GENERAL ASSEMBLY

NETWORK ANALYSIS

Matt Brems

DSI+

LEARNING OBJECTIVES

- By the end of this lesson, students should be able to:
  - Describe the relevance of networks.
  - Identify undirected, directed, cyclic, acyclic, connected, disconnected networks, and multigraphs.
  - Identify the importance of degree and adjacency matrix.
  - Use the NetworkX package to do basic network analysis.

NETWORKS

correlated data

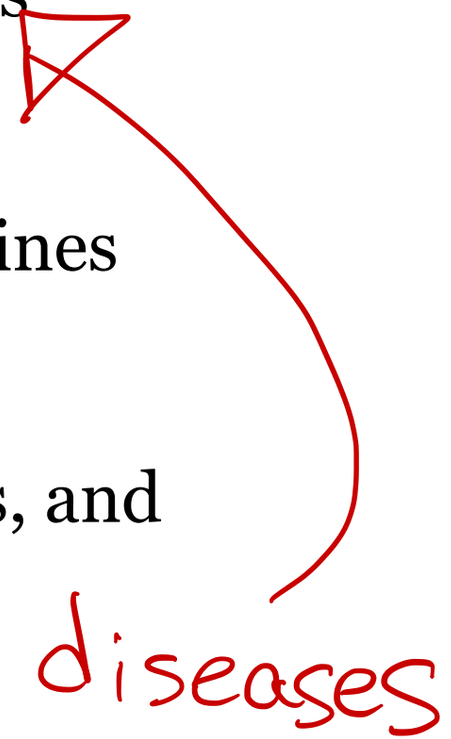
massive amounts

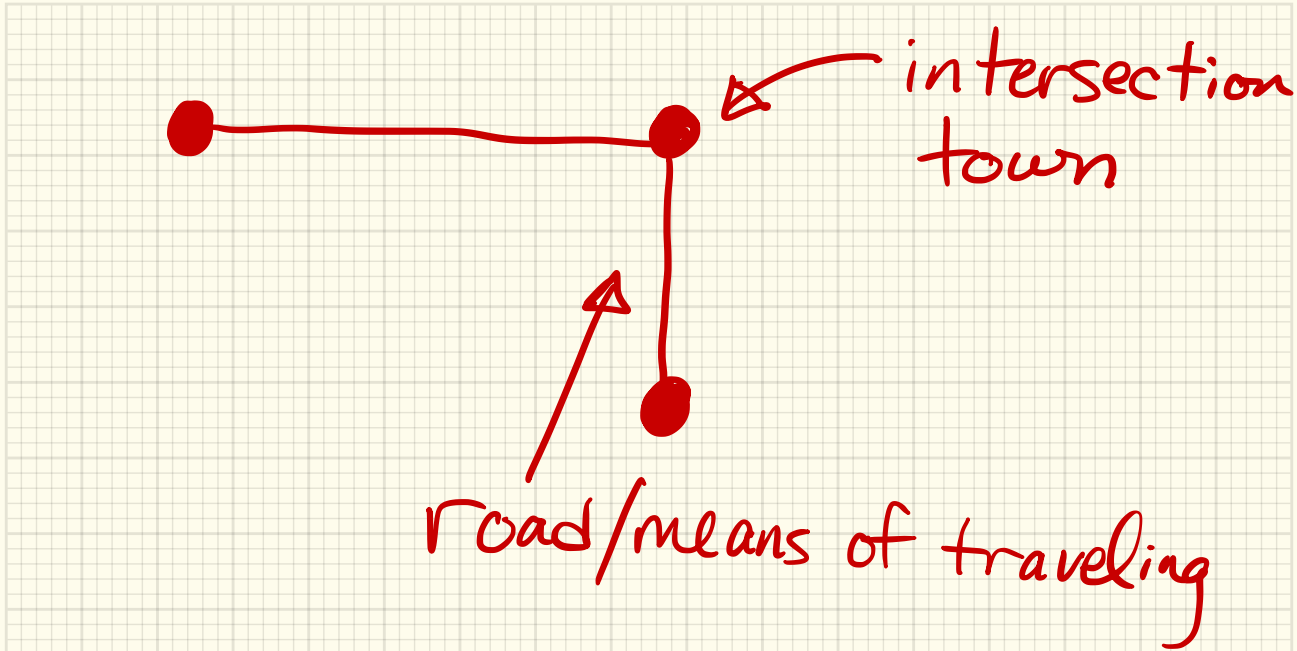
"I think the next century will be the century of complexity."
– Stephen Hawking

↳ numbers
↳ categorical
↳ text
↳ images
↳ videos

EXAMPLES OF NETWORKS

- Social network: professional, friendship, and family ties
- Utility network: power grid, generators, transmission lines
- Biological network: interactions among genes, proteins, and metabolites.
- Trade network: exchanging goods and services



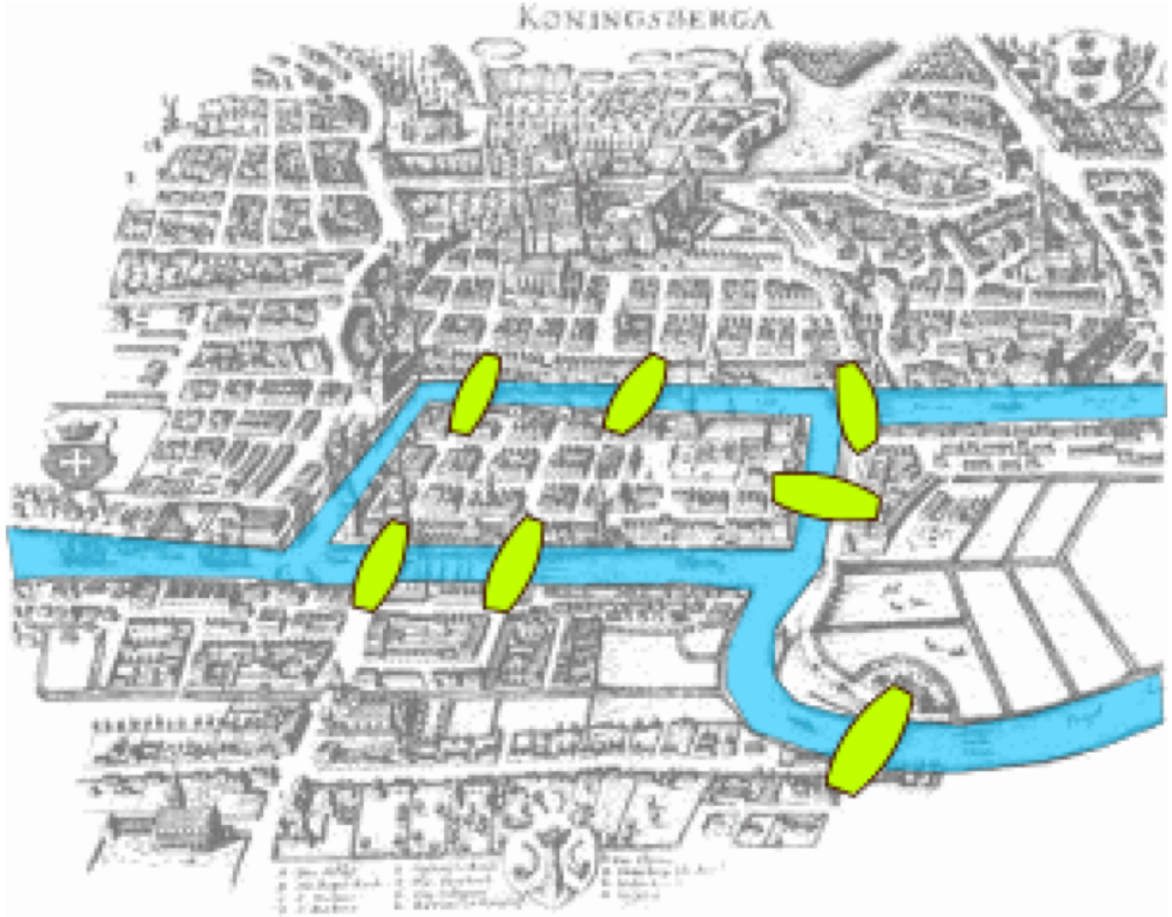

intersection
town
road/means of traveling

NETWORK ANALYSIS

- Can we take these examples above, boil them down to very basic components, and come up with a set of rules that apply for all networks, regardless of application?
    - This is, roughly, the area of network science.
    - You may also hear the terms “graph theory” and “topology,” which are very closely related to “network science.”

not like "Bueller"

SEVEN BRIDGES OF KÖNIGSBERG

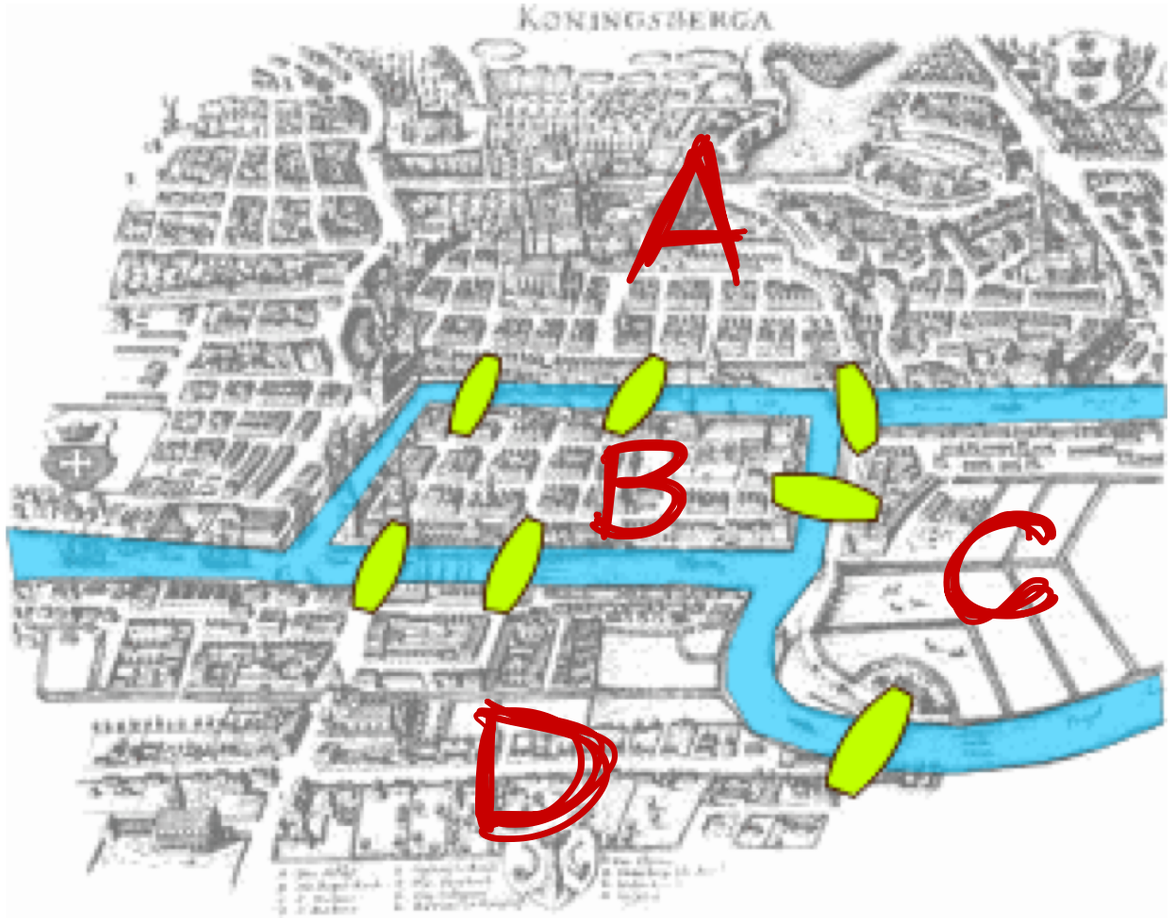


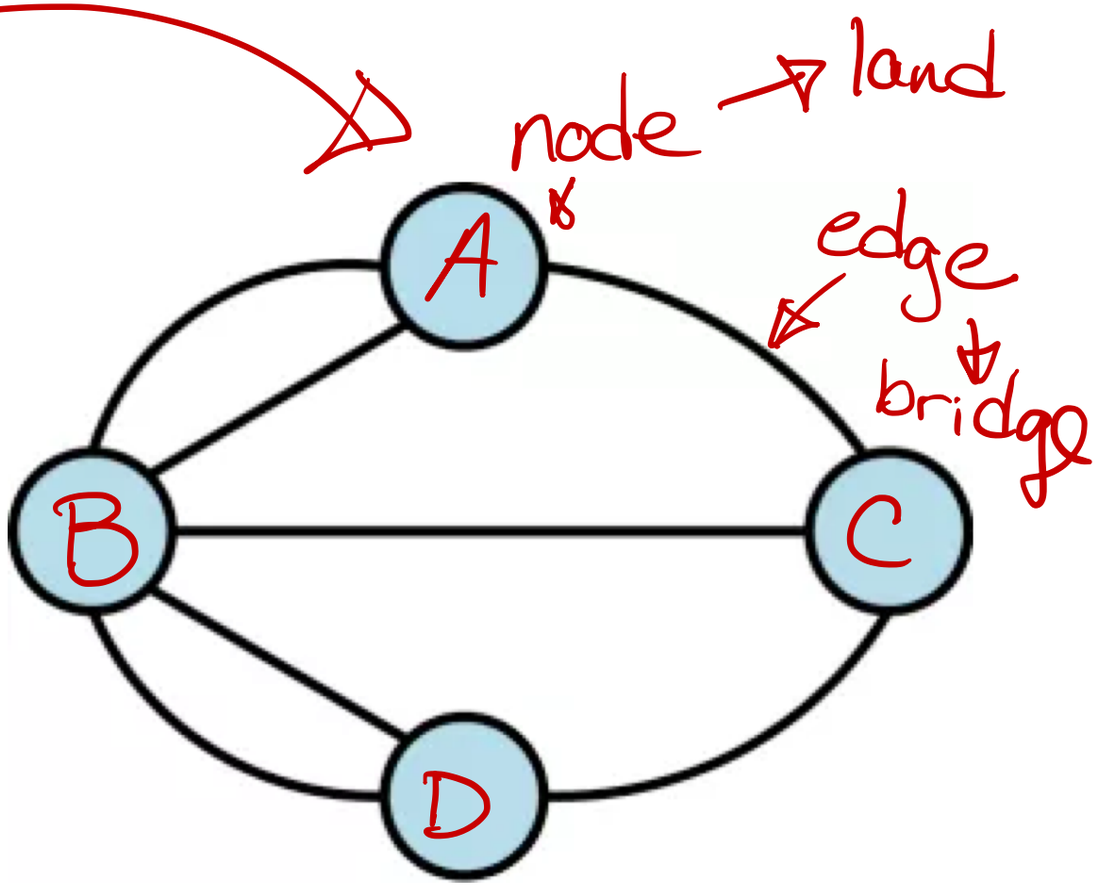
Euler → "oiler"

- Can we traverse all seven bridges exactly once? No!

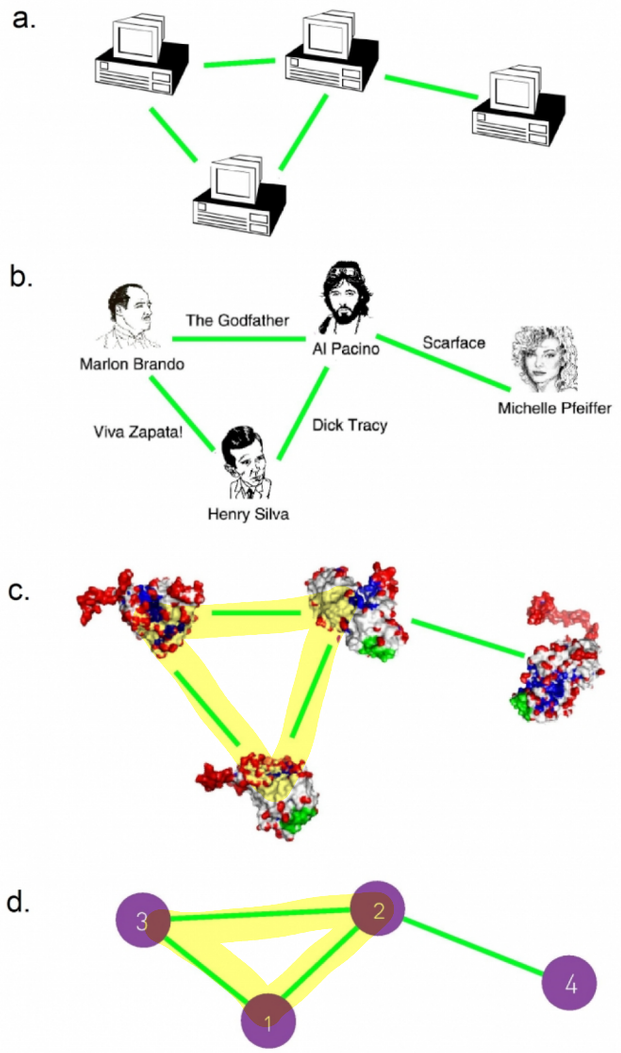
← water

SEVEN BRIDGES OF KÖNIGSBERG





NETWORKS



- By representing complex systems with **nodes** and **edges**, we can directly compare different systems.
    - If a rule applies to one network with a given structure, that rule should apply to all other networks with the same structure!

TYPES OF NETWORKS

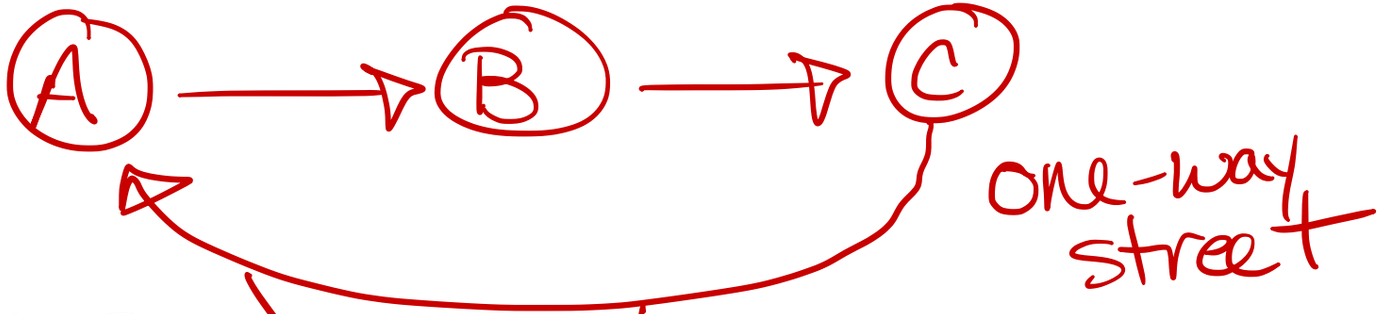
- **Undirected:** An undirected network is one where the connection extends in both directions.

two-way streets
most airports

Facebook friends

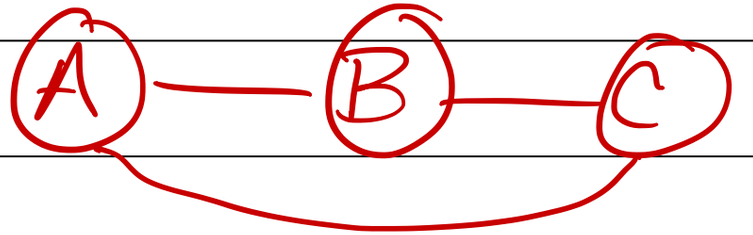
- **Directed:** A directed network is one where the connection may only flow in one direction.

love triangles
blood flow (fluid in pipe)
most rivers



Twitter / Instagram
electrical current

TYPES OF NETWORKS



- **Cyclic:** A cyclic network is one that contains at least one cycle.
  - A cycle exists when a node can be connected to itself by traversing at least one edge.

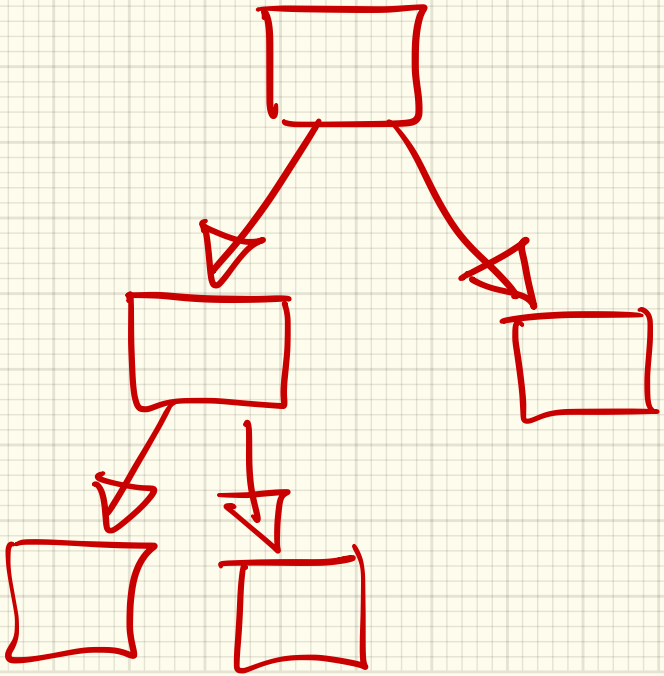
roads christmas tree lights blood flow

- **Acyclic:** An acyclic network is one that contains no cycles.

arteries digestive system time

P → VP → Speaker → PPT → State → → →
Senate

Decision Trees



directed acyclic graphs

TYPES OF NETWORKS

- **Multigraph:** A multigraph may have multiple links connecting the same pair of nodes.

Seven Bridges of Königsberg

DC Metro

Metro Center — McPherson Sq.

Slack

nodes: accounts

edges: channels

MB

EC instr.

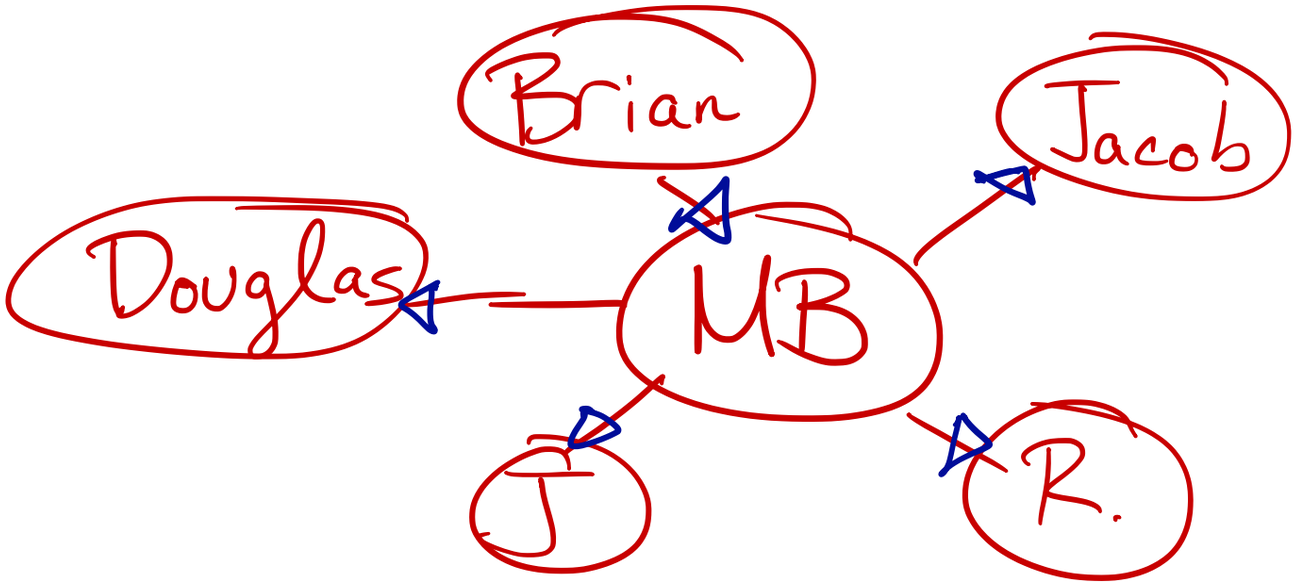
EC instr.

DM

Boom

OTHER NETWORK TERMINOLOGY

- **Degree:** The degree of a node in an undirected graph is given by the number of links of that node.
  - **In-degree:** The number of links heading **into** that node.
  - **Out-degree:** The number of links heading **out of** that node.



deg(MB) = 5

out-degree(MB) = 4

in-degree(MB) = 1

NETWORK SUMMARY STATISTICS

- What does the average degree of a network tell us?

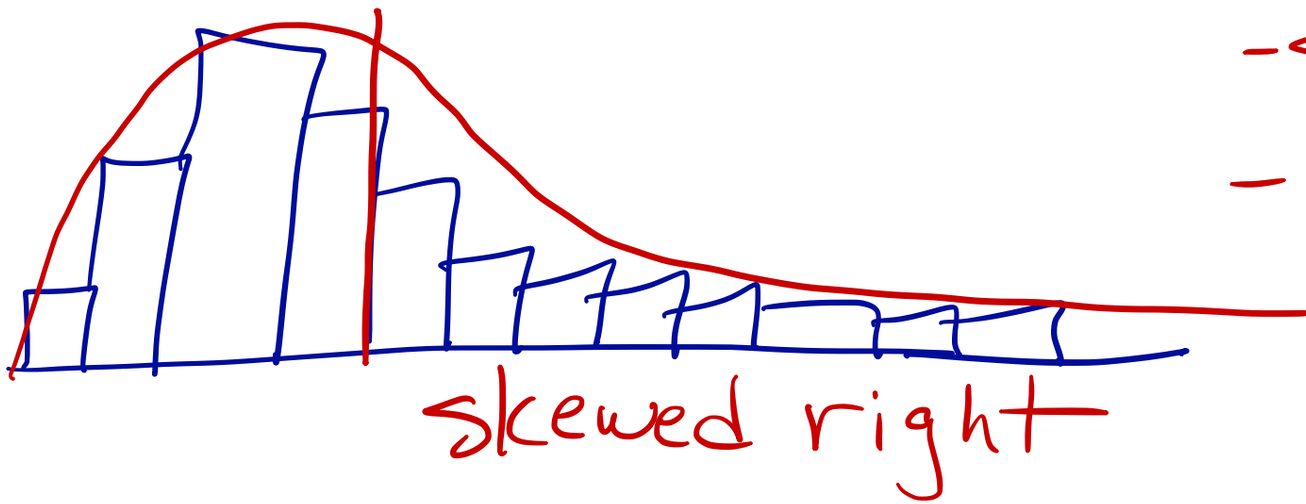
avg. degree = 10

avg. degree = 1,000

connectedness

how quickly/easily

- What does the degree distribution of a network tell us?



- set of all values of deg.
- how frequently we observe each.

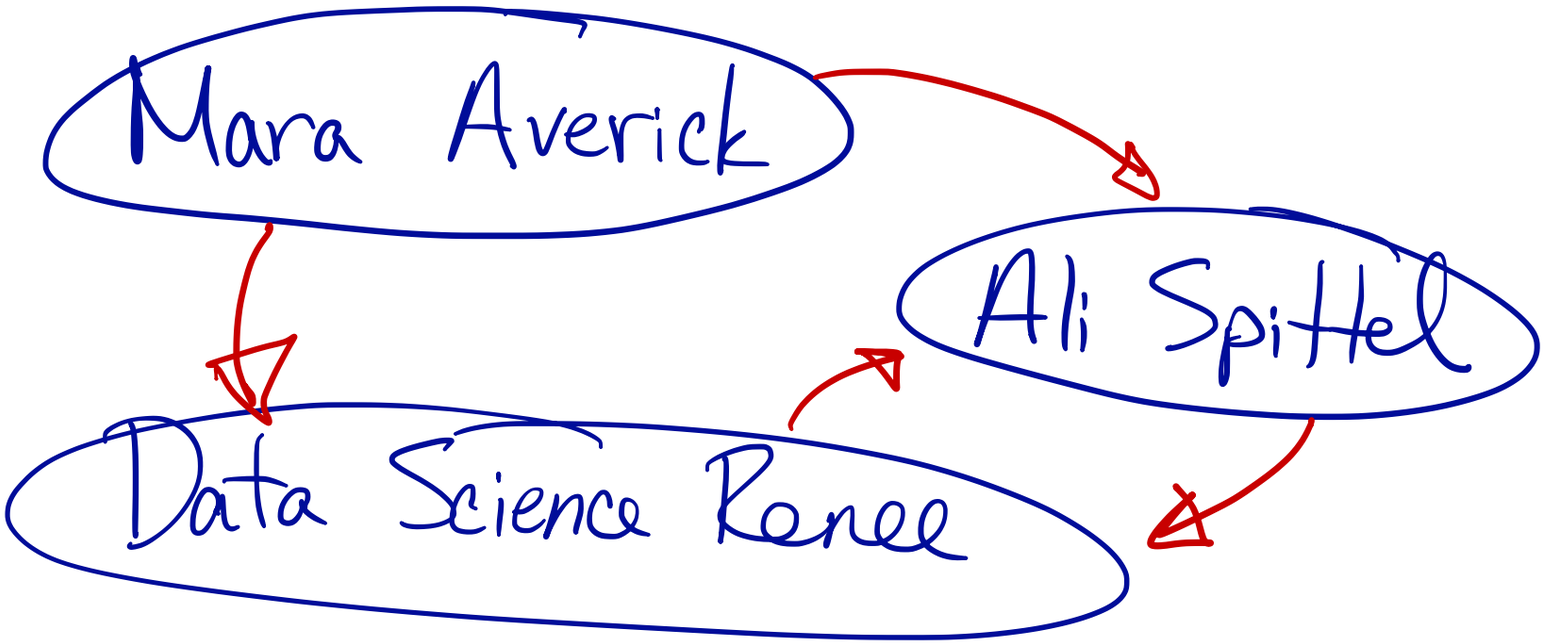
"Dataclysm" OkCupid

the more interconnected two partners were with each others' friends, the MORE LIKELY they were to split up.

REPRESENTING A NETWORK

- Visually representing a network can be difficult. (We'll get into that later.)

- We often represent a network with an **adjacency matrix**.



	MA	AS	DR
MA	0	1	1
AS	0	0	1
DR	0	1	0

USES OF ADJACENCY MATRICES

- Our computers and modeling techniques do well with matrices.

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

- If we take our adjacency matrix A and raise it to the n^{th} power, we can see how many paths of length n exist from node to node.

$A^3 \rightarrow$ how many paths of length 3 exist from node i to j

- It's useful for detecting communities in networks.
  - Leading eigenvectors can be used to detect communities.

↳ groups of friends

FINAL TERMS

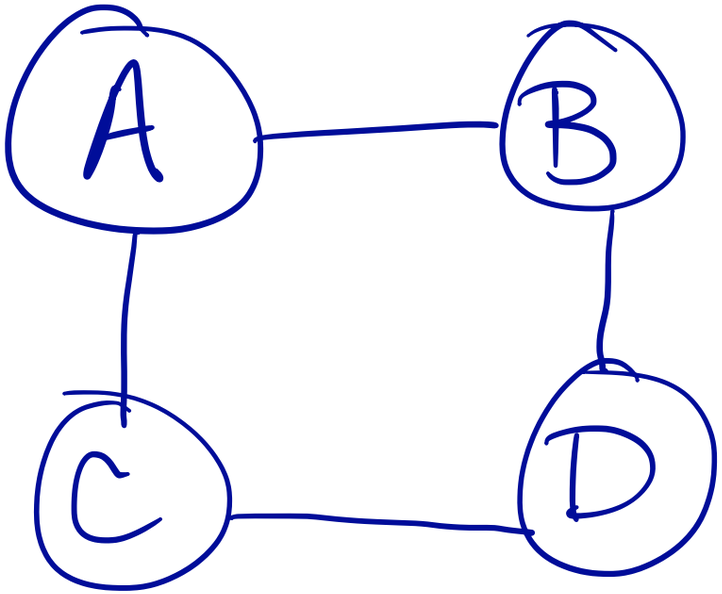
- **Minimum Path Length:** The minimum number of edges required to traverse to connect node *i* and node *j*.

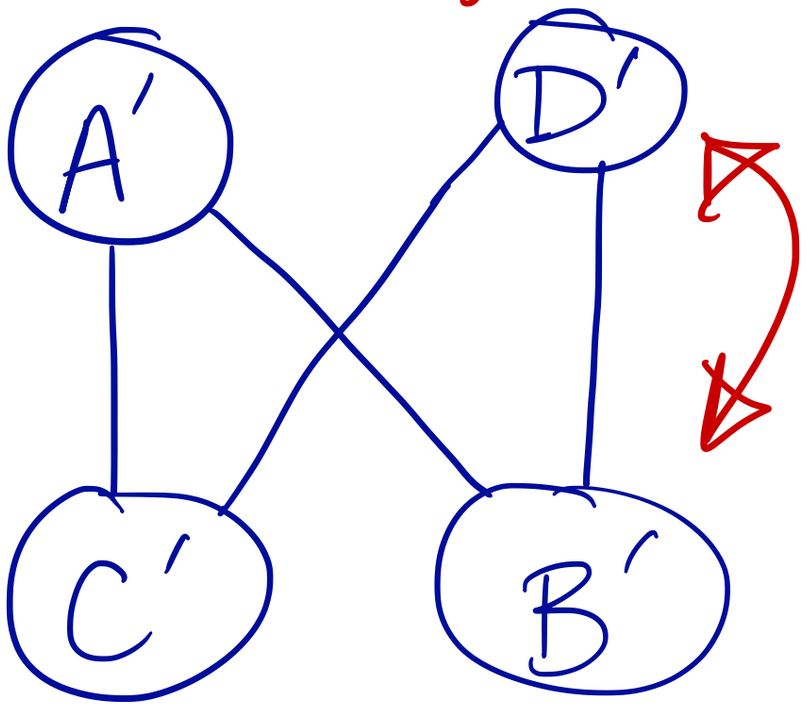
- **Diameter:** The maximum shortest path length of a network.
  ↳ six degrees of separation (Kevin Bacon)
- **Connected:** Two nodes *i* and *j* are connected if there exists a path from *i* to *j*.

- **Isomorphic:** Informally, we say two graphs are isomorphic if we can twist one graph to look exactly like the other without cutting or gluing.

EXAMPLE OF ISOMORPHIC GRAPH

These networks are isomorphic.





UNSUPERVISED LEARNING

- **Unsupervised learning** is where we have, as part of our training data, no observed Y values.

- **Supervised learning** is where we have observed Y values as part of our training data.

↳ Regression

↳ Classification

UNSUPERVISED LEARNING APPLICATIONS ON NETWORKS

1) What communities exist in a network?

↳ fraud ring ↳ terrorist activity

↳ friends / colleagues

2) Anomaly detection.

DIFFICULTIES WITH NETWORK ANALYSIS

- Storage
- Inference
- Visualization

DIFFICULTIES WITH NETWORK ANALYSIS

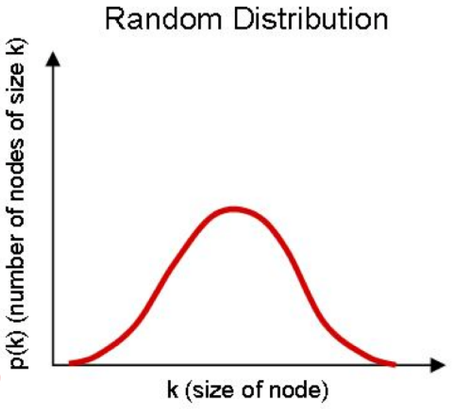
Graph QL → network database

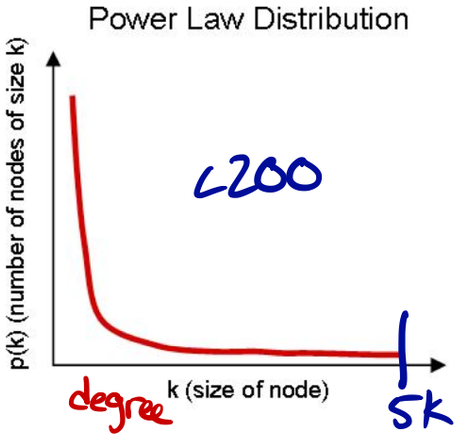
Networks

- Storage

↳ adjacency matrix

↳ 2.4 billion monthly active users





2,400,000,000 × 2,400,000,000

↳ incredibly sparse! (lots of zeroes)

DIFFICULTIES WITH NETWORK ANALYSIS

- Inference
  - ↳ by necessity, we need to sample
    - ↳ sample nodes (100,000 individuals)
    - ↳ sample edges (100k friendships)
    - ↳ snowball forest fire sampling

DIFFICULTIES WITH NETWORK ANALYSIS

- Visualization