Working with Data in a Connected World: the Power of Graph Data Science

Clair J. Sullivan, PhD
Data Science Advocate
Twitter: @CJLovesData1
Medium: https://medium.com/@cj2001

working-with-data-in-a-connected-world-the-power-of-graph-data-science

github.com/cj2001/pydata2021

Question 1: Where are you at in your data science journey? (Use an emoji reaction to vote)

❤️ I am not a data scientist but I manage them
💛 I am not a data scientist but I work with them
💚 I am presently studying data science
💙 I am presently looking for a job in data science
🤍 I am presently working as a professional data scientist (edited)

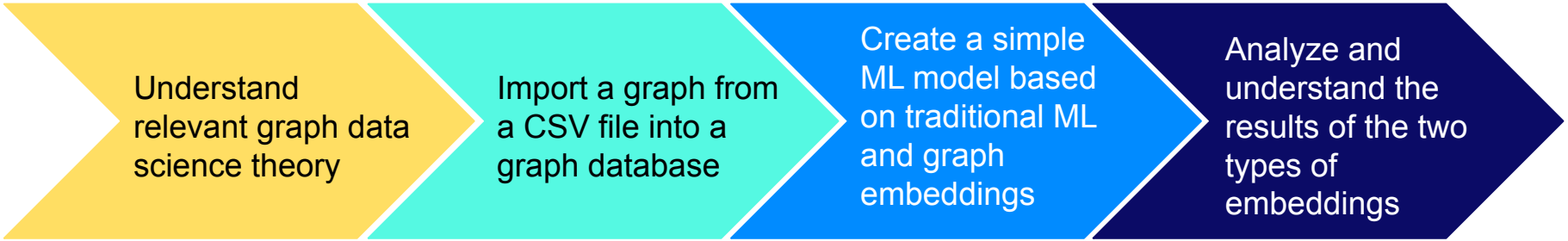
🤍 1

Question 2: What is your current knowledge of graphs?

🔴 What are graphs?
🟡 I have heard of them but not worked with them before
🟢 I have created some test projects with graphs or otherwise tinkered with them
🔵 I work with graphs occasionally in my education / professional career
⚪ I work with graphs on a daily basis in my education / professional career

🔴 1

By the end of this tutorial you will be able to...



What we are going to do today

- Why graphs? Why not just SQL?
- How do I know if I have a “graph-y” problem?
- A very brief introduction to graph theory
- Graph machine learning (ML)
- Dive into code!
- Wrap up

working-with-data-in-a-connected-world-the-power-of-graph-data-science

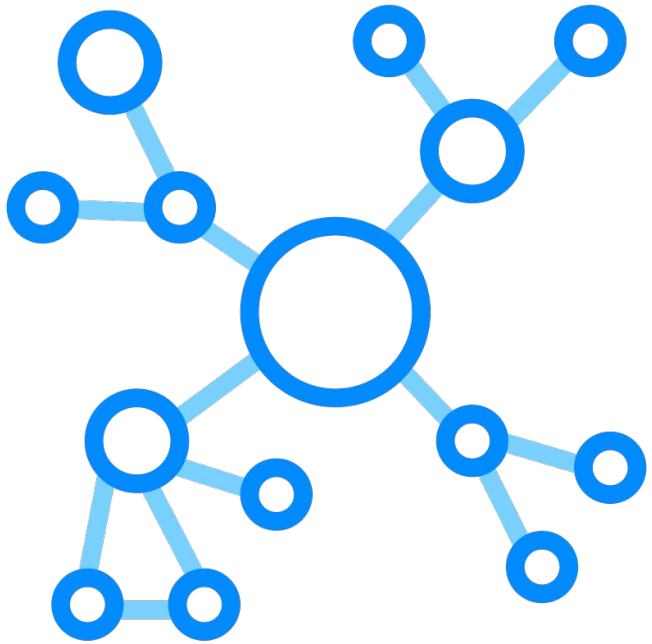
github.com/cj2001/pydata2021

Two Key Concepts

1. It can be good to break the assumption that individual data points are independent

2. Modeling relationships can result in models that are less noisy, more accurate

What is a graph?



Common examples

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

Columnar data for churn prediction

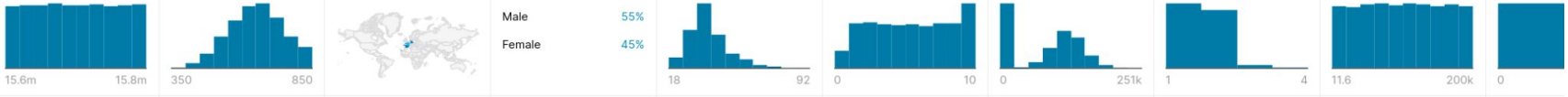
Churn_Modelling.csv (684.86 kB)

Detail | Compact | Column

10 of 14 columns

About this file

Based upon data of employees of a bank we calculate whether a employee stands a chance to stay in the company or not.



CustomerId	CreditScore	Geography	Gender	Age	Tenure	Balance	NumOfProducts	EstimatedSalary	Exited
The unique customer id	Their credit score	Which Country they belong to	Their Gender	Age	The time of bond with company	The amount left with them	The products they own.	Their estimated salary	Whether the or leave
15634602	619	France	Female	42	2	0	1	101348.88	1
15647311	608	Spain	Female	41	1	83807.86	1	112542.58	0
15619304	502	France	Female	42	8	159660.8	3	113931.57	1
15701354	699	France	Female	39	1	0	2	93826.63	0
15737888	850	Spain	Female	43	2	125510.82	1	79084.1	0
15574012	645	Spain	Male	44	8	113755.78	2	149756.71	1
15592531	822	France	Male	50	7	0	2	10062.8	0
15656148	376	Germany	Female	29	4	115046.74	4	119346.88	1
15792365	501	France	Male	44	4	142051.07	2	74940.5	0
15592389	684	France	Male	27	2	134603.88	1	71725.73	0
15767821	528	France	Male	31	6	102016.72	2	80181.12	0
15737173	497	Spain	Male	24	3	0	2	76390.01	0

A graph model of churn prediction

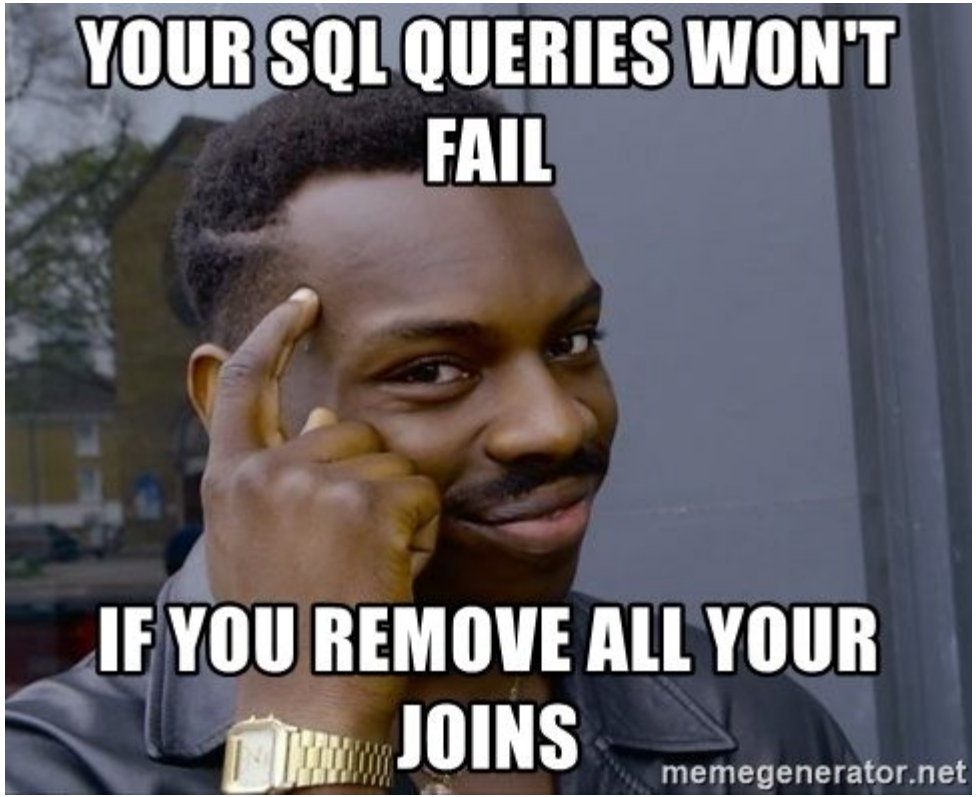
A recommendation engine problem

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

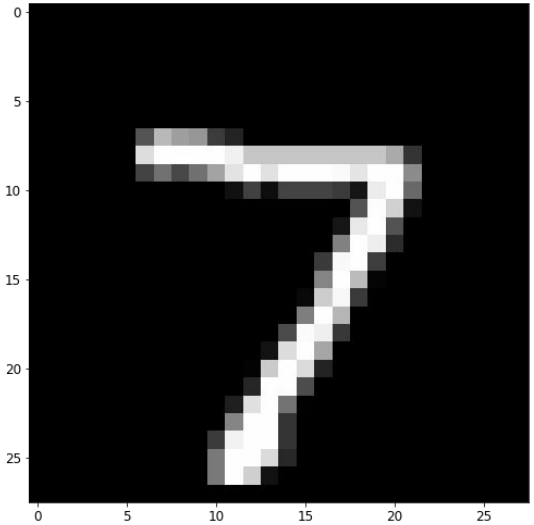
How do you know if you have a graph-y problem?

Rule of thumb:

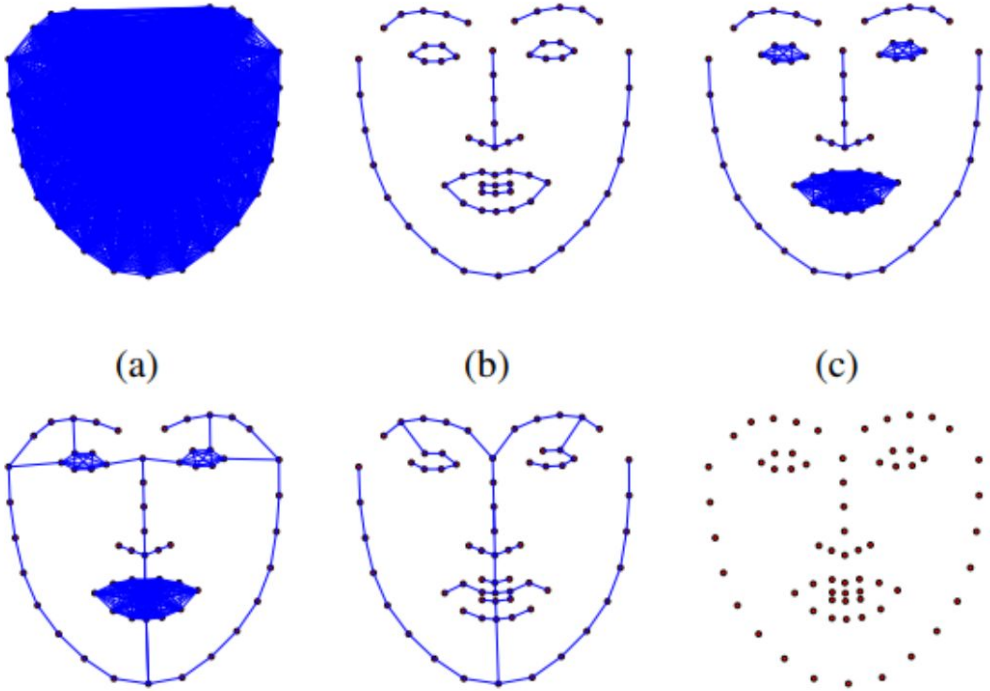
If you have to do more than a couple SQL JOINs then suspect you have a graph-y problem



Will it graph: MNIST

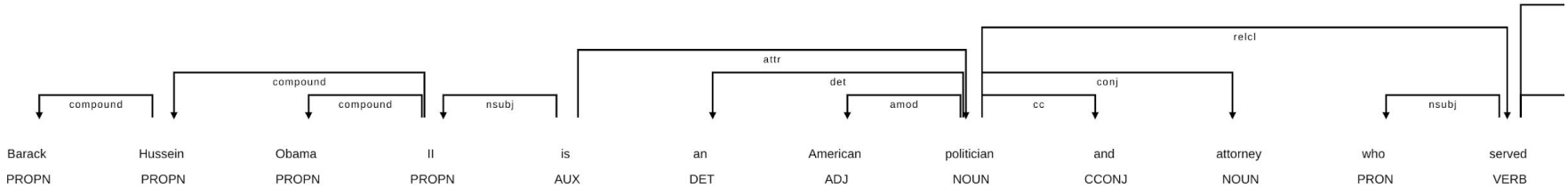


Will it graph: facial recognition



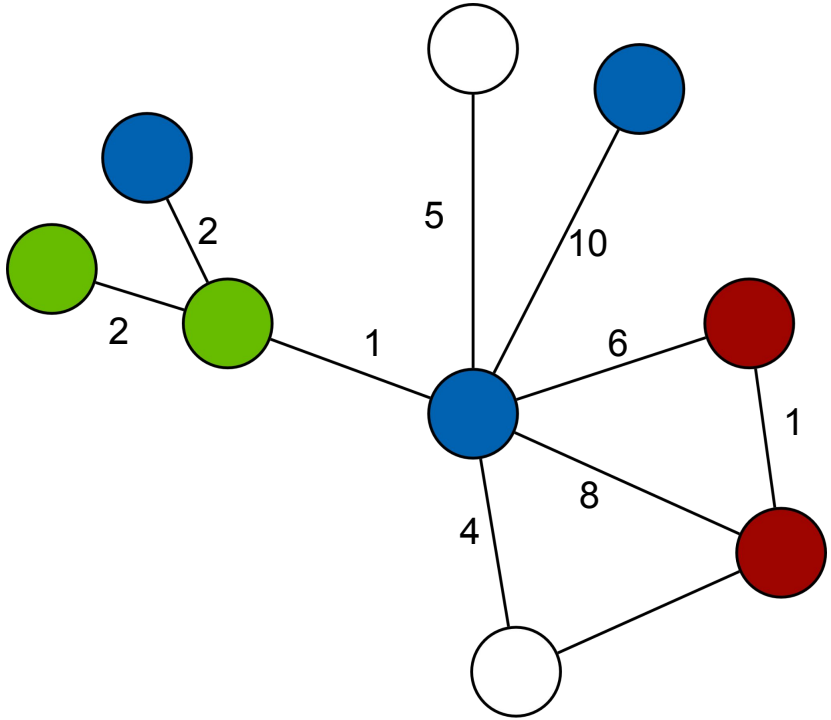
https://medium.com/@BorisAKnyazev/tutorial-on-graph-neural-networks-for-computer-vision-and-beyond-part-1-3d9fada3b80d

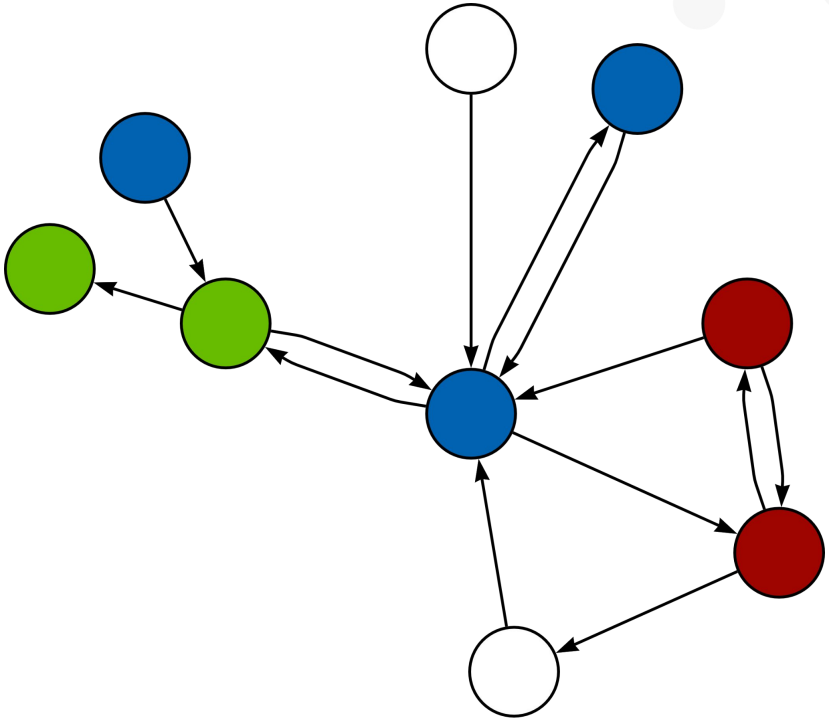
Will it graph: natural language processing



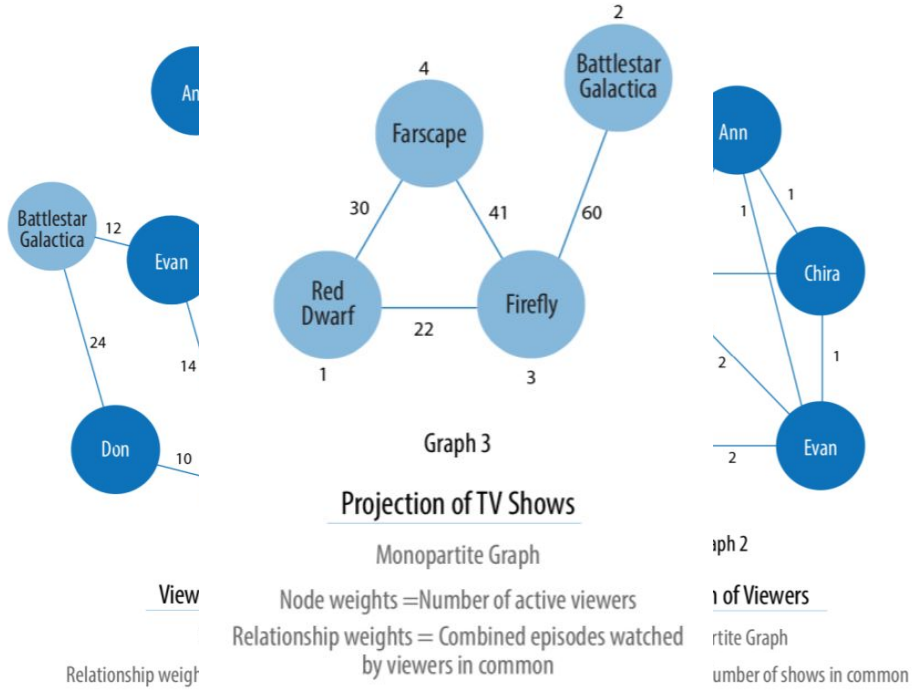
Some basic graph theory

Directed vs. Undirected vs. Weighted



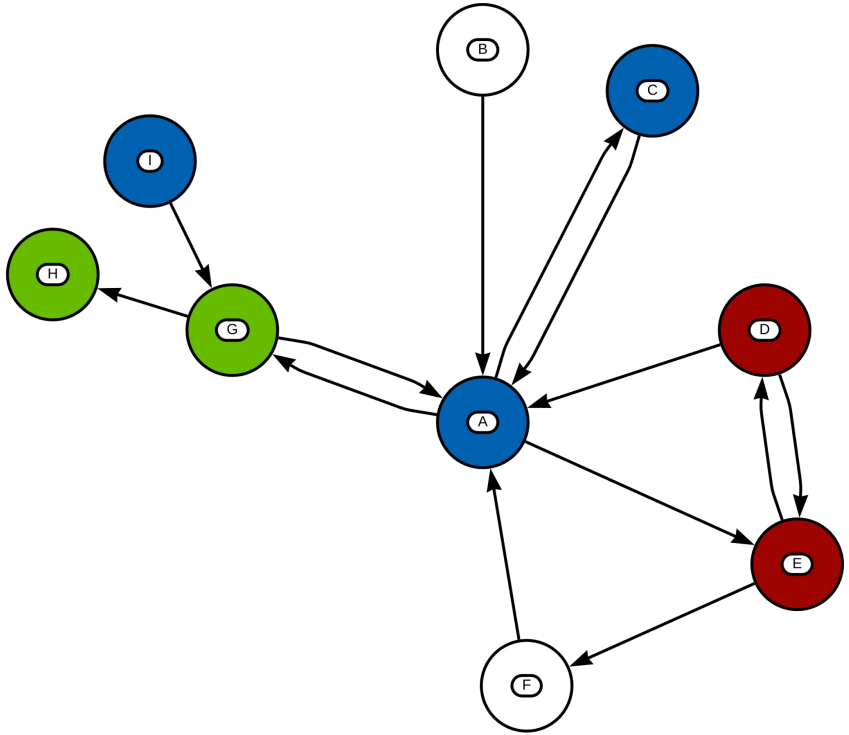


Monopartite vs. Bipartite



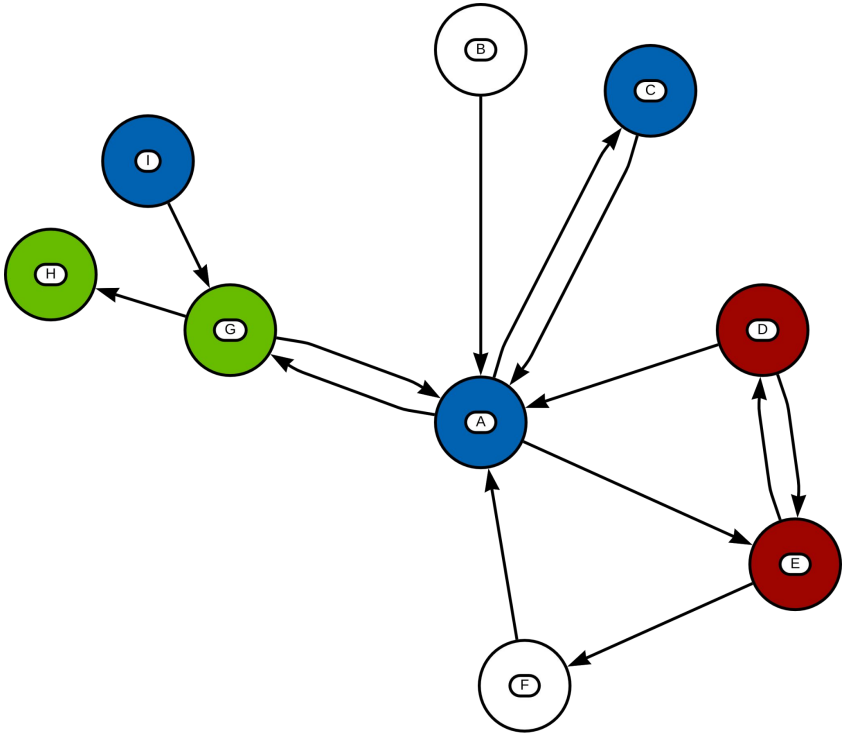
Graph Algorithms: Practical Examples in Apache Spark and Neo4j, M. Needham and A.E. Hodler (2019)

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

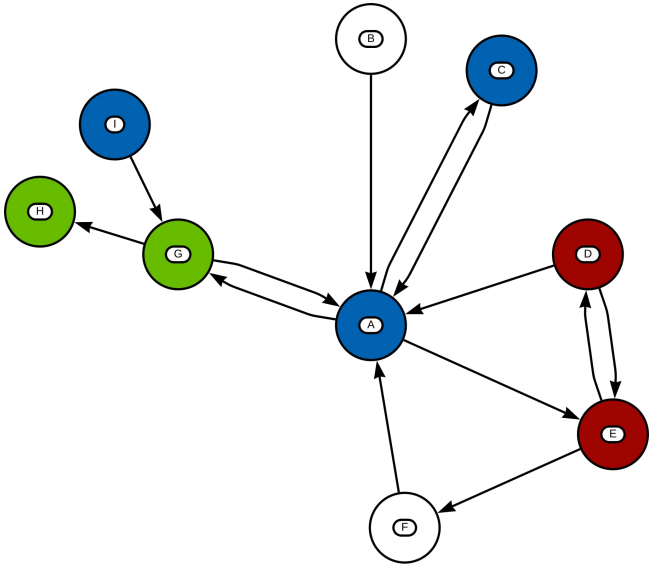
Degree

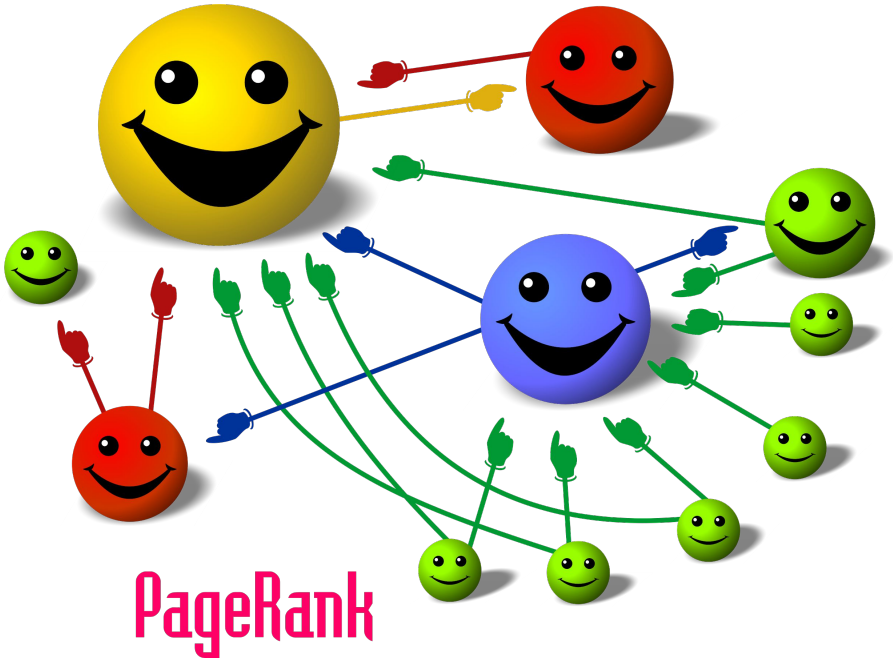


	A	B	C	D	E	F	G	H	I
A	8	0	0	0	0	0	0	0	0
B	0	1	0	0	0	0	0	0	0
C	0	0	2	0	0	0	0	0	0
D	0	0	0	3	0	0	0	0	0
E	0	0	0	0	4	0	0	0	0
F	0	0	0	0	0	2	0	0	0
G	0	0	0	0	0	0	4	0	0
H	0	0	0	0	0	0	0	1	0
I	0	0	0	0	0	0	0	0	1

Node importance (centrality algorithms)

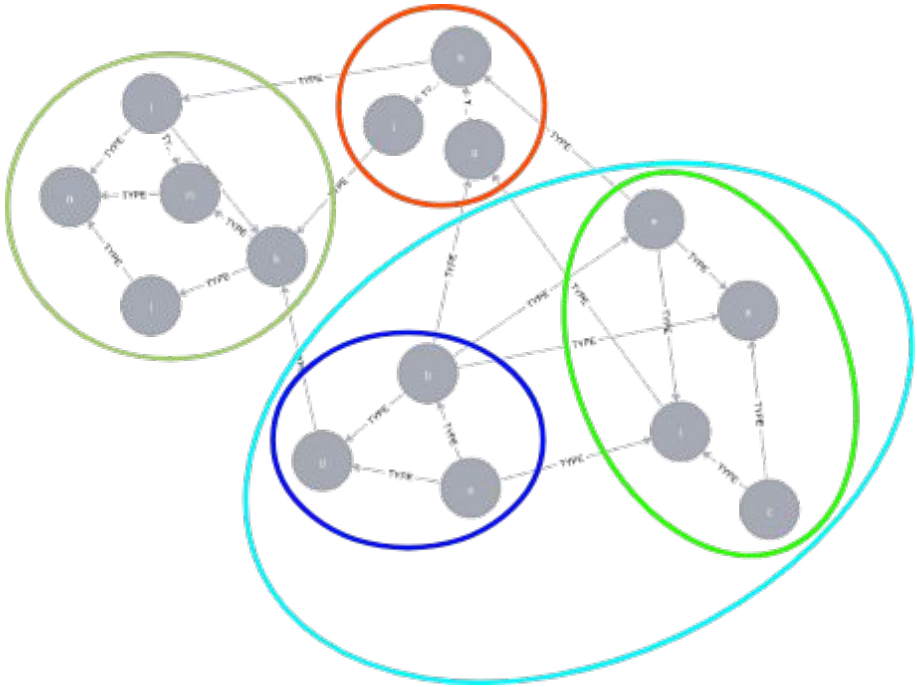
- Degree Centrality
- Betweenness centrality
- PageRank and friends





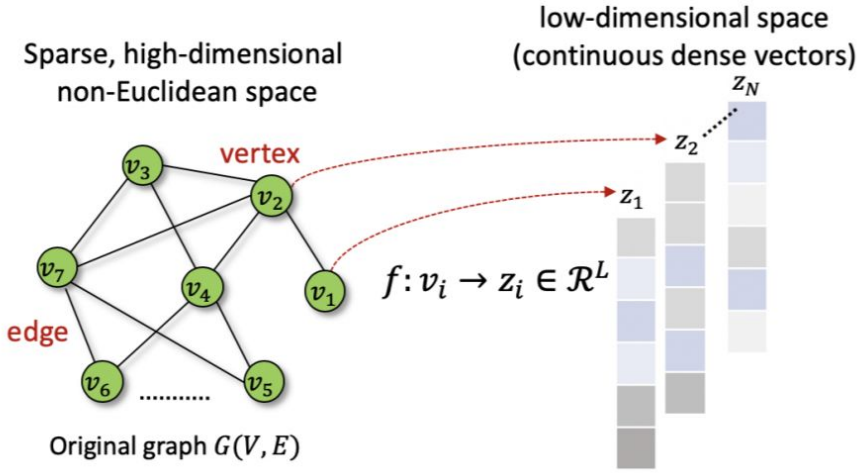
Community detection

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



Graph embeddings

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



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

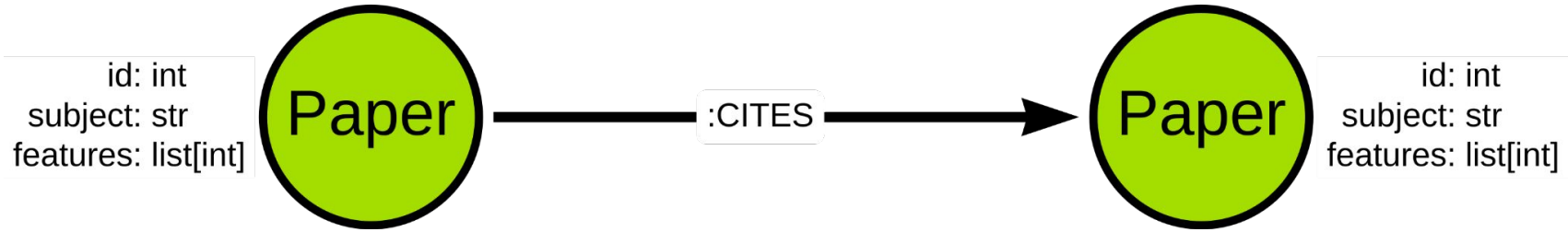
All of the same ML models can be run using graph embeddings!

- Classification (binary, multi-class, multi-label)
- Regression
- Clustering
- Dimensionality reduction
- Similarity
- Plus more that are unique to graphs!
  - Link prediction
  - (Sub)graph-level structural similarity

Let’s do some coding!!!

CORA Dataset

- 2708 scientific publications in data science
- 7 classes
- 5429 citation relationships
- Abstracts one-hot encoded to a vocabulary of 1433 words



What tools you will need to code along

- Google Colab: https://colab.research.google.com
- Neo4j Sandbox: https://sandbox.neo4j.com

https://github.com/cj2001/pydata2021

working-with-data-in-a-connected-world-the-power-of-graph-data-science

github.com/cj2001/pydata2021

What we have done today

- Why graphs? Why not just SQL?
- How do I know if I have a “graph-y” problem?
- A very brief introduction to graph theory
- Graph machine learning (ML)
- Dive into code!
- Wrap up

Two Key Concepts

1. It can be good to break the assumption that individual data points are independent

2. Modeling relationships can result in models that are less noisy, more accurate

Thank you!

@CJLovesData1