Data Mining and Analysis

Graph Mining: 1

CSCE 676 :: Fall 2019
Texas A&M University
Department of Computer Science & Engineering
Prof. James Caverlee

Resources

Networks, Crowds, and Markets. Chapter 2 and Chapter 3

MMDS Chapter 10.1, 10.2, 10.4 (ignore 10.4.4)

Louvain method (Wikipedia)

DMTT Chapter 17.4

Agenda

Today

Basics, Frequent itemsets —> graph mining, Finding Important Nodes

Wednesday

Social Networks, Community Detection

Friday

Community Detection

Later in the semester —> graph/node embeddings!

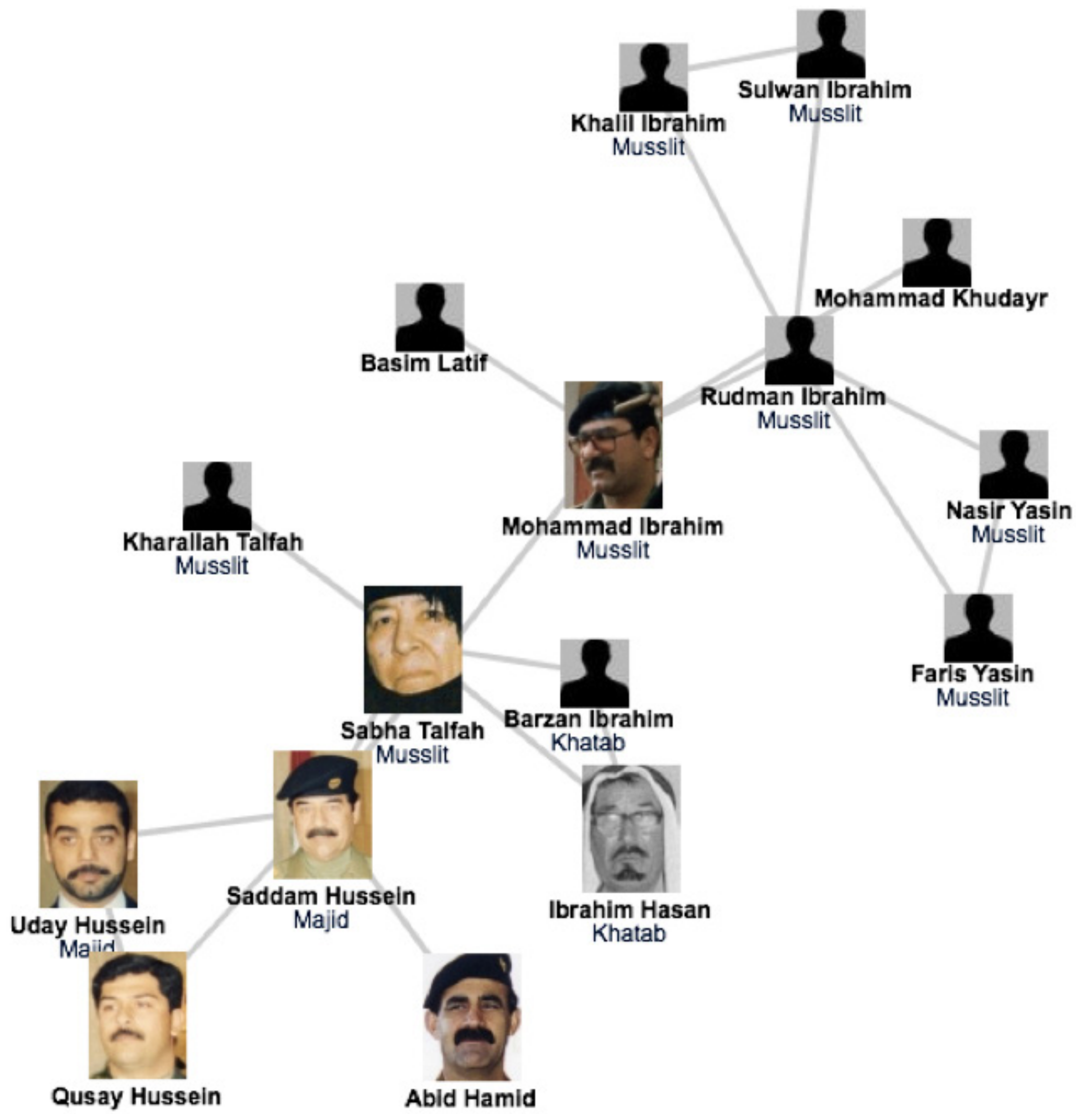


Image 1.2b

The network of Saddam Hussein.

The Social Network. A small region of the social network reconstructed by the US forces in the process of searching for Saddam Hussein. The map represents the relationship between individuals in Saddam's inner circle.

Basic concepts

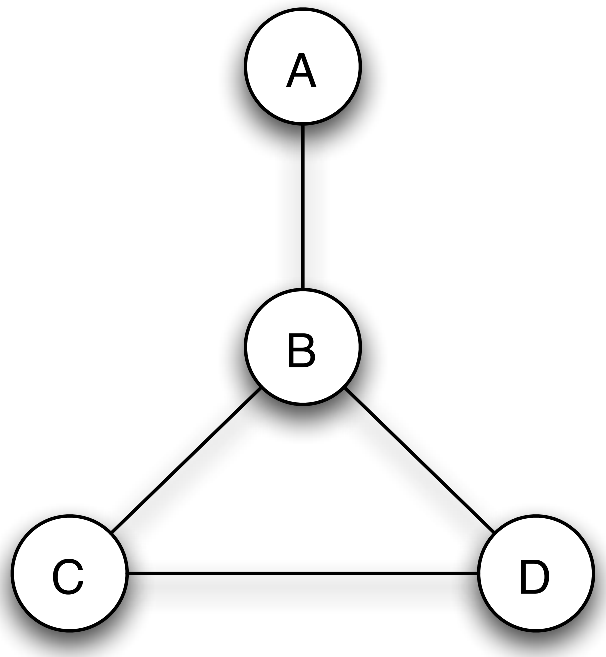
nodes

edges (directed or undirected)

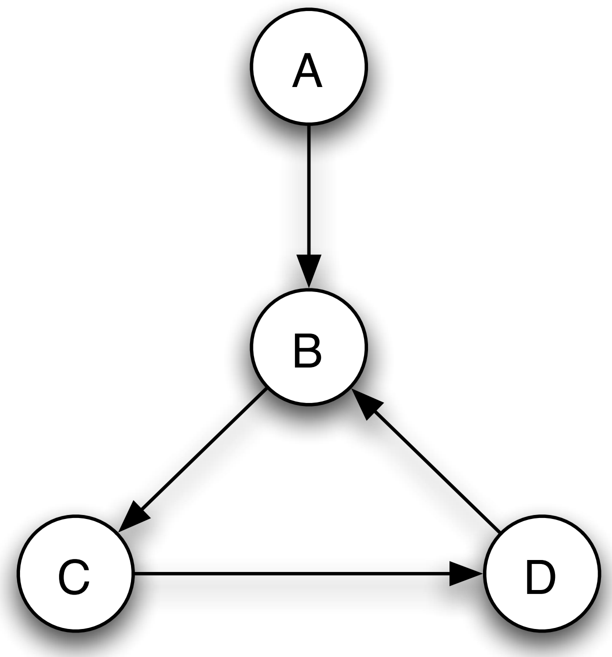
paths

cycles

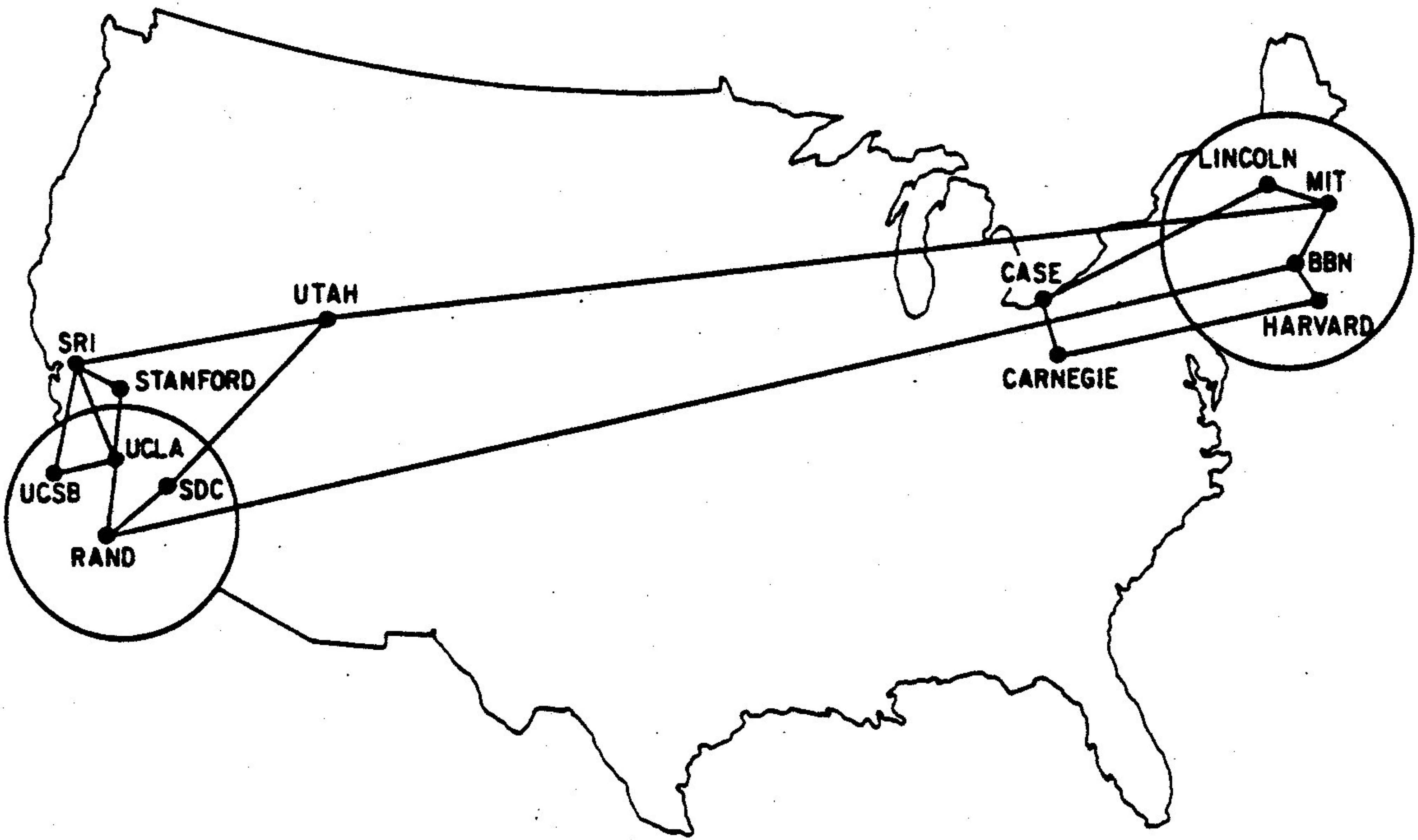
components (and giant components)



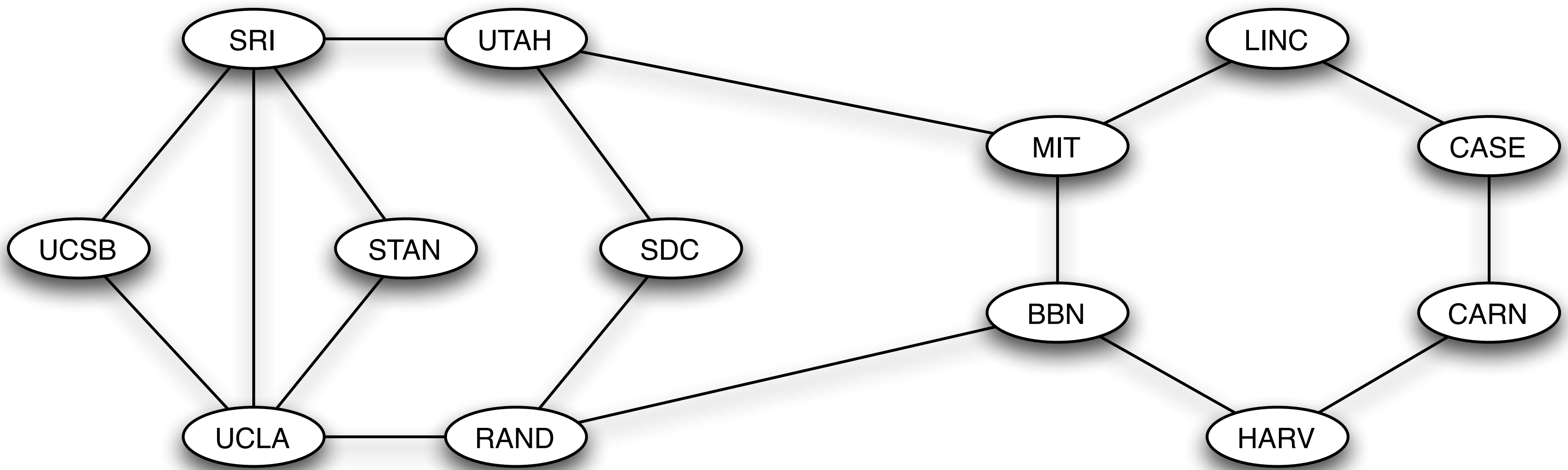
(a) *A graph on 4 nodes.*



(b) *A directed graph on 4 nodes.*



LINCOLN
MIT
BBN
HARVARD
CASE
CARNEGIE
UTAH
SRI
STANFORD
UCLA
UCSB
SDC
RAND


SRI
UTAH
LINC
MIT
CASE
UCSB
STAN
SDC
BBN
CARN
UCLA
RAND
HARV

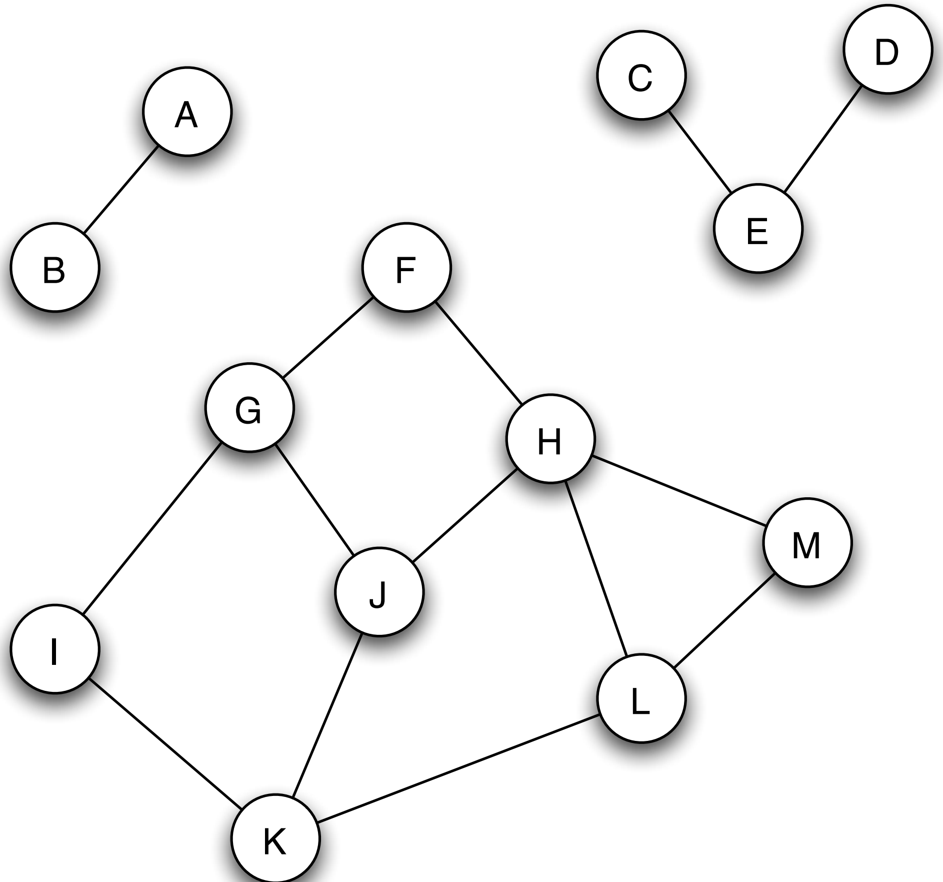


Figure 2.5: A graph with three connected components.

In practice, most graphs have a single giant component. But consider the Western hemisphere vs Europe (at the age of exploration) … human diseases evolved independently, technology

Network Datasets

Collaboration graphs

Who-talks-to-whom graphs

Information linkage graphs

Technological networks

Networks in the natural world

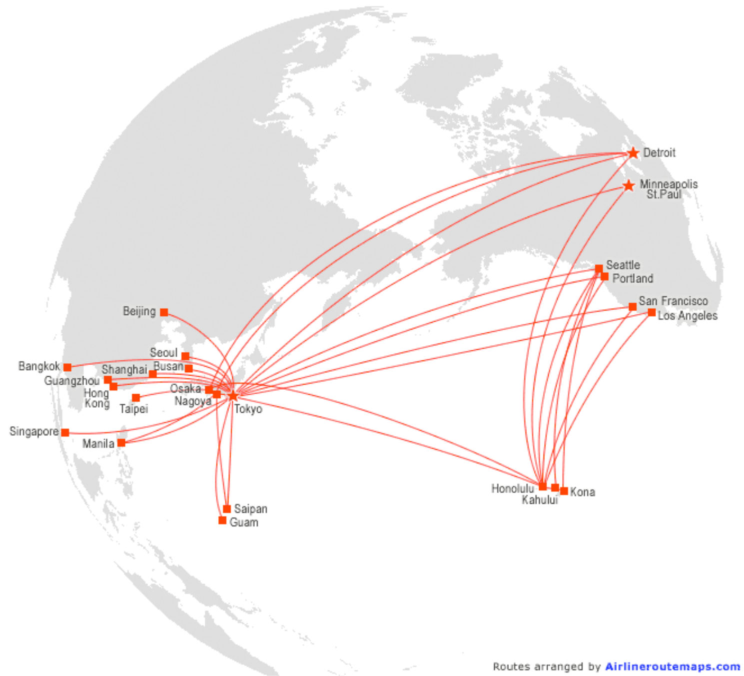
Some Network Resources

https://snap.stanford.edu/data/index.html

https://aws.amazon.com/datasets/marvel-universe-social-graph/

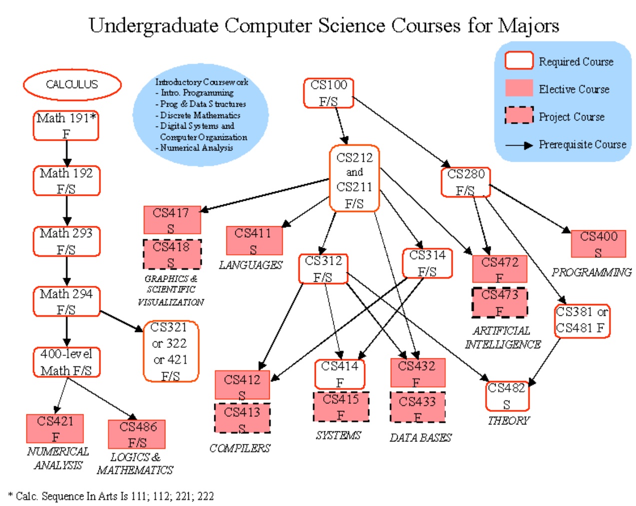
http://www-personal.umich.edu/~mejn/netdata/

https://networkdata.ics.uci.edu/index.html



(a) *Airline routes*

(b) *Subway map*



(c) *Flowchart of college courses*

(d) *Tank Street Bridge in Brisbane*

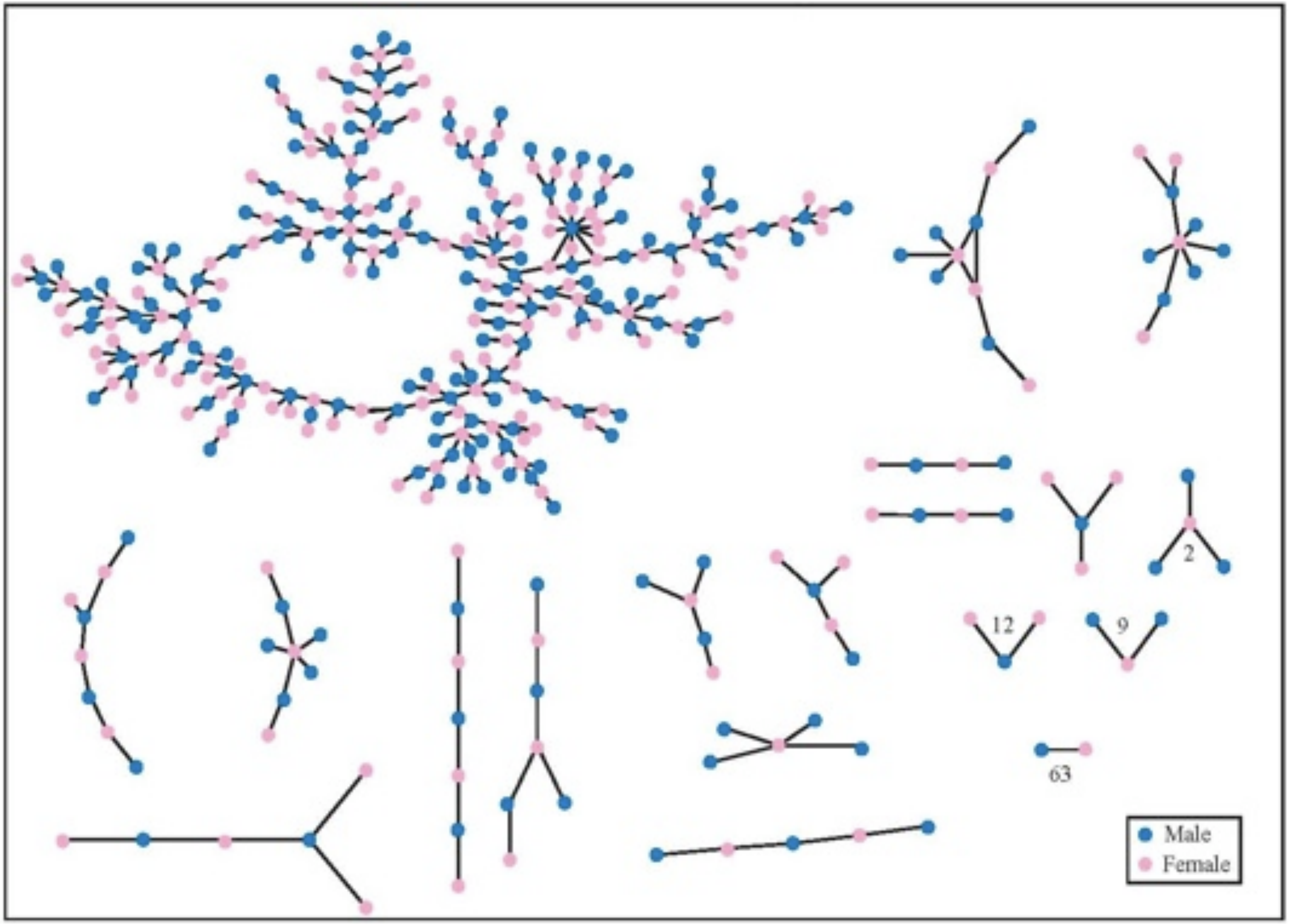
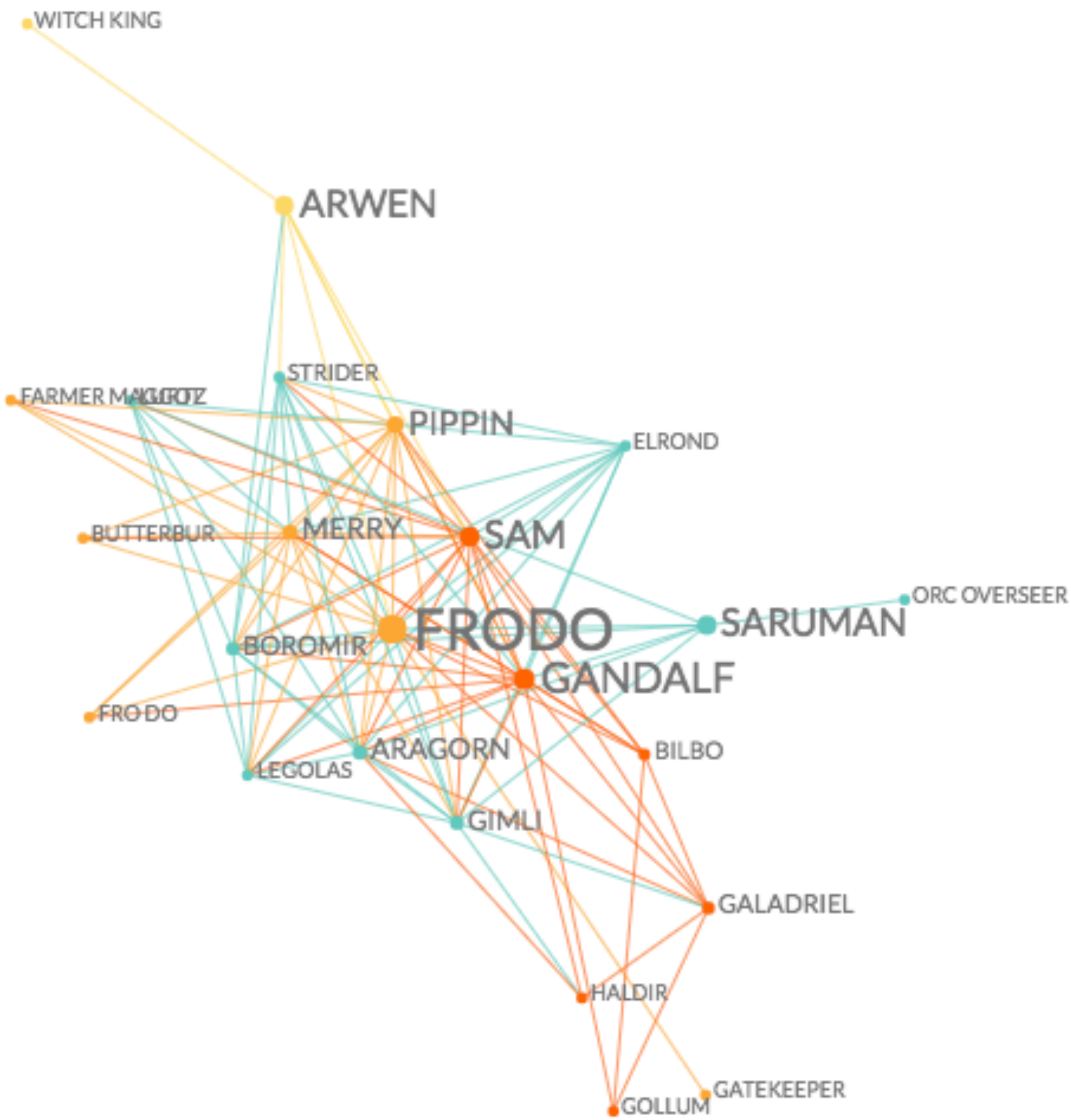
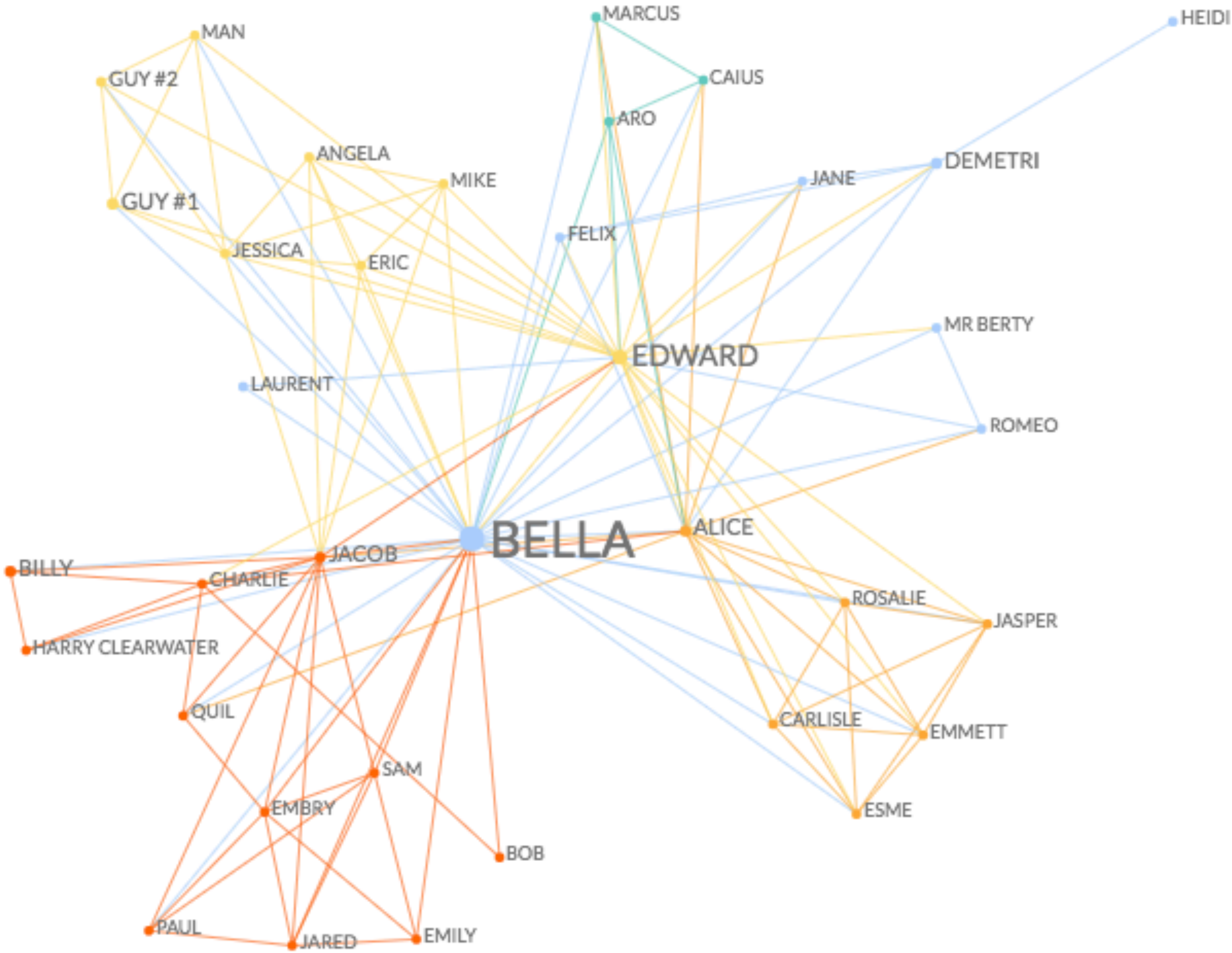
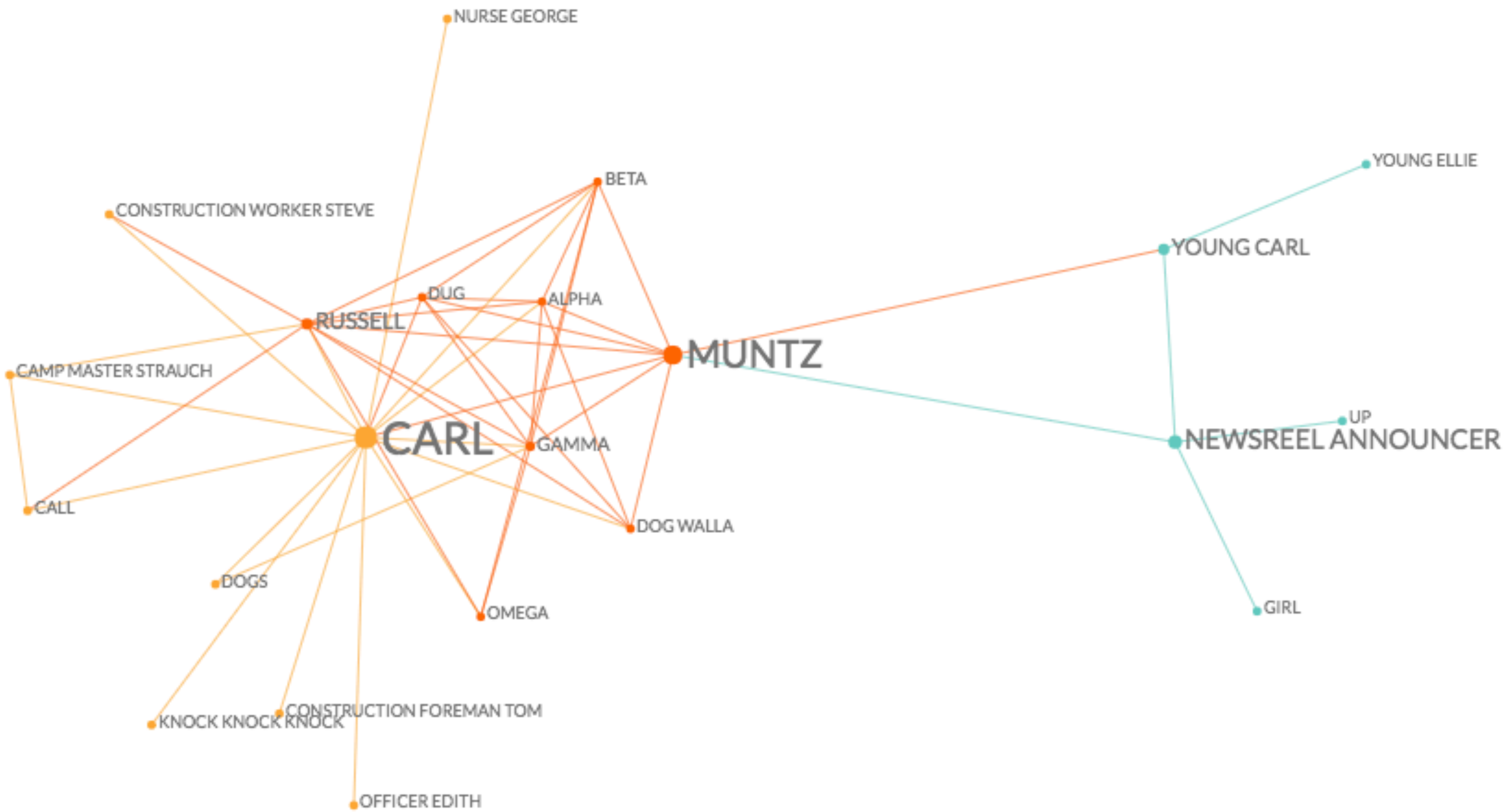


Figure 2.7: A network in which the nodes are students in a large American high school, and an edge joins two who had a romantic relationship at some point during the 18-month period in which the study was conducted [49].


WITCH KING
ARWEN
STRIDER
FARMER MAGGOT
PIPPIN
ELROND
BUTTERBUR
MERRY
SAM
ORC OVERSEER
FRODO
SARUMAN
BOROMIR
GANDALF
FRO DO
ARAGORN
BILBO
LEGOLAS
GIMLI
GALADRIEL
HALDIR
GATEKEEPER
GOLLUM


MARCUS
HEIDI
MAN
CAIUS
GUY #2
ARO
DEMETRI
ANGELA
JANE
MIKE
GUY #1
FELIX
JESSICA
ERIC
MR BERTY
EDWARD
LAURENT
ROMEO
BELLA
ALICE
JACOB
BILLY
CHARLIE
ROSALIE
JASPER
HARRY CLEARWATER
QUIL
CARLISLE
EMMETT
SAM
EMBRY
ESME
BOB
PAUL
JARED
EMILY


PILOT
KRISTI
MEGAN
PAUSE
ARON
RANA
ERIK
MOM
WAYNE

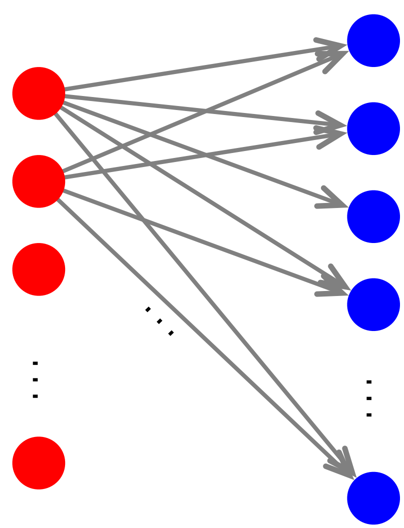

NURSE GEORGE
BETA
CONSTRUCTION WORKER STEVE
YOUNG ELLIE
YOUNG CARL
DUG
ALPHA
RUSSELL
MUNTZ
CAMP MASTER STRAUCH
UP
CARL
GAMMA
NEWSREEL ANNOUNCER
CALL
DOG WALLA
DOGS
GIRL
OMEGA
CONSTRUCTION FOREMAN TOM
KNOCK KNOCK KNOCK
OFFICER EDITH

Connecting Graph Mining to Frequent Item Sets

Idea 1: Trawling [Kumar ’99]

Searching for small communities in the Web graph

What is the signature of a community / discussion in a Web graph?



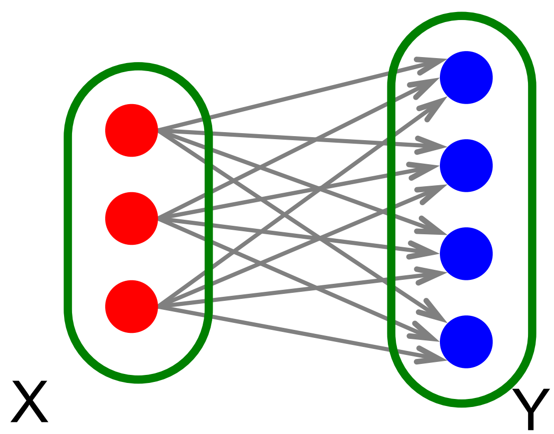
Intuition: Many people all talking about the same things

Dense 2-layer graph

Searching for Small Communities

A more well-defined problem:
Enumerate complete bipartite subgraphs $K_{s,t}$

Where $K_{s,t}$: s nodes on the "left" where each links to the same t other nodes on the "right"



$K_{3,4}$

$|X| = s = 3$
$|Y| = t = 4$

Fully connected

Frequent Itemset Enumeration

Market Basket Analysis!

Universe U of n items

Baskets: m subsets of U: $S_1, S_2, \ldots, S_m \subseteq U$
(S_i is a set of items one person bought)

Support: Frequency threshold f

Goal:

Find all subsets T s.t. $T \subseteq S_i$ of at least f sets S_i
(items in T were bought together at least f times)

What's the connection between the itemsets and complete bipartite graphs?

双边的

From Itemsets to Bipartite $K_{s,t}$

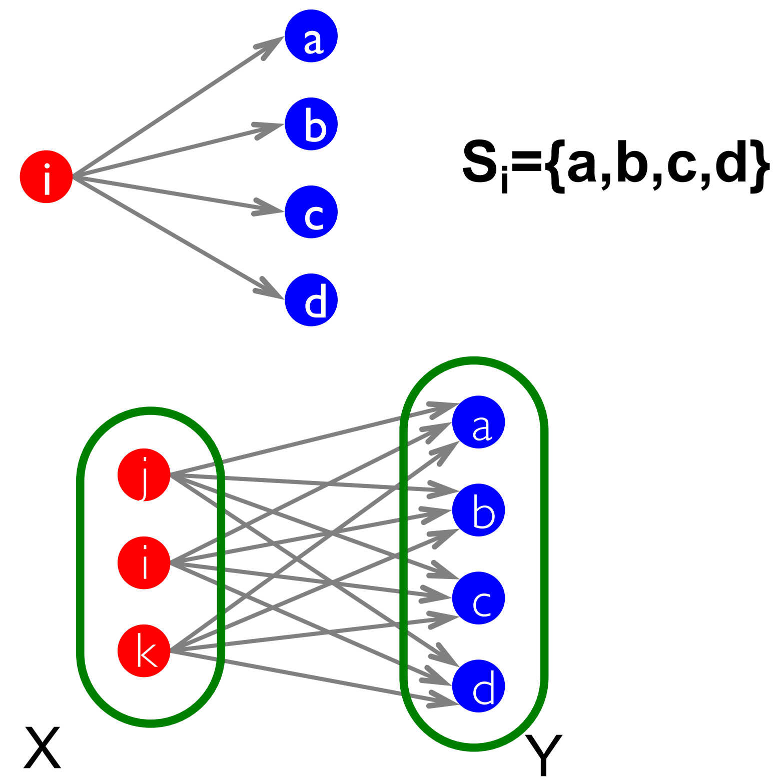
Frequent itemsets = complete bipartite graphs!

How?

View each node i as a set S_i of nodes i points to

$K_{s,t}$ = a set Y of size t that occurs in s sets S_i

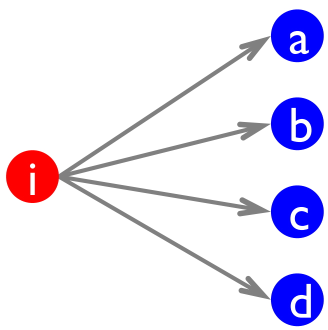
Looking for $K_{s,t}$ —> set of frequency threshold to s and look at layer t – all frequent sets of size t



s … minimum support (|X|=s)
t … itemset size (|Y|=t)

From Itemsets to Bipartite $K_{s,t}$

View each node i as a set S_i of nodes i points to



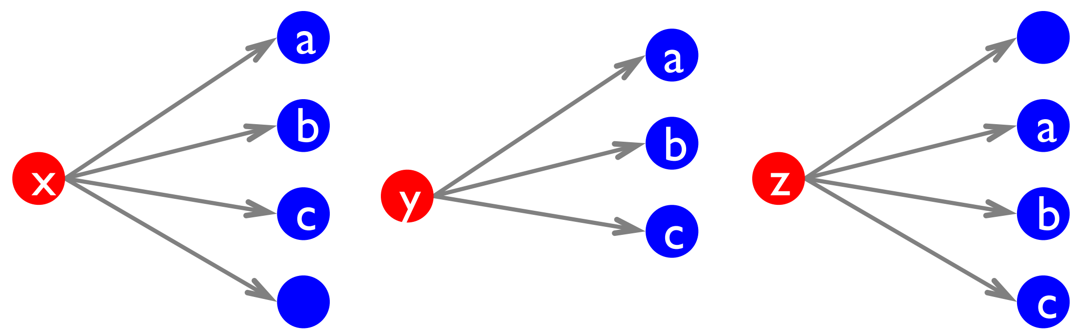
S_i={a,b,c,d}

Find frequent itemsets:

s … minimum suppor

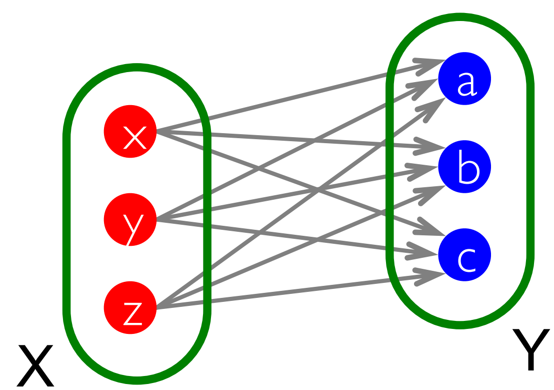
t … itemset size

Say we find a frequent itemset Y={a,b,c} of supp s
So, there are s nodes that link to all of {a,b,c}:

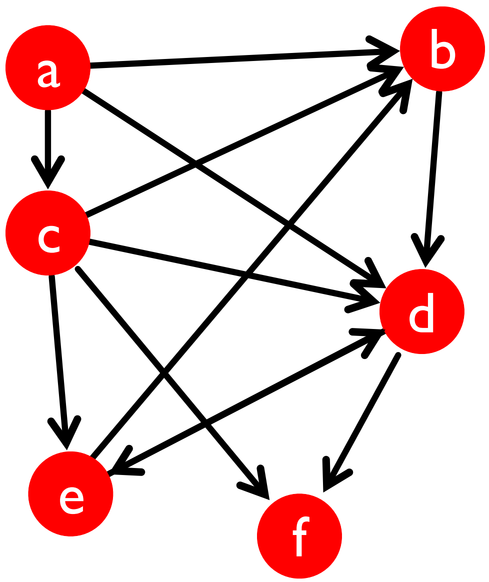


We found Ks,t!

$K_{s,t}$ = a set Y of size t that occurs in s sets S_i



Example



Itemsets:

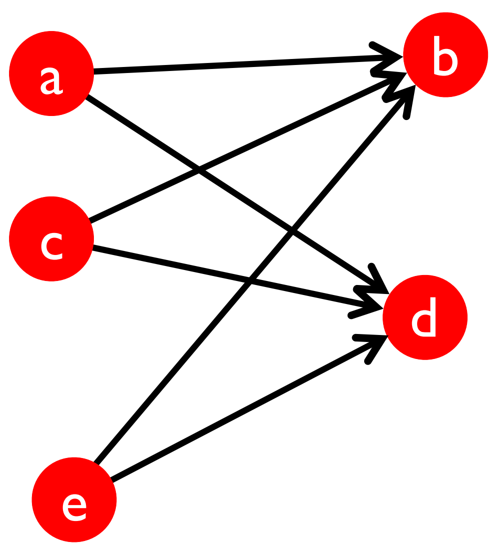
a = {b,c,d}
b = {d}
c = {b,d,e,f}
d = {e,f}
e = {b,d}
f = {}

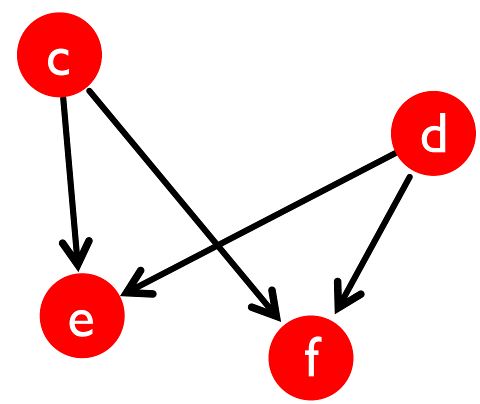
Support threshold s=2

{b,d}: support 3

{e,f}: support 2

And we just found 2 bipartite subgraphs:





Example

A community of Australian fire brigades

Authorities	Hubs
NSW Rural Fire Service Internet Site	New South Wales Fir...ial Australian Links
NSW Fire Brigades	Feuerwehrlinks Australien
Sutherland Rural Fire Service	FireNet Information Network
CFA: County Fire Authority	The Cherrybrook Rur...re Brigade Home Page
"The National Cente...ted Children's Ho...	New South Wales Fir...ial Australian Links
CRAFTI Internet Connexions-INFO	Fire Departments, F... Information Network
Welcome to Blackwoo... Fire Safety Serv...	The Australian Firefighter Page
The World Famous Guestbook Server	Kristiansand brannv...dens brannvesener...
Wilberforce County Fire Brigade	Australian Fire Services Links
NEW SOUTH WALES FIR...ES 377 STATION	The 911 F,P,M., Fir...mp; Canada A Section
Woronora Bushfire Brigade	Feuerwehrlinks Australien
Mongarlowe Bush Fire – Home Page	Sanctuary Point Rural Fire Brigade
Golden Square Fire Brigade	Fire Trails "l...ghters around the...
FIREBREAK Home Page	FireSafe – Fire and Safety Directory
Guises Creek Volunt...fficial Home Page...	Kristiansand Firede...departments of th...

[Kumar, Raghavan, Rajagopalan, Tomkins: Trawling the Web for emerging cyber-communities 1999]

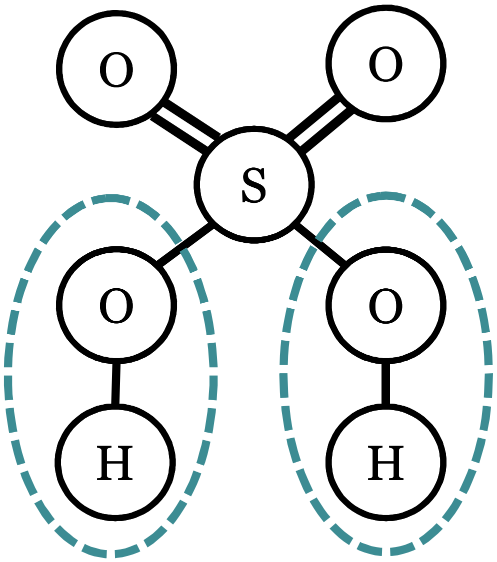
Idea 2: Frequent Subgraph Mining

Instead of finding frequent itemsets, lets look for frequent subgraphs

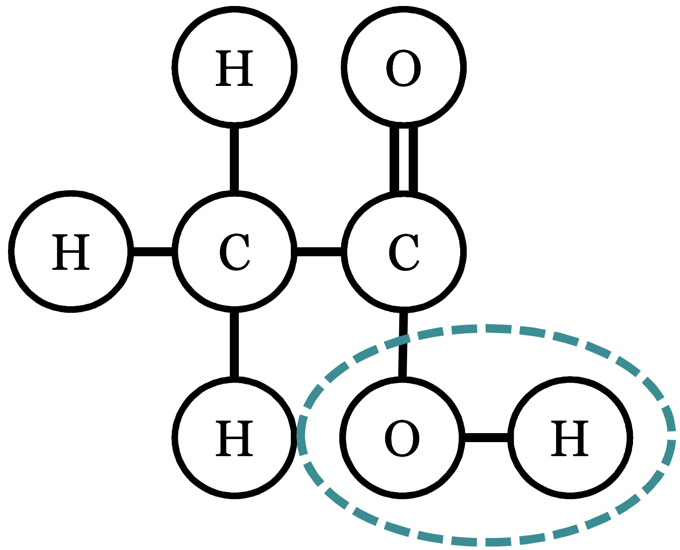
Frequent subgraph mining:

Discovery of graph structures that occur a significant number of times across a set of graphs

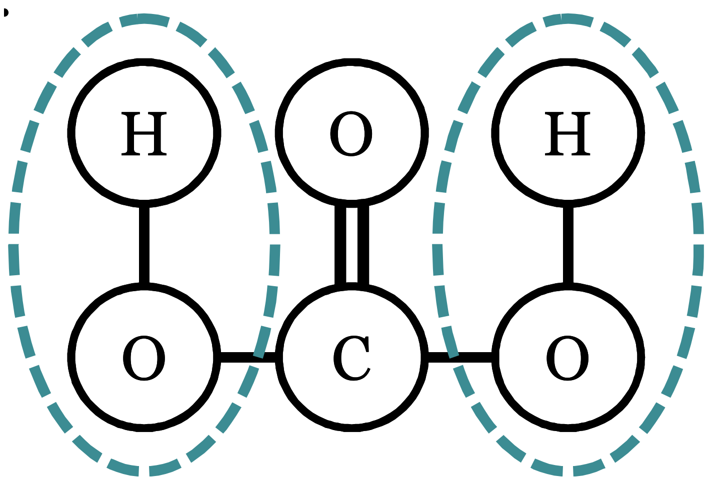
Sulfuric Acid

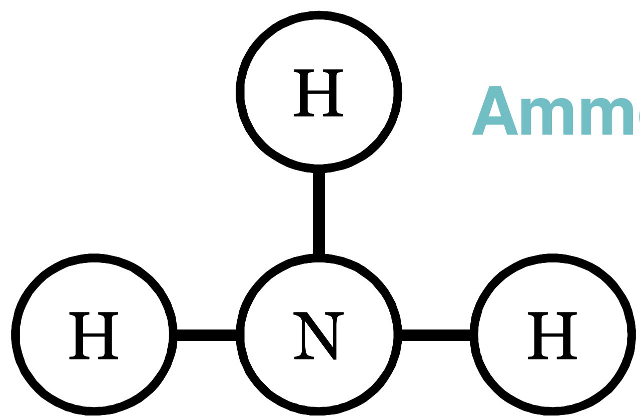


Acetic Acid



Carbonic Acid

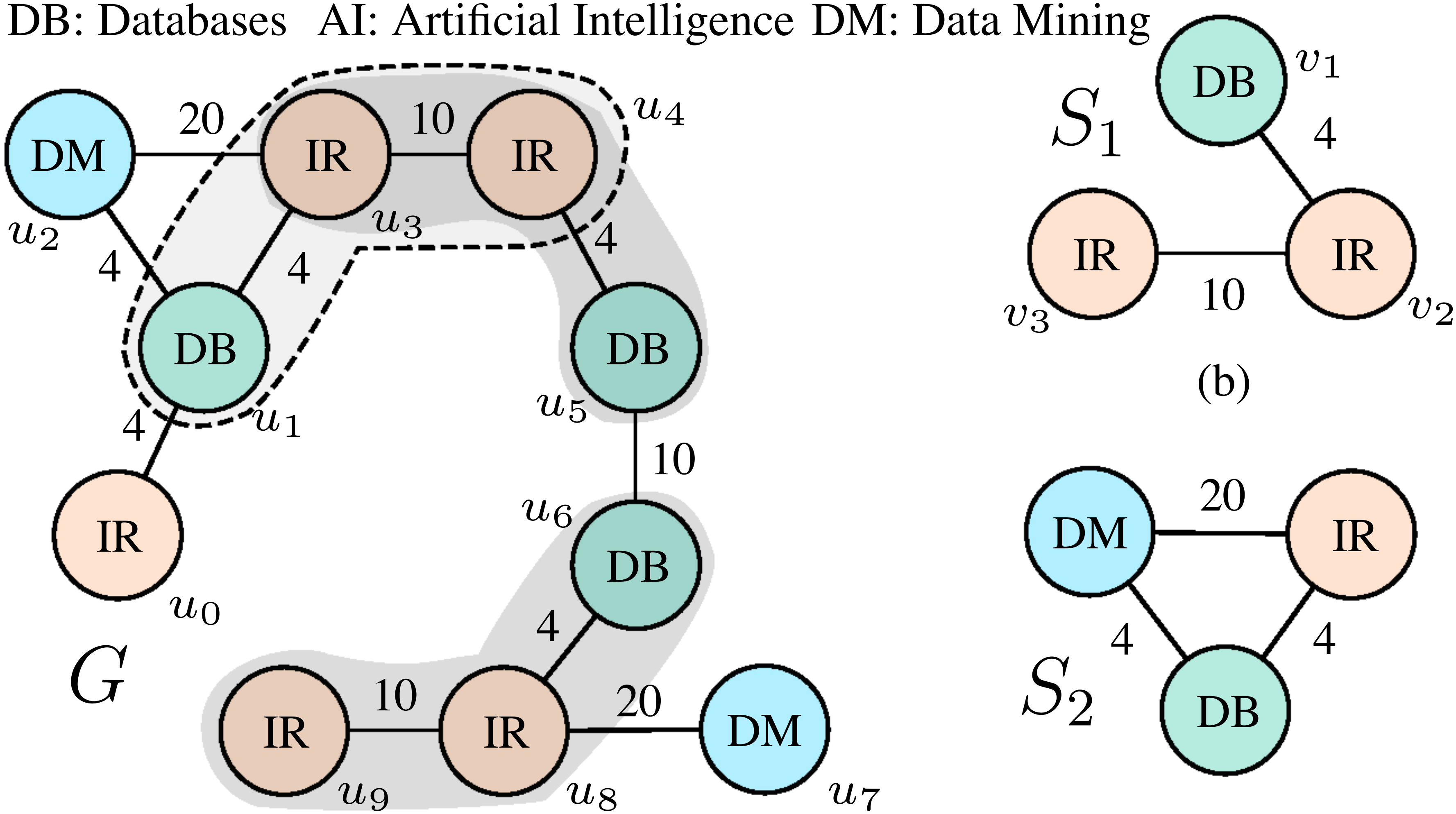




Ammonia

https://www.csc2.ncsu.edu/faculty/nfsamato/practical-graph-mining-with-R/slides/pdf/Frequent_Subgraph_Mining.pdf

DB: Databases AI: Artificial Intelligence DM: Data Mining

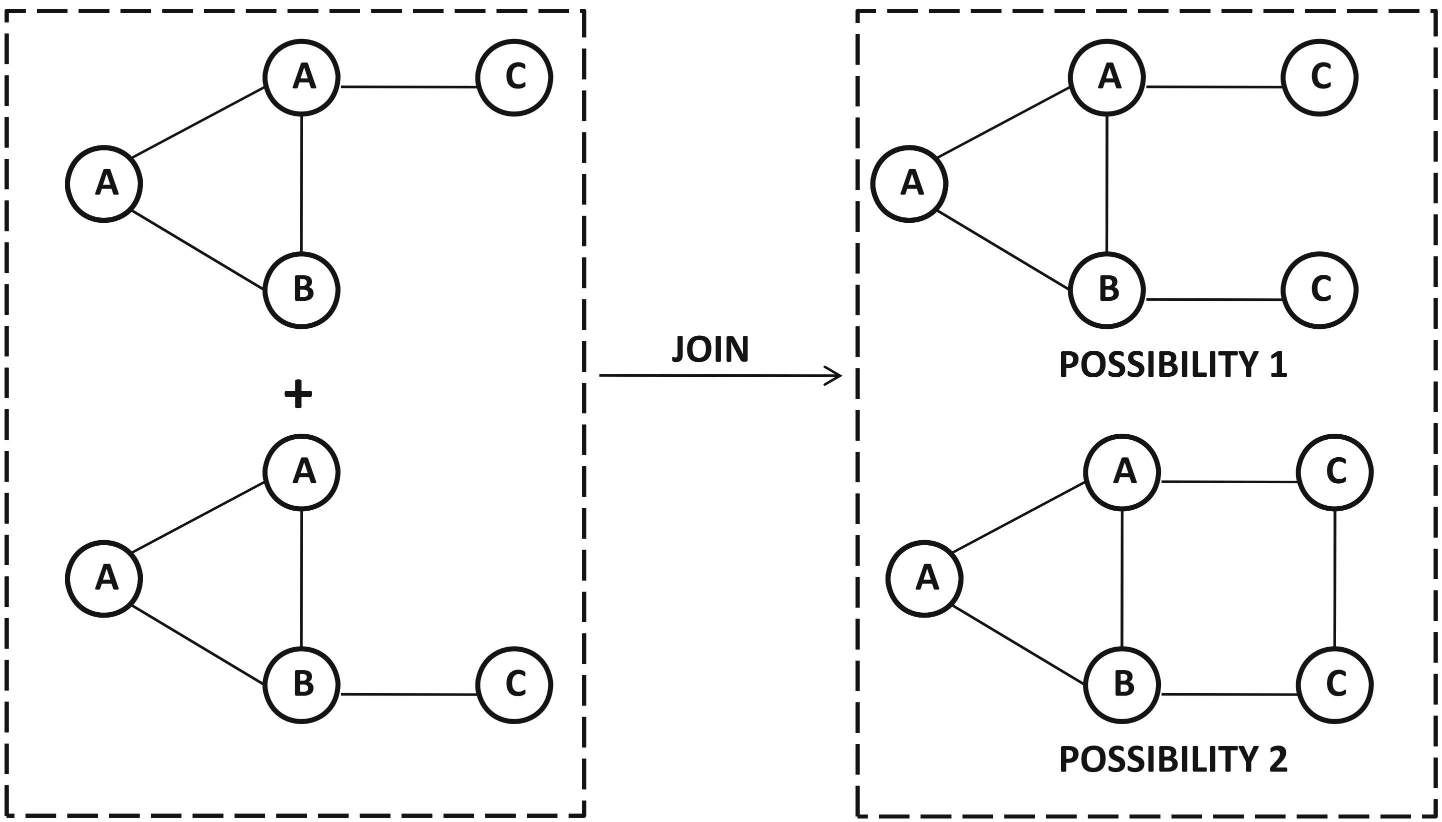


(a) (b) (c)

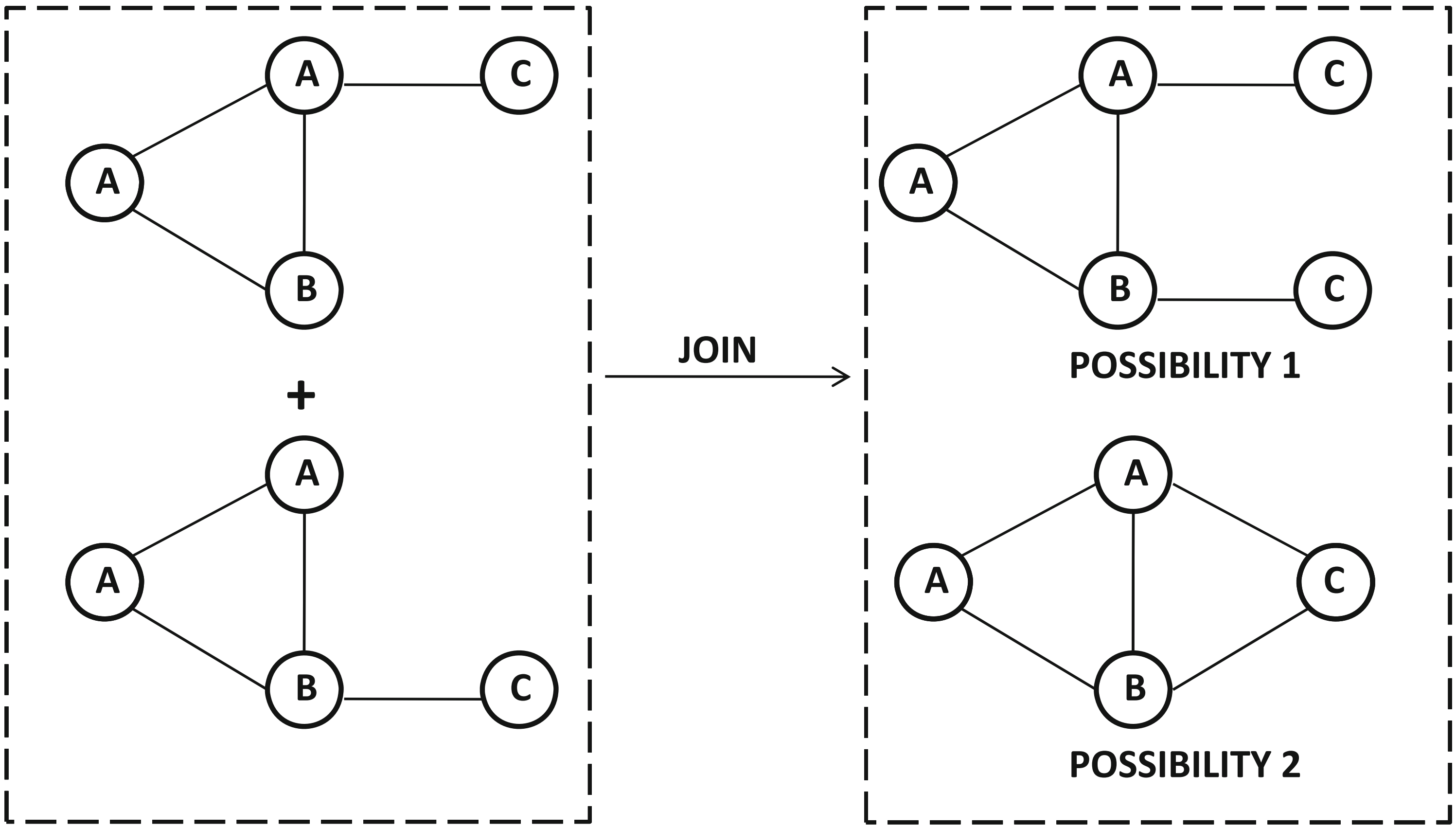
GRAMI: Frequent Subgraph and Pattern Mining in a Single Large Graph, VLDB 2014

Algorithm *GraphApriori*(Graph Database: $\mathcal{G}$,
Minimum Support: *minsup*);
begin
$\mathcal{F}_1$ = { All Frequent singleton graphs };
$k = 1$;
while $\mathcal{F}_k$ is not empty **do begin**
1 Generate $\mathcal{C}_{k+1}$ by joining pairs of graphs in $\mathcal{F}_k$ that share a subgraph of size $(k-1)$ in common;
2 Prune subgraphs from $\mathcal{C}_{k+1}$ that violate downward closure;
3 Determine $\mathcal{F}_{k+1}$ by support counting on $(\mathcal{C}_{k+1}, \mathcal{G})$ and retaining subgraphs from $\mathcal{C}_{k+1}$ with support at least *minsup*;
$k = k + 1$;
end;
return$(\cup_{i=1}^{k} \mathcal{F}_i)$;
end

Node-based join



Edge-based Join



Finding Important Nodes

Which nodes are important?

Why would we care?

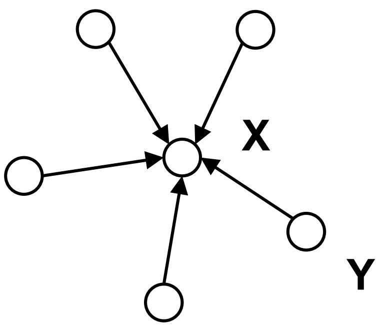
…

How would we find?

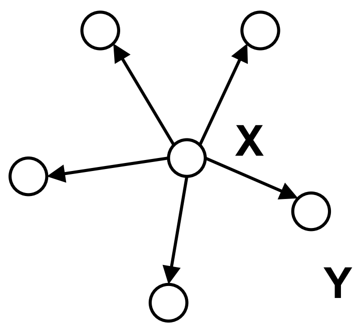
…

Q: How to Measure Centrality?

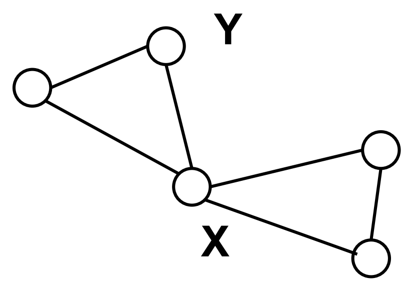
In each of the following networks, X has higher centrality than Y according to a particular measure



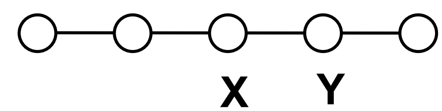
indegree



outdegree



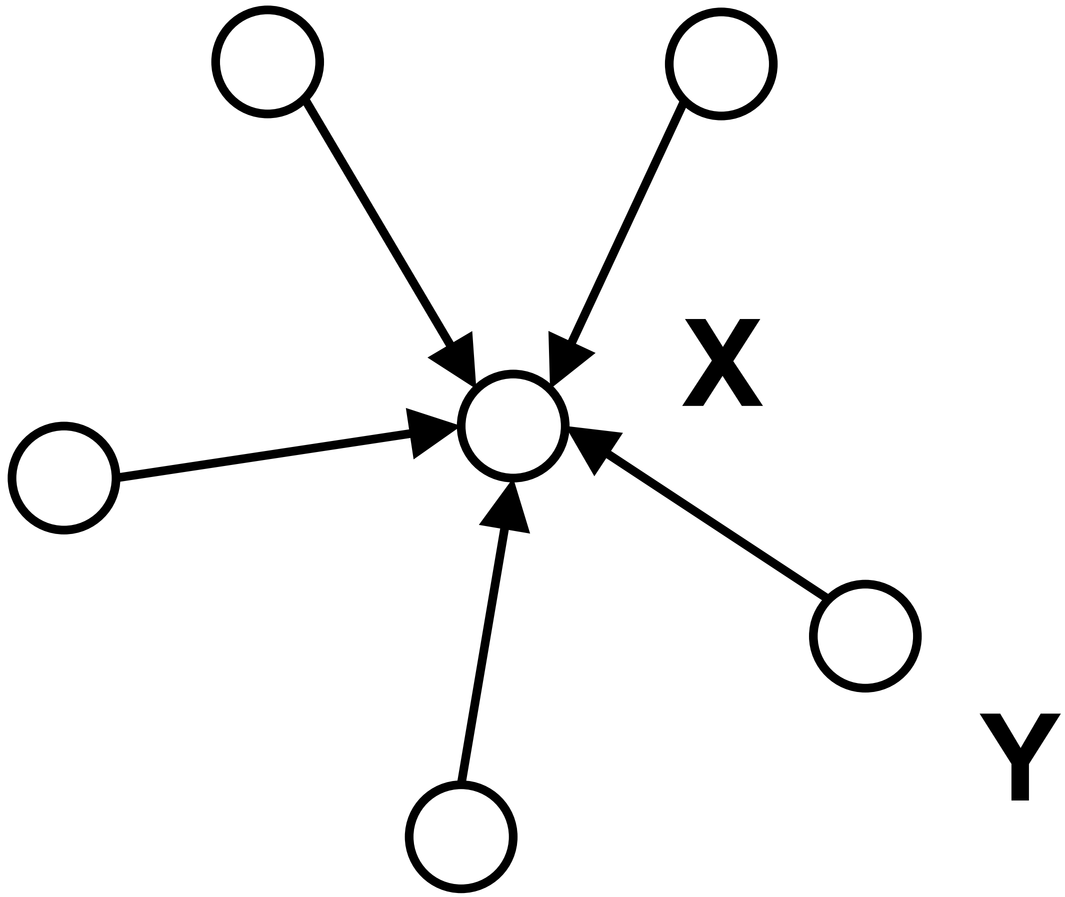
betweenness



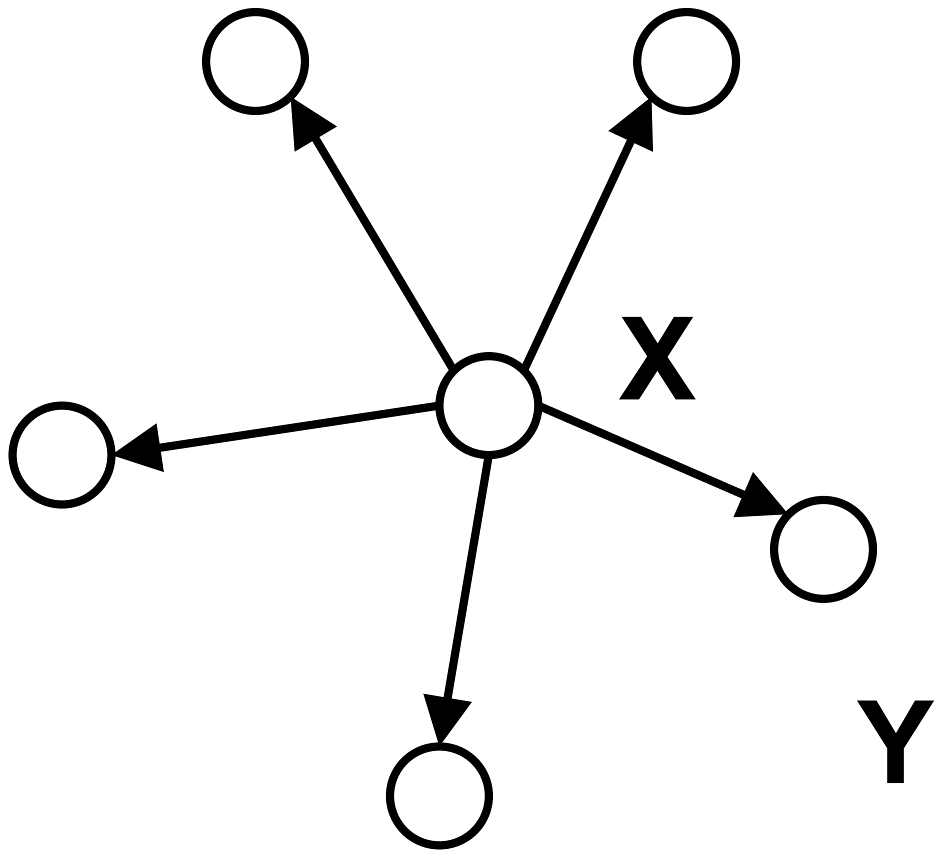
closeness

From Lada Adamic

In-degree centrality

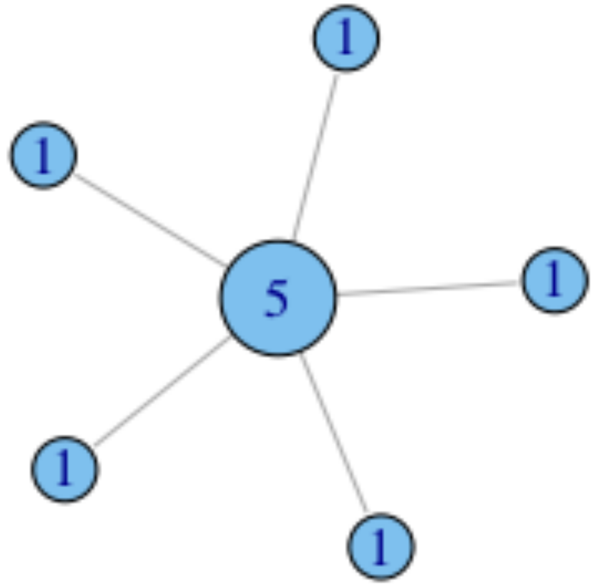


Out-degree centrality



Undirected degree centrality

Nodes with more friends are more central

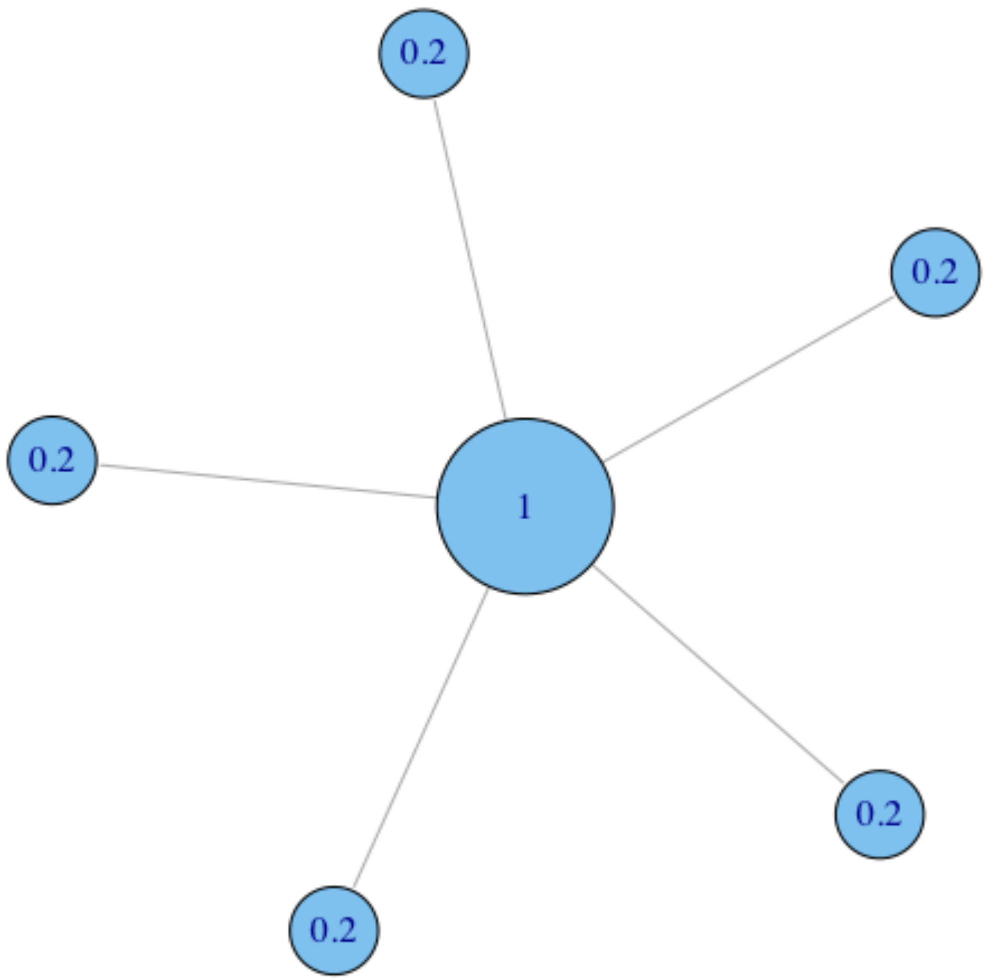


Key assumption: the connections that your friends have are unimportant; all that matters is what your friends can do directly for you

Examples?

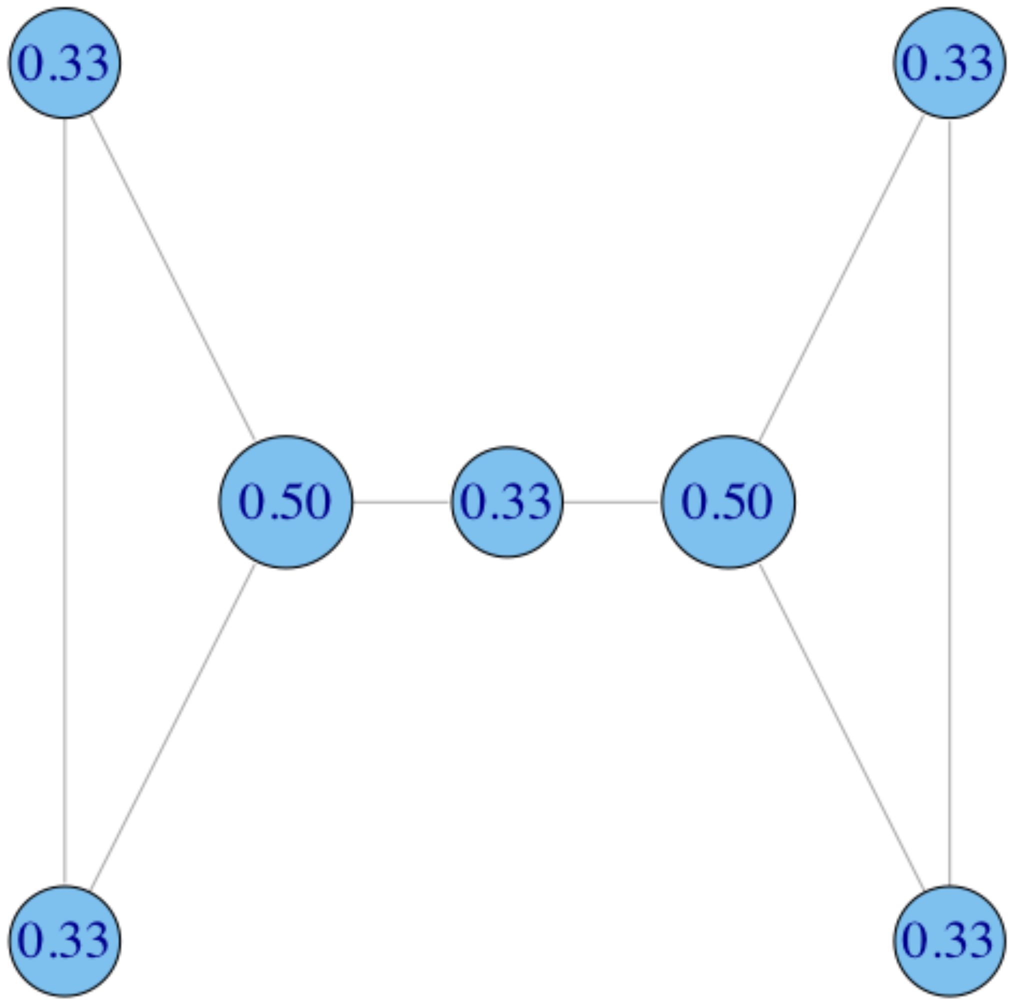
Normalization

divide degree by max (N-1)

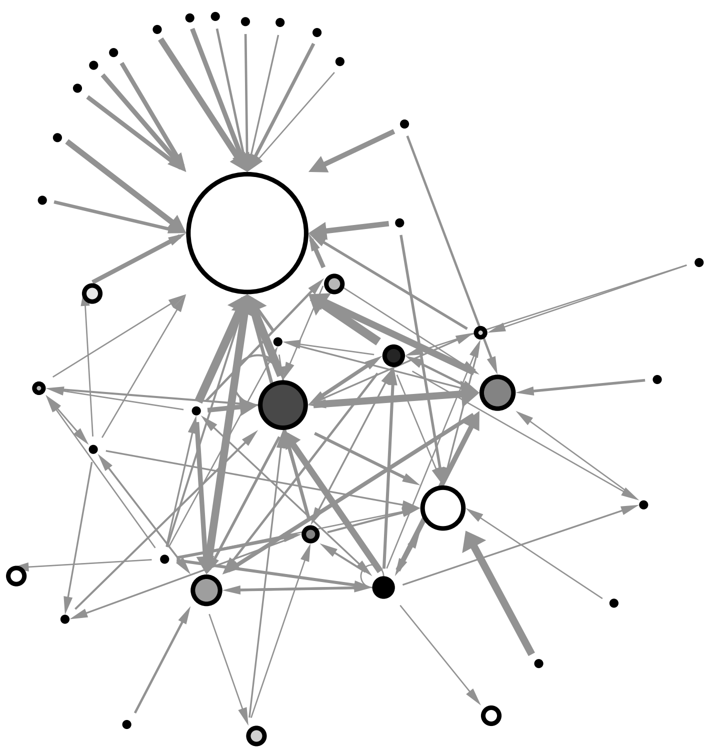


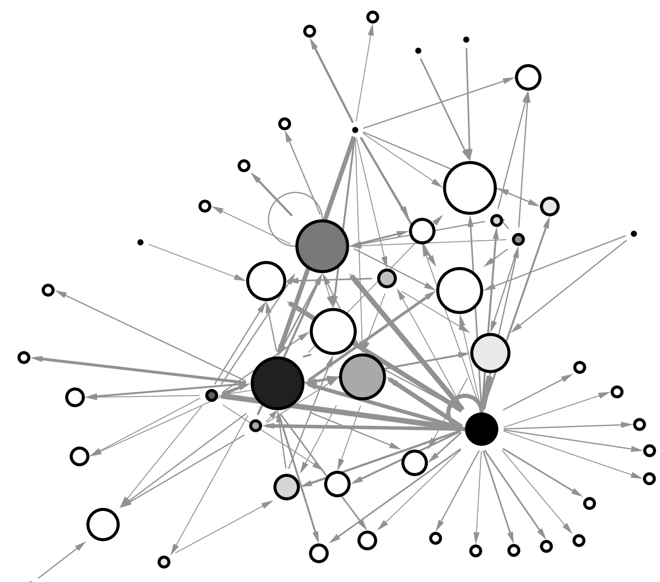
Normalization

divide degree by max (N-1)



Example: financial trading



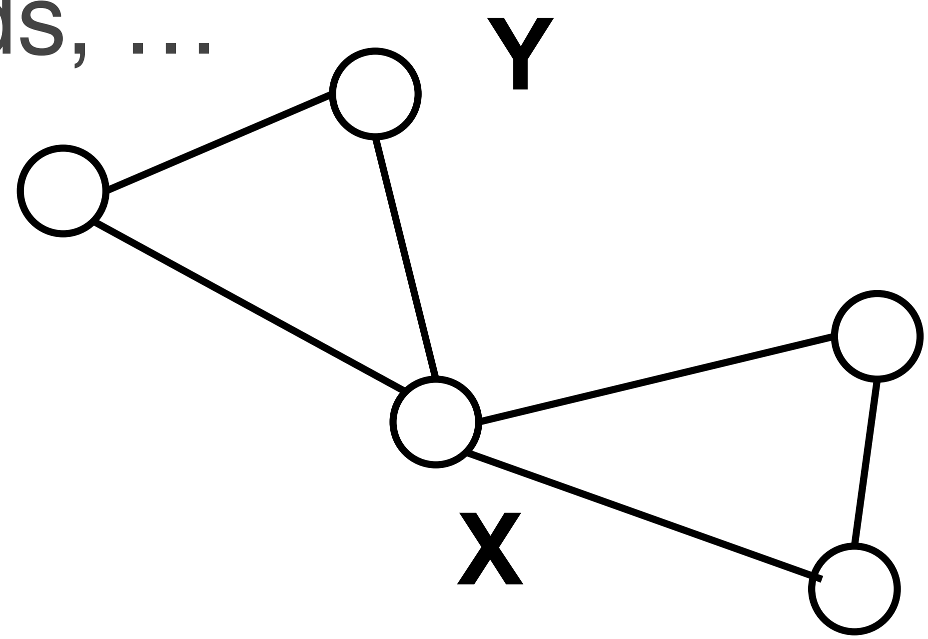


Q: What are these degree-based centrality measures missing?

Brokerage!

Connecting me to others

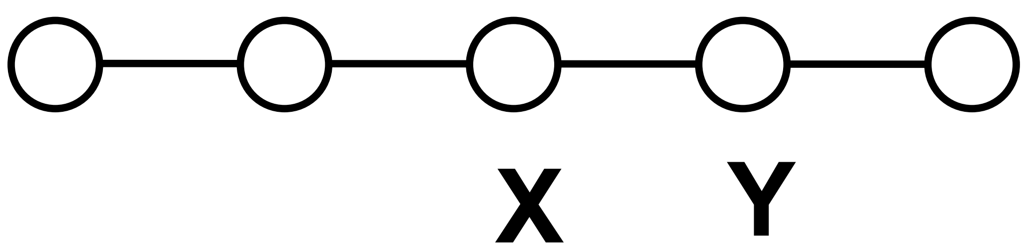
friends-of-friends, …



Betweenness centrality

Betweenness: Capturing brokerage in a centrality measure

Intuition: how many pairs of individuals would have to go through you in order to reach one another in the minimum number of hops?

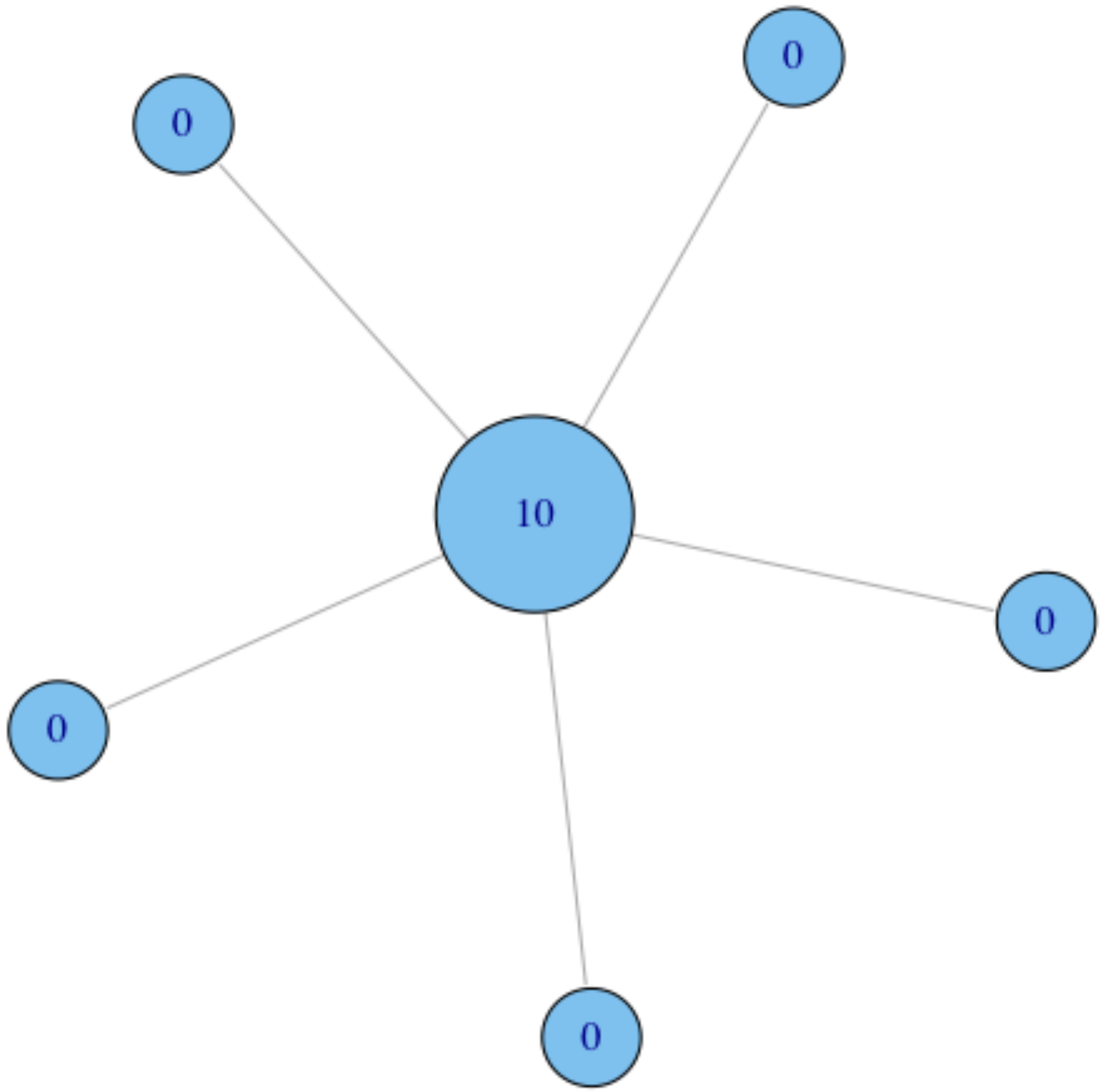


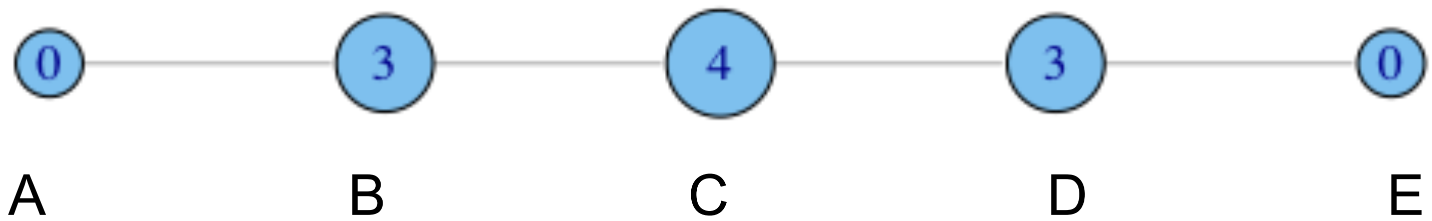
Betweenness centrality

$$g(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

where σ_{st} is the total number of shortest paths from node s to node t and $\sigma_{st}(v)$ is the number of those paths that pass through v.

Usually normalized by total number of possible vertex pairs (excluding itself)


0
0
10
0
0
0



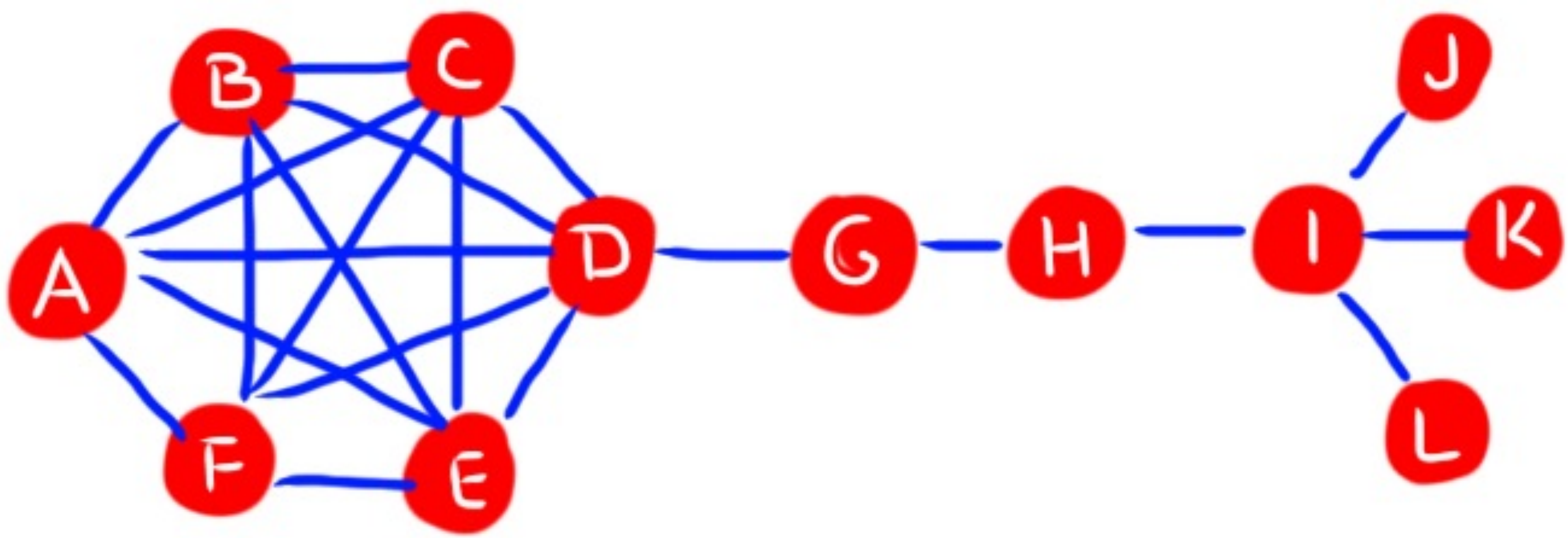
A lies between no two other vertices

B lies between A and 3 other vertices: C, D, and E

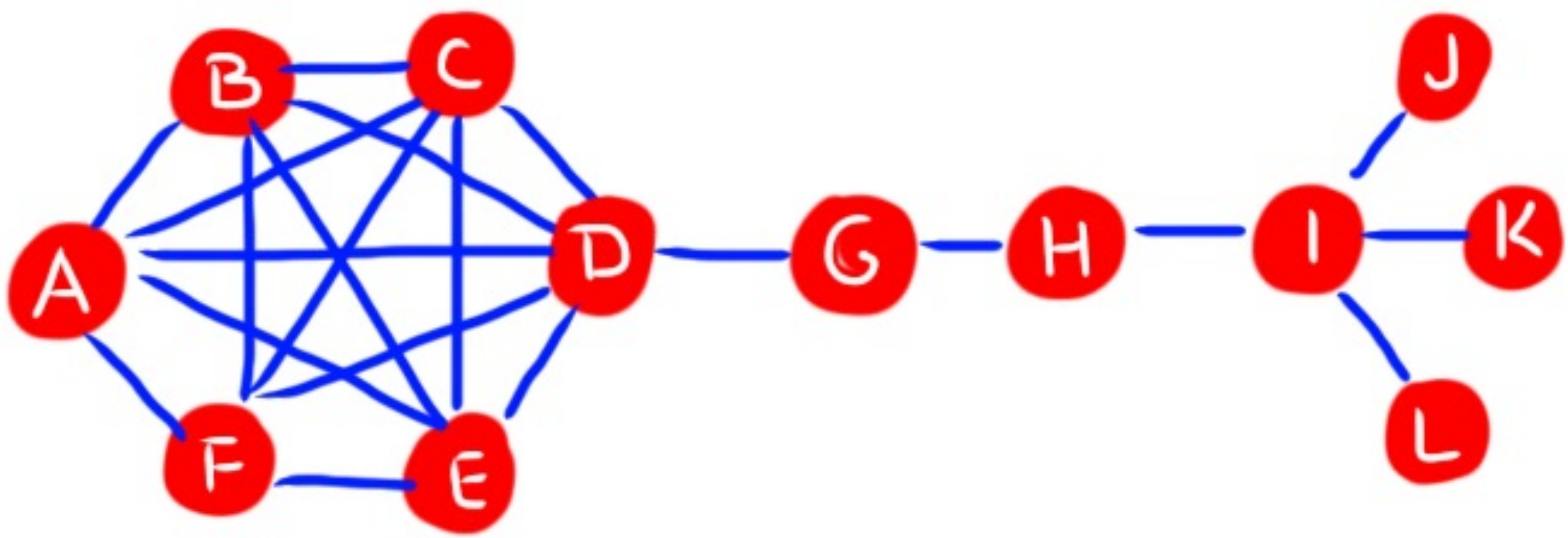
C lies between 4 pairs of vertices (A,D),(A,E),(B,D),(B,E)

note that there are no alternate paths for these pairs to take, so C gets full credit

Q: Find a node with high betweenness but low degree



Q: Find a node with low betweenness but high degree

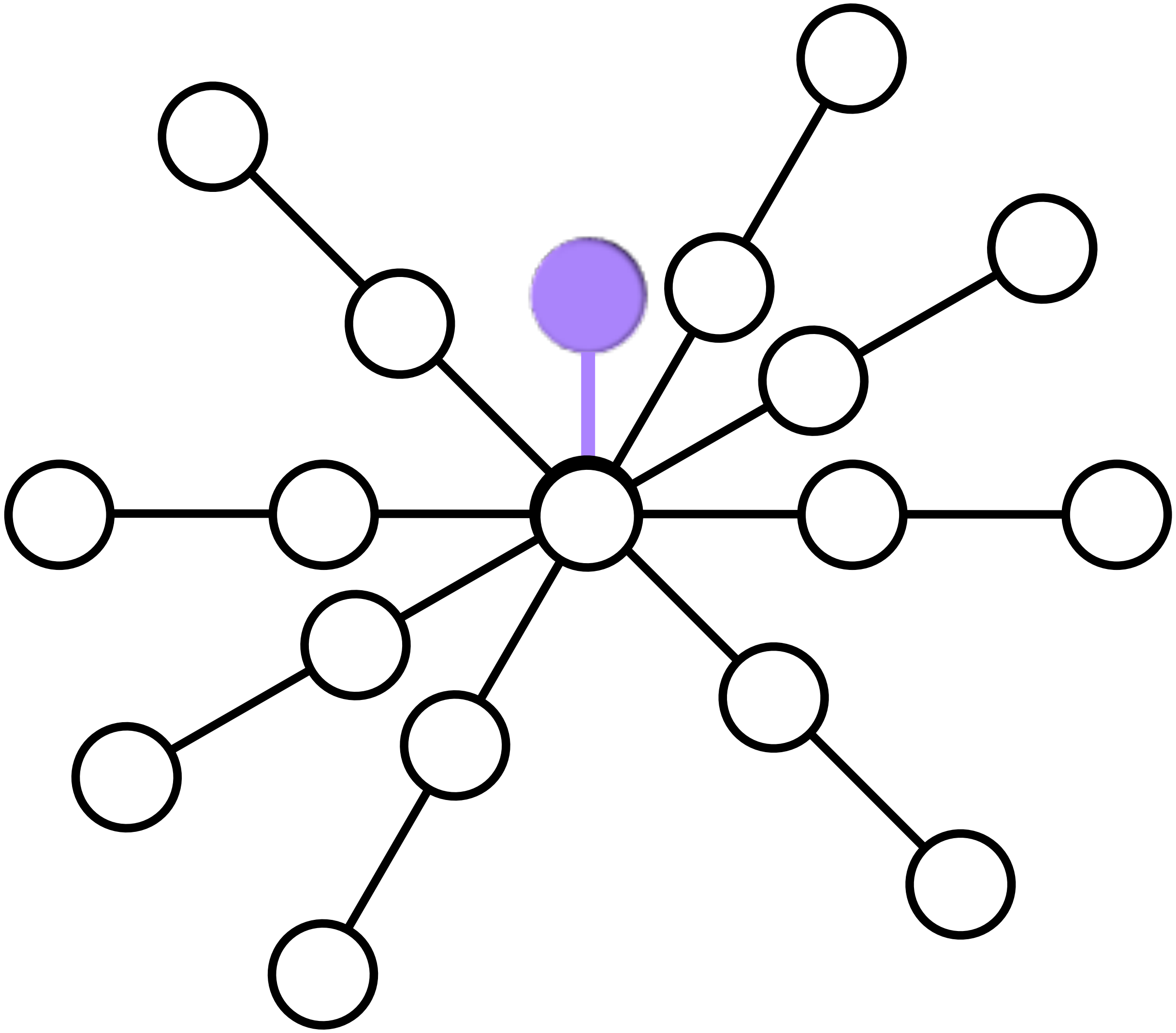


Closeness centrality

What if it’s not so important to have many direct friends?

Or be “between” others

But one still wants to be in the “middle” of things, not too far from the center



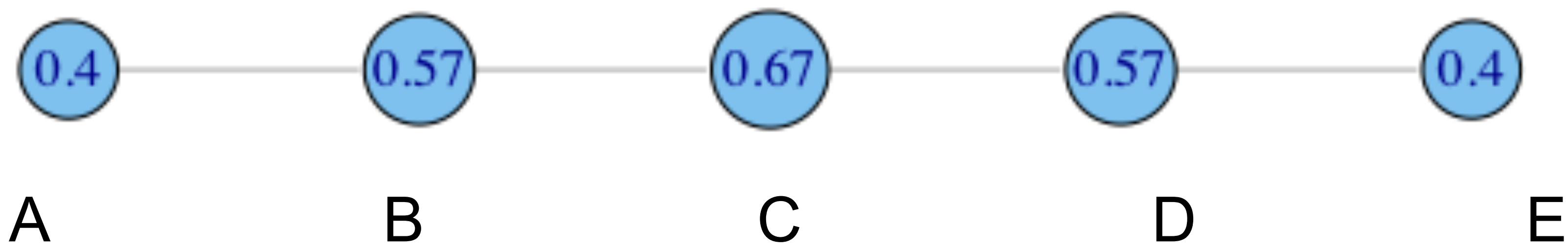
Closeness centrality

Closeness is based on the length of the average shortest path between a node and all other nodes in the network

$$C_c(i) = \left[\sum_{j=1}^{N} d(i,j) \right]^{-1}$$

Normalized:

$$C_C^{'}(i) = (C_C(i))/(N-1)$$



$$C_c^{'}(A) = \left[\frac{\sum_{j=1}^{N} d(A,j)}{N-1}\right]^{-1} = \left[\frac{1+2+3+4}{4}\right]^{-1} = \left[\frac{10}{4}\right]^{-1} = 0.4$$

Q: node with high degree but low closeness?