Activity: Review assumptions and expectations

- What were your assumptions about this class?
- What were your expectations?

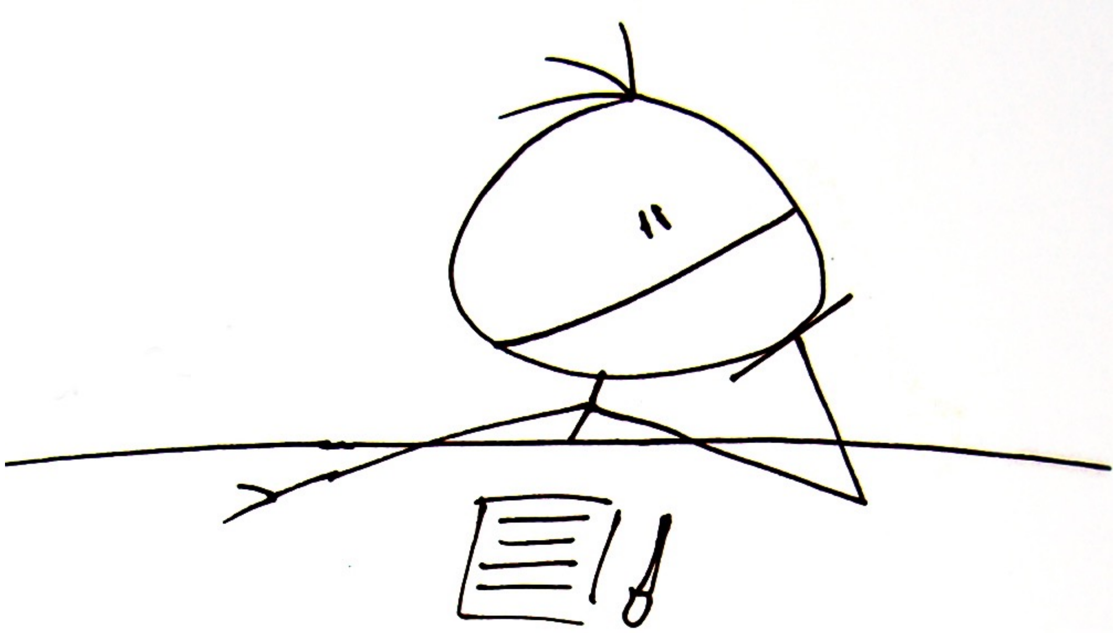
<<Instructor>> aggregates responses on the board

Activity: Write down your expectations

- Write down your expectations for the remaining classes

- Send me an email with your expectations

--> This will help shape my coverage of topics



Activity: think and write

What you want to know more about?

For example,

Basic visualization use in Python

- https://jakevdp.github.io/PythonDataScienceHandbook/04.01-simple-line-plots.html

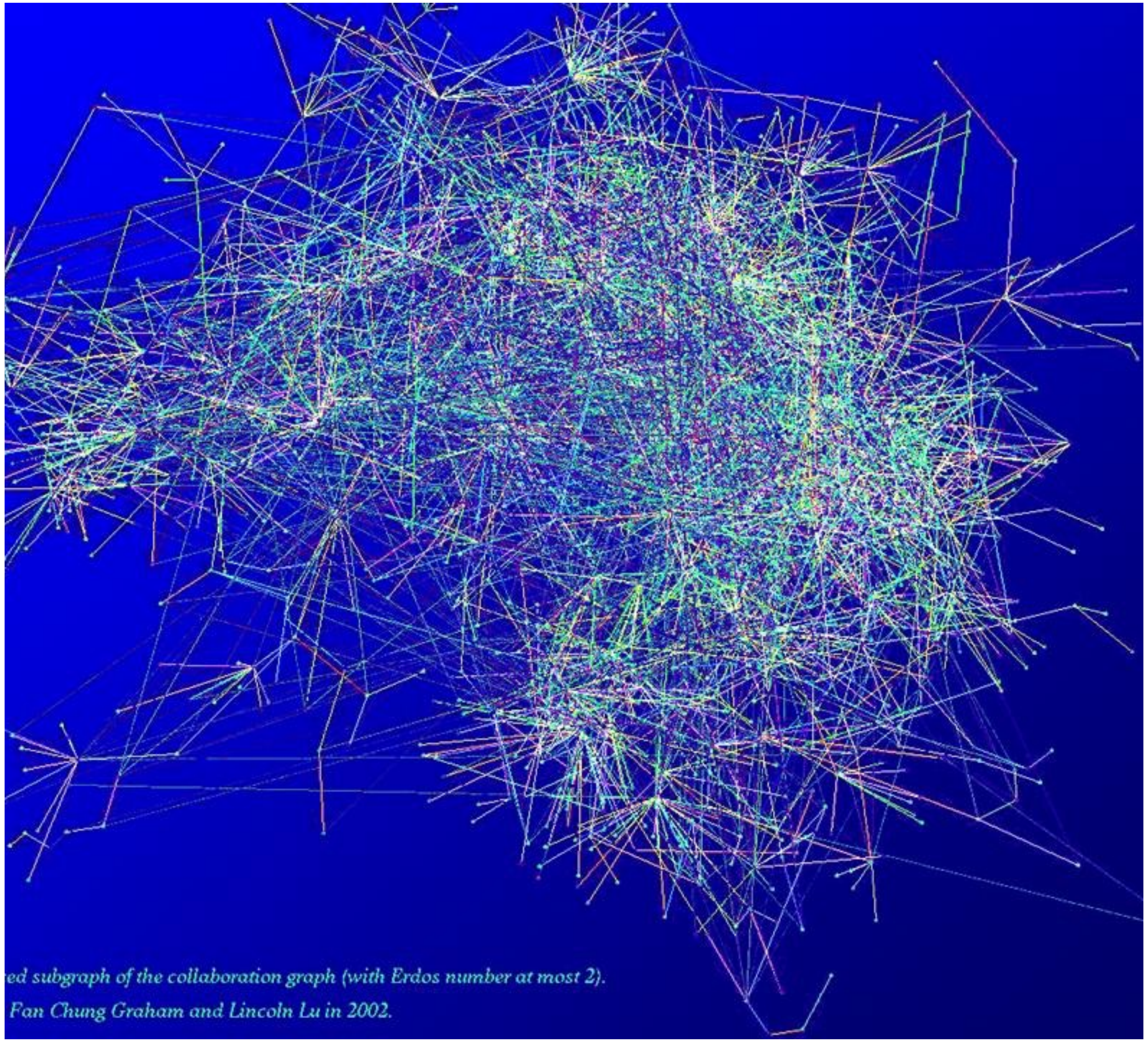
Gallery of what's possible in visualizations

- https://matplotlib.org/gallery.html

Fancy visualizations

- https://www.makeovermonday.co.uk/makeovers/

AND NOW BACK TO OUR REGULARLY SCHEDULED PROGRAMMING



ed subgraph of the collaboration graph (with Erdos number at most 2).
Fan Chung Graham and Lincoln Lu in 2002.

Tables, Graphs, and Property Graphs

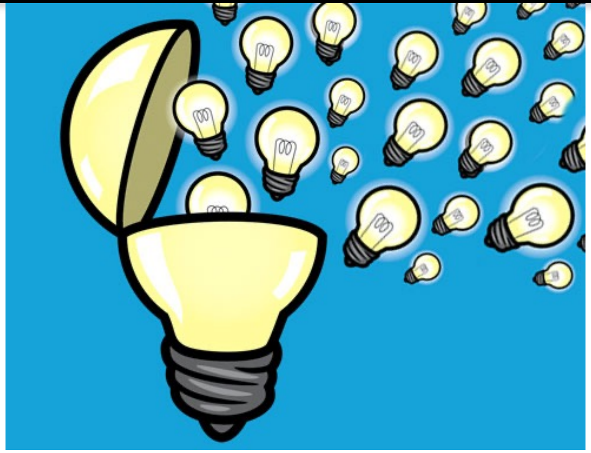
Source: https://www.math.ucsd.edu/~fan/complex/

By the end of this session,

You should be able to

- Explain when to use a list, table, and graph
- Describe the advantages of using graph data structures
- Create a visualization of a graph
- Show how to query information stored in a graph

- ~~Addressing points from previous week~~
- Introduction to Graphs
- Directed Graphs
- Elevator Pitch
- Weighted Graphs
- Property Graphs
- Homework



Data Structures covered in Data 601

- Scalars (int, float):
  - 4
  - 524.52
  - -934
- Strings:
  - ”mehmet"
  - "data 601"

- Lists:
  - ["ben", 4, -934, "data 601"]
- Tuples:
  - ("ben", "bob")
  - (3, 4, 9)
- Dictionaries:
  - { "instructor":”mehmet", "number of cats": 4, (3,4,9): -934, "class":"data 601}
- Tables (CSV, Pandas, Excel) with rows and columns

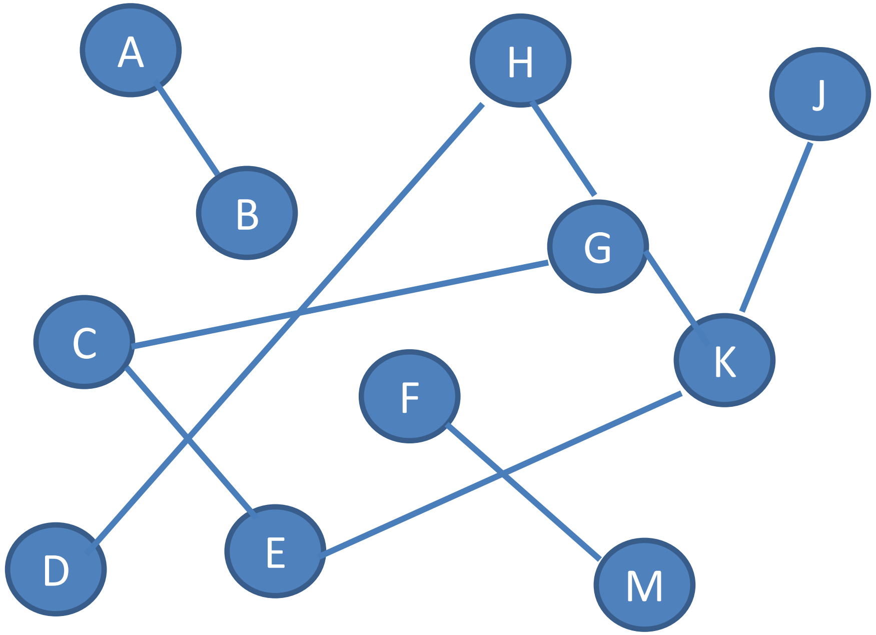
Table 1 describing relations among entities

Person	Knows	Duration in years
Bob	Anna	1
Bob	Kate	4
Bob	James	2
Anna	Kate	4
Max	Jim	1
Kate	Jim	8
Angela	Anna	2

Table 2 describing relations among entities

Company	Person	Position	Friend count
Pepsi	Bob	Delivery	142
Acme	Kate	Sales	47
Heavy Industry Inc.	Bob	Research	124
FunTymes	Anna	Coordinator	634
Roboflex	Max	Manager	152
HR Support	Kate	Tester	89
UMBC	Angela	Data Scientist	252

Graph: nodes and edges



Node = vertex = an entity

Edge = link = a relation or connection between entities

Nodes: A, B, C, D, E, F, H, J, K, G, M

Edges: (A,B), (C,G), (C,E), (D,H), (E,K), (F,M), (H,G), (J,K), (G,K)

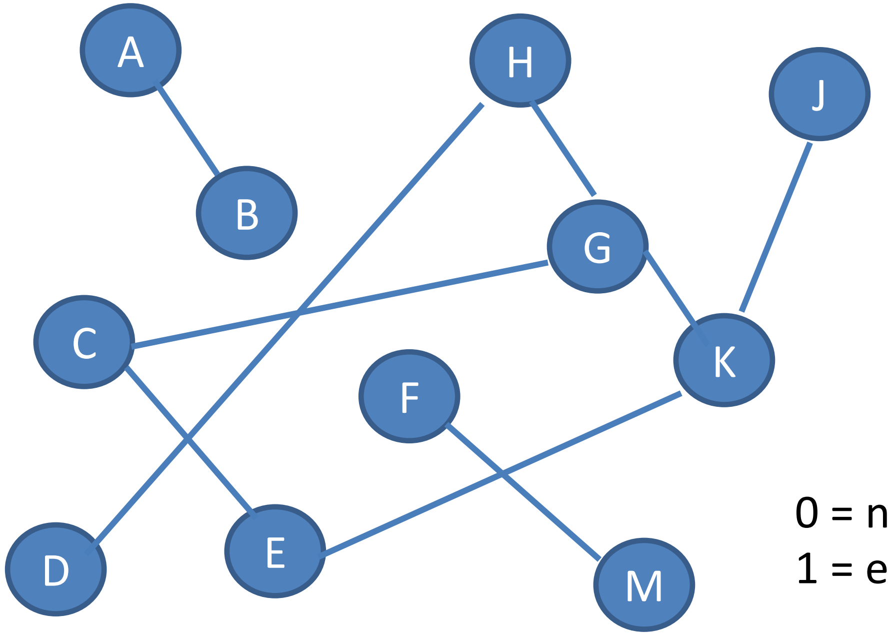
Caveat: this graph is unrelated to the tables in the previous slide

Conventional distinction

- Networks are the systems of interrelated objects (in the real world)
- Graphs are the mathematical model for representing networks

source

Relation between graph and adjacency matrix



0 = no connection
1 = edge exists

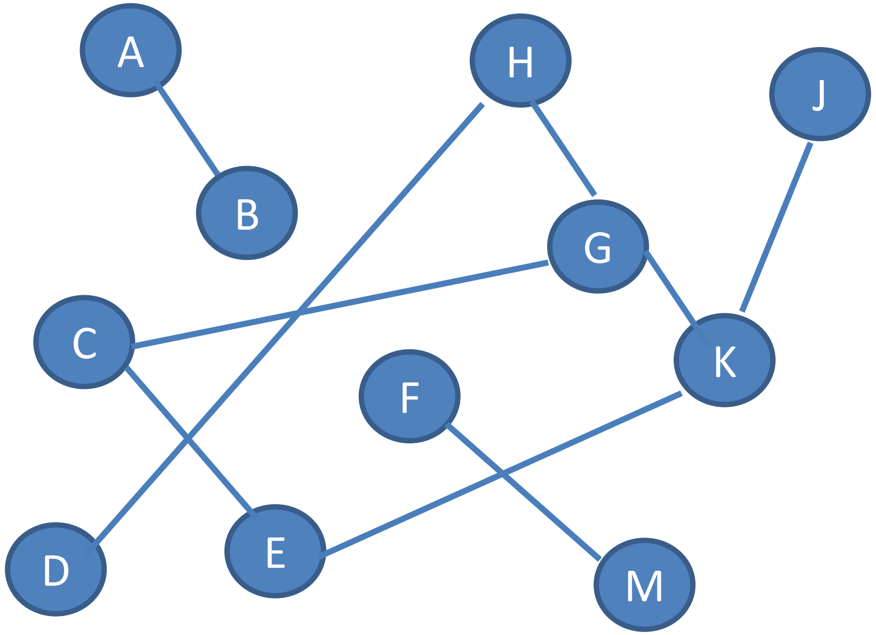
	A	B	C	D	E	F	G	H	J	K	M
A	0	1	0	0	0	0	0	0	0	0	0
B	1	0	0	0	0	0	0	0	0	0	0
C	0	0	0	0	1	0	1	0	0	0	0
D	0	0	0	0	0	0	0	1	0	0	0
E	0	0	1	0	0	0	0	0	0	1	0
F	0	0	0	0	0	0	0	0	0	0	1
G	0	0	1	0	0	0	0	1	0	1	0
H	0	0	0	1	0	0	1	0	0	0	0
J	0	0	0	0	0	0	0	0	0	1	0
K	0	0	0	0	1	0	1	0	1	0	0
M	0	0	0	0	0	1	0	0	0	0	0

Nodes: A, B, C, D, E, F, H, J, K, G, M

Edges: (A,B), (C,G), (C,E), (D,H), (E,K), (F,M), (H,G), (J,K), (G,K)

Number of entries with "1" is double the number of edges

Matrix for Undirected graph has diagonal symmetry



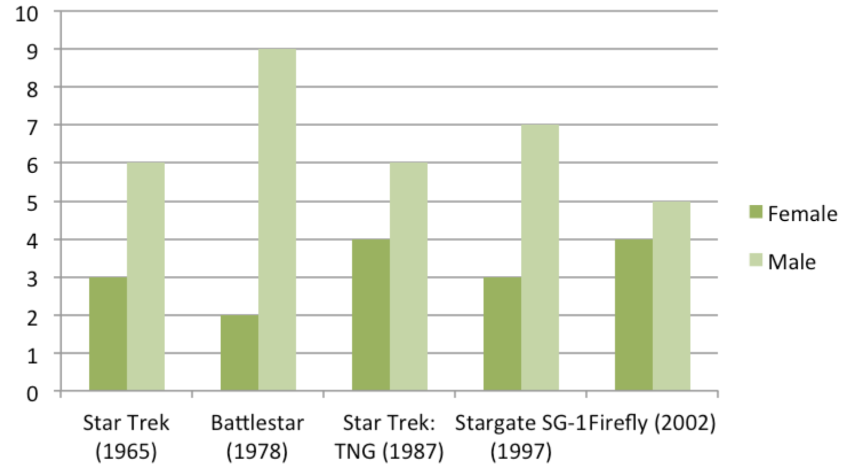
Nodes: A, B, C, D, E, F, H, J, K, G, M

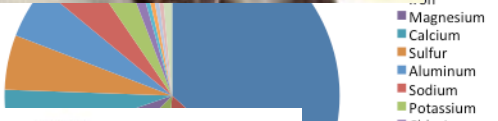
Edges: (A,B), (C,G), (C,E), (D,H), (E,K), (F,M), (H,G), (J,K), (G,K)

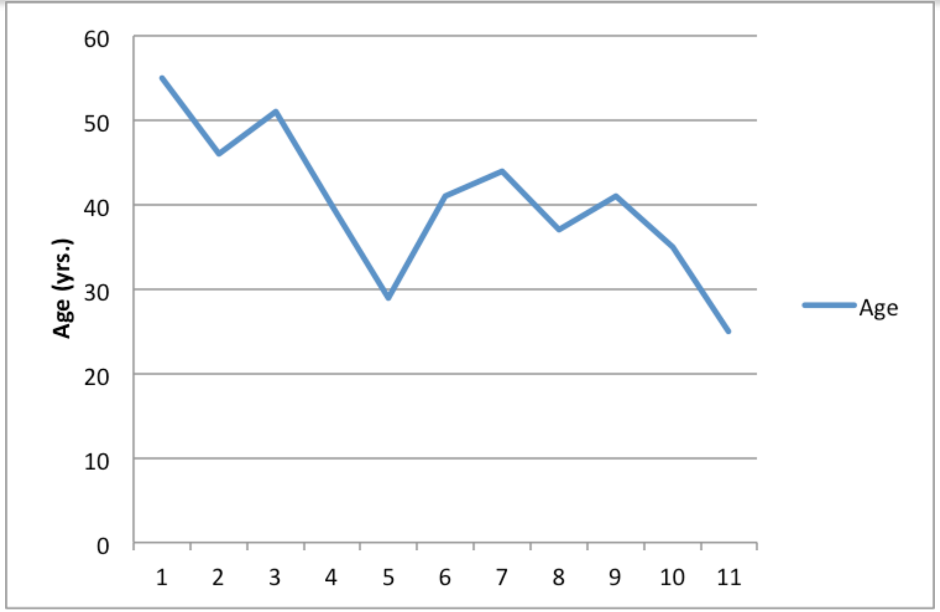
	A	B	C	D	E	F	G	H	J	K	M
A	0	1	0	0	0	0	0	0	0	0	0
B	1	0	0	0	0	0	0	0	0	0	0
C	0	0	0	0	1	0	1	0	0	0	0
D	0	0	0	0	0	0	0	1	0	0	0
E	0	0	1	0	0	0	0	0	0	1	0
F	0	0	0	0	0	0	0	0	0	0	1
G	0	0	1	0	0	0	0	1	0	1	0
H	0	0	0	1	0	0	1	0	0	0	0
J	0	0	0	0	0	0	0	0	0	1	0
K	0	0	0	0	1	0	1	0	1	0	0
M	0	0	0	0	0	1	0	0	0	0	0

Number of entries with "1" is double the number of edges

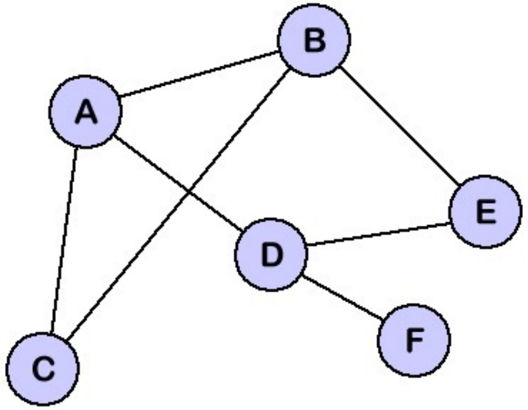
Difference between graph, plot, chart, figure?



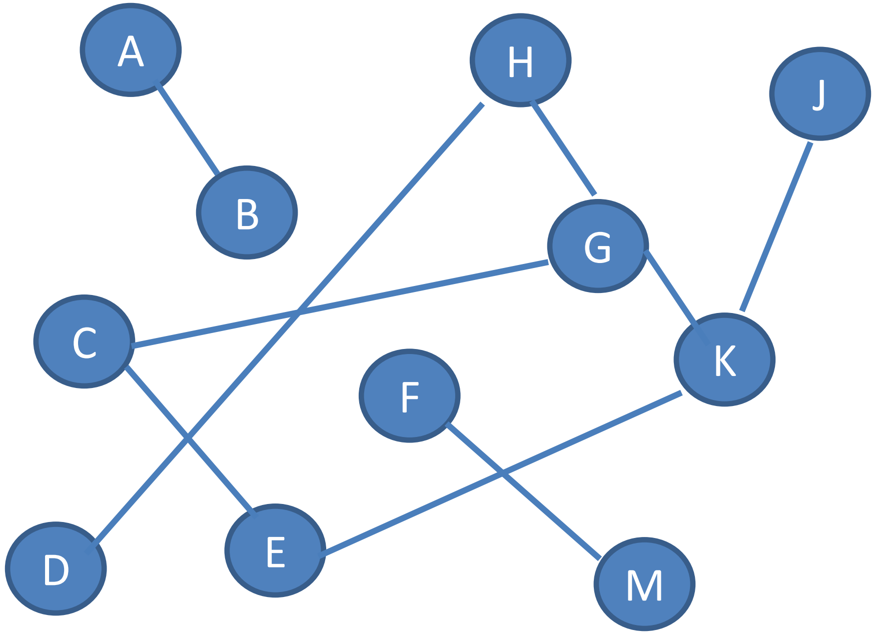




All are referred to as {graph, chart, plot, figure}



Why add graphs if tables are equivalent?



	A	B	C	D	E	F	G	H	J	K	M
A	0	1	0	0	0	0	0	0	0	0	0
B	1	0	0	0	0	0	0	0	0	0	0
C	0	0	0	0	1	0	1	0	0	0	0
D	0	0	0	0	0	0	0	1	0	0	0
E	0	0	1	0	0	0	0	0	0	1	0
F	0	0	0	0	0	0	0	0	0	0	1
G	0	0	1	0	0	0	0	1	0	1	0
H	0	0	0	1	0	0	1	0	0	0	0
J	0	0	0	0	0	0	0	0	0	1	0
K	0	0	0	0	1	0	1	0	1	0	0
M	0	0	0	0	0	1	0	0	0	0	0

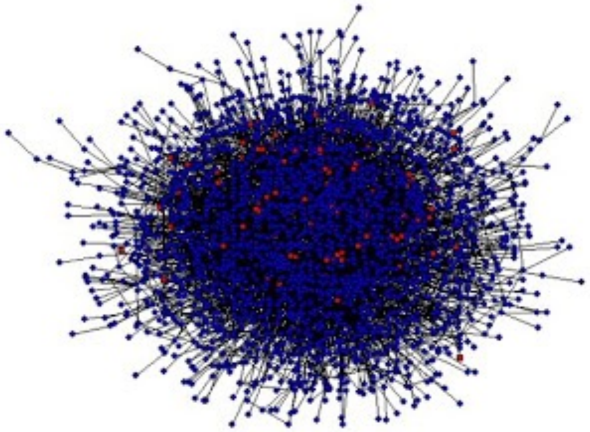
Use graphs when focused on relations among entities

Benefits:

- Different way of thinking about the problem
- Useful for questions where relations are complex
- Defining all values (ie in a table) is not required
- Potentially more efficient representation when compared to a table
- Representation is intuititive
- Makes pretty pictures

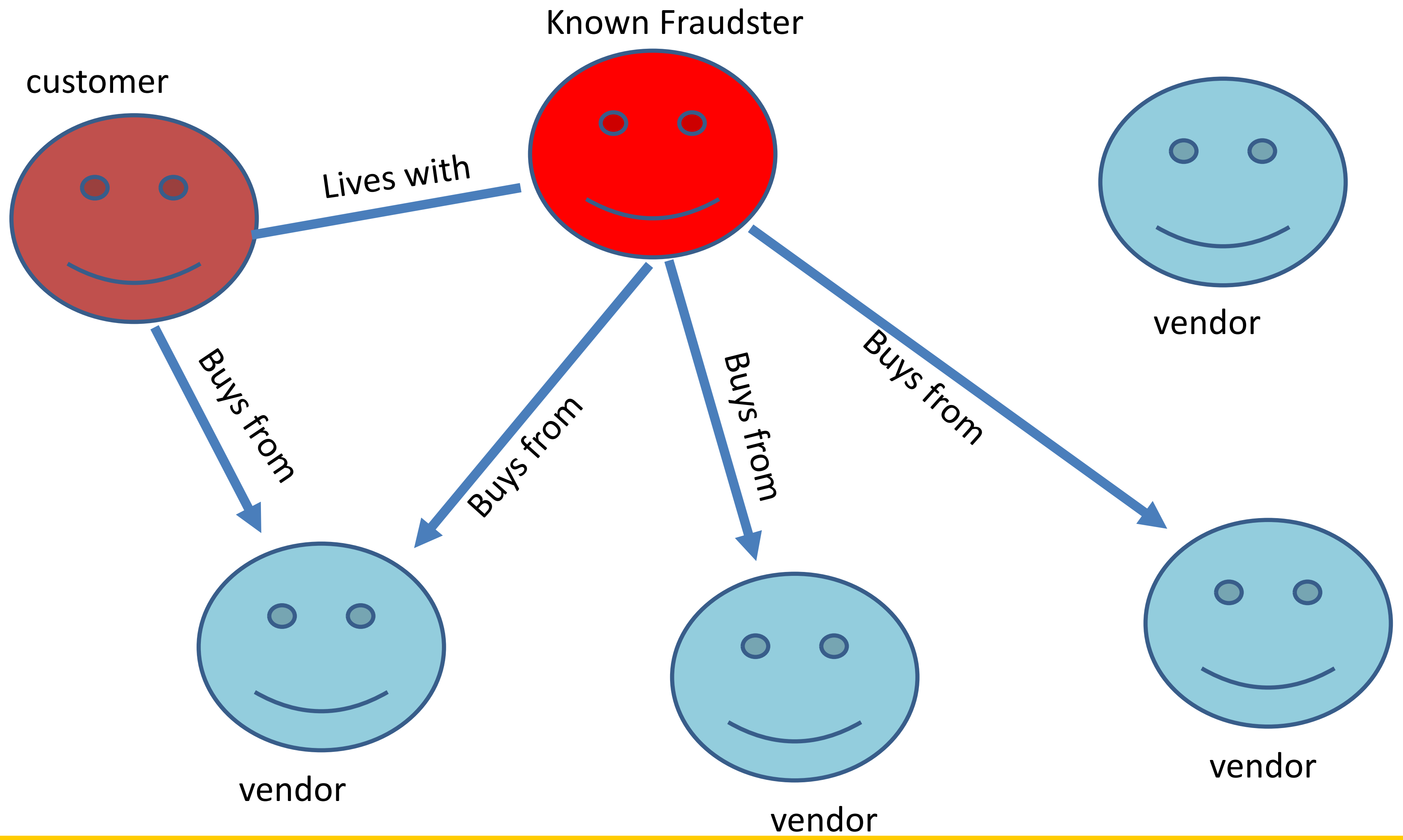
Costs of using graph to represent data

- Requires different algorithms (compared to table-based data)
- Learning
- Visual Rendering of large graphs is slow
- Visual rendering of large graphs is useless: "hairball"



Use case for graphs

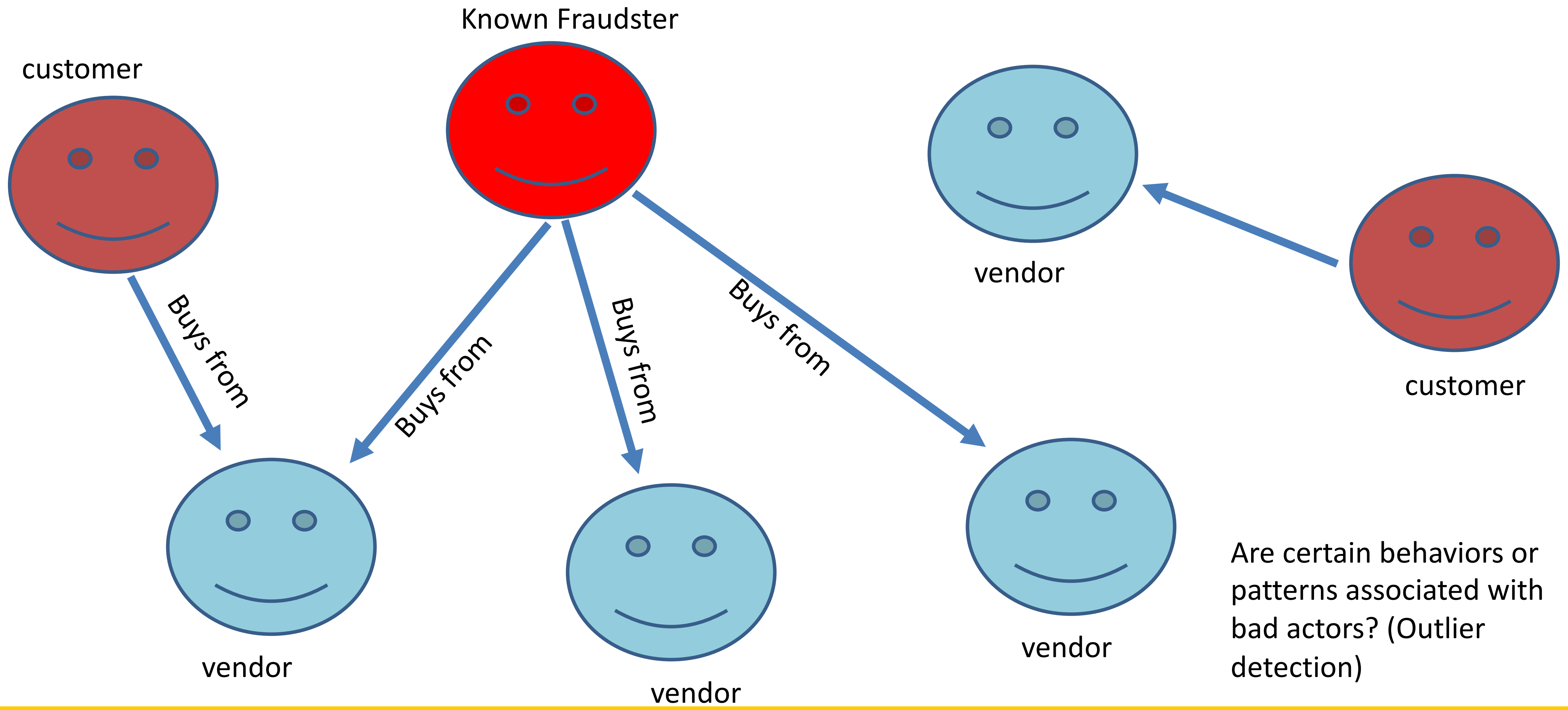
Fraud detection for banking and for credit card transactions



Based on activities of known bad actor, do other (related) people have similar behavior?

Use case for graphs:

Fraud detection for banking and for credit card transactions



Are certain behaviors or patterns associated with bad actors? (Outlier detection)

Use case for graphs:
Natural Language Processing

https://medium.com/@aneesha/beyond-bag-of-words-using-pytextrank-to-find-phrases-and-summarize-text-f736fa3773c5

- Words are added to the graph as nodes
- An edge is added between words that co-occur within N words of each other

This is then used as a feature for machine learning algorithms

Social Network Analysis applied in advertising

Companies sell graph database of consumers

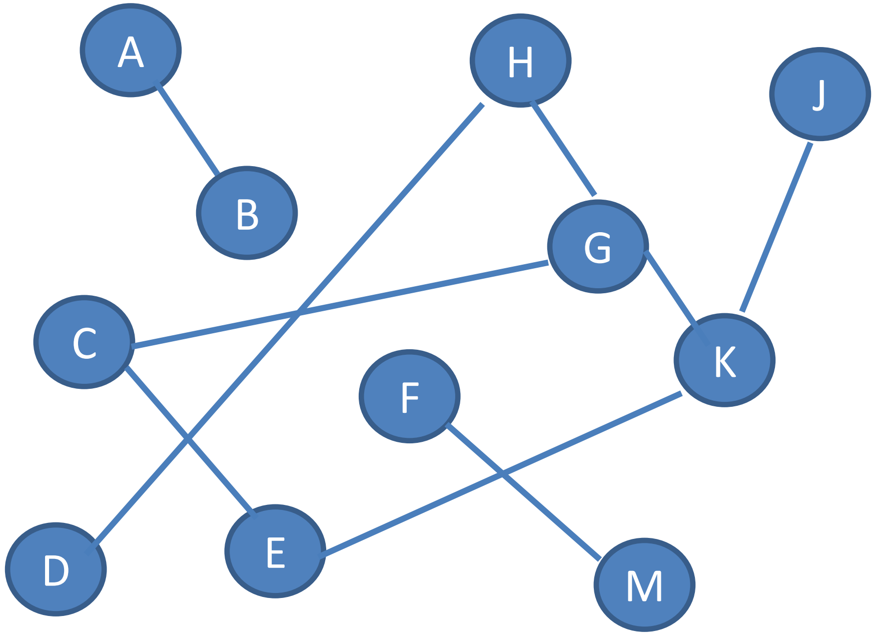
- https://www.faraday.io/technology/faraday-identity-graph

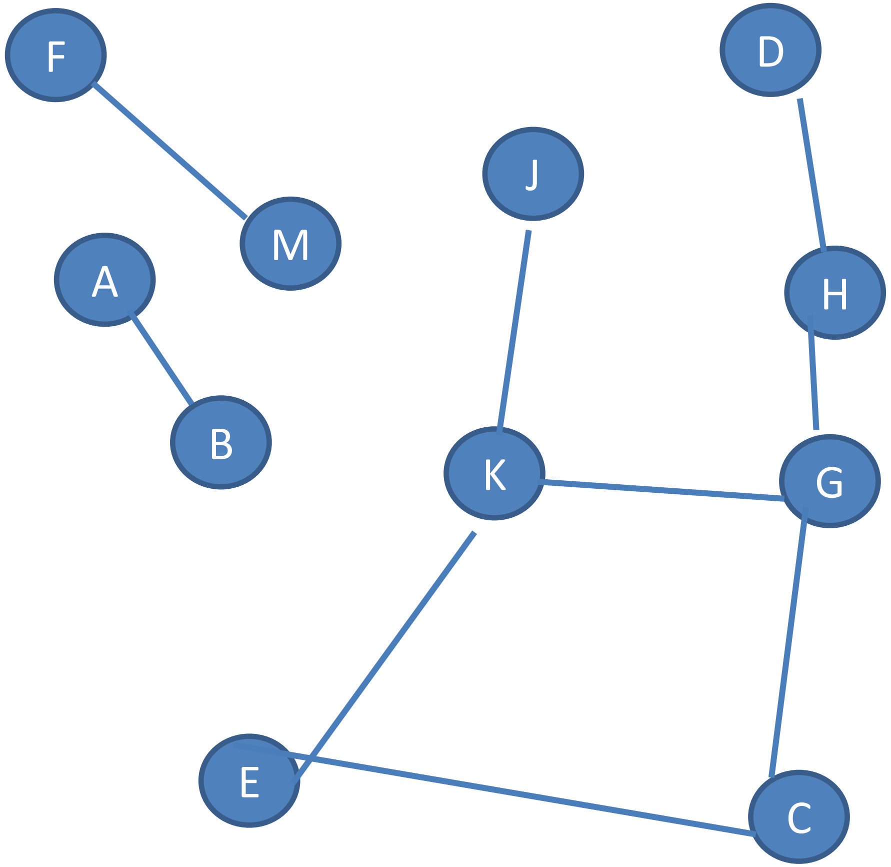
Social Network Analysis applied in advertising

How the graph gets used by advertisers:

- https://blog.liveintent.com/identity-graphs-power-the-next-era-of-marketing/
- https://www.signal.co/blog/6-things-about-id-graphs/
- https://www.zdnet.com/article/microsoft-touts-linkedin-graph-api-and-ai-foundations-of-its-new-bing-ad-service/

Relations and Layout are unrelated





Positions of nodes is irrelevant

Length of edges is irrelevant

Static Visualization of Graphs

Graphviz is a stand-alone graph rendering tool with a domain specific language, "dot," for describing graphs

- http://www.graphviz.org/

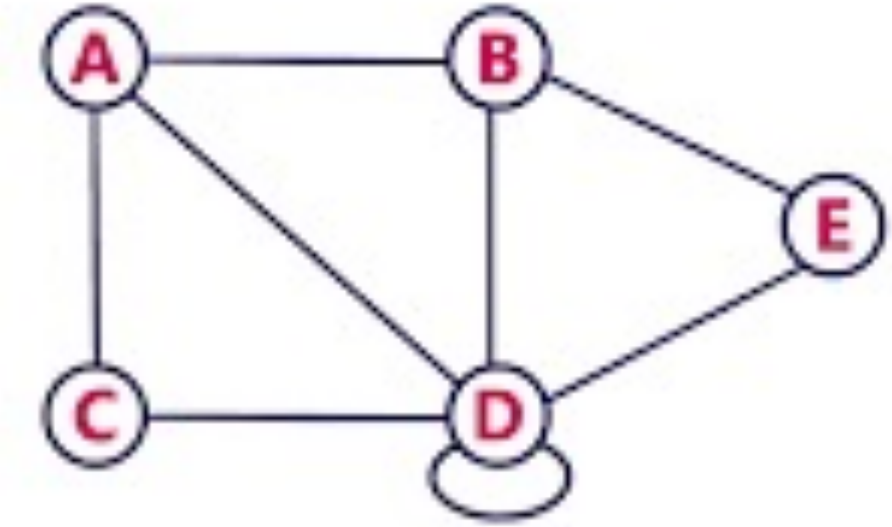
There's a Python module:

- https://pypi.org/project/graphviz/

Demo: `graphviz_for_static_graph_visualization.ipynb`

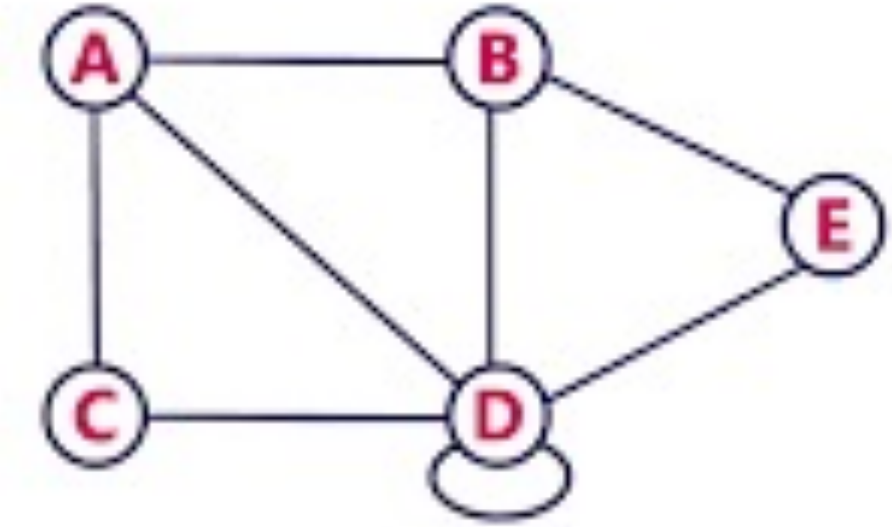
KEEP CALM
ITS
QUIZ TIME

Vote for the missing entry in the table



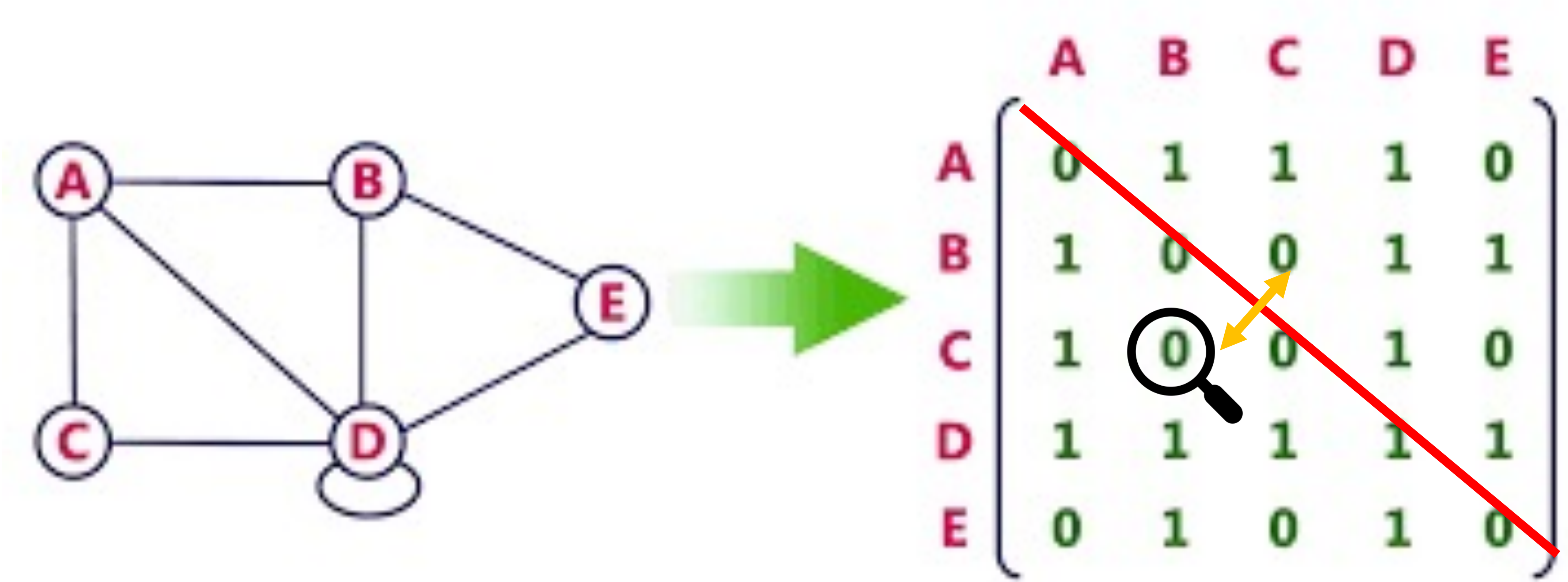
$$\begin{array}{c|ccccc} & A & B & C & D & E \\ \hline A & 0 & 1 & 1 & 1 & 0 \\ B & 1 & 0 & 0 & 1 & 1 \\ C & 1 & \square & 0 & 1 & 0 \\ D & 1 & 1 & 1 & 1 & 1 \\ E & 0 & 1 & 0 & 1 & 0 \end{array}$$

Answer for the missing entry in the table



$$\begin{array}{c|ccccc|}
 & A & B & C & D & E \\
A & 0 & 1 & 1 & 1 & 0 \\
B & 1 & 0 & 0 & 1 & 1 \\
C & 1 & 0 & 0 & 1 & 0 \\
D & 1 & 1 & 1 & 1 & 1 \\
E & 0 & 1 & 0 & 1 & 0
\end{array}$$

Explanation for the missing entry in the table



$$\begin{array}{c|ccccc} & A & B & C & D & E \\ \hline A & 0 & 1 & 1 & 1 & 0 \\ B & 1 & 0 & 0 & 1 & 1 \\ C & 1 & 0 & 0 & 1 & 0 \\ D & 1 & 1 & 1 & 1 & 1 \\ E & 0 & 1 & 0 & 1 & 0 \end{array}$$

Use symmetry of adjacency matrix to identify missing value

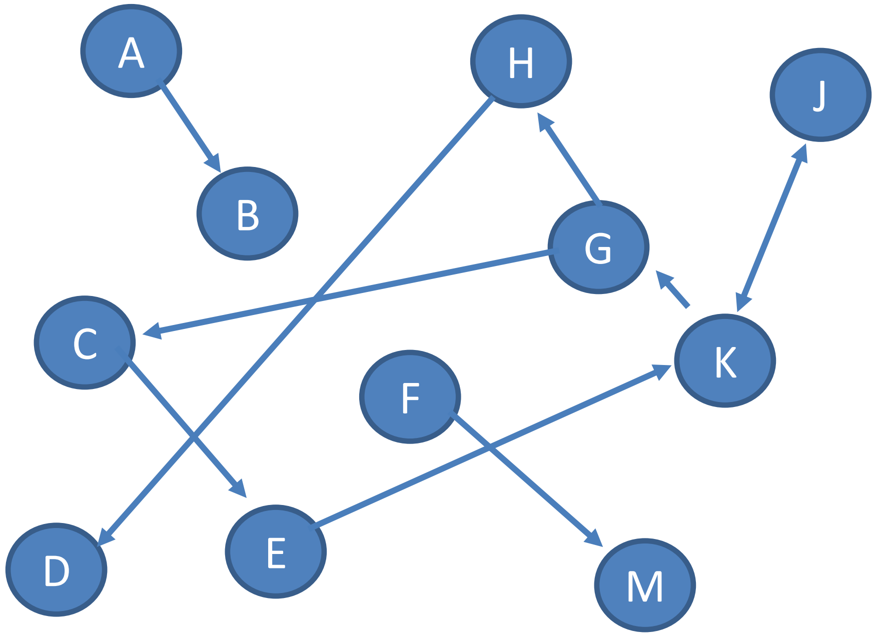
Another Python package for graphs: networkx

Demo: networkx.ipynb

- https://en.wikipedia.org/wiki/Six_Degrees_of_Kevin_Bacon
- https://en.wikipedia.org/wiki/Shortest_path_problem#Undirected_graph
- https://en.wikipedia.org/wiki/Floyd%E2%80%93Warshall_algorithm
- https://networkx.github.io/documentation/stable/reference/algorithms/shortest_paths.html

- ~~Addressing points from week 11~~
- ~~Introduction to Graphs~~
- ~~Graph visualization~~
- Directed Graphs
- Elevator pitch
- Weighted Graphs
- Property Graphs
- Homework

Directed Graph: edges with direction



Edges can be

- Directionless
- single direction
- both directions

What is "correct" depends on what the graph is describing

Example application of directed graph

- Folders and files
  - *Demo*: `Pycallgraph.ipynb`
  - *Demo*: `files_and_folders.ipynb`

Use case for a graph of web links

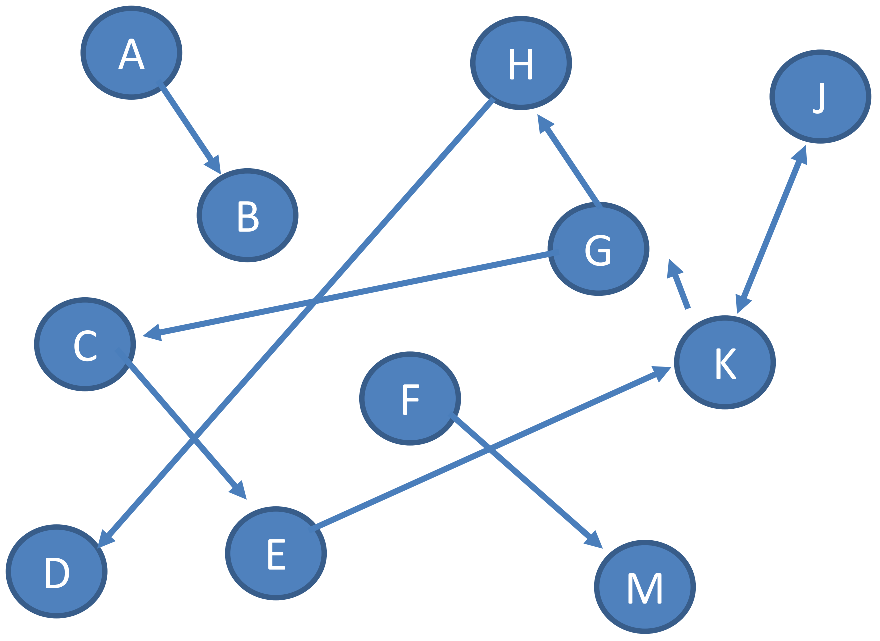
- https://en.wikipedia.org/wiki/PageRank

Websites to make a graph of

- https://scrapethissite.com/
- http://toscrape.com/

You have the tools – wget, curl, scrapy

Relation between table and directed graph



Destination node (to)

Source Node (From)

	A	B	C	D	E	F	G	H	J	K	M
A	0	1	0	0	0	0	0	0	0	0	0
B	0	0	0	0	0	0	0	0	0	0	0
C	0	0	0	0	1	0	0	0	0	0	0
D	0	0	0	0	0	0	0	0	0	0	0
E	0	0	0	0	0	0	0	0	0	1	0
F	0	0	0	0	0	0	0	0	0	0	1
G	0	0	1	0	0	0	0	1	0	0	0
H	0	0	0	1	0	0	0	0	0	0	0
J	0	0	0	0	0	0	0	0	0	1	0
K	0	0	0	0	0	0	1	0	1	0	0
M	0	0	0	0	0	0	0	0	0	0	0

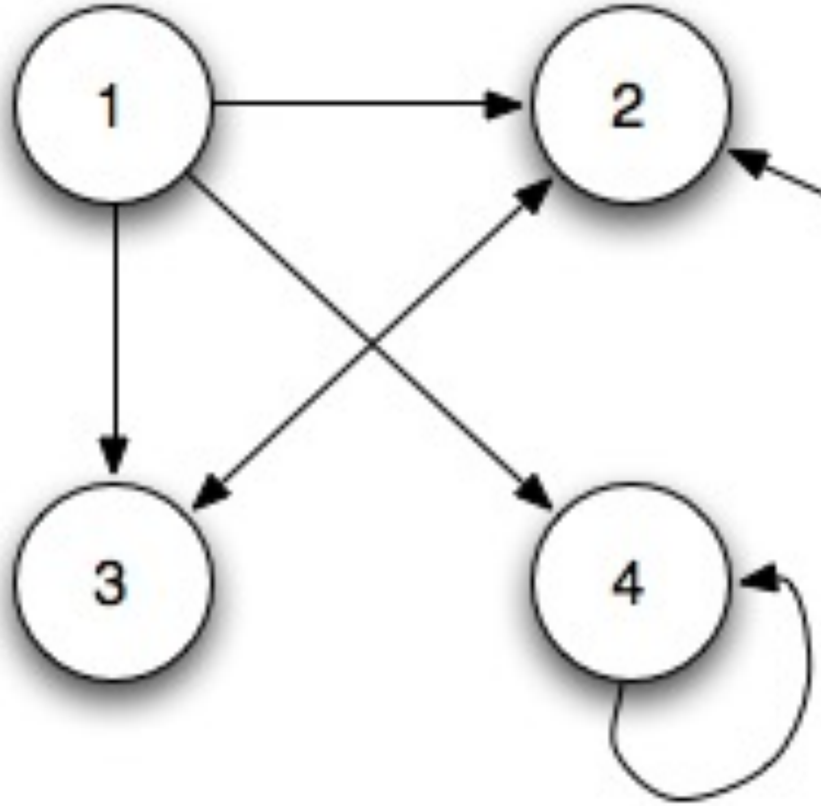
Sanity check: Number of entries with "1" should equal number of arrow heads

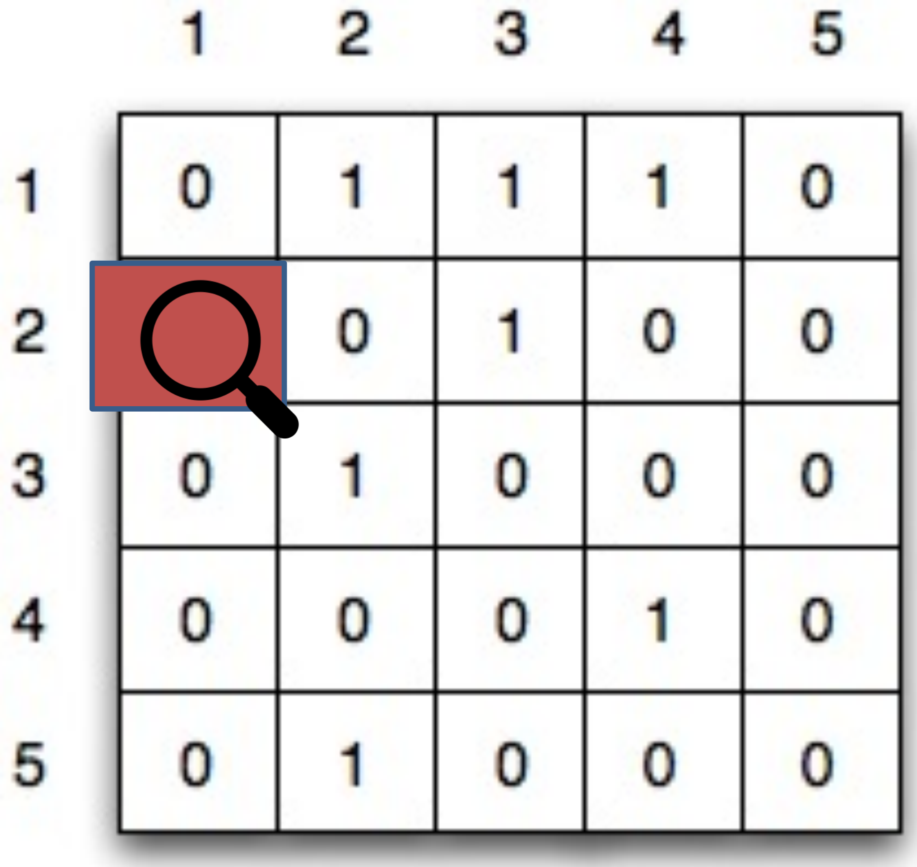
Vote for the missing entry in the adjacency matrix

Objective of this quiz is to evaluate your understanding of how well you understand relation between directed graph and adjacency matrix

- Quizzes 2 and 3 of 4

Vote for the missing entry in the table

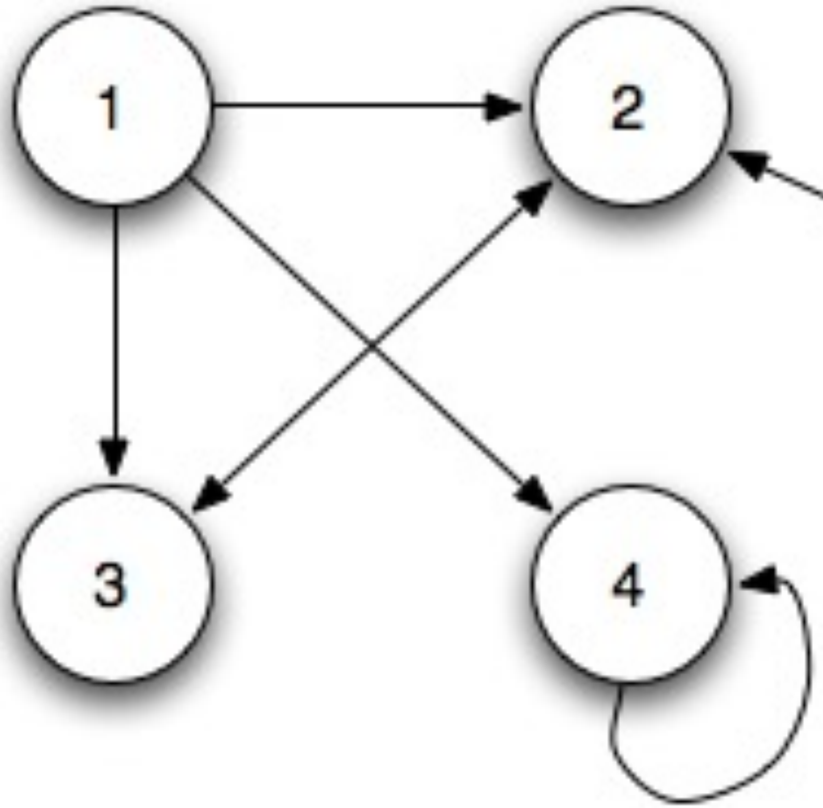




	1	2	3	4	5
1	0	1	1	1	0
2	🔍	0	1	0	0
3	0	1	0	0	0
4	0	0	0	1	0
5	0	1	0	0	0

Source: http://faculty.ycp.edu/~dbabcock/PastCourses/cs360/lectures/lecture15.html

Answer for the missing entry in the table



	1	2	3	4	5
1	0	1	1	1	0
2	0	0	1	0	0
3	0	1	0	0	0
4	0	0	0	1	0
5	0	1	0	0	0

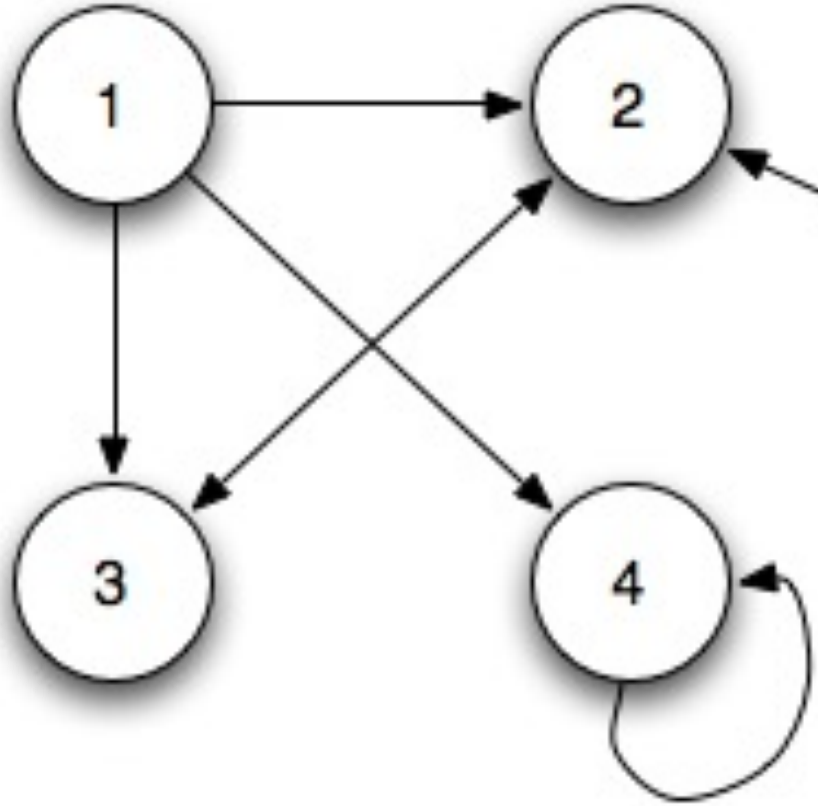
Source: http://faculty.ycp.edu/~dbabcock/PastCourses/cs360/lectures/lecture15.html

Explanation for the missing entry in the table

Directed graphs are not symmetric

1 --> 2 and 1 --> 3 and 1 --> 4; first row captures that

Nothing goes to 1

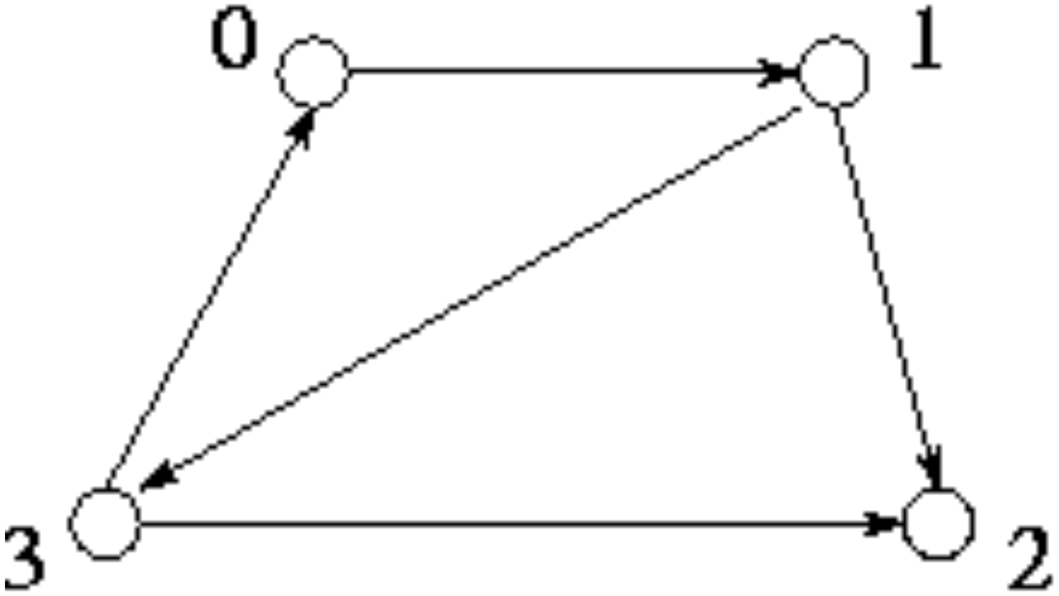


To

From

	1	2	3	4	5
1	0	1	1	1	0
2	0	0	1	0	0
3	0	1	0	0	0
4	0	0	0	1	0
5	0	1	0	0	0

Source: http://faculty.ycp.edu/~dbabcock/PastCourses/cs360/lectures/lecture15.html



G

$$\begin{array}{c|cccc} & \mathit{0} & \mathit{1} & \mathit{2} & \mathit{3} \\ \hline \mathit{0} & 0 & 1 & 0 & 0 \\ \mathit{1} & 0 & 0 & 1 & 1 \\ \mathit{2} & 0 & 0 & 0 & 0 \\ \mathit{3} & 🔍 & 0 & 1 & 0 \end{array}$$

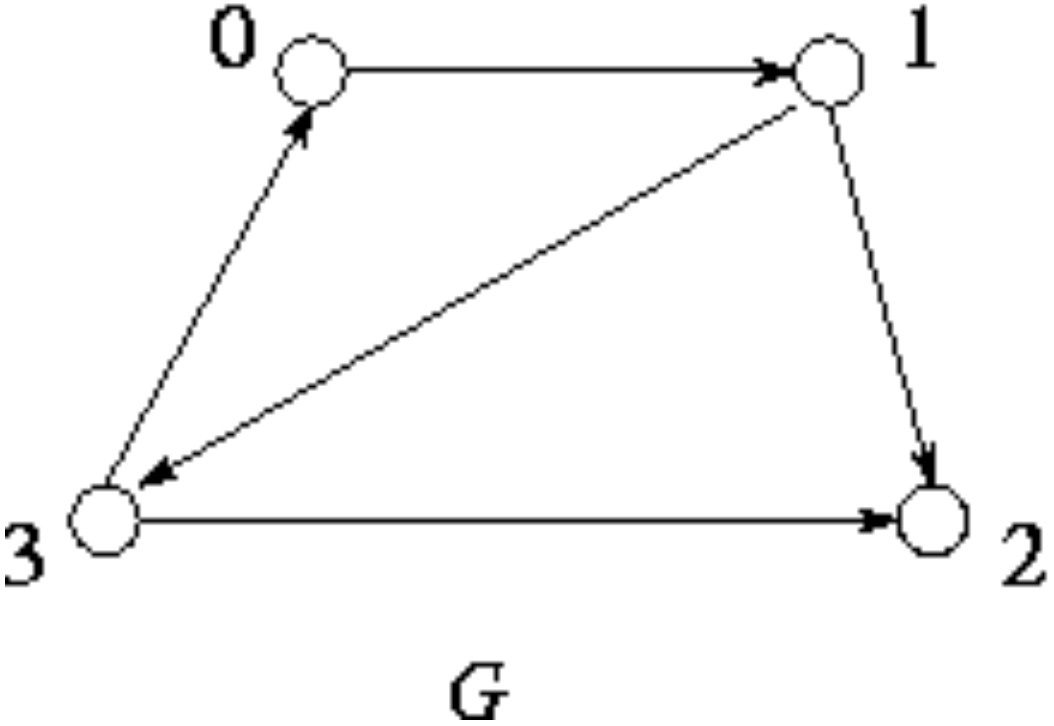
A

Answer for missing entry in the table

Directed graphs are not symmetric

1 --> 2 and 3 --> 2; column "2" captures that

From 3 to 0 means the lower-left value is 1



From

$$
\begin{array}{c|cccc}
 & 0 & 1 & 2 & 3 \\
\hline
0 & 0 & 1 & 0 & 0 \\
1 & 0 & 0 & 1 & 1 \\
2 & 0 & 0 & 0 & 0 \\
3 & 1 & 0 & 1 & 0
\end{array}
$$

A

- ~~Addressing points from previous week~~
- ~~Introduction to Graphs~~
- ~~Directed Graphs~~
- Elevator pitch
- Weighted Graphs
- Property Graphs
- Homework

How I use short exposure to leaders

Where: hallway encounters

- Leader's exposure to low-level truth is limited
- [Good] leaders don't want to micromanage

- What does leadership expect to hear?
  - Not enough resources (people, money, time, computers, data)
  - Everything is going great
  - I will get that report to you by Thursday

How I use short exposure to leaders

- Can I describe something novel?
  - What are you planning to do next?
  - Rumors need to be vetted before sharing
  - Duplicative efforts with other teams

- No one likes delivering bad news
  - (Project) is going to be late
  - (Project) is unlikely to succeed

--> This is the ground truth leaders want

Mid-level managers dislike delivering bad news to leaders

- ~~Addressing points from previous week~~
- ~~Introduction to Graphs~~
- ~~Directed Graphs~~
- ~~Elevator Pitch~~
- Weighted Graphs
- Property Graphs
- Homework

Use case for graphs: Logistics

Routing of

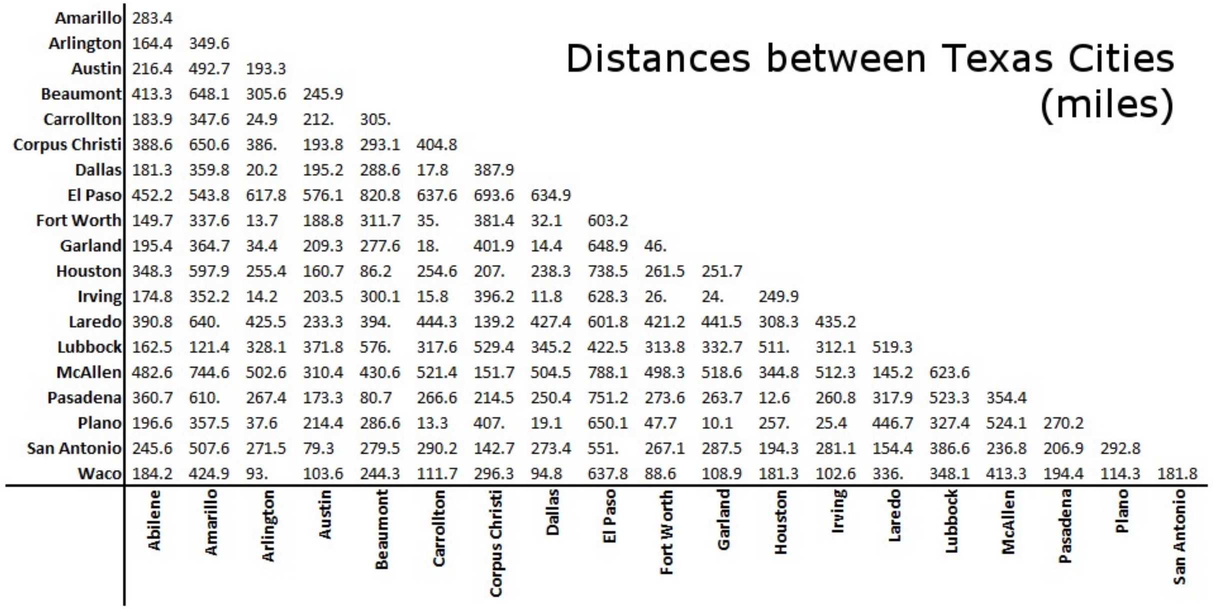
- Flights (American, Delta, Southwest, United, Spirit)
- Delivery trucks (Amazon, USPS, FedEx, UPS, DHL)
- Electrical power

Each physical location is separated by a distance --> time, money

Which choice of routes is cheapest or fastest?

Example application of weighted graph

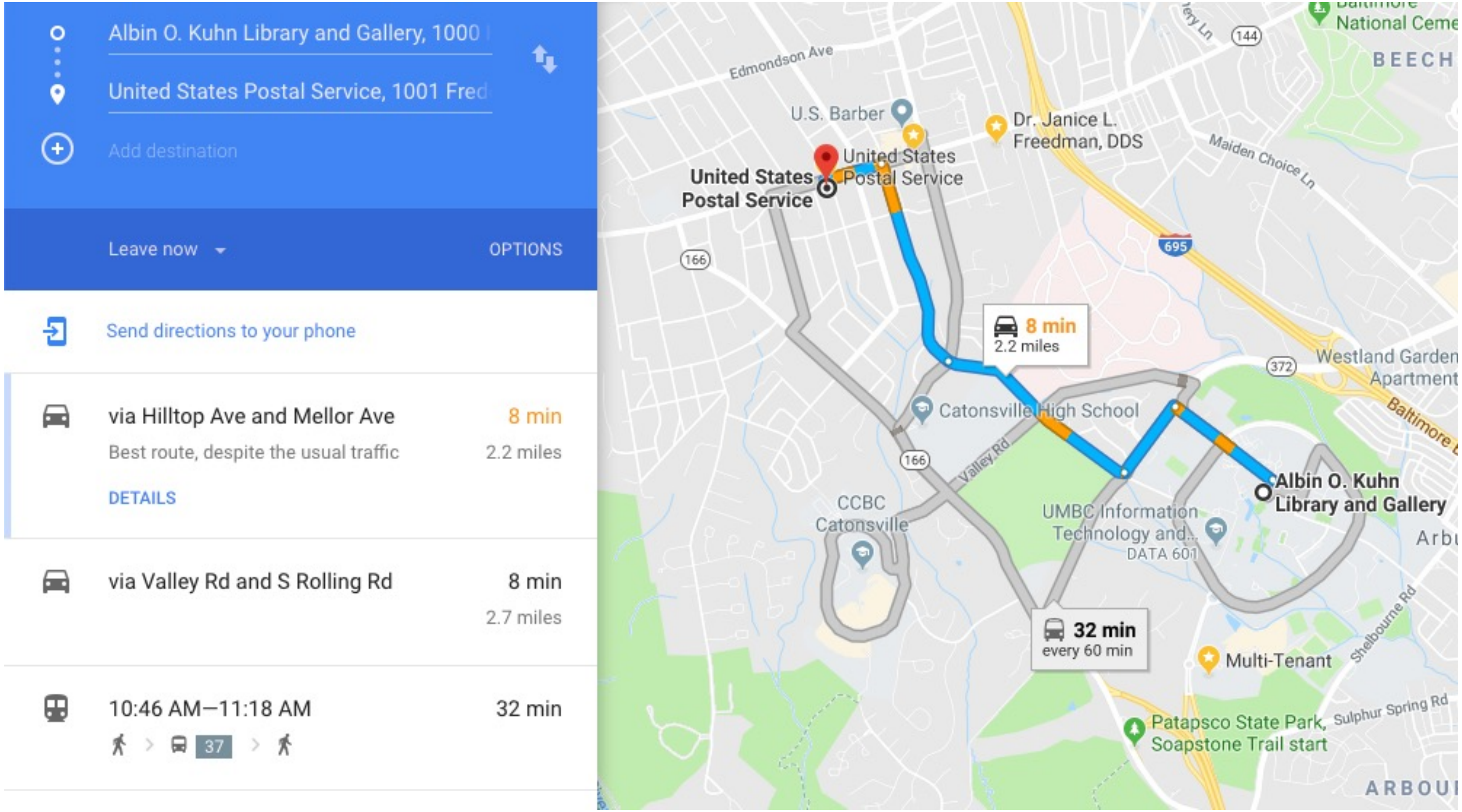
- Travelling salesperson problem: "Given a list of cities and the distances between each pair of cities, what is the shortest possible route that visits each city and returns to the origin city?"



Distances between Texas Cities (miles)

	Abilene	Amarillo	Arlington	Austin	Beaumont	Carrollton	Corpus Christi	Dallas	El Paso	Fort Worth	Garland	Houston	Irving	Laredo	Lubbock	McAllen	Pasadena	Plano	San Antonio
Amarillo	283.4																		
Arlington	164.4	349.6																	
Austin	216.4	492.7	193.3																
Beaumont	413.3	648.1	305.6	245.9															
Carrollton	183.9	347.6	24.9	212.	305.														
Corpus Christi	388.6	650.6	386.	193.8	293.1	404.8													
Dallas	181.3	359.8	20.2	195.2	288.6	17.8	387.9												
El Paso	452.2	543.8	617.8	576.1	820.8	637.6	693.6	634.9											
Fort Worth	149.7	337.6	13.7	188.8	311.7	35.	381.4	32.1	603.2										
Garland	195.4	364.7	34.4	209.3	277.6	18.	401.9	14.4	648.9	46.									
Houston	348.3	597.9	255.4	160.7	86.2	254.6	207.	238.3	738.5	261.5	251.7								
Irving	174.8	352.2	14.2	203.5	300.1	15.8	396.2	11.8	628.3	26.	24.	249.9							
Laredo	390.8	640.	425.5	233.3	394.	444.3	139.2	427.4	601.8	421.2	441.5	308.3	435.2						
Lubbock	162.5	121.4	328.1	371.8	576.	317.6	529.4	345.2	422.5	313.8	332.7	511.	312.1	519.3					
McAllen	482.6	744.6	502.6	310.4	430.6	521.4	151.7	504.5	788.1	498.3	518.6	344.8	512.3	145.2	623.6				
Pasadena	360.7	610.	267.4	173.3	80.7	266.6	214.5	250.4	751.2	273.6	263.7	12.6	260.8	317.9	523.3	354.4			
Plano	196.6	357.5	37.6	214.4	286.6	13.3	407.	19.1	650.1	47.7	10.1	257.	25.4	446.7	327.4	524.1	270.2		
San Antonio	245.6	507.6	271.5	79.3	279.5	290.2	142.7	273.4	551.	267.1	287.5	194.3	281.1	154.4	386.6	236.8	206.9	292.8	
Waco	184.2	424.9	93.	103.6	244.3	111.7	296.3	94.8	637.8	88.6	108.9	181.3	102.6	336.	348.1	413.3	194.4	114.3	181.8

source



Albin O. Kuhn Library and Gallery, 1000

United States Postal Service, 1001 Fred

Add destination

Leave now OPTIONS

Send directions to your phone

via Hilltop Ave and Mellor Ave 8 min
Best route, despite the usual traffic 2.2 miles
DETAILS

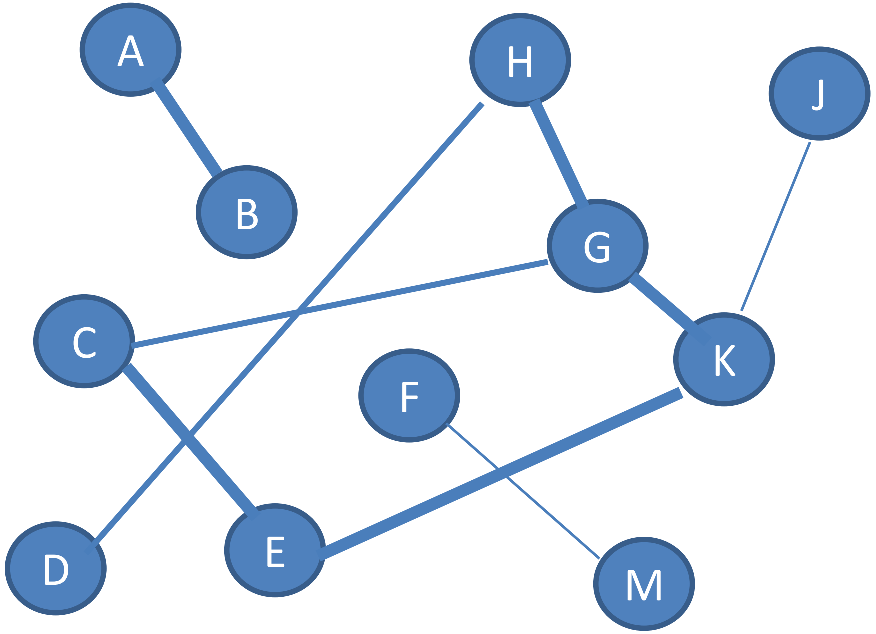
via Valley Rd and S Rolling Rd 8 min
2.7 miles

10:46 AM—11:18 AM 32 min
37

https://blogs.cornell.edu/info2040/2011/09/14/google-maps-its-just-one-big-graph/

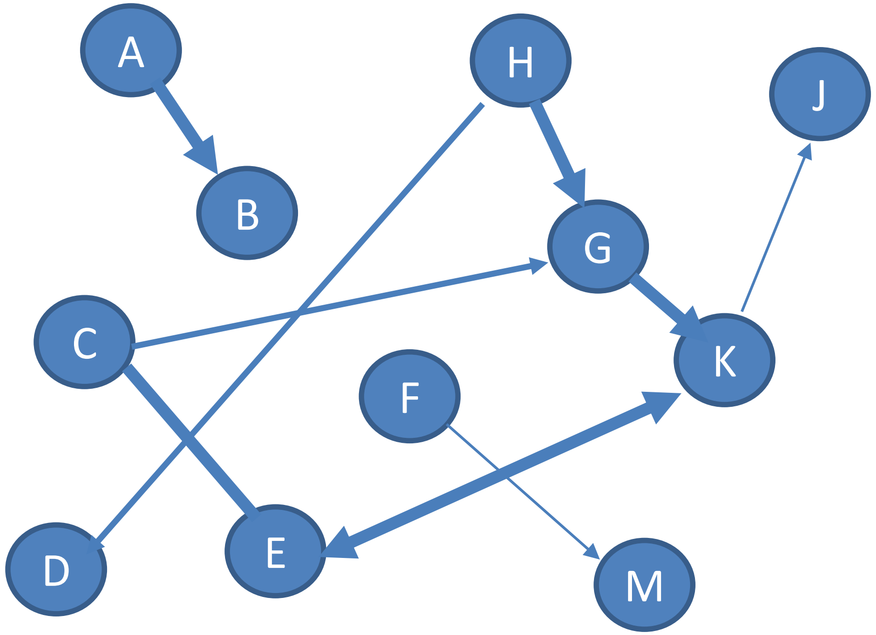
https://motherboard.vice.com/en_us/article/4x3pp9/the-simple-elegant-algorithm-that-makes-google-maps-possible

weighted graph and adjacency matrix



	A	B	C	D	E	F	G	H	J	K	M
A	0	4	0	0	0	0	0	0	0	0	0
B	4	0	0	0	0	0	0	0	0	0	0
C	0	0	0	0	4	0	2	0	0	0	0
D	0	0	0	0	0	0	0	2	0	0	0
E	0	0	4	0	0	0	0	0	0	4	0
F	0	0	0	0	0	0	0	0	0	0	1
G	0	0	2	0	0	0	0	4	0	4	0
H	0	0	0	2	0	0	4	0	0	0	0
J	0	0	0	0	0	0	0	0	0	1	0
K	0	0	0	0	4	0	4	0	1	0	0
M	0	0	0	0	0	1	0	0	0	0	0

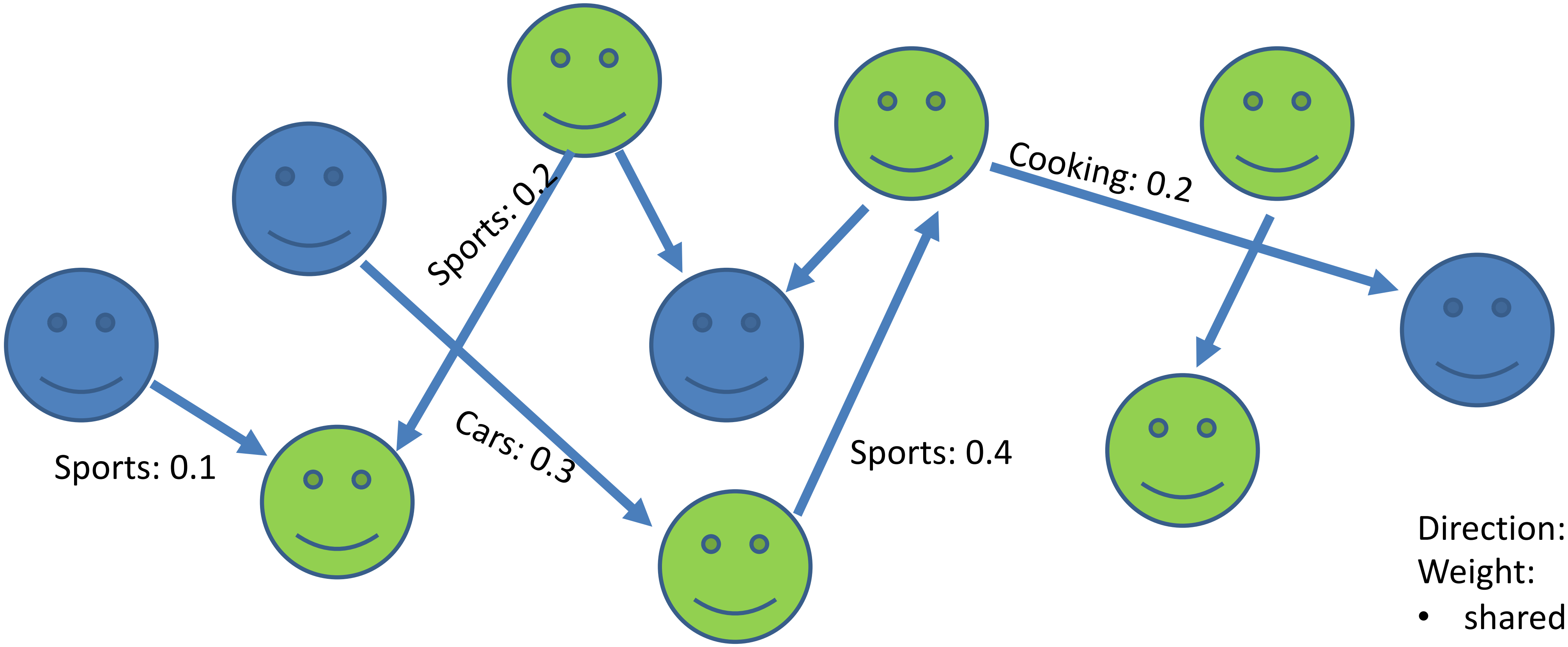
Directed weighted graph and adjacency matrix



	A	B	C	D	E	F	G	H	J	K	M
A	0	4	0	0	0	0	0	0	0	0	0
B	4	0	0	0	0	0	0	0	0	0	0
C	0	0	0	0	4	0	2	0	0	0	0
D	0	0	0	0	0	0	0	2	0	0	0
E	0	0	4	0	0	0	0	0	0	4	0
F	0	0	0	0	0	0	0	0	0	0	1
G	0	0	2	0	0	0	0	4	0	4	0
H	0	0	0	2	0	0	4	0	0	0	0
J	0	0	0	0	0	0	0	0	0	1	0
K	0	0	0	0	4	0	4	0	1	0	0
M	0	0	0	0	0	1	0	0	0	0	0

Example application of Weighted directed graphs

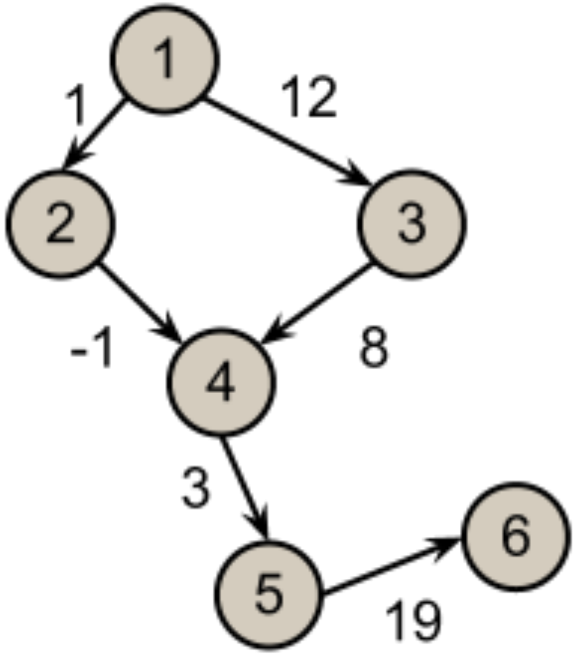
Social interactions, e.g., dating



Direction: interested in
Weight:
- shared interests
- distance

Vote for the missing entry in the table

Weighted Directed Graph & Adjacency Matrix



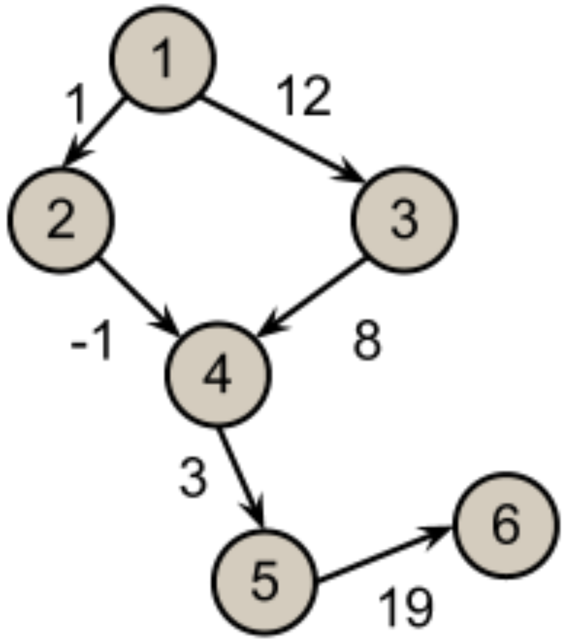
Weighted Directed Graph

	1	2	3	4	5	6
1	0	1	12	0	0	0
2	-1	0	0	-1	0	0
3	-12	0	0	8	0	0
4	0	1	-8	0	3	0
5	0	🔍	0	-3	0	19
6	0	0	0	0	-19	0

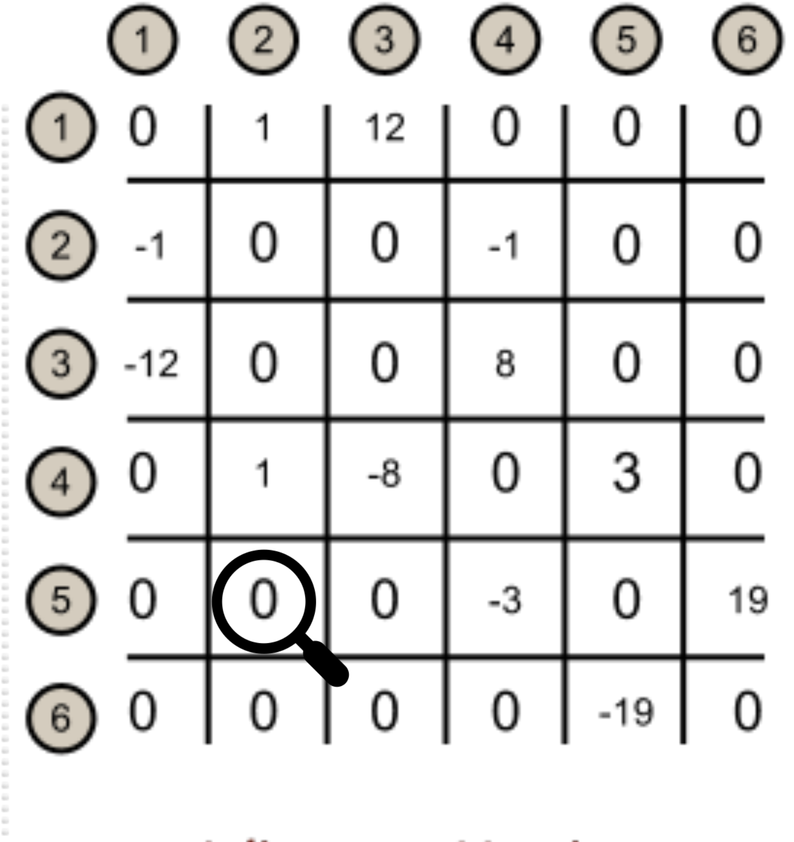
Adjacency Matrix

Weighted Directed Graph & Adjacency Matrix

No direct connection between 2 and 5



Weighted Directed Graph



	1	2	3	4	5	6
1	0	1	12	0	0	0
2	-1	0	0	-1	0	0
3	-12	0	0	8	0	0
4	0	1	-8	0	3	0
5	0	0	0	-3	0	19
6	0	0	0	0	-19	0

Adjacency Matrix

Standard problems when using graphs

- Fully connected graph has little value
- Visualization may not add value
- Visualization can be computationally expensive for large graphs
- Visualization can be slow to render for large graphs

- ~~Addressing points from previous week~~
- ~~Introduction to Graphs~~
- ~~Directed Graphs~~
- ~~Elevator Pitch~~
- ~~Weighted Graphs~~
- Property Graphs
- Homework

Property graph = graph + attributes

- Let a node represent a person
- An edge is a relationship between people

- Person:Ben   is_friend_of(duration_years:3)   Person:Mary
- Person:Ben   is_friend_of(duration_years:1)   Person:Alex
- Person:Ben(born_year:1923)
- Person:Mary(born_year:1958)
- Person:Alex(born_year:1982)

Neo4j

- Domain Specific Language: CYPHER
- Open source (free) and paid versions
- Web interface for queries and visualization
- Use py2neo for Python interaction

Graphs and Machine Learning

- The nodes, node attributes, edges, and edge attributes are not always provided in a dataset
- Machine Learning can be used to predict missing edges and other information needed
- Large graphs have more features to base the training on

https://towardsdatascience.com/graph-representation-learning-dd64106c9763

LinkedIn

- You can leverage the connections you've made in this class!

https://www.linkedin.com/in/msarica/

- ~~Addressing points from previous week~~
- ~~Introduction to Graphs~~
- ~~Directed Graphs~~
- ~~Elevator Pitch~~
- ~~Weighted Graphs~~
- ~~Property Graphs~~
- Homework

Activity: Know/Don't know/Want to know

Send me an email with the following:

- 1 thing you already knew that was covered in this session
- 1 thing covered in this session you didn't previously know
- 1 thing you wanted to know but wasn't covered in this session

Reading assignment

"Mining massive datasets"

http://infolab.stanford.edu/~ullman/mmds/book0n.pdf

- page 163 to 168 (181 to 186 in the PDF)
- plus the summary of chapter 5 on page 196-197 (214 to 215 in pdf)

Homework assignment: parse .ipynb files

- Create a notebook that finds all other notebooks used for Data 601 on your computer and generates a list of the modules used
- Lines of code that start with either "`from`" or "`import`"
- Your notebook should analyze at least 8 notebooks
- Submit your python notebook (.ipynb file) containing the code for the implementation and the results. Results are a list of modules used.

- *Optional bonus question*: How many lines of Python code have you created for Data 601?

Log data

- https://www.loganalyzer.net/log-analysis-tutorial/log-file-sample-explain.html
- https://www.jafsoft.com/searchengines/log_sample.html
- https://www.blendo.co/blog/clickstream-data-mining-techniques-introduction/

- http://www.herongyang.com/Windows/Web-Log-File-IIS-Apache-Sample.html
- https://dataplatform.cloud.ibm.com/docs/content/wsj/streaming-pipelines/clickstream_example_pipeline.html

Wikipedia clickstream

- https://old.datahub.io/dataset/wikipedia-clickstream
- https://en.wikipedia.org/wiki/Click_path
- https://ewulczyn.github.io/Wikipedia_Clickstream_Getting_Started/
- https://wikimediafoundation.org/news/2018/01/16/wikipedia-rabbit-hole-clickstream/
- https://wikimediafoundation.org/news/2018/01/16/wikipedia-rabbit-hole-clickstream/
- http://databricks-michael.s3.amazonaws.com/Wikipedia%20Clickstream%20Data.html