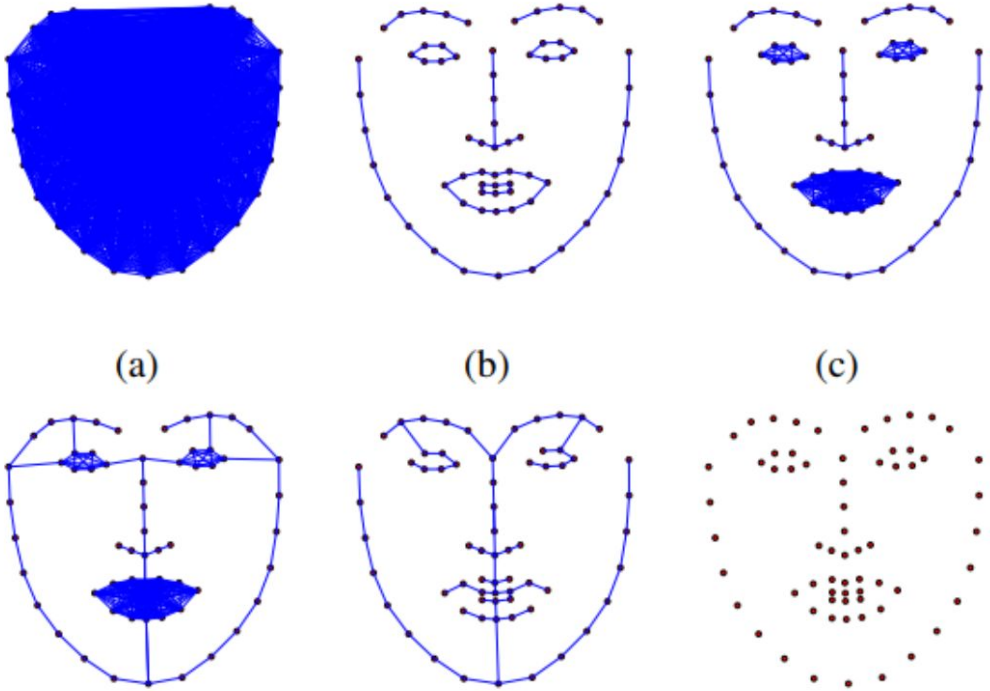
Neo4j, Inc. All rights reserved 2021

Will it graph: MNIST



Neo4j, Inc. All rights reserved 2021

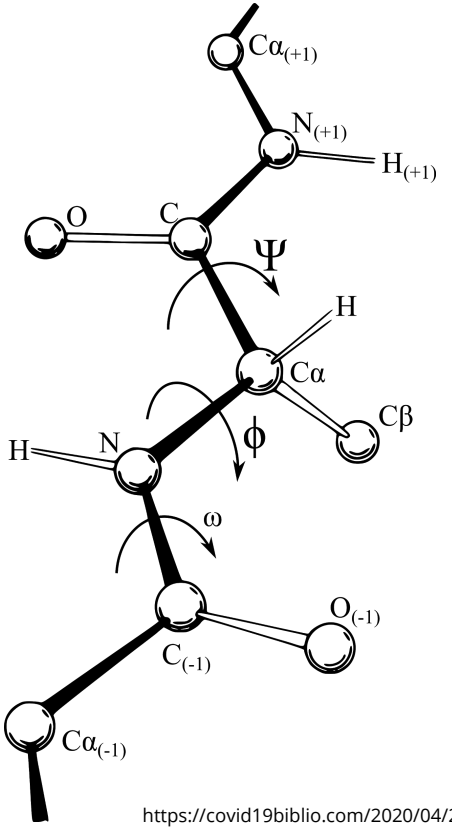
Will it graph: facial recognition

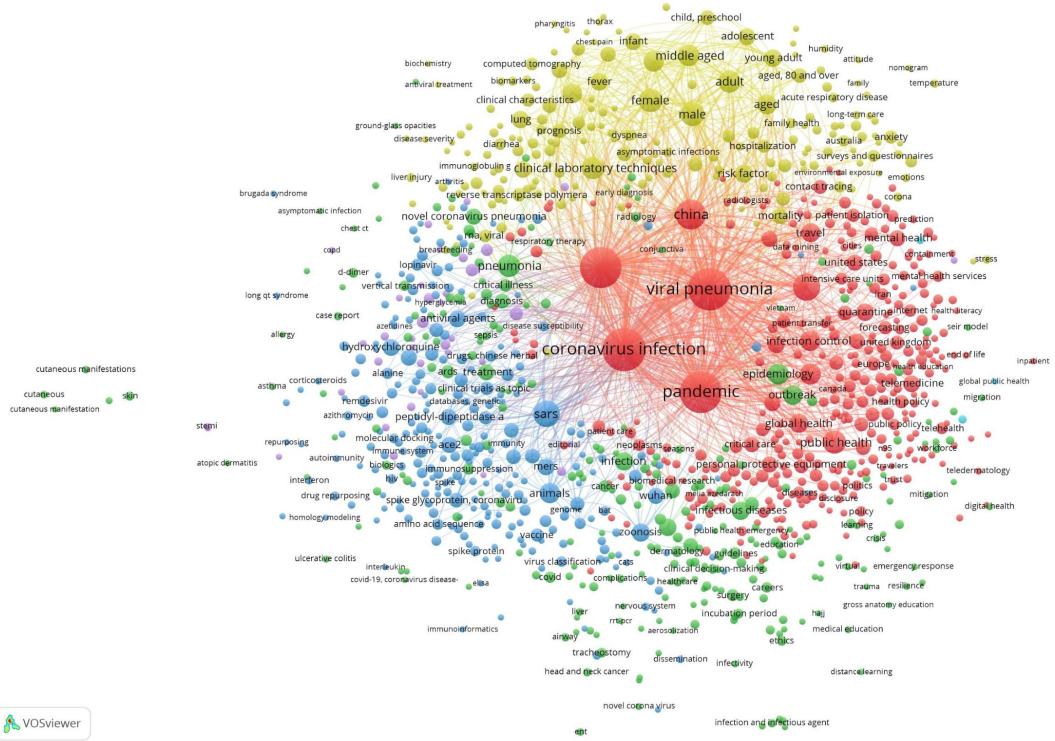


https://medium.com/@BorisAKnyazev/tutorial-on-graph-neural-networks-for-computer-vision-and-beyond-part-1-3d9fada3b80d

Neo4j, Inc. All rights reserved 2021

Will it graph: drug discovery

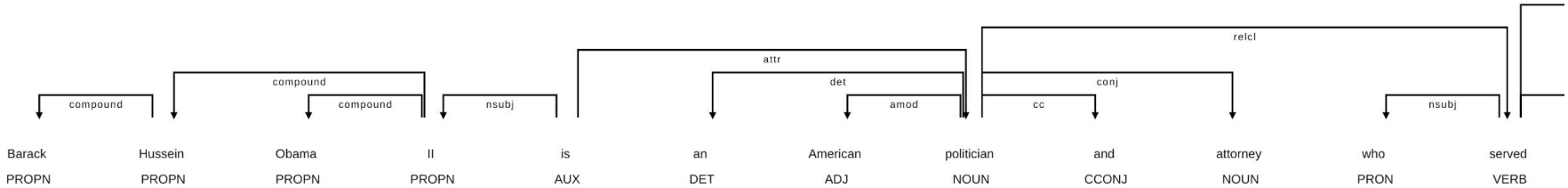




https://covid19biblio.com/2020/04/28/keyword-co-occurrence-network-graph-for-the-overall-research-field-on-covid-19-up-to-april-27th-2020/

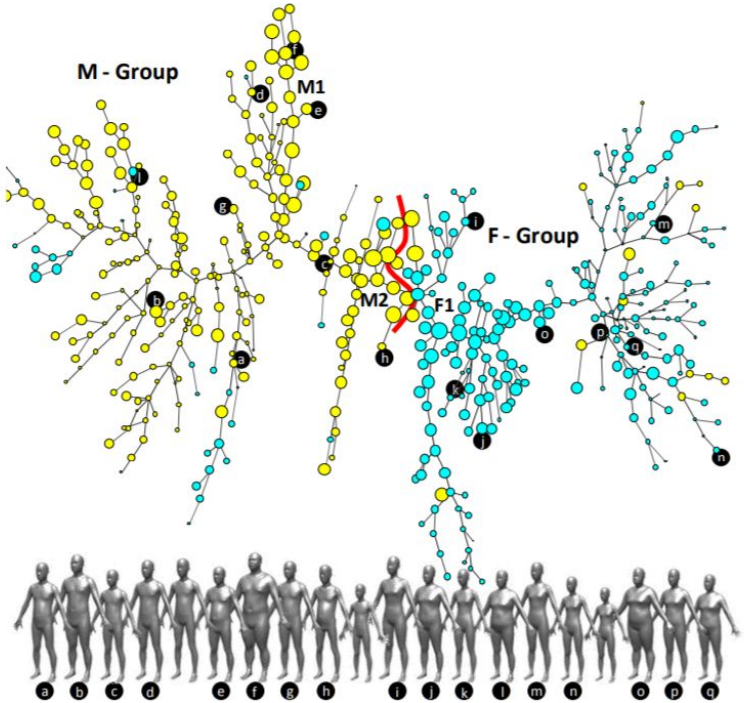
Neo4j, Inc. All rights reserved 2021

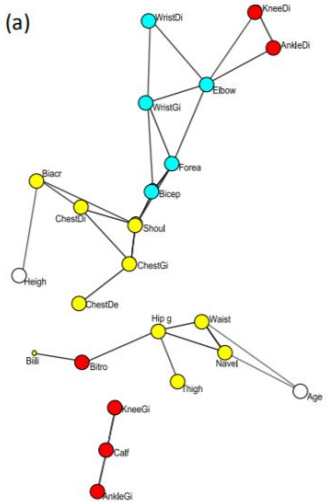
Will it graph: natural language processing

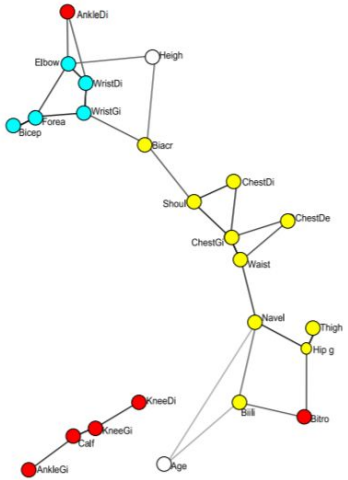


Neo4j, Inc. All rights reserved 2021

Will it graph: body composition analysis







Lower limbs | Upper limbs | Torso | General

Neo4j, Inc. All rights reserved 2021

C.N. Bezuidenhout and R.R. Domleo, *Sci. Res. Essays* (2013) 1274-1281.

How do you create a graph?

Neo4j, Inc. All rights reserved 2021

Typically you will need...

- Node list(s)
  - Unique identifiers help
  - Node properties, if available
- Edge list(s)
  - Edge names help
  - Weights, if available
- A graph model
- Answer the in-memory versus database question
  - Most of what I am going to show today can be done in either NetworkX or Neo4j
  - Choice depends on infrastructure and scalability needs

Neo4j, Inc. All rights reserved 2021

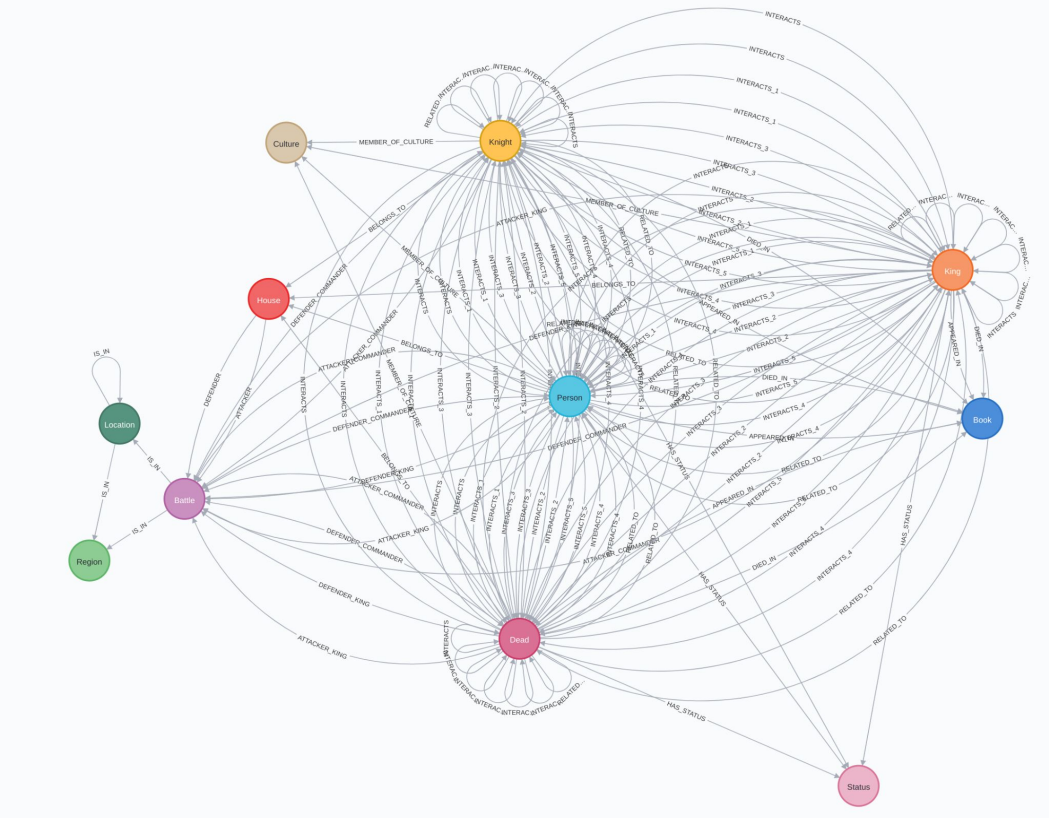
Node and edge lists

S.No	actual	pred	alive	plod	name	title	male	culture
1	0	0	0.054	0.946	Viserys II Targaryen		1	
2	1	0	0.387	0.613	Walder Frey	Lord of the Crossing	1	Rivermen
3	1	0	0.493	0.507	Addison Hill	Ser	1	
4	0	0	0.076	0.924	Aemma Arryn	Queen	0	
5	1	1	0.617	0.383	Sylva Santagar	Greenstone	0	Dornish
6	1	0	0.021	0.979	Tommen Baratheon		1	
7	0	0	0.014	0.986	Valarr Targaryen	Hand of the King	1	Valyrian
8	0	0	0.036	0.964	Viserys I Targaryen		1	
9	0	1	0.724	0.276	Wilbert	Ser	1	
10	1	0	0.391	0.609	Wilbert Osgrey	Ser	1	

name	year	battle_number	attacker_king	defender_king	attac
Battle of the Golden Tooth	298	1	Joffrey Baratheon/Tommen Baratheon	Robb Stark	Lanni
Battle at the Mummer's Ford	298	2	Joffrey Baratheon/Tommen Baratheon	Robb Stark	Lanni
Battle of Riverrun	298	3	Joffrey Baratheon/Tommen Baratheon	Robb Stark	Lanni
Battle of the Green Fork	298	4	Robb Stark	Joffrey Baratheon/Tommen Baratheon	Stark
Battle of the Whispering Wood	298	5	Robb Stark	Joffrey Baratheon/Tommen Baratheon	Stark
Battle of the Camps	298	6	Robb Stark	Joffrey Baratheon/Tommen Baratheon	Stark
Sack of Darry	298	7	Joffrey Baratheon/Tommen Baratheon	Robb Stark	Lanni
Battle of Moat Cailin	299	8	Balon Greyjoy/Euron Greyjoy	Robb Stark	Greyj
Battle of Deepwood Motte	299	9	Balon Greyjoy/Euron Greyjoy	Robb Stark	Greyj

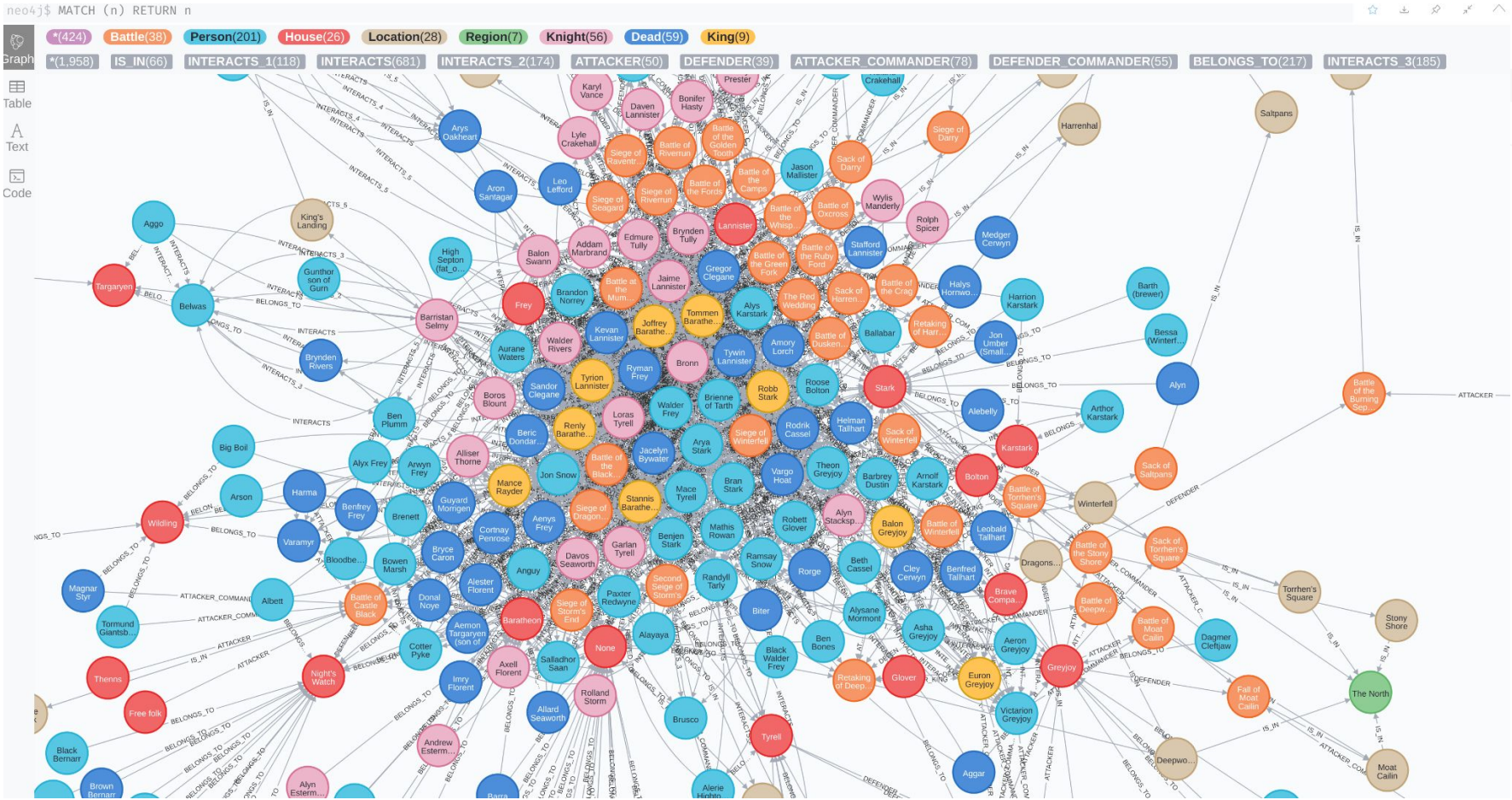
Neo4j, Inc. All rights reserved 2021
https://github.com/tomasonjo/neo4j-game-of-thrones

The importance of graph modeling



Neo4j, Inc. All rights reserved 2021

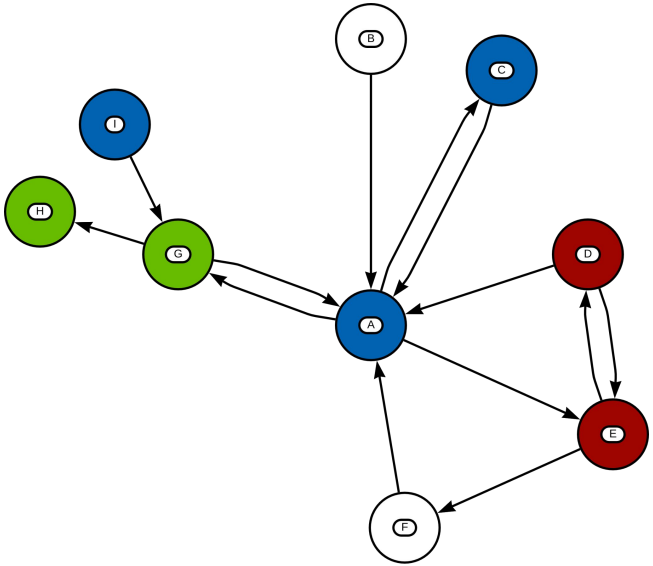
https://github.com/tomasonjo/neo4j-game-of-thrones

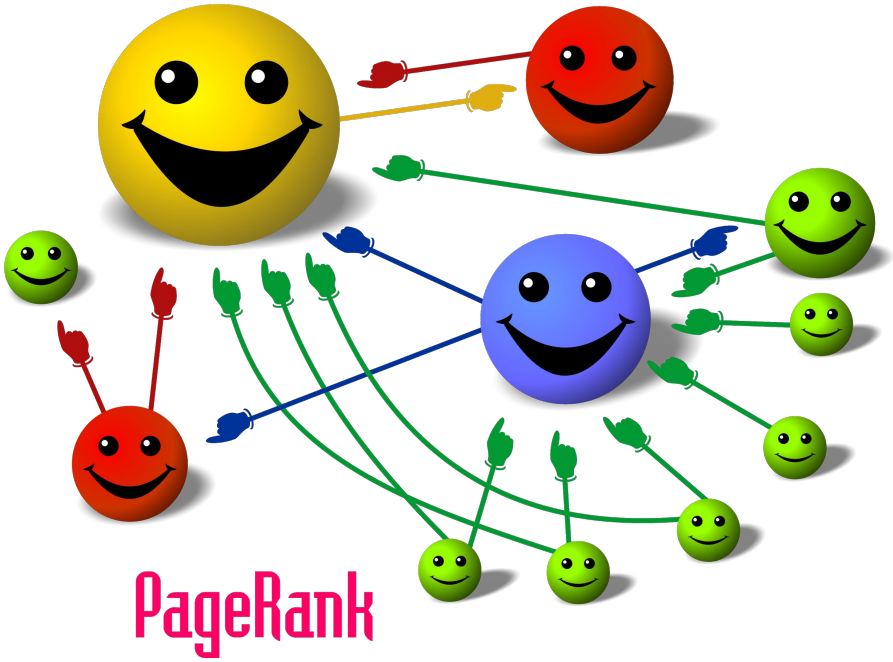


Neo4j, Inc. All rights reserved 2021

Node importance (centrality algorithms)

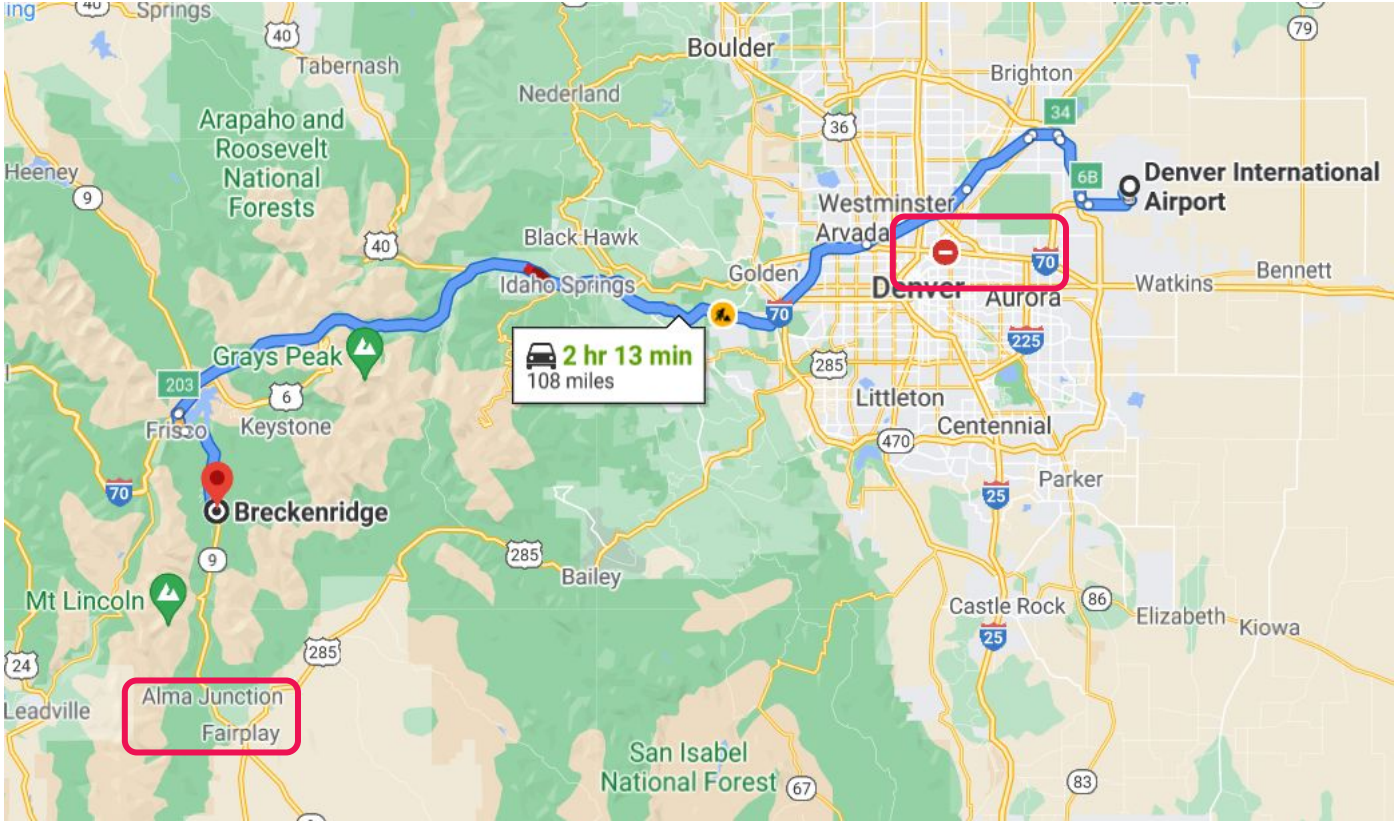
- Degree Centrality
- Betweenness centrality
- PageRank and friends





Neo4j, Inc. All rights reserved 2021

Path finding

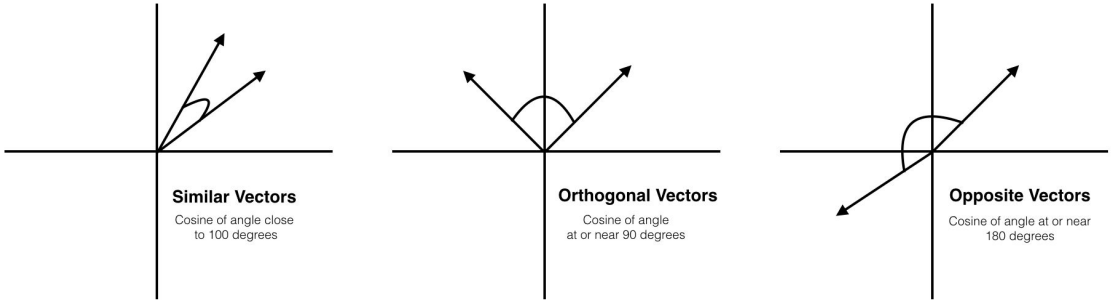


Neo4j, Inc. All rights reserved 2021

Node similarity

- Jaccard
- Cosine
- Euclidean distance
- K-Nearest Neighbors (KNN)

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$$

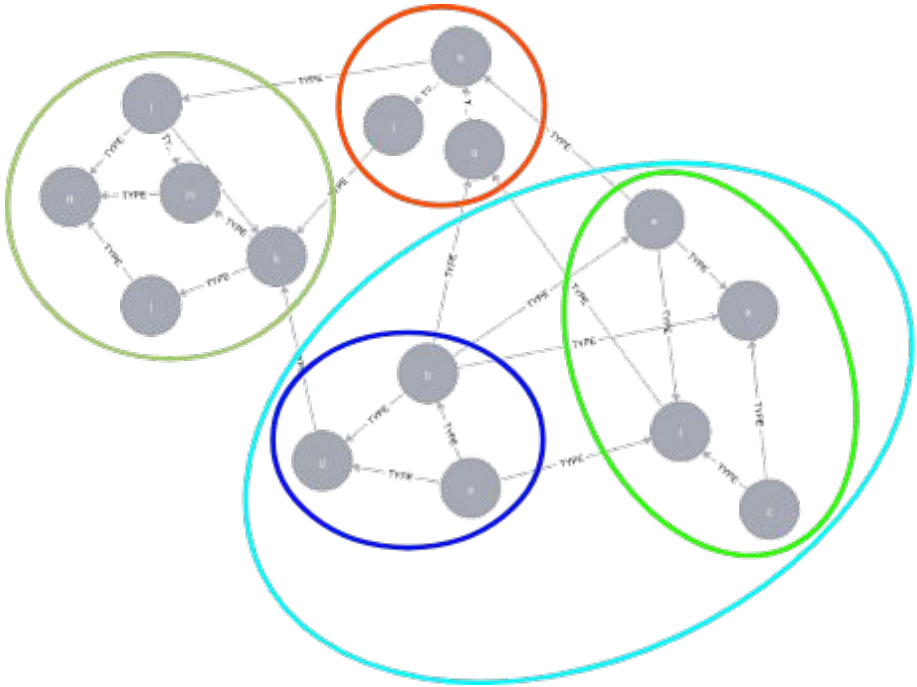


https://learning.oreilly.com/library/view/mastering-machine-learning/9781785283451/ba8bef27-953e-42a4-8180-cea152af8118.xhtml

Neo4j, Inc. All rights reserved 2021

Community detection

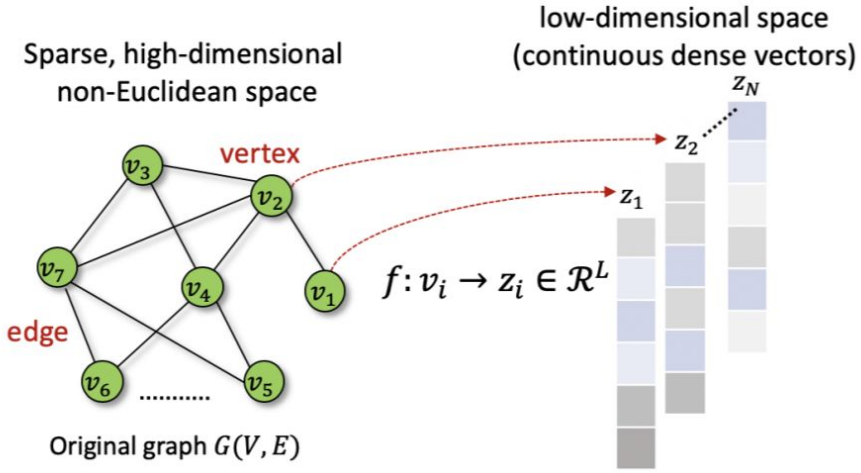
- Connected components (union find)
- Label propagation
- Speaker listener label propagation
- Louvain modularity



Neo4j, Inc. All rights reserved 2021

Graph embeddings

- Transductive
- Inductive
- Matrix factorization
- Methods based on random walks
  - FastRP
  - node2vec
- Methods based on neural networks



M. Xu (2020) *arXiv:2012.08019v1*

Neo4j, Inc. All rights reserved 2021

Let’s create a database!

dev.neo4j.com/sandbox

Neo4j, Inc. All rights reserved 2021

What can you do with a graph once you have it? (Including machine learning.)

Let’s transition to Jupyter notebook...

Neo4j, Inc. All rights reserved 2021

Some helpful references

- The code for this talk
  - dev.neo4j.com/data_science_graph_intro_repo
- Docker container for JupyterLab + Neo4j
  - dev.neo4j.com/docker_neo_jupyter
- Create a free database via Sandbox
  - dev.neo4j.com/sandbox
- Create a free database via Aura
  - dev.neo4j.com/aura
  - Does not include GDS
- Graph Data Science Library
- Graph Algorithms book

Neo4j, Inc. All rights reserved 2021

Thank you!

@CJLovesData1

Neo4j, Inc. All rights reserved 2021