[2024 Winter HGU Bio+AI Workshop]

AI-BASED ANALYSIS OF SOCIAL AND BIOLOGICAL NETWORKS

Lecturer: Hankyu Jang

Date: 2024/02/07 – 2024/02/08

About me

PC Member & Reviewer

AAAI
Association for the Advancement of Artificial Intelligence

epiDAMIK @ KDD 2022

SOCIAL NETWORK ANALYSIS AND MINING

Education

2009-2016
BS in Computer Science and Management

HANDONG GLOBAL UNIVERSITY

2016-2018
MS in Data Science

INDIANA UNIVERSITY
BLOOMINGTON

2018-2023
PhD in Computer Science

THE UNIVERSITY OF IOWA

Industry Experience

2021
Machine Learning and Data Science Intern

AMFAM

Data Validation
Graph Neural Networks

2023
Machine Learning Intern

PIVOT BIO

Explainable AI

2022
Applied Scientist Intern

amazon

Fraud Community Detection
Graph Neural Networks

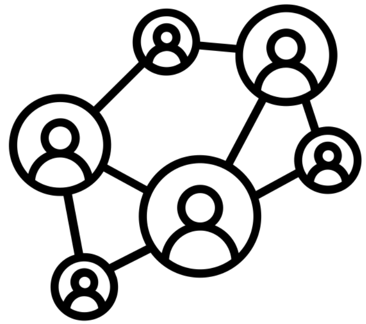
2023
Applied Scientist

amazon

Fraud Detection

Agenda

- Part1: Network science basics by examples
  - Netflix: movie recommendation
  - Facebook: friend recommendation, viral marketing
  - Google: web search
  - Network biology & network medicine
- Part2: Applying machine learning to graphs
  - Node classification
  - Link prediction
  - Network embedding



Part1: Network Science Basics by Examples

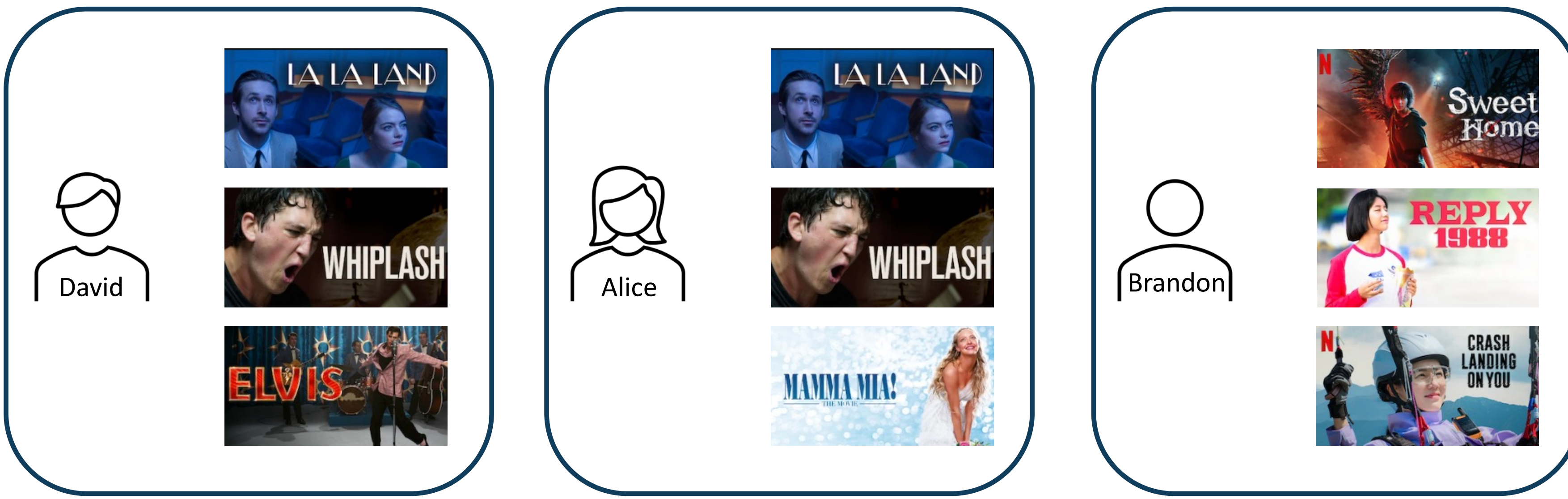
ML internship interview with Netflix

Interview question: write a recommendation algorithm

- that finds ***similar users*** with you
- and recommends TV content that they watched

Which user Alice | Brandon is similar with David?

Then, which TV content would you recommend to David?



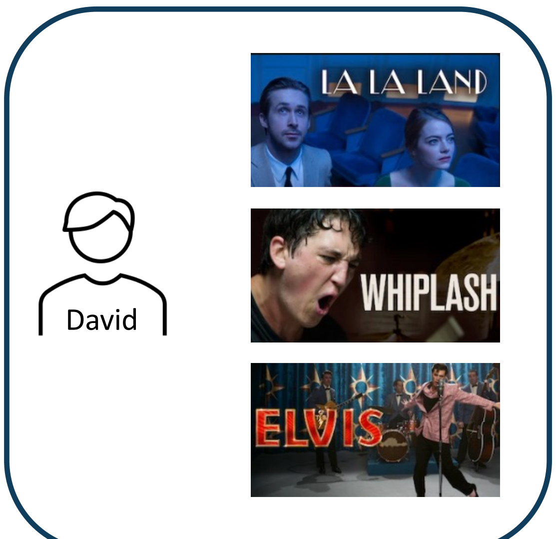
ML internship interview with Netflix

Two users are similar if the overlapping number of TV content is large

- M_{David} = {LaLaLand, Whiplash, Elvis}
- M_{Alice} = {LaLaLand, Whiplash, MaMaMia}
- $M_{Brandon}$ = {SweetHome, Reply1988, CrashLandingOnYou}
- $M_{David} \cap M_{Alice} > M_{David} \cap M_{Brandon}$

Any issue with this algorithm?

ML internship interview with Netflix



Need to re-define ‘similar users’

ML internship interview with Netflix

Two users are similar if the overlapping number of TV content is large, yet the total TV contents they watched is small

- $S(\text{David, Alice}) = M_{David} \cap M_{Alice} / M_{David} \cup M_{Alice} = 2 / 4 = 0.5$
- $S(\text{David, Cavin}) = M_{David} \cap M_{Cavin} / M_{David} \cup M_{Cavin} = 3 / 12 = 0.25$
- S(David, Alice) > S(David, Cavin) , so recommend the TV content that Alice watched, MaMaMia, to David!

Connection to Network Science

This ML Internship Interview question with Netflix is a ***Network Science*** problem:

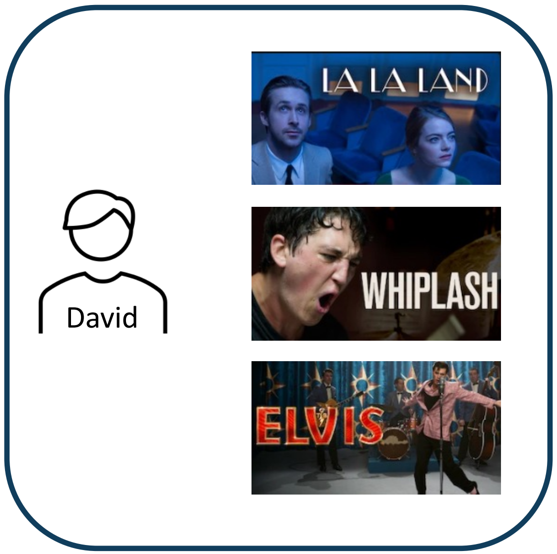
- **Problem**: Given a graph with user nodes and TV content nodes and edges (e.g., user watching a TV content), design an algorithm that recommends a content to a user
- **Solution**: Find similar user nodes via Jaccard similarity coefficient, and recommend TV content nodes connected with the similar user

What is a graph? Nodes? Edges? Jaccard similarity coefficient?

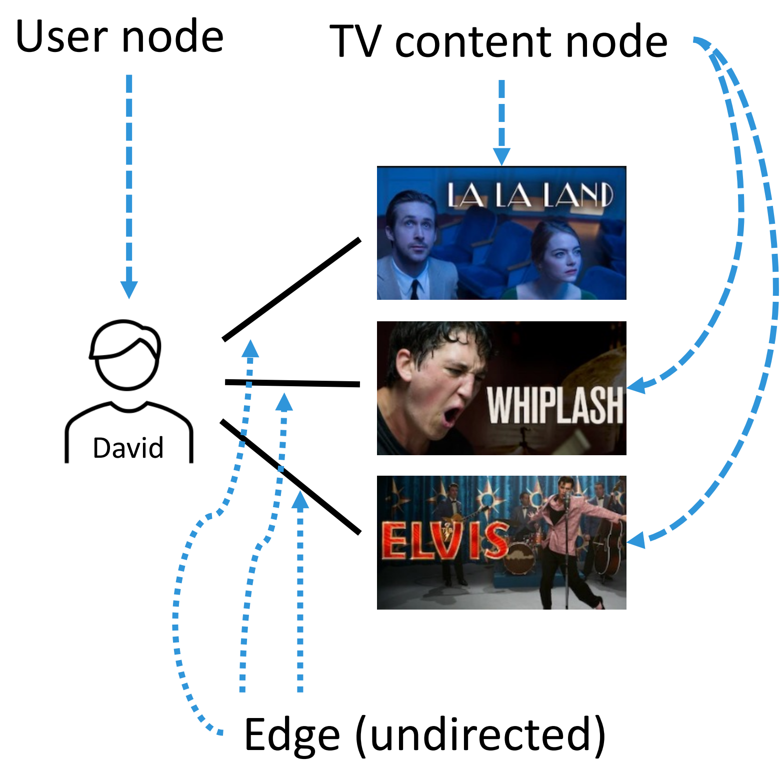
Graph

A **graph** (network) is made up of **nodes** (vertices) and **edges** (links)

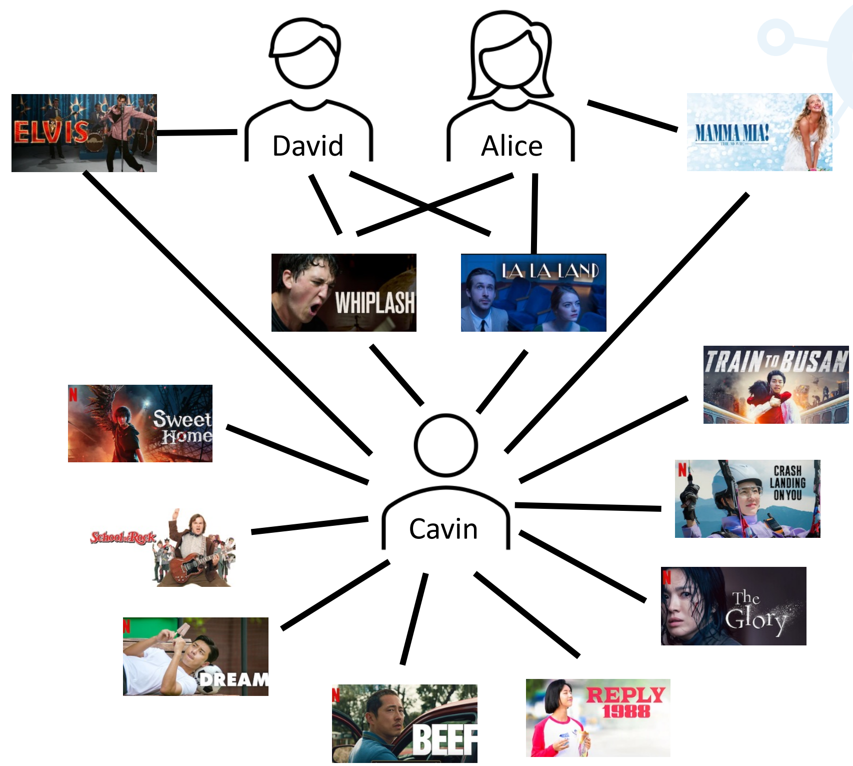
- **Simple graph**: one type of node. Undirected edge
- **Bipartite graph**: 2 types of nodes. Edges connect nodes with different types



We can represent this information as a graph (on the right)

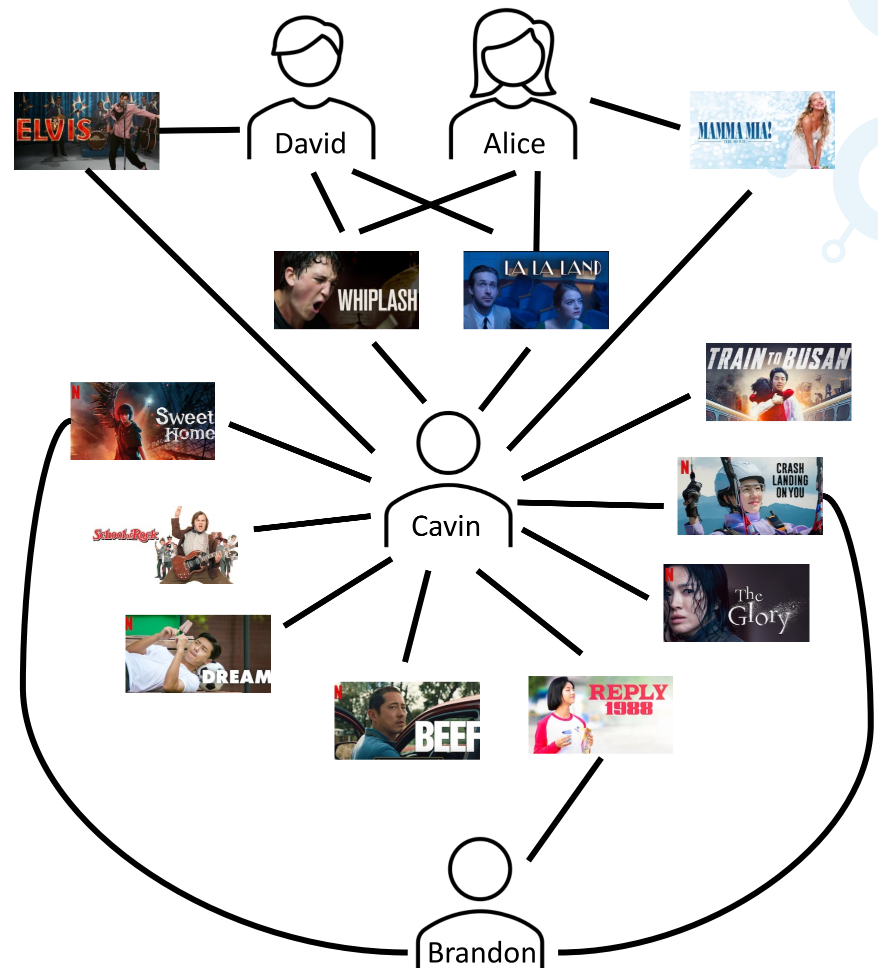


Graph



We can represent this information as a graph (on the right)

Graph



Visualizing a large graph is hard!

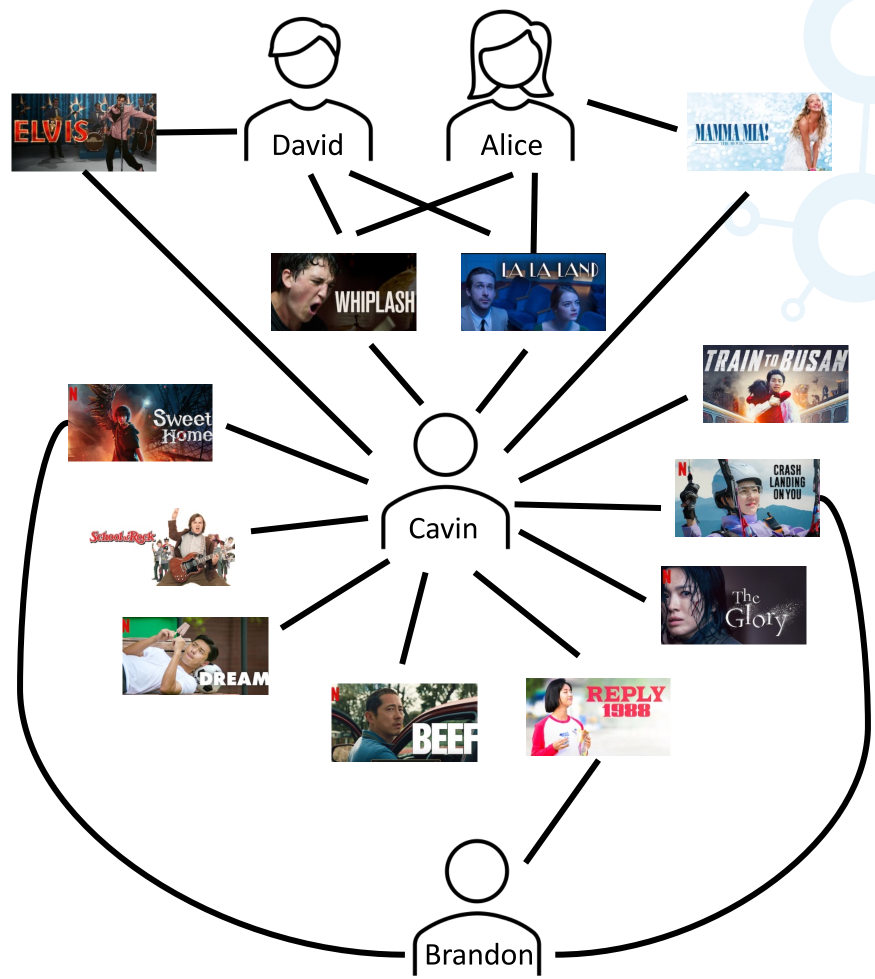
Graph terminology

A graph $G = (V, E)$

- V : a set of nodes
- E : a set of edges

Two nodes are ***neighbors*** if they are connected with an edge

- $\Gamma(u)$: a set of neighbors of node u
- $\deg(u)$: degree of u, that is, $|\Gamma(u)|$



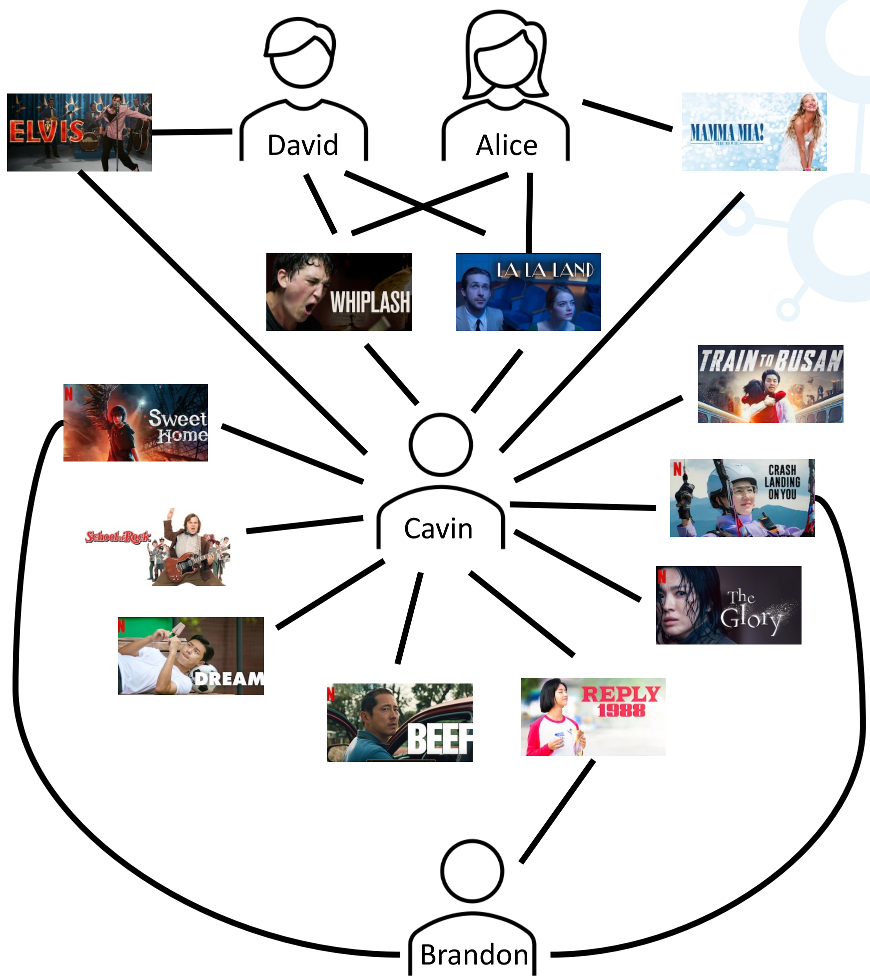
Graph terminology

Two nodes are ***neighbors*** if they are connected with an edge

- $\Gamma(u)$: a set of neighbors of node u
- $\deg(u)$: degree of u, that is, $|\Gamma(u)|$

Question: What is $\Gamma(David)$?

Question: What is $deg(David)$?



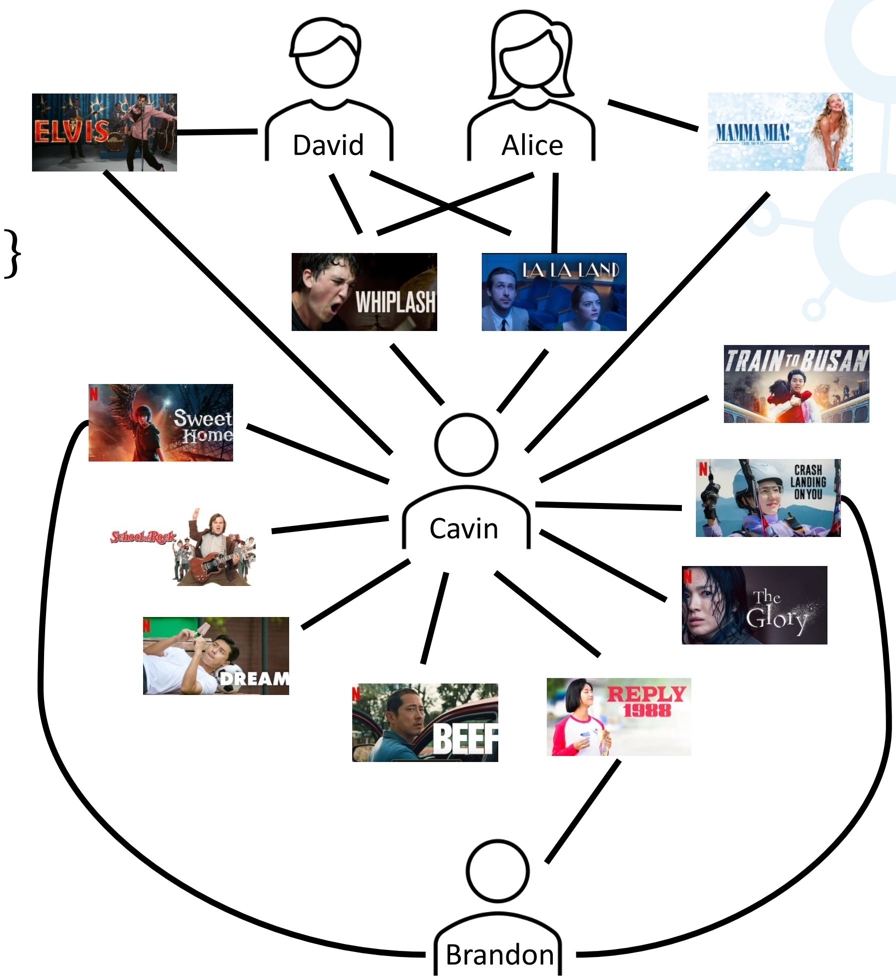
Graph terminology

$\Gamma(David)$
$= \{\, Elvis, Whiplash, LaLaLand \,\}$

$\Gamma(Alice)$
$= \{\, MaMaMia, Whiplash, LaLaLand \,\}$

Common neighbors of node u and v are the set of nodes that are neighbors of both u and v

Question: Common neighbors of David and Alice?



Back to our solution to Netflix interview question

Solution 1: Define similar users in terms of common neighbors

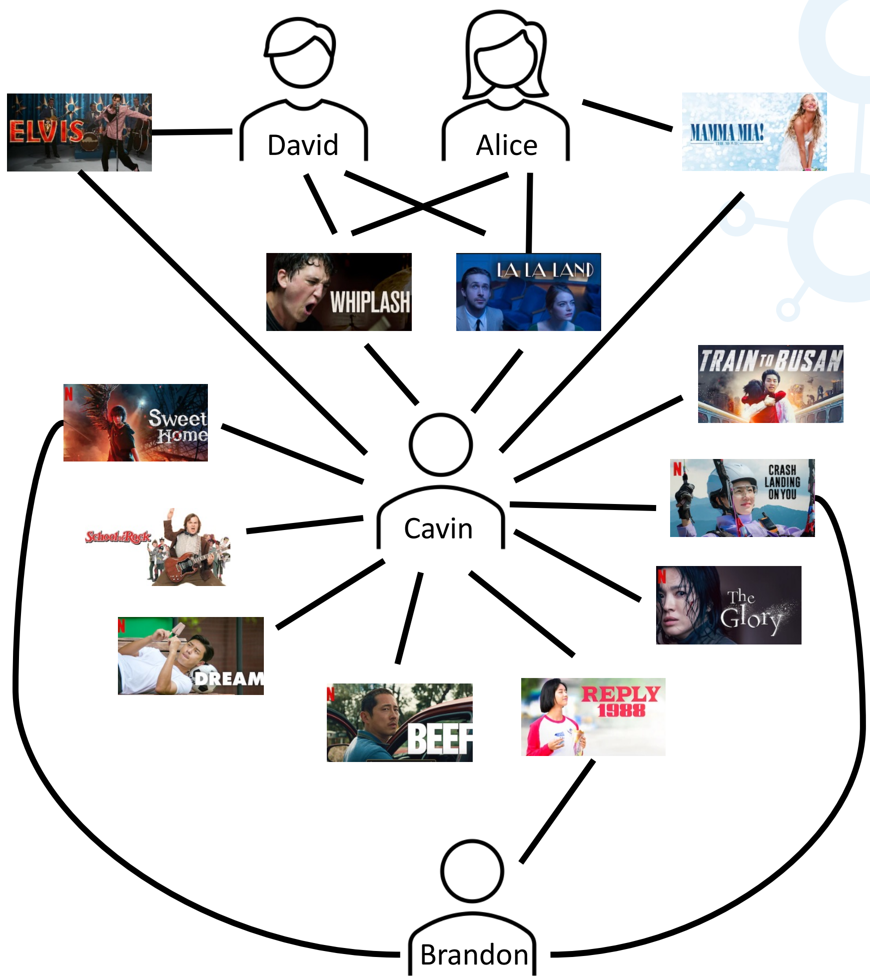
- $S_{CN}(A,B) = |\Gamma(A) \cap \Gamma(B)|$

Solution 2: Define similar users in *Jaccard similarity coefficient*

- $S_J(A,B) = \dfrac{|\Gamma(A) \cap \Gamma(B)|}{|\Gamma(A) \cup \Gamma(B)|}$

Then recommend TV contents that the similar user watched

Different definition of similarity leads to different TV content recommendation!



D. Liben-Nowell, J. Kleinberg. The Link Prediction Problem for Social Networks (2004). http://www.cs.cornell.edu/home/kleinber/link-pred.pdf

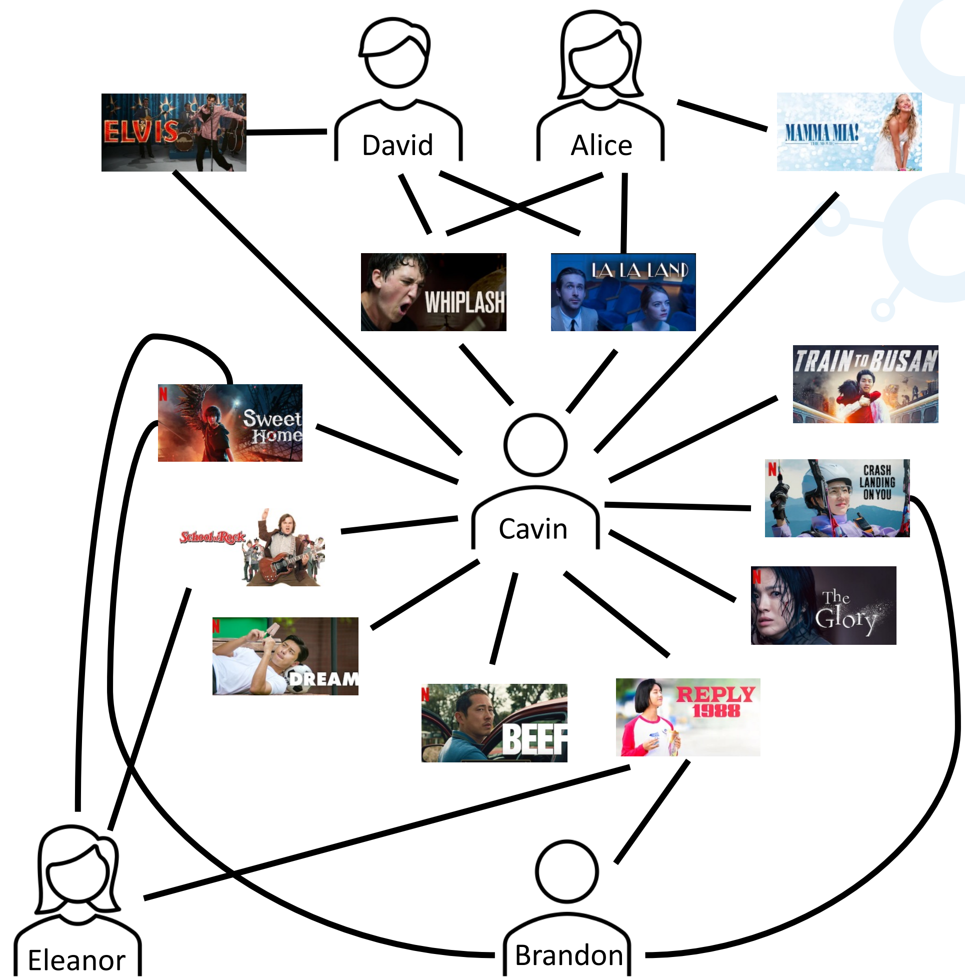
Recommend a TV content to Eleanor

Define similar users in ***Jaccard similarity coefficient***

- $S_J(A,B) = \dfrac{|\,\Gamma(A) \cap \Gamma(B)\,|}{|\,\Gamma(A) \cup \Gamma(B)\,|}$

David and Alice has no common neighbors with Eleanor

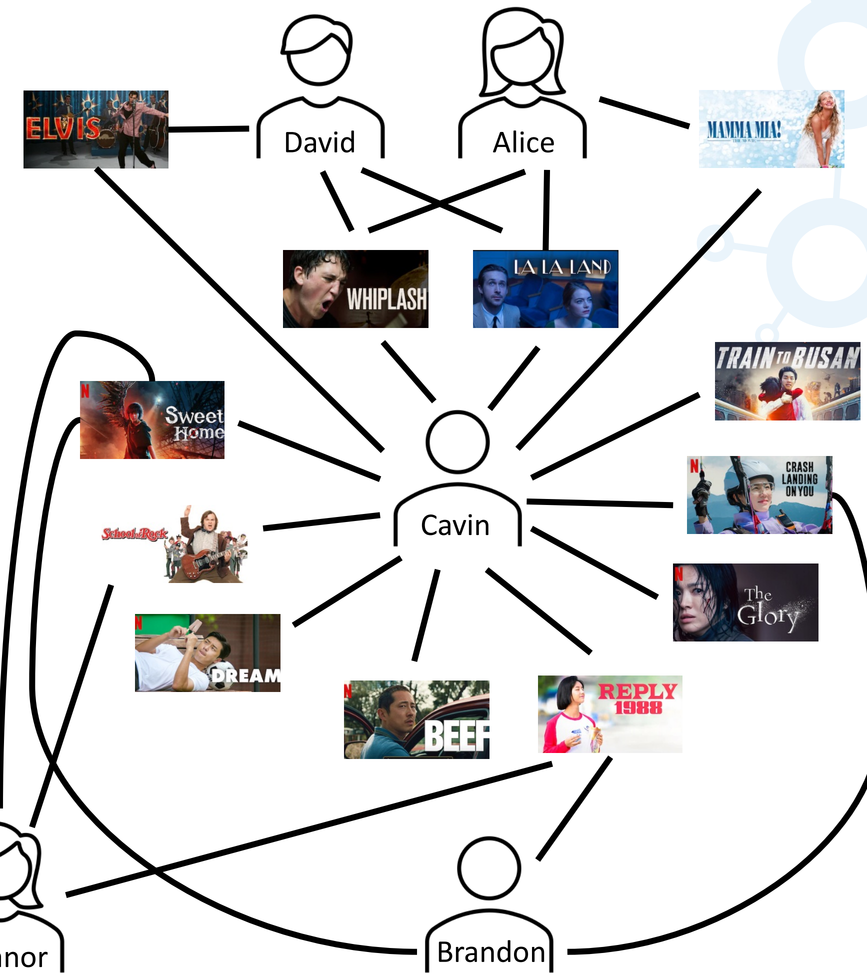
- $S_J(Elanor, Cavin)$= ?
- $S_J(Elanor, Brandon)$= ?



Similarity matrix

A similarity matrix composes of similarity values computed for all possible node pairs

$$
\begin{array}{cc}
 & \begin{array}{ccccc} \text{Alice} & \text{Brandon} & \text{Cavin} & \text{David} & \text{Eleanor} \end{array} \\
\begin{array}{c} \text{Alice} \\ \text{Brandon} \\ \text{Cavin} \\ \text{David} \\ \text{Eleanor} \end{array} &
\begin{bmatrix}
 & 0 & 0.25 & 0.5 & 0 \\
 & & 0.25 & 0 & 0.5 \\
 & & & 0.25 & 0.25 \\
 & & & & 0 \\
 & & & &
\end{bmatrix}
\end{array}
$$

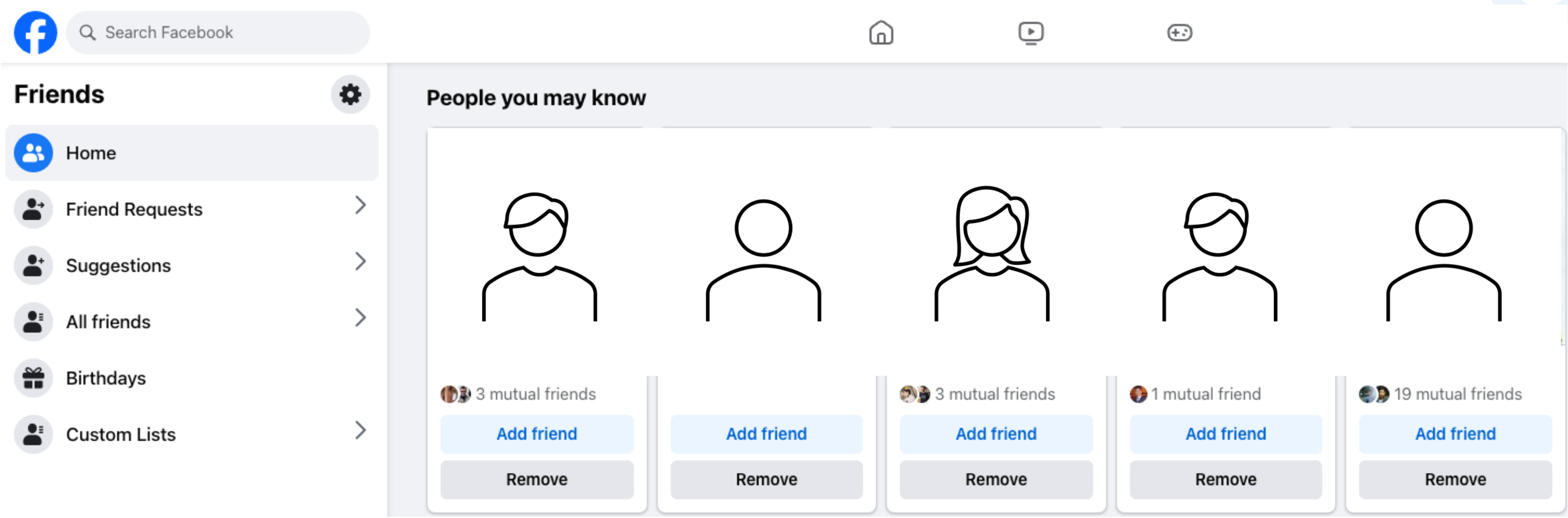


The matrix gets really large, if we have a large number of users

What about friend recommendation in Facebook?

How does Facebook recommend these people to you?

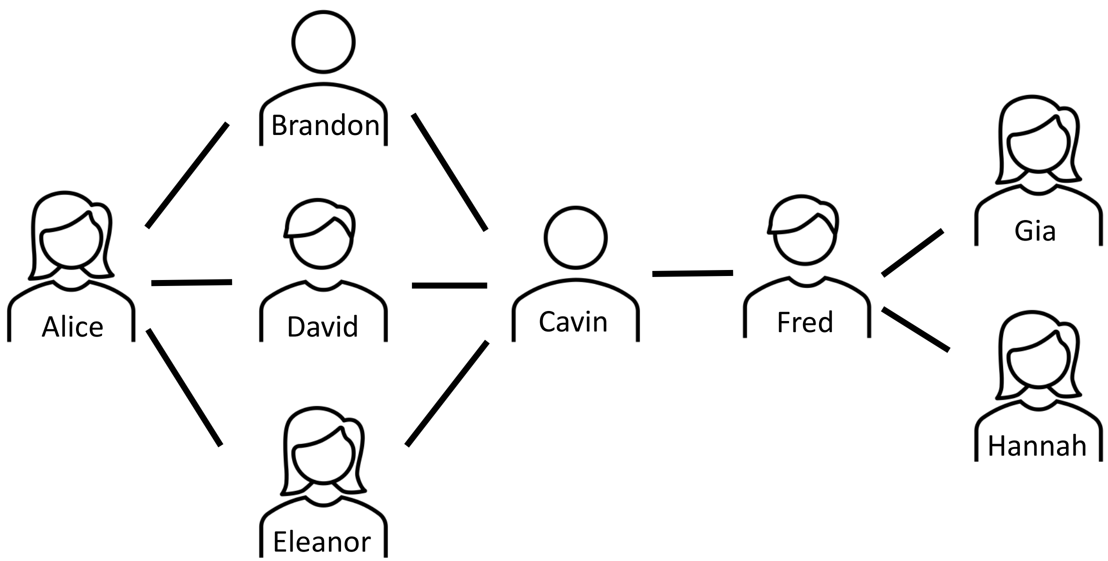
The core technology is again, ***Network Science***



Social Network

Node: Facebook user

Edge: Friendship



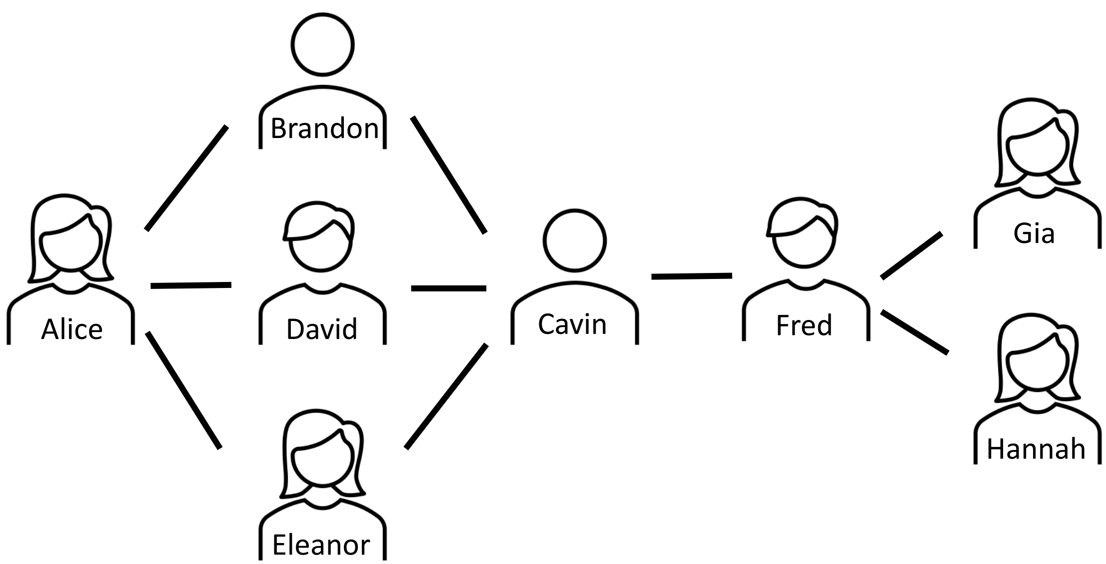
People you may know…

Question: Write a recommendation algorithm

- that finds *similar users* with you
- and recommends them

Based on what we learned so far, how would you approach this problem?

E.g., who would you recommend to Alice?

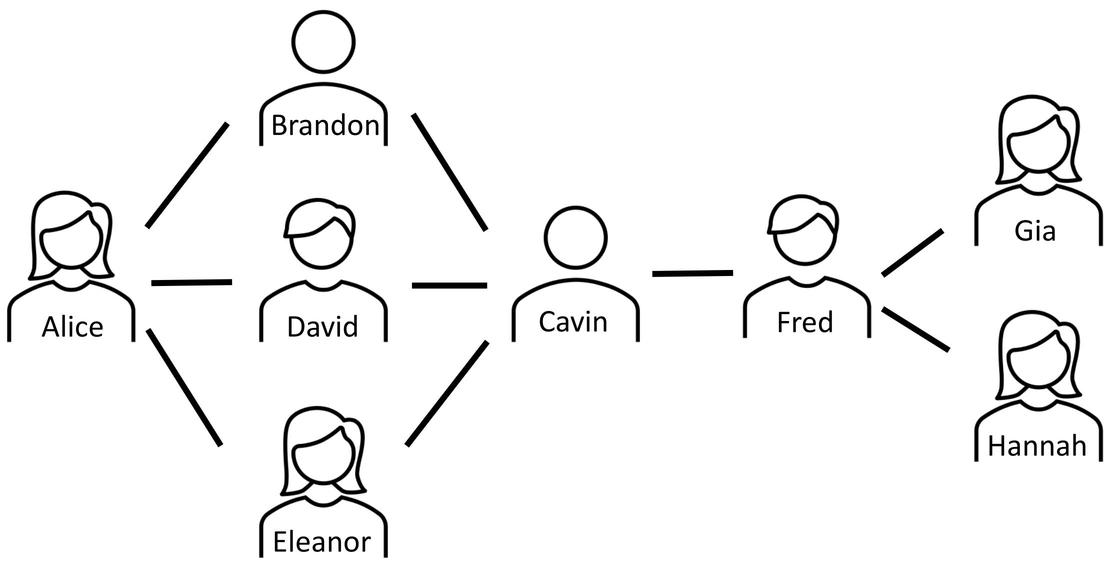


Neighborhood based recommendation

For Alice, compute similarity score with Cavin, Fred, Gia, and Hannah

- S_{CN}(Alice, Cavin) = 3
- Recommend Cavin to Alice

What would happen if Facebook keep recommending friends this way?



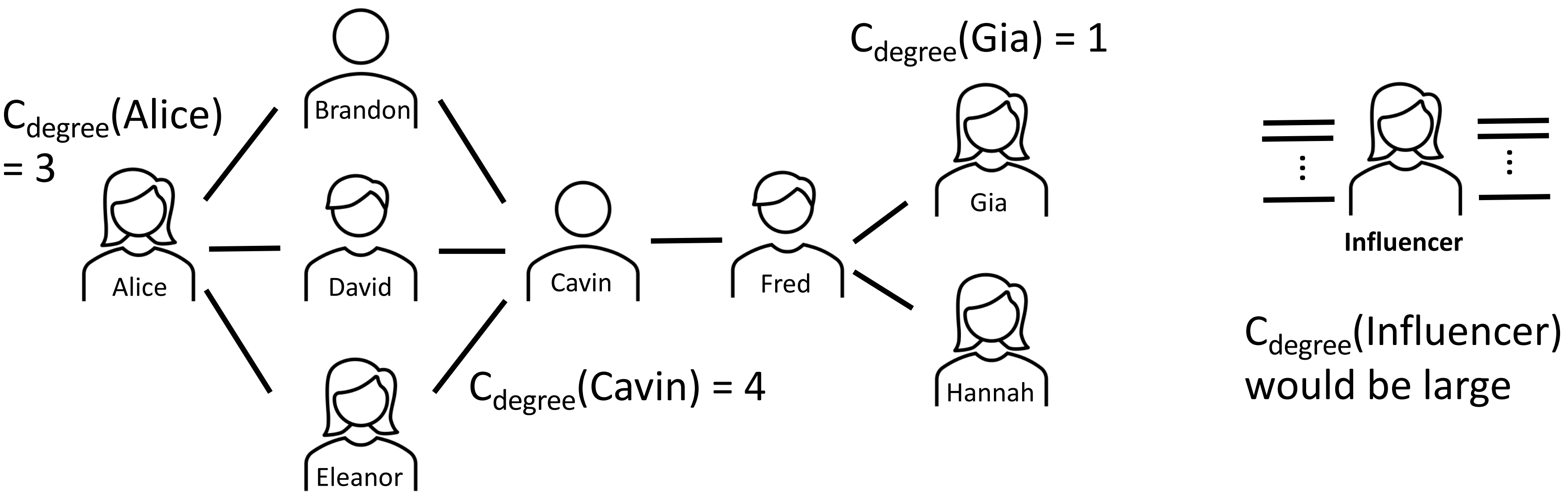
Recommending influencers

Users may want to be connected with famous figures, like influencers

How to find these influential nodes?

Network centrality is a problem of finding "central" nodes in a graph

- Degree centrality (C_{degree}) of a node: degree of the node



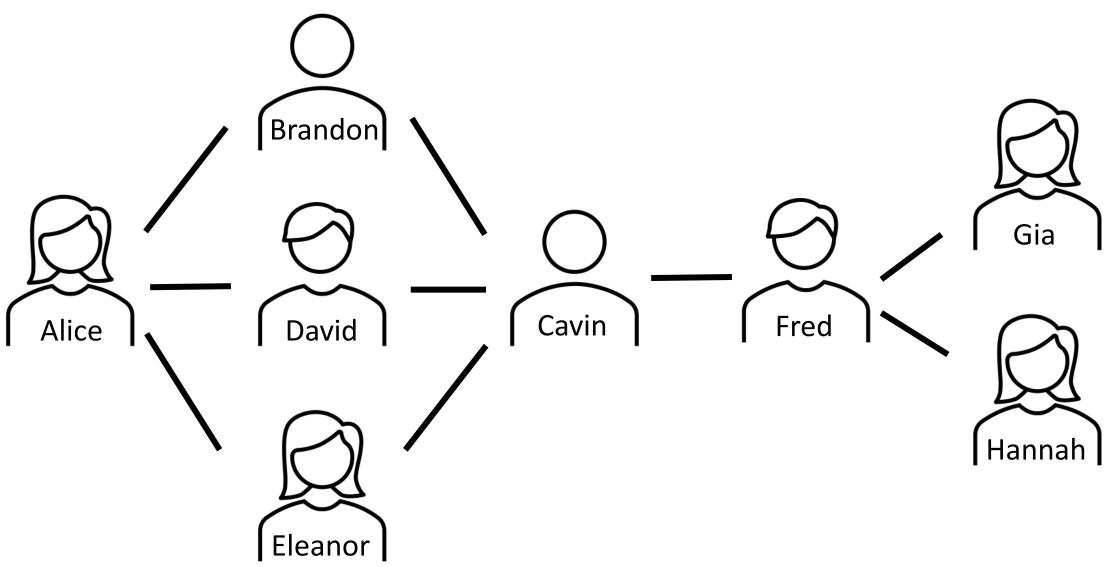
Whom to choose for viral marketing?

Question: You want to promote a product in this group of people. You have budget to let 1 person try your product. Who would you choose?

You need to find central node, such that word will spread fast in this community

- Assumption: word spreads only via edges

Cavin seems to be close to everyone, so maybe choose Cavin!



More graph terminologies

Path: sequence of nodes, connected via edges. No repetition allowed

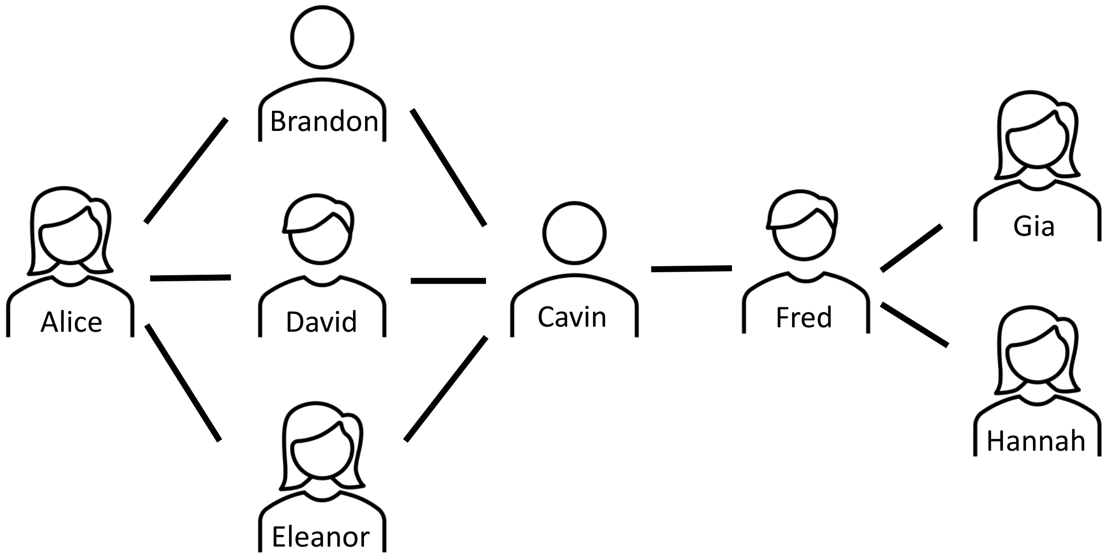
There are 3 paths from Alice to David

- Alice-David | Alice-Brandon-Cavin-David | Alice-Eleanor-Cavin-David

Shortest Path from Alice to David is Alice-David

Cavin is in ***2-hop neighborhood*** from Alice

Fred is in ***3-hop neighborhood*** from Alice, because ***shortest path distance*** is 3



Shortest-path based centrality measures

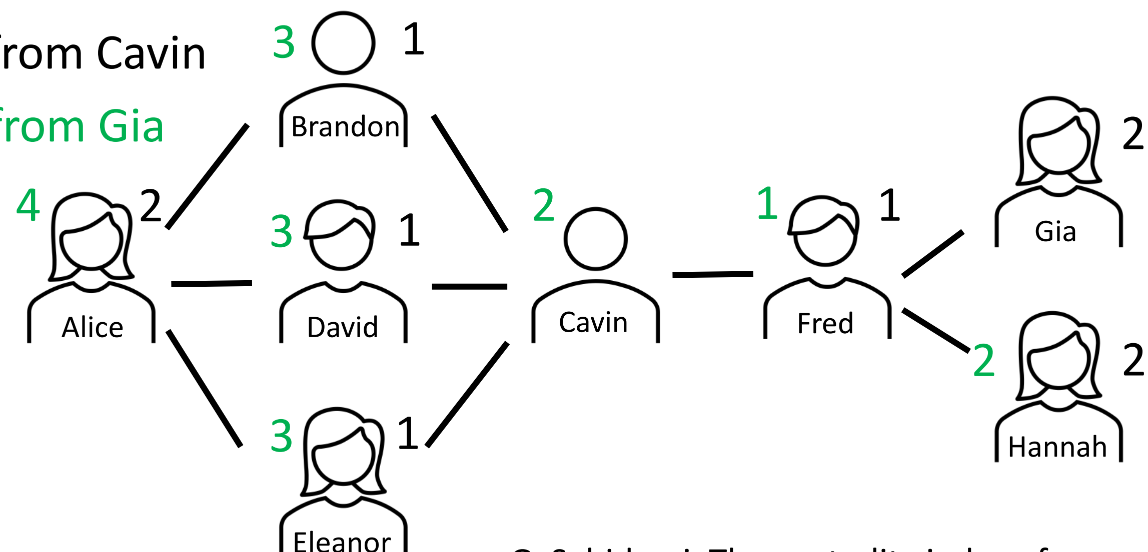
Closeness centrality ($C_{closeness}$): The more central a node is, the closer it is to all other nodes

- $C_{closeness}$ of a node: (total nodes - 1) / (sum of shortest path distances to all)
- $C_{closeness}$(Cavin) = 7 / (2 + 1 + 1 + 1 + 1 + 2 + 2) = 7/10 = 0.7
- $C_{closeness}$(Gia) = 7 / (4 + 3 + 3 + 3 + 2 + 1 + 2) = 7/18 = 0.39

So, compute $C_{closeness}$ for all the nodes, and select the node with largest centrality

Hop distance from Cavin

Hop distance from Gia



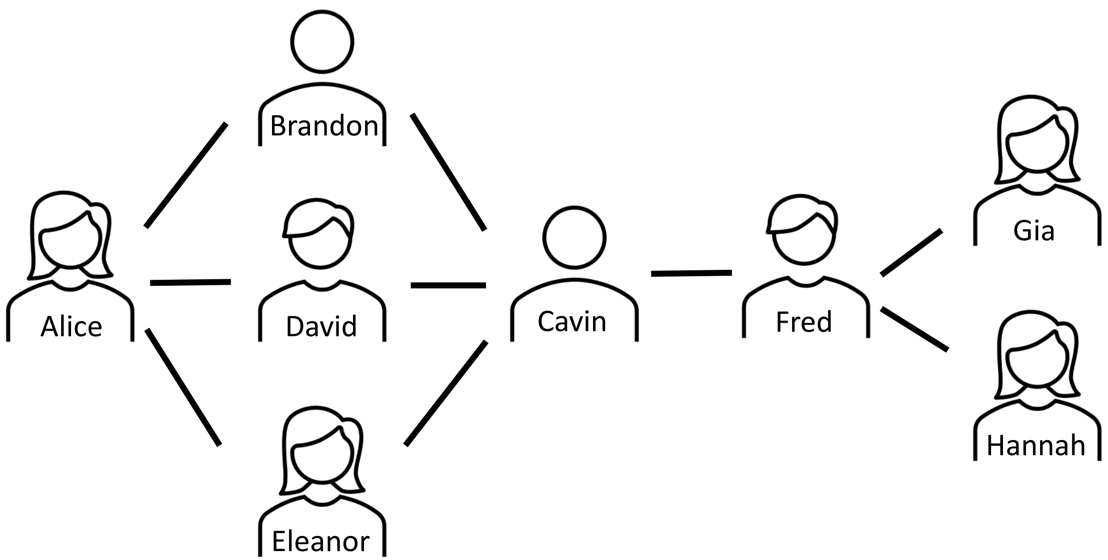
G. Sabidussi, The centrality index of a graph, Psychometrika, 1966

Shortest-path based centrality measures

Betweenness Centrality ($C_{betweenness}$): A node is central if it appears the most, in shortest paths for all pairs of nodes

- $C_{betweenness}(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}}$
- σ_{st}: total number of shortest paths from node s to node t
- $\sigma_{st}(v)$: total number of those, that pass through v

Cavin has the largest $C_{betweenness}$



L. C. Freeman, A set of measures of centrality based on betweenness, Sociometry, 1977

Google search – rank pages

The core business of Google is in web search

The search engine ranks web pages, and show the most relevant ones on the top

How is this being done?

This is a problem of finding ***central nodes*** in a graph of web pages

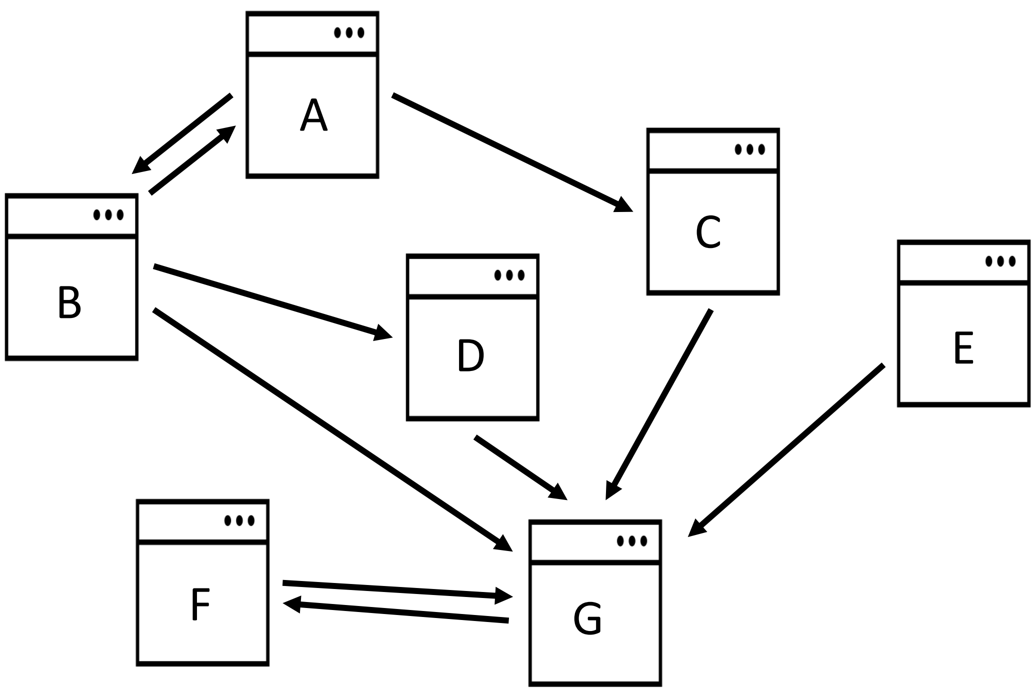
Graph of web pages

Node: Web page

in-edge: Incoming hyperlink from other web pages

out-edge: Outgoing hyperlink to other web pages

Question: which webpage looks ***central***?



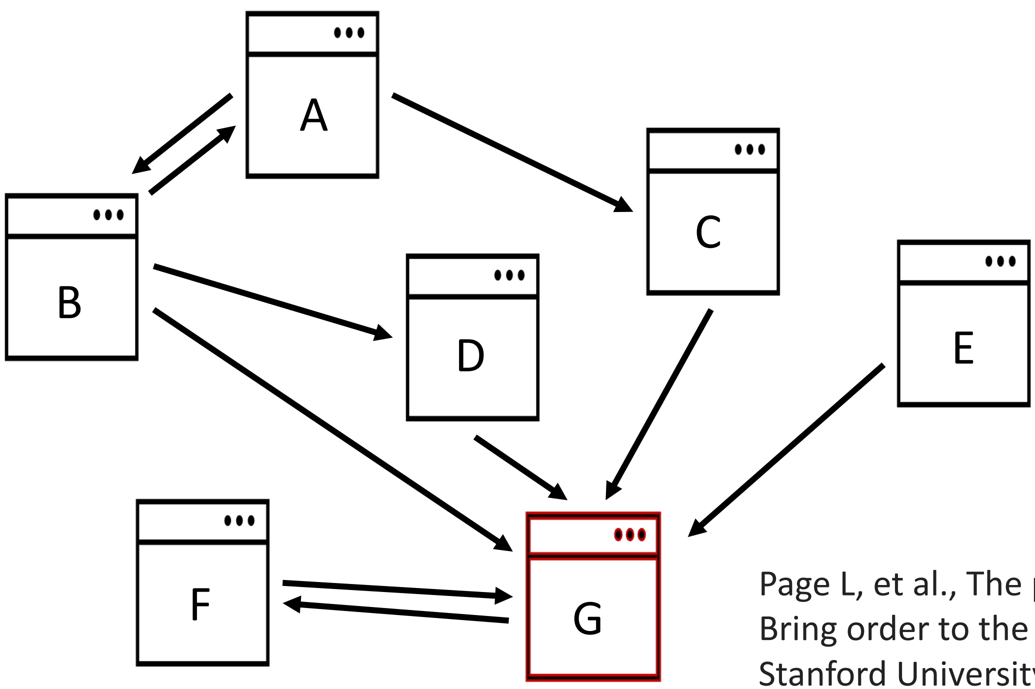
PageRank centrality – this is how Google started

PageRank is developed in 1996 at Stanford University as a research project

Assumption: More important websites are likely to receive more links

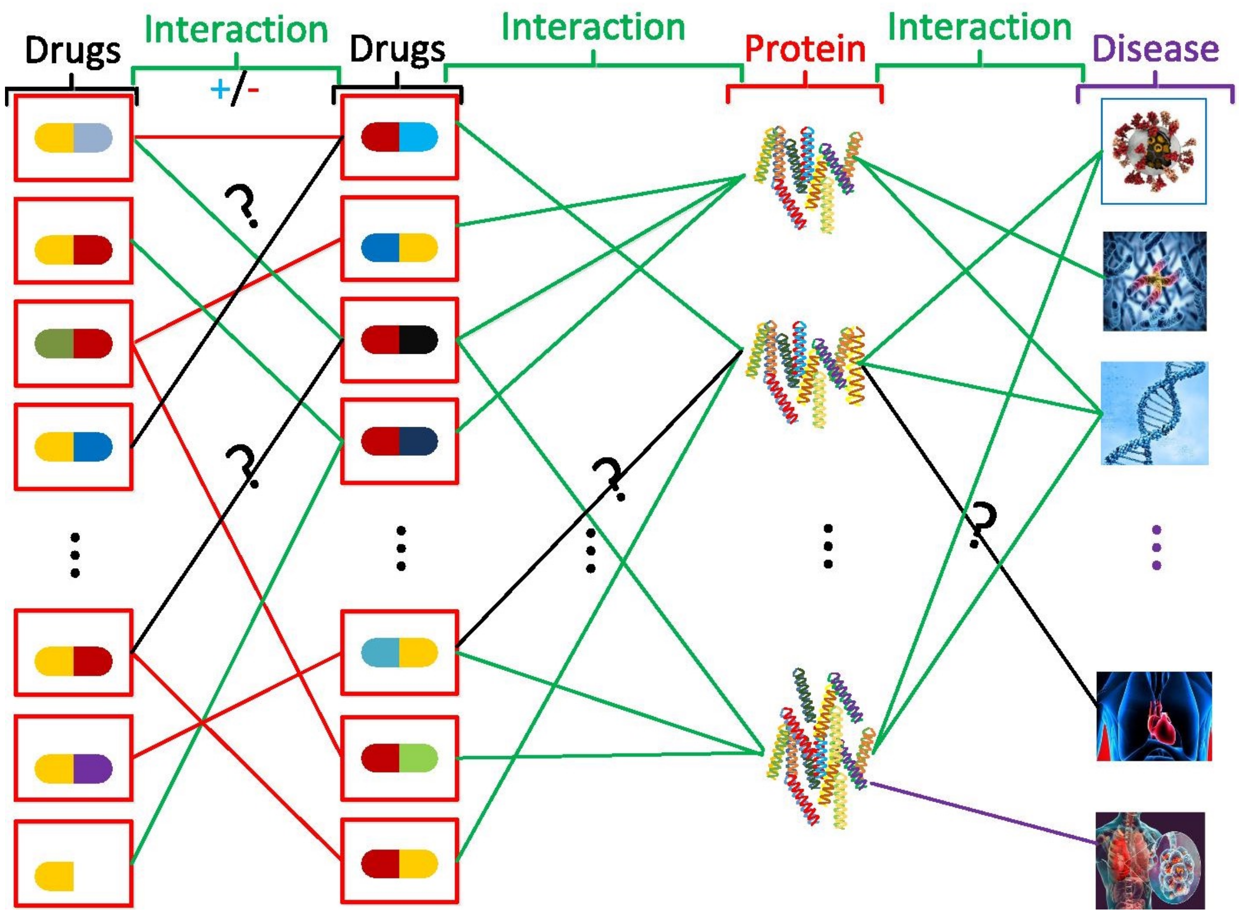
How the centrality is computed in PageRank:

- Let there be random web-surfers. They randomly click links over and over
- Websites that are visited more, have higher PageRank centrality than others



Page L, et al., The pagerank citation ranking: Bring order to the web. Technical report, Stanford University; 1998

Network biology & network medicine



Abbas K, Abbasi A, Dong S, Niu L, Yu L, Chen B, Cai SM, Hasan Q. Application of network link prediction in drug discovery. BMC bioinformatics. 2021

Network biology & network medicine

Drug - target protein prediction

- Predict which drug will affect which unknown proteins

Required data

- Drugs-Protein bipartite network

Abbas K, Abbasi A, Dong S, Niu L, Yu L, Chen B, Cai SM, Hasan Q. Application of network link prediction in drug discovery. BMC bioinformatics. 2021

Network biology & network medicine

Drug - disease prediction

- Find drugs with similar ***chemical structure.*** Similar drugs can be used to treat same disease

Required data

- Chemical structure network
- Drug–Disease bipartite network

Abbas K, Abbasi A, Dong S, Niu L, Yu L, Chen B, Cai SM, Hasan Q. Application of network link prediction in drug discovery. BMC bioinformatics. 2021

Network biology & network medicine

Drug-Drug reaction prediction

- From known combination of drugs that cause adverse side effects, predict reaction of unknown combination of drugs

Required data

- Combination of drugs that cause adverse side effects (e.g., headache, vomit)
- This is graph of drugs, with side effect information on edges

Abbas K, Abbasi A, Dong S, Niu L, Yu L, Chen B, Cai SM, Hasan Q. Application of network link prediction in drug discovery. BMC bioinformatics. 2021

Network biology & network medicine

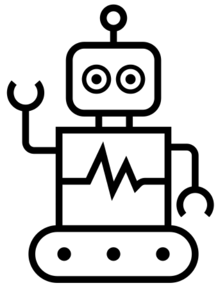
Disease-Gene association prediction

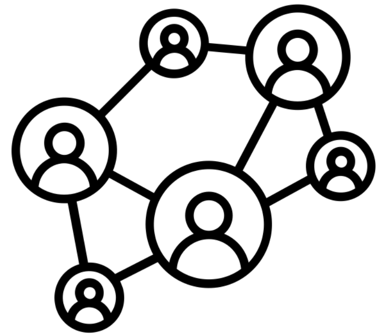
- Use known disease-gene association to find unknown associations
- This is known as network approach for ***genomic data analysis***

Required data

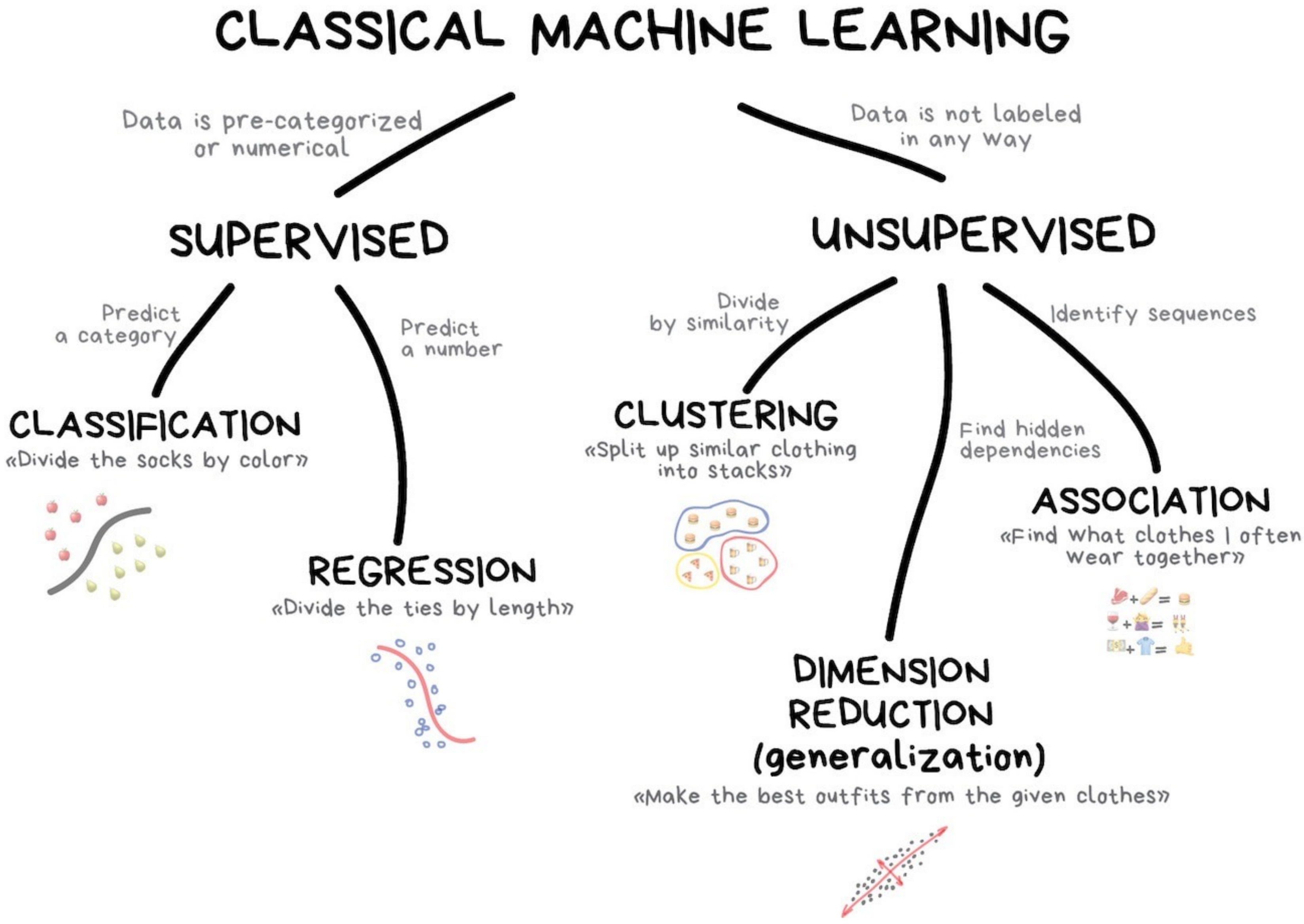
- Disease-Gene association bipartite network

Abbas K, Abbasi A, Dong S, Niu L, Yu L, Chen B, Cai SM, Hasan Q. Application of network link prediction in drug discovery. BMC bioinformatics. 2021





Part2: Applying Machine Learning to Graphs



https://vas3k.com/blog/machine_learning/?fbclid=IwAR0NjjOJlZt

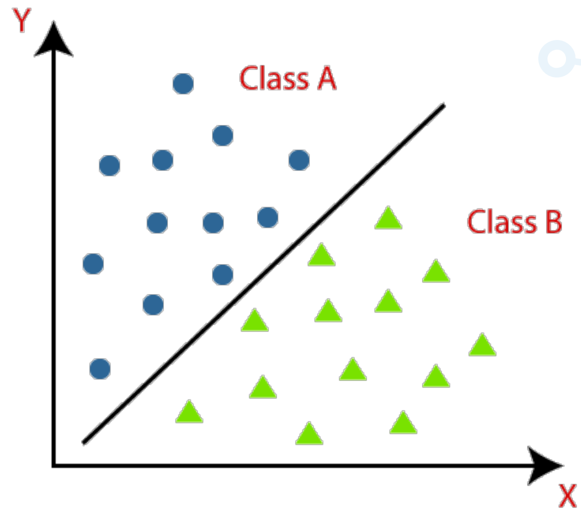
Classification

Supervised learning technique to identify the category of new observations

- Classify an email by looking at the content within the email

zxcv@xxx.com

https://penplusbytes.org/strategies-for-dealing-with-e-mail-spam/



https://www.javatpoint.com/classification-algorithm-in-machine-learning

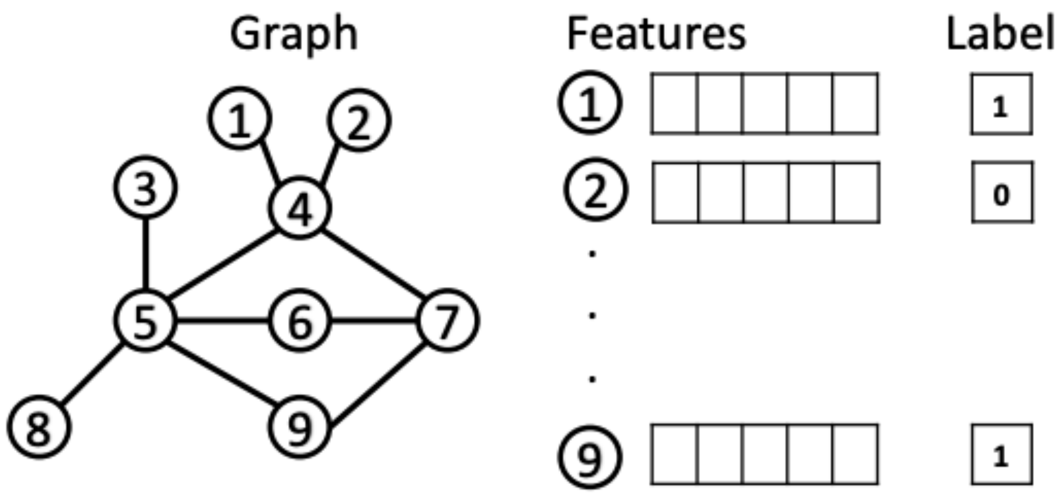
What if, we have some additional information? How to use this information?

- Email from asdf@xxx.com was previously flagged as ‘Spam’
- Contents sent by asdf@xxx.com and zxcv@xxx.com are similar

Node classification

When training a classification model, we use

- Features and label for each node (e.g., a common dataset) and
- The connectivity of the nodes (represented as a graph)

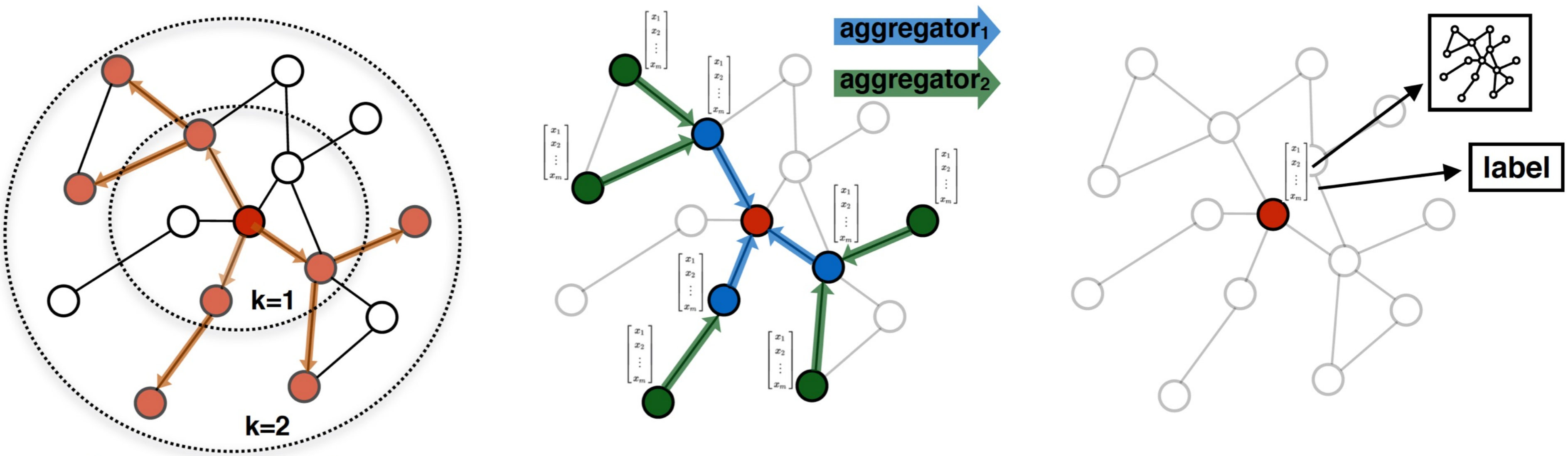


Graph neural networks (GNNs)

Idea comes from convolutional neural network (CNN) architecture

- Nearby pixels in an image are similar, so use nearby pixels when training

Nearby nodes are similar, so use nearby nodes' features when training



1. Sample neighborhood

2. Aggregate feature information from neighbors

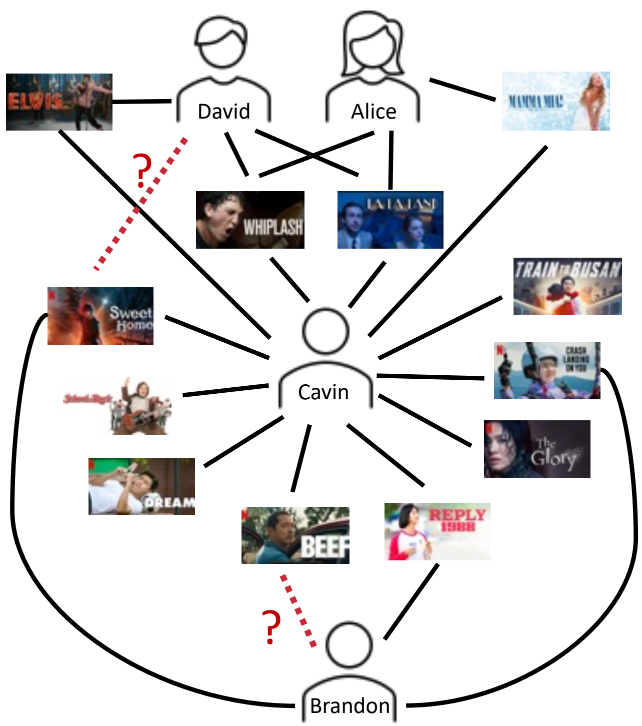
3. Predict graph context and label using aggregated information

https://snap.stanford.edu/graphsage/

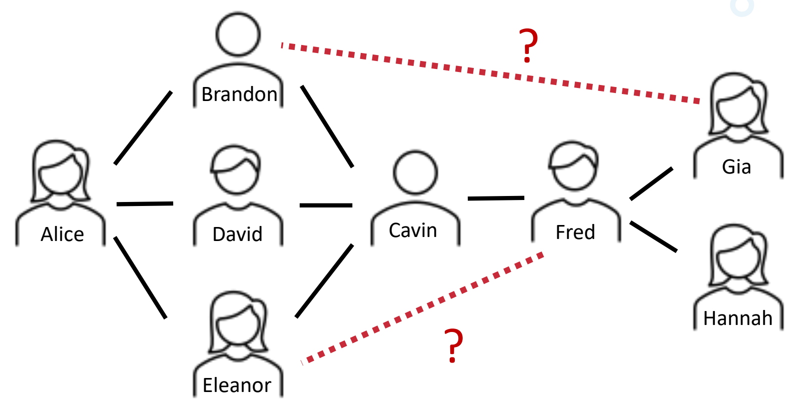
Hamilton W, Ying Z, Leskovec J. Inductive representation learning on large graphs. NIPS 2017

Link prediction

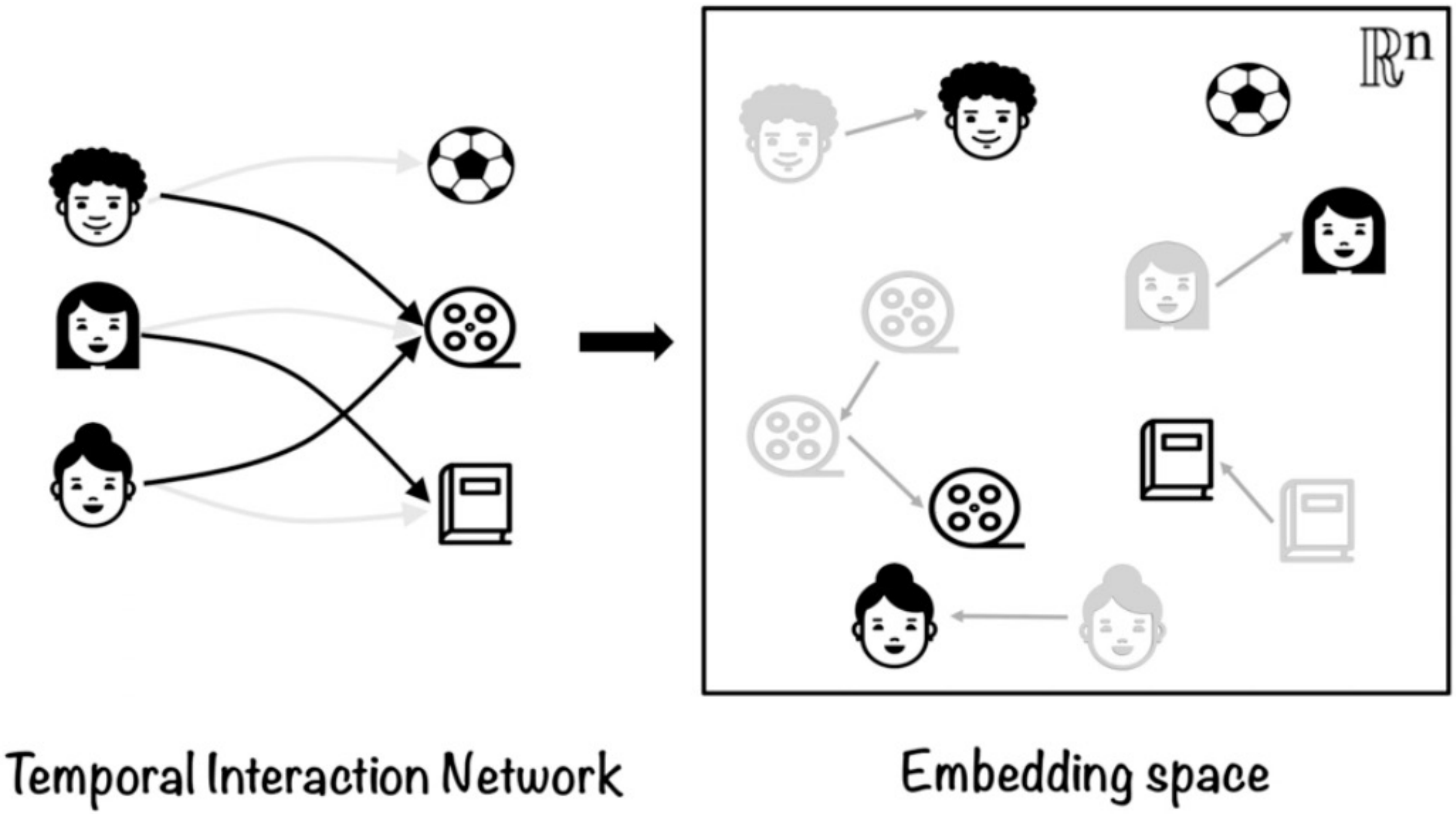
- Recommending items to users



- Recommending friends in SNS



JODIE: Dynamic link prediction method



https://snap.stanford.edu/jodie/#paper

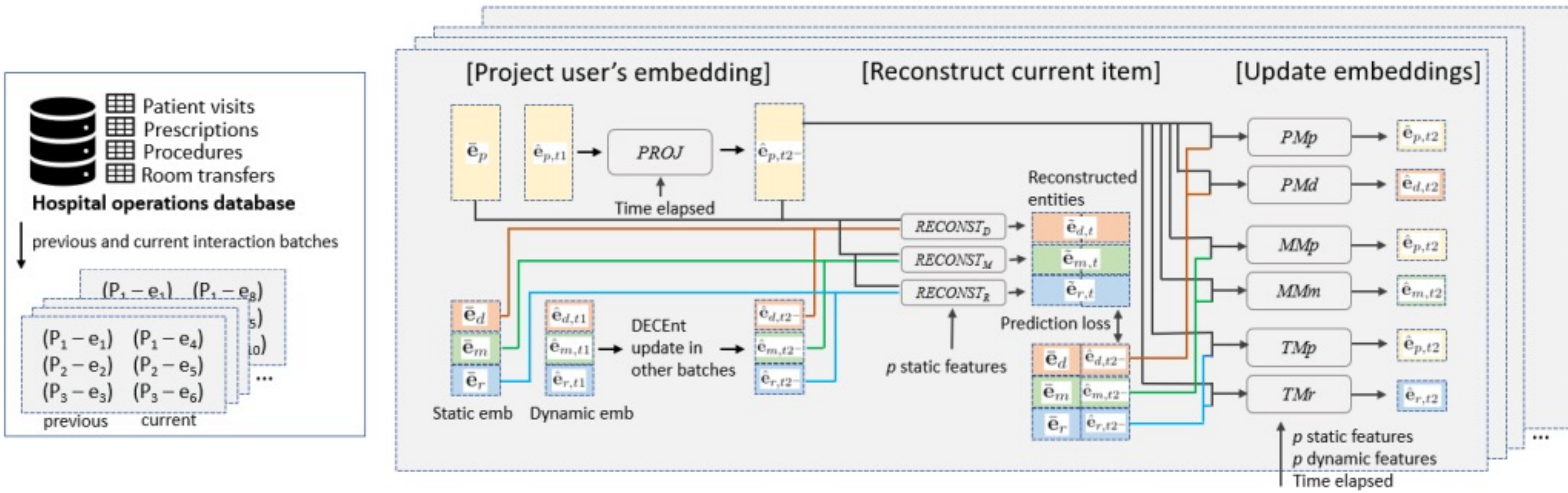
Kumar S, Zhang X, Leskovec J. Predicting dynamic embedding trajectory in temporal interaction networks. KDD 2019

Network embedding

Neural network models naturally learns ‘hidden representation’ of each input

- GNN based models for node classification
- Temporal graph network based models for link prediction

This hidden representation is powerful, to use as ‘features’ for other tasks



Patient embedding is learned using patient – healthcare entity interactions

Patient embeddings were predictive in many healthcare modeling tasks

Hankyu Jang, et al. Dynamic Healthcare Embeddings for Improving Patient Care. IEEE/ACM ASONAM 2022

Tutorial

- Open https://colab.research.google.com/
- Upload HGU_Bio_AI_workshop_Tutorial.ipynb