

Network Data

Manoel Horta Ribeiro
manoel@cs.princeton.edu



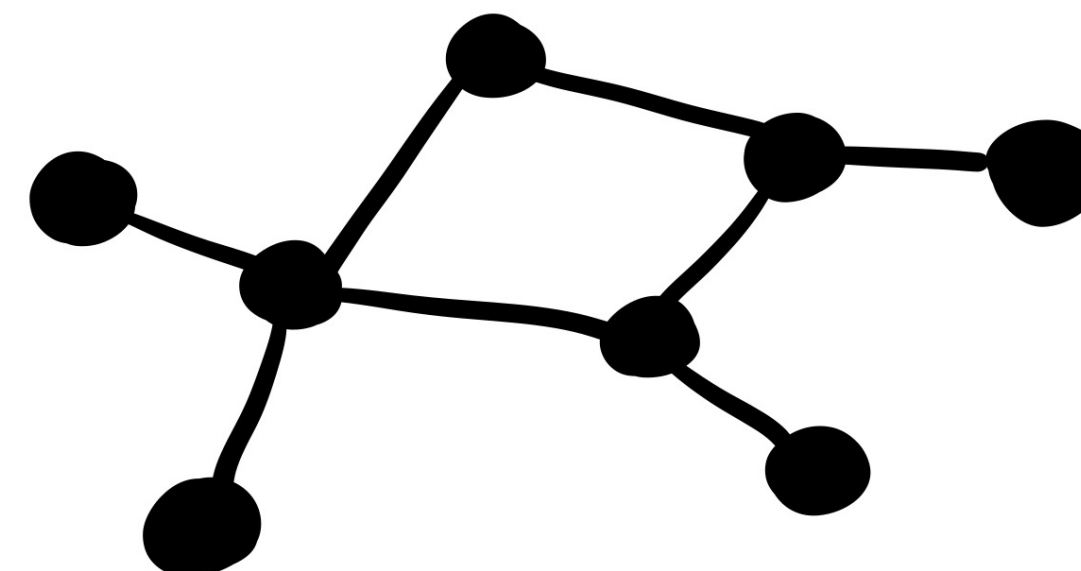
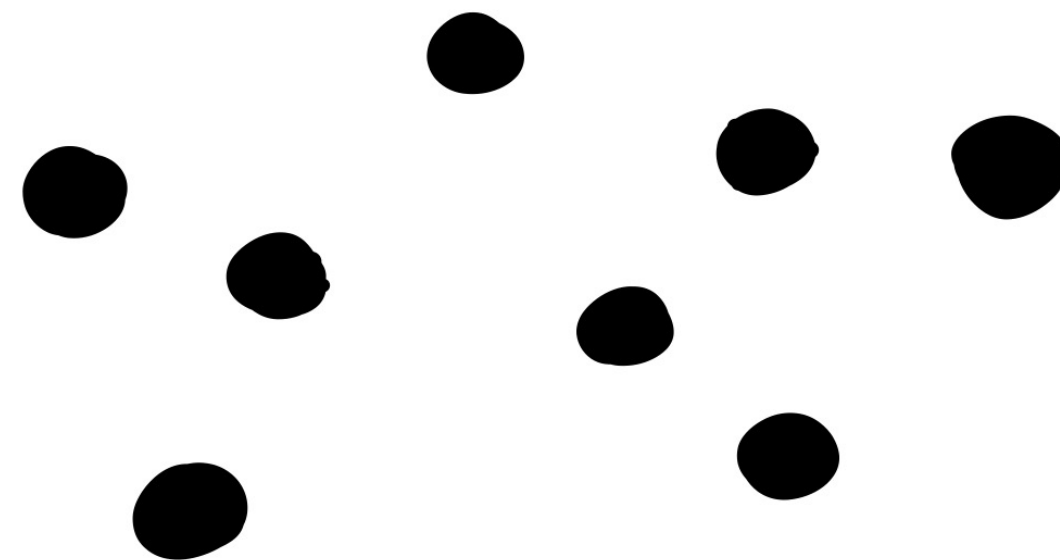
Quizz Time!



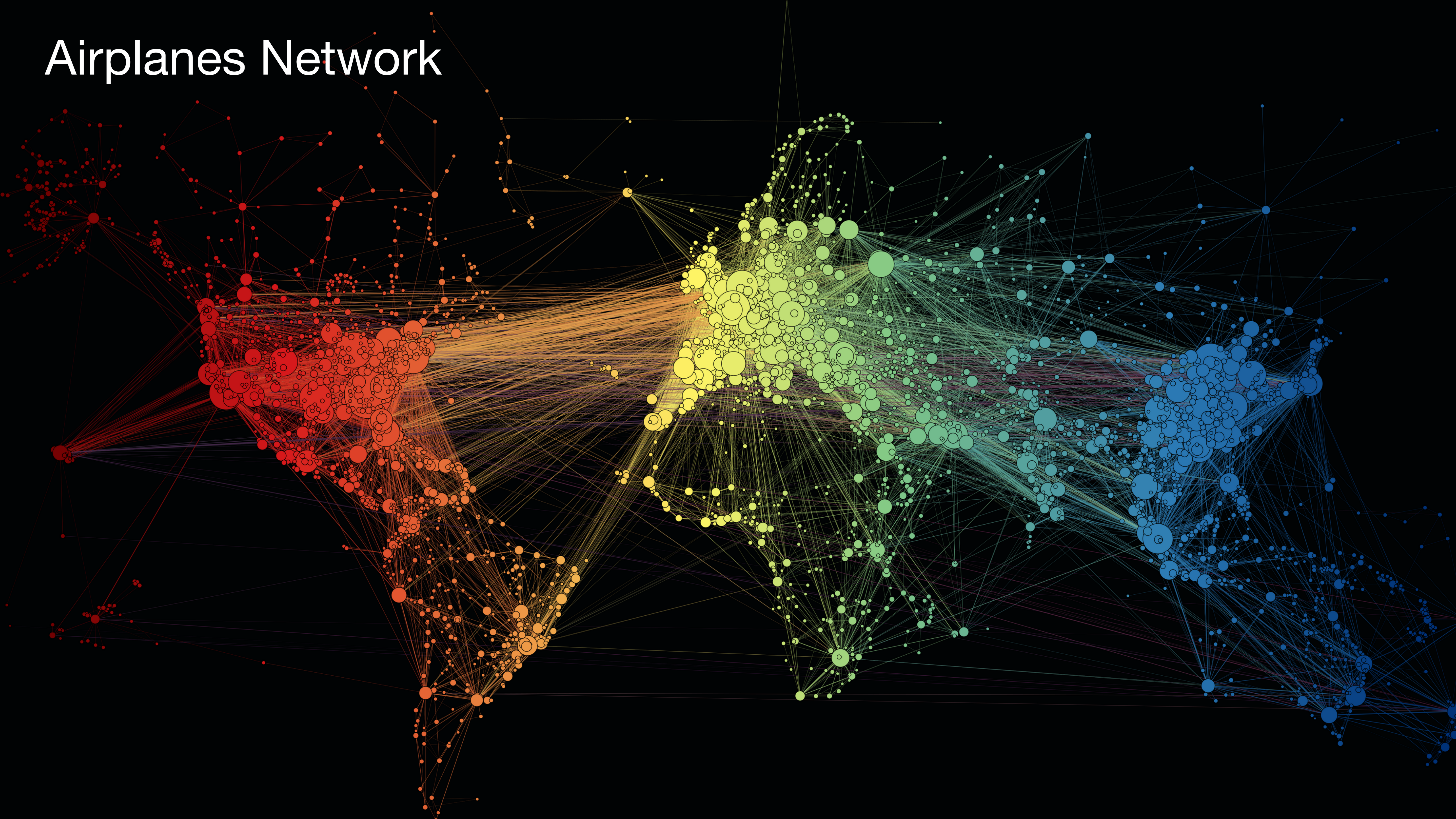
Beyond Flat Tables

Id	Name	Age
1	Manoel	32
2	Willy	27
3	Joanne	36
...	...	...

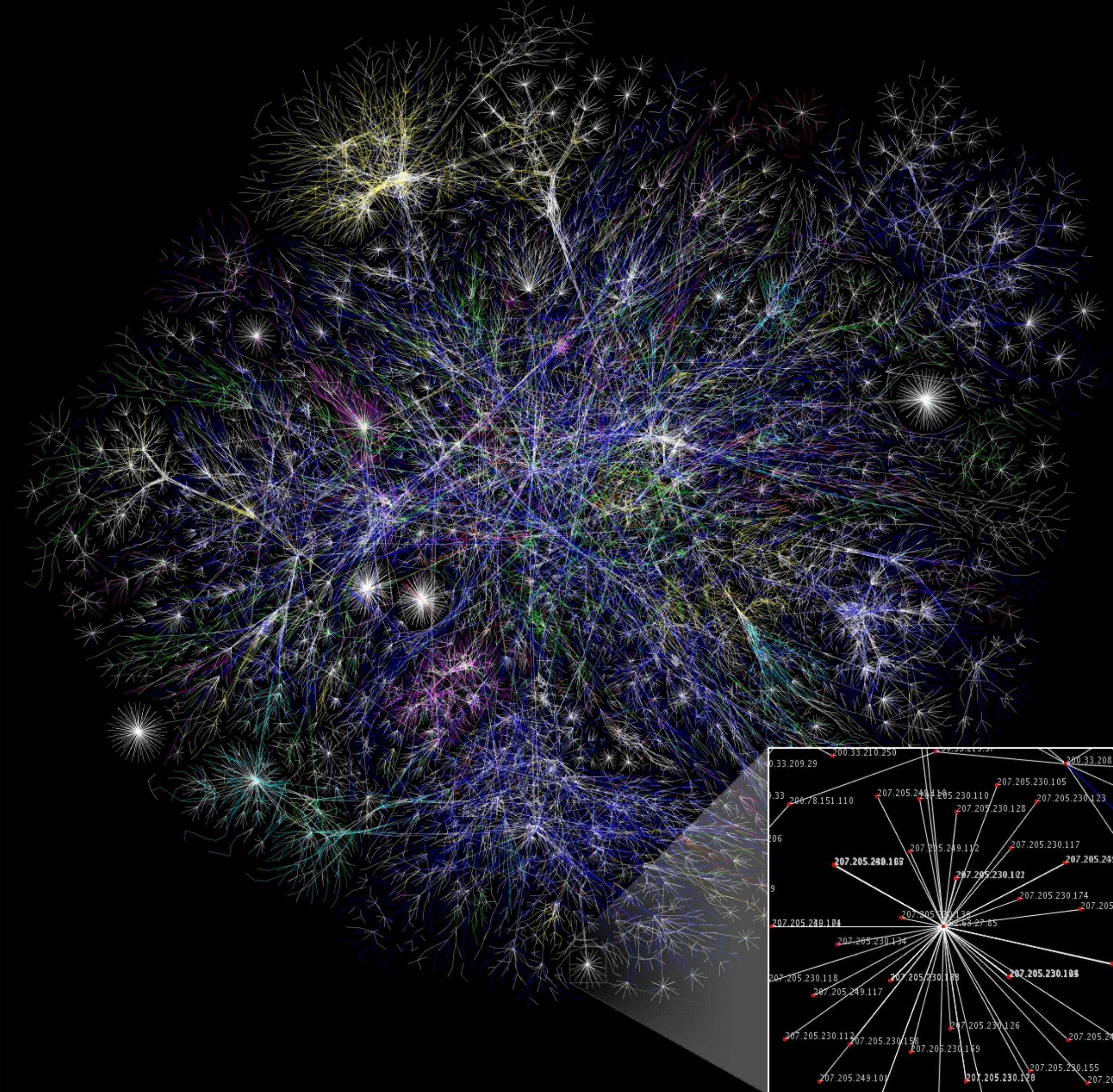
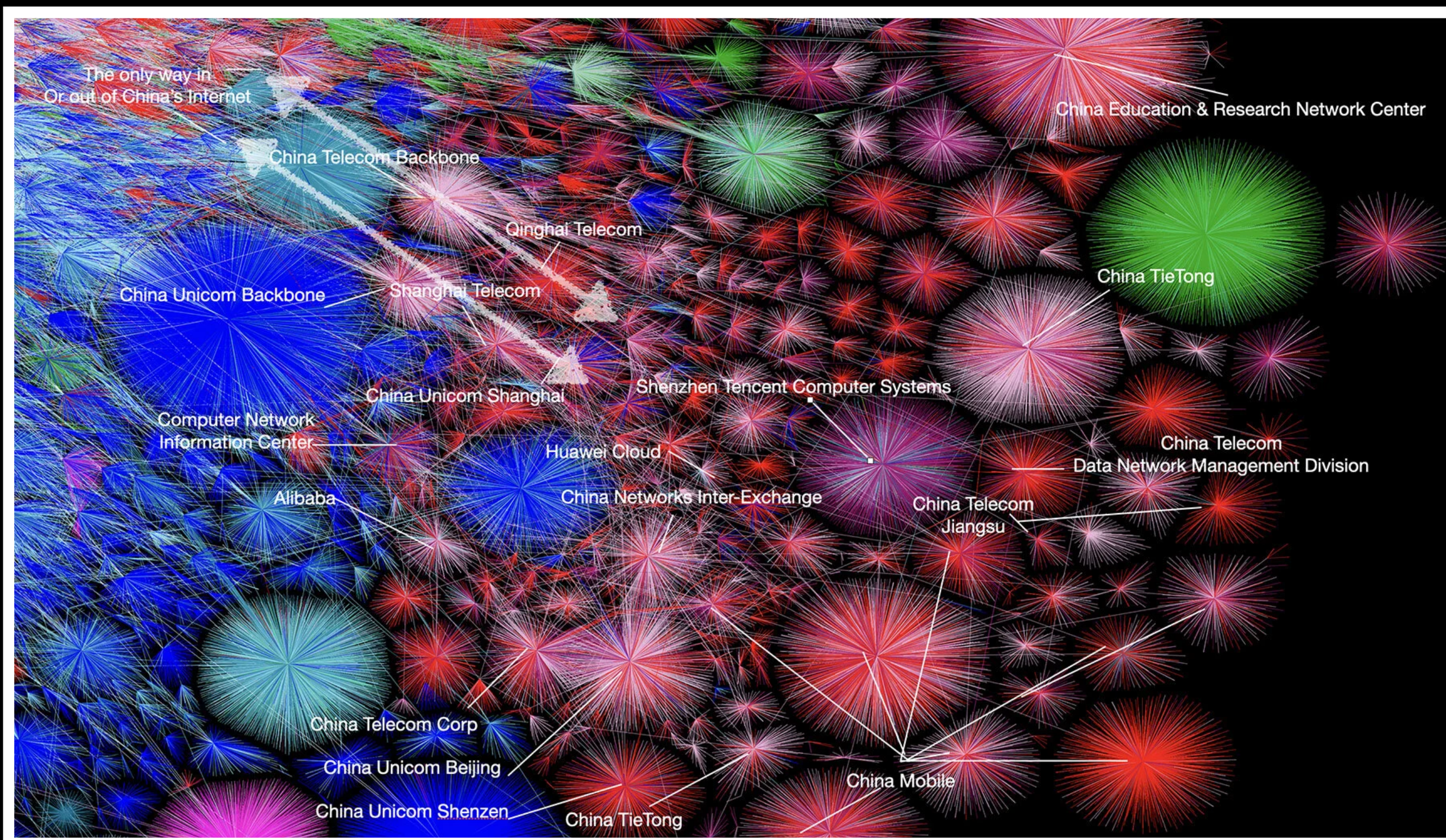
id1	Id2
1	6
6	7
2	7
...	...



Airplanes Network



The Internet *circa* 2005





MENU

APPETIZER

Graphs Refresher

MAIN COURSE

Network Properties

DESSERT

Node Importance



MENU

APPETIZER

Graphs Refresher

MAIN COURSE

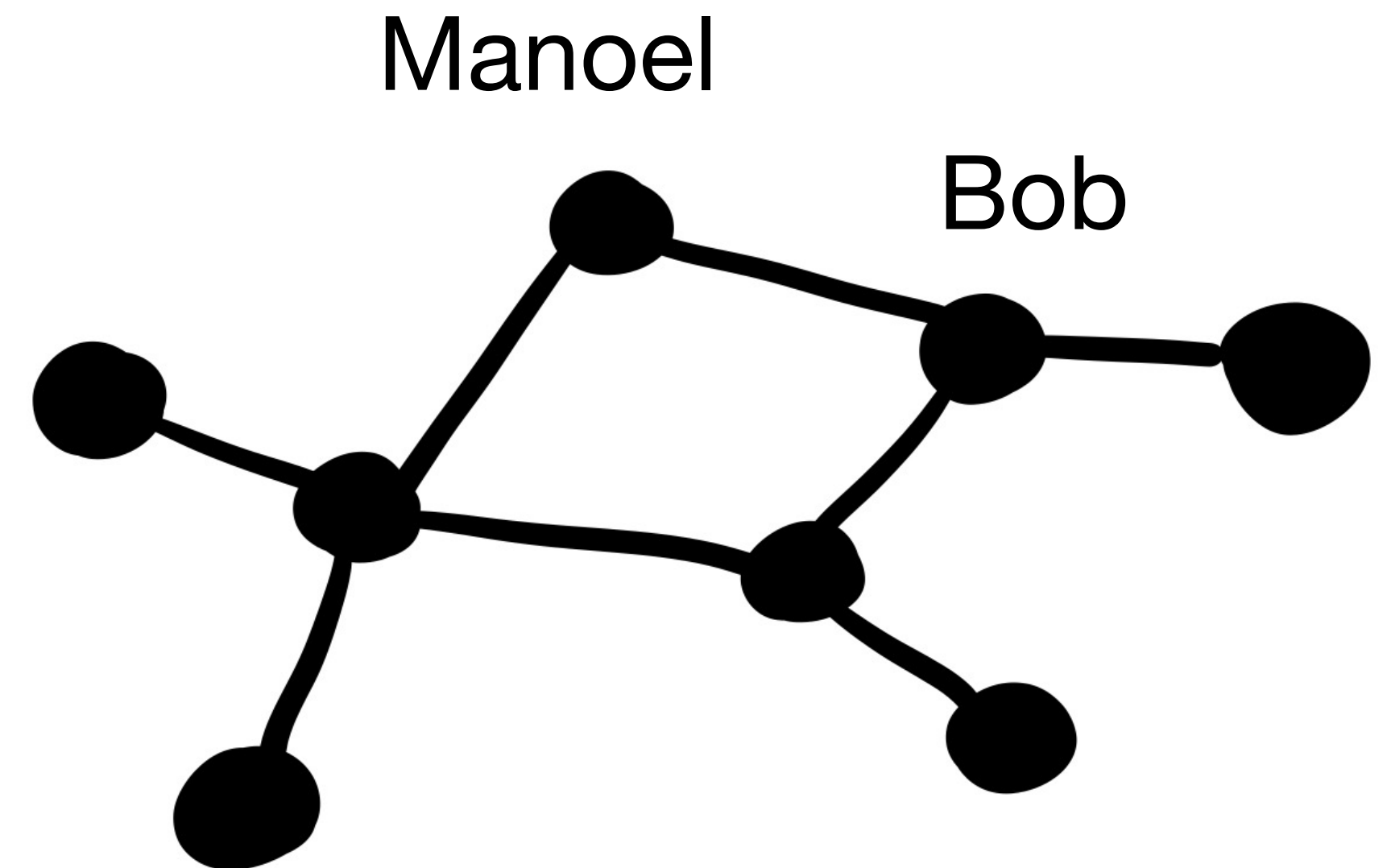
Network Properties

DESSERT

Node Importance

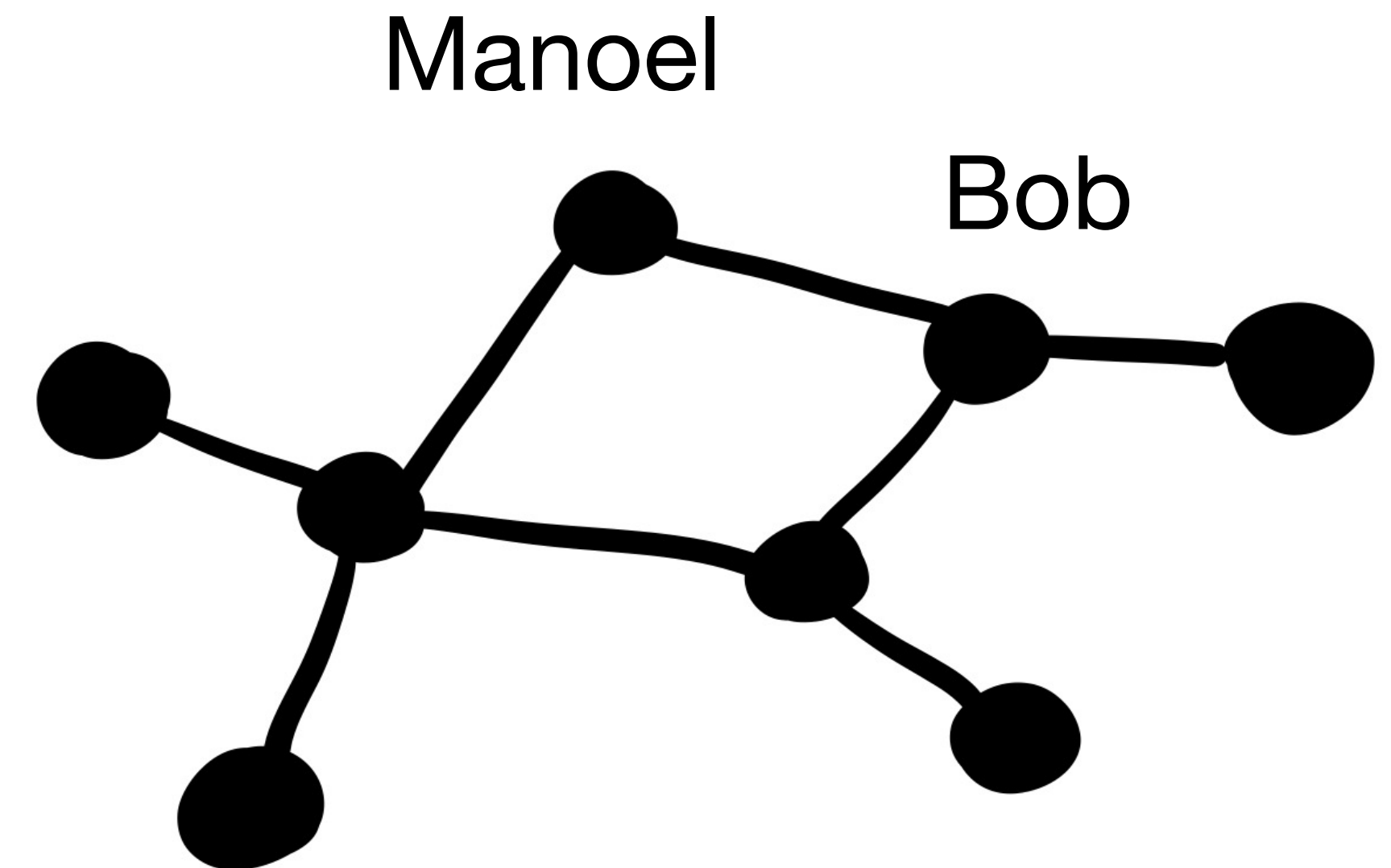
Undirected Graph

- A graph $G = (V, E)$
- **Entities:** Nodes or *Vertices*
- **Relationships:** Links or *Edges*

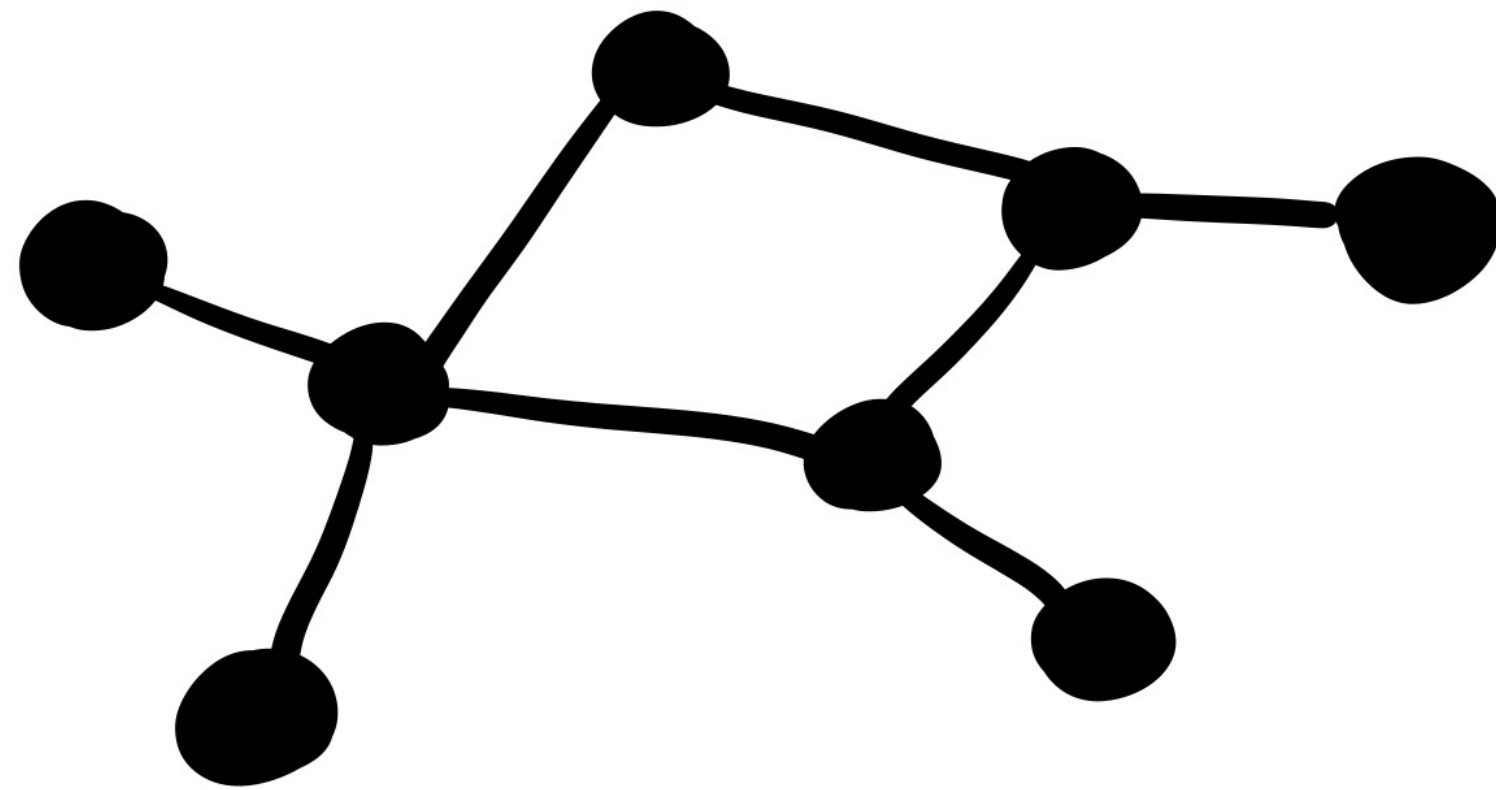


Node Degree ($\deg v$)

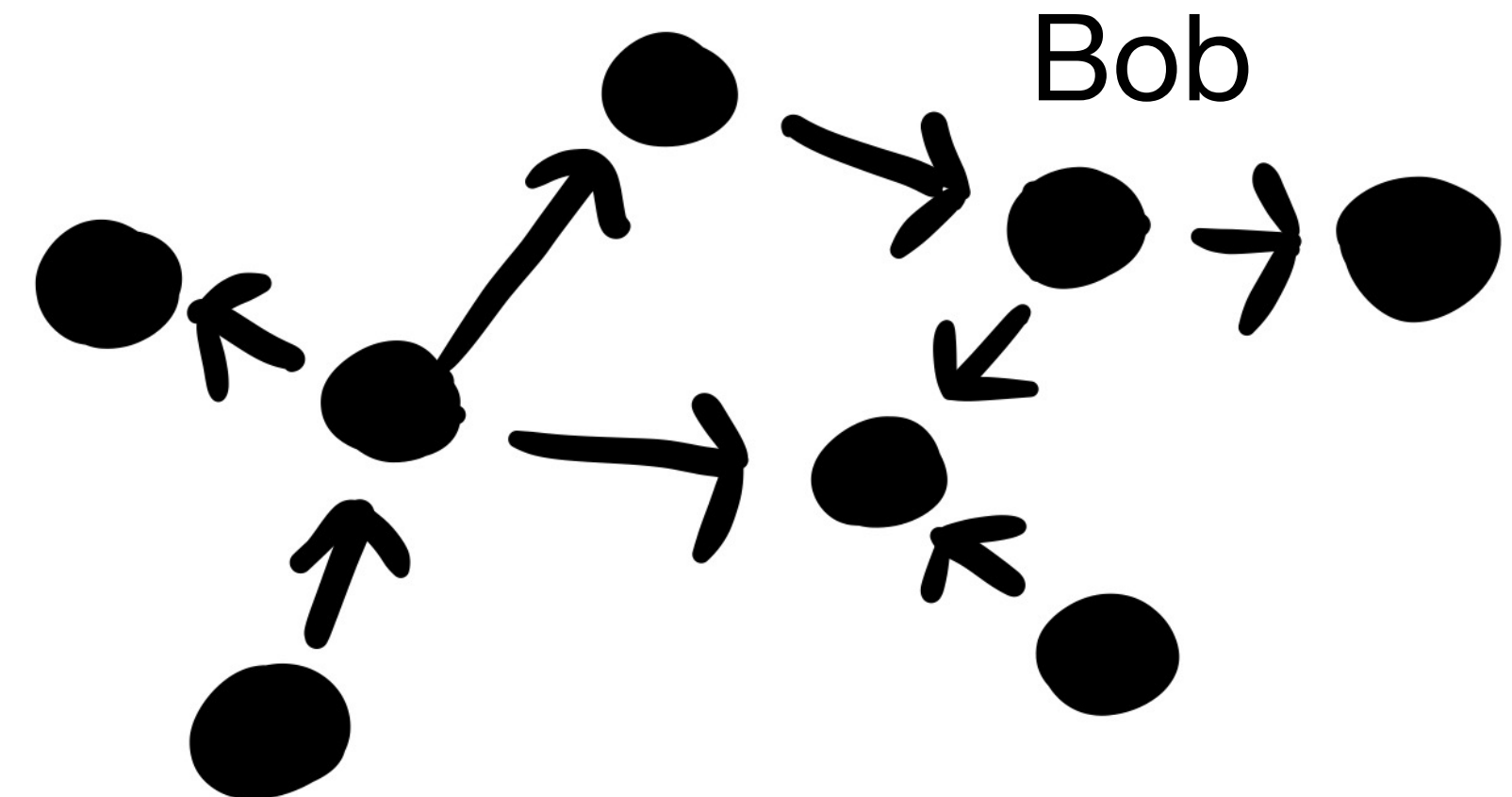
- The number of adjacent edges to a given node.
 - $\deg \text{Manoel} = 2$
 - $\deg \text{Bob} = 2$



Directed Graphs



Instagram Friends

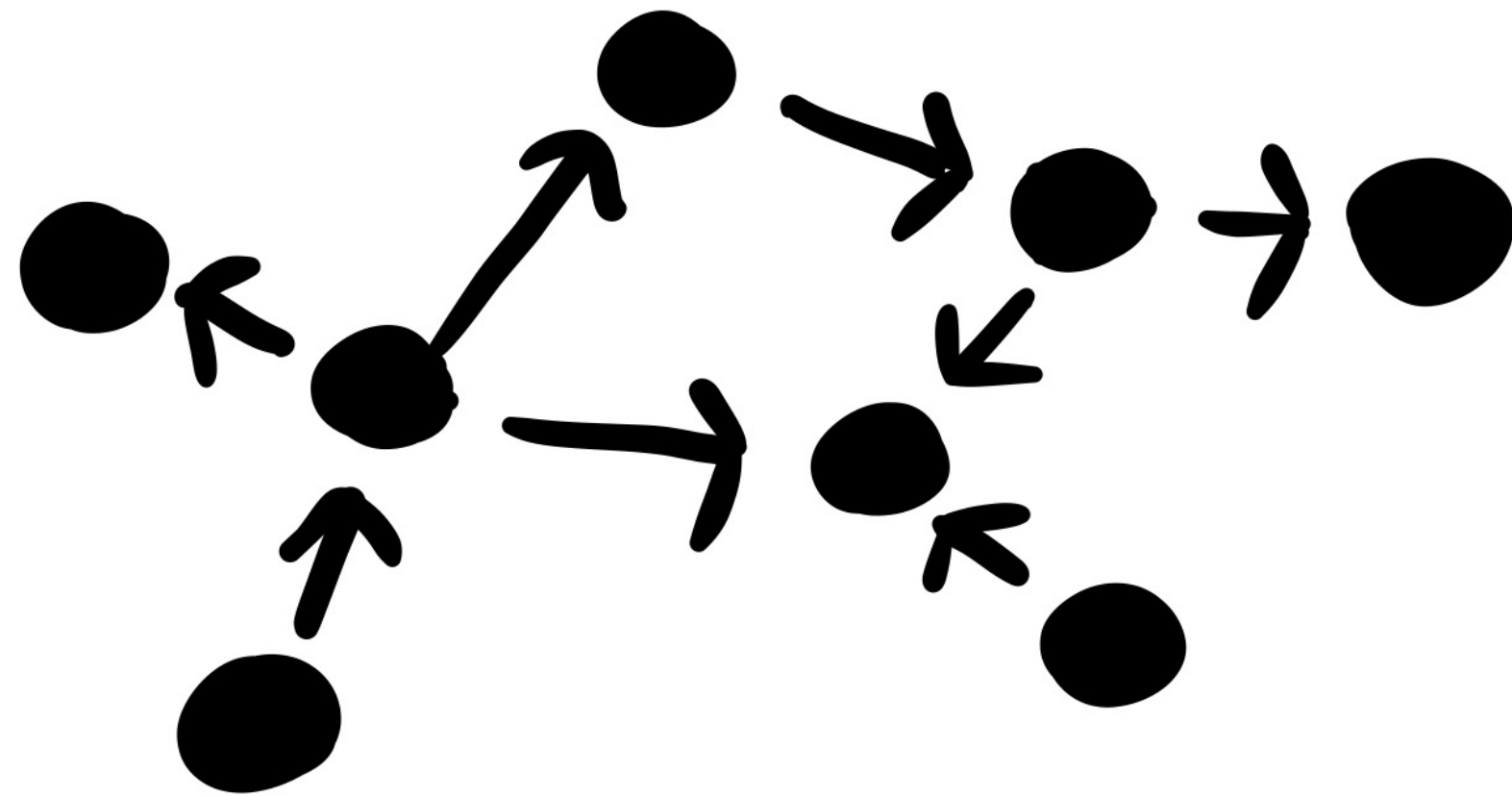


Twitter Followers

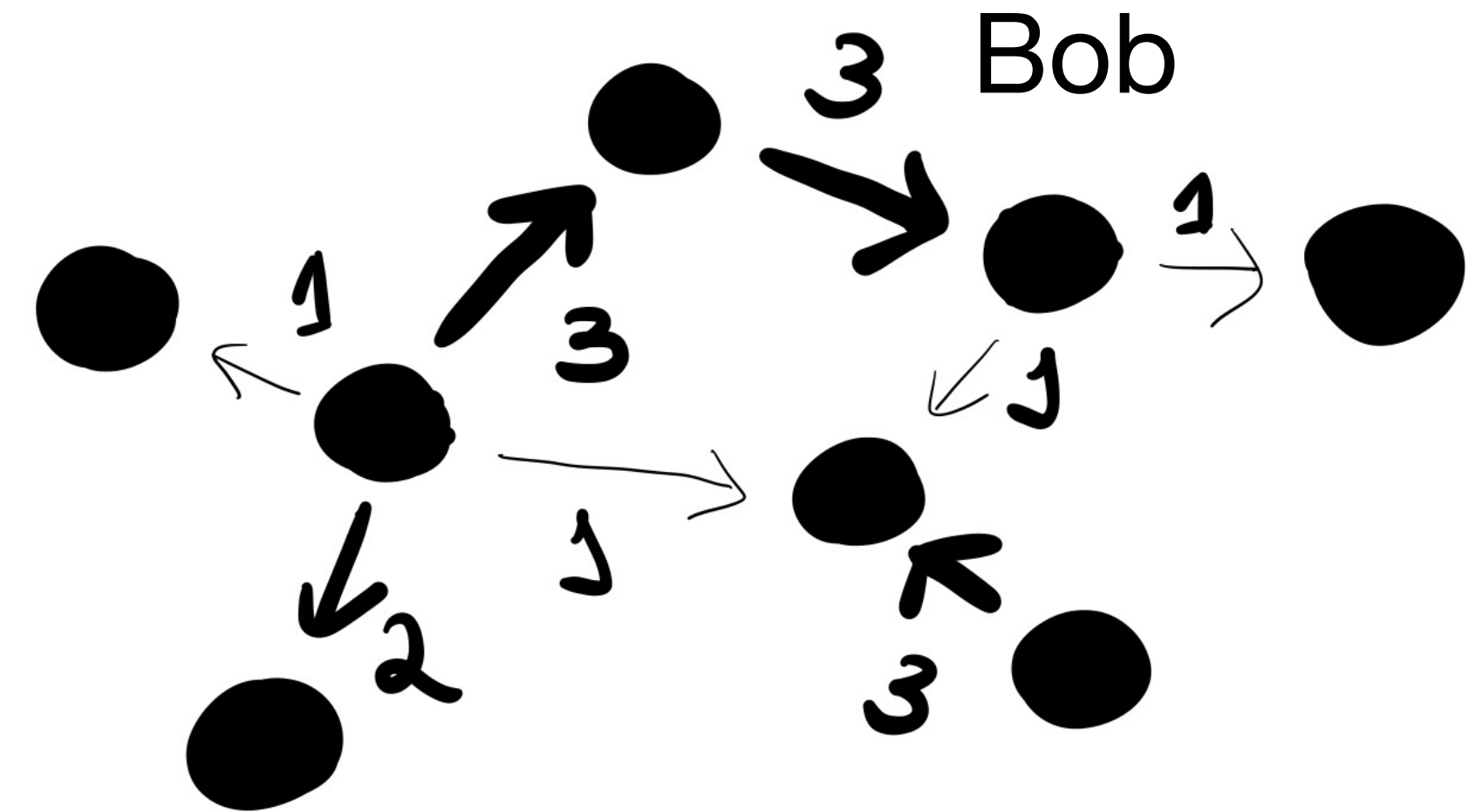
Indegree: $\deg^+ \text{ Bob} = 1$

Outdegree: $\deg^- \text{ Bob} = 2$

Weighted Graphs



Twitter Followers

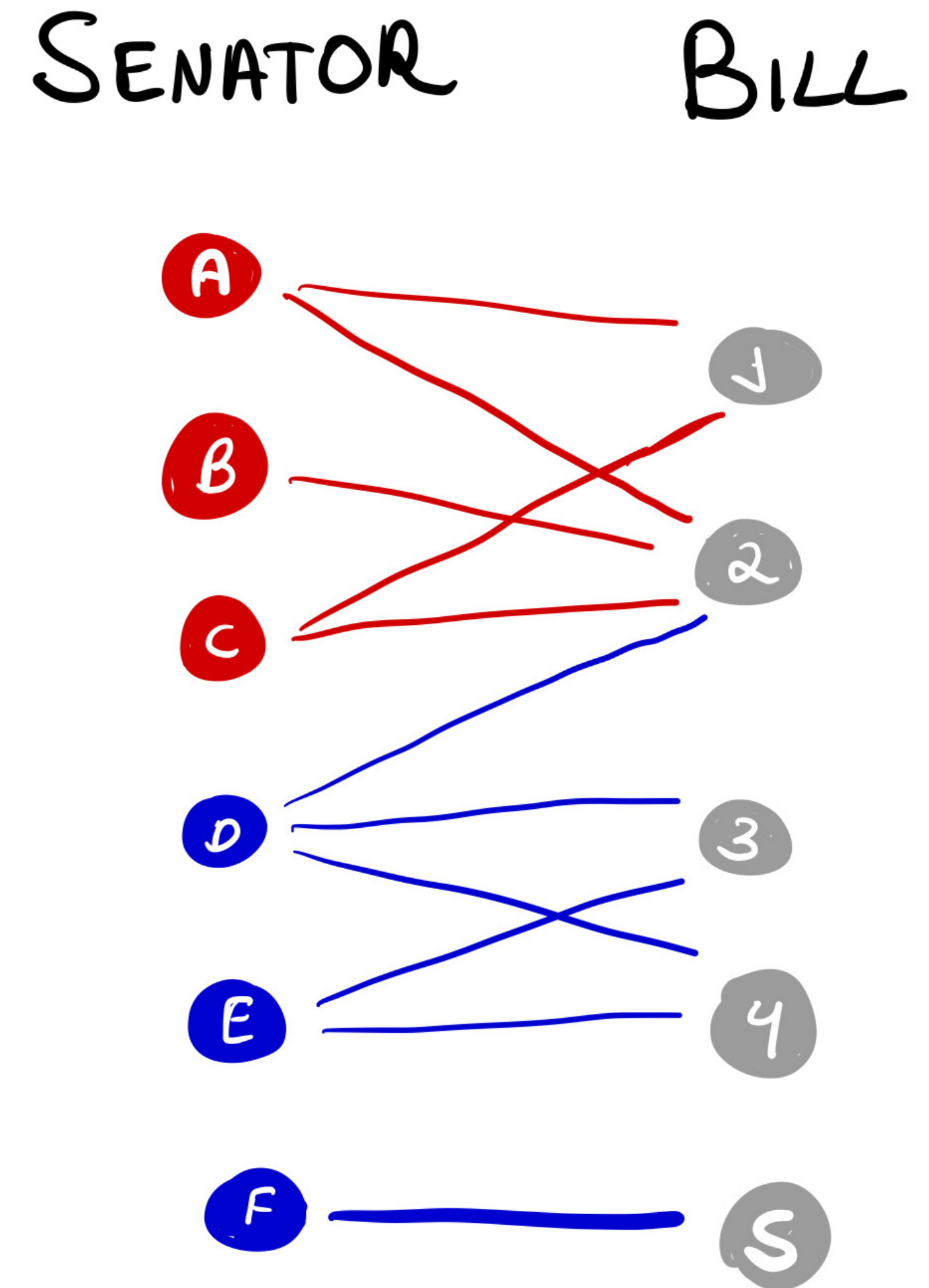


Weekly phone calls

Weighted Indegree: $\deg^+ \text{ Bob} = 3$
Weighted Outdegree: $\deg^- \text{ Bob} = 2$

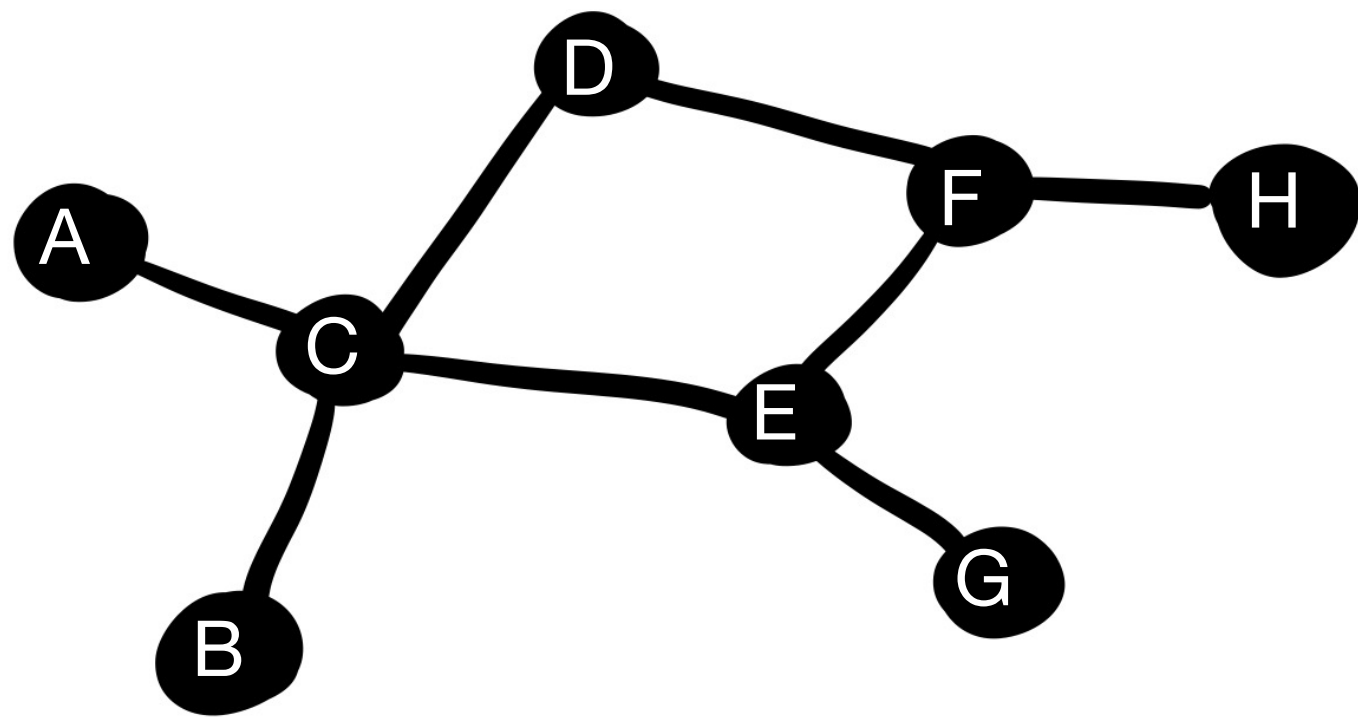
Bipartite Graphs

- Nodes can be divided into sets $\{U, V\}$ s.t. each edge connects a node in U to one in V
- Examples:
 - *Senators and Bills; Edges = support*
 - *Actors and Movies; Edges = appearance*
 - *Users to Products; Edges = ratings*
 - *Recipes to ingredients; Edges = contain*



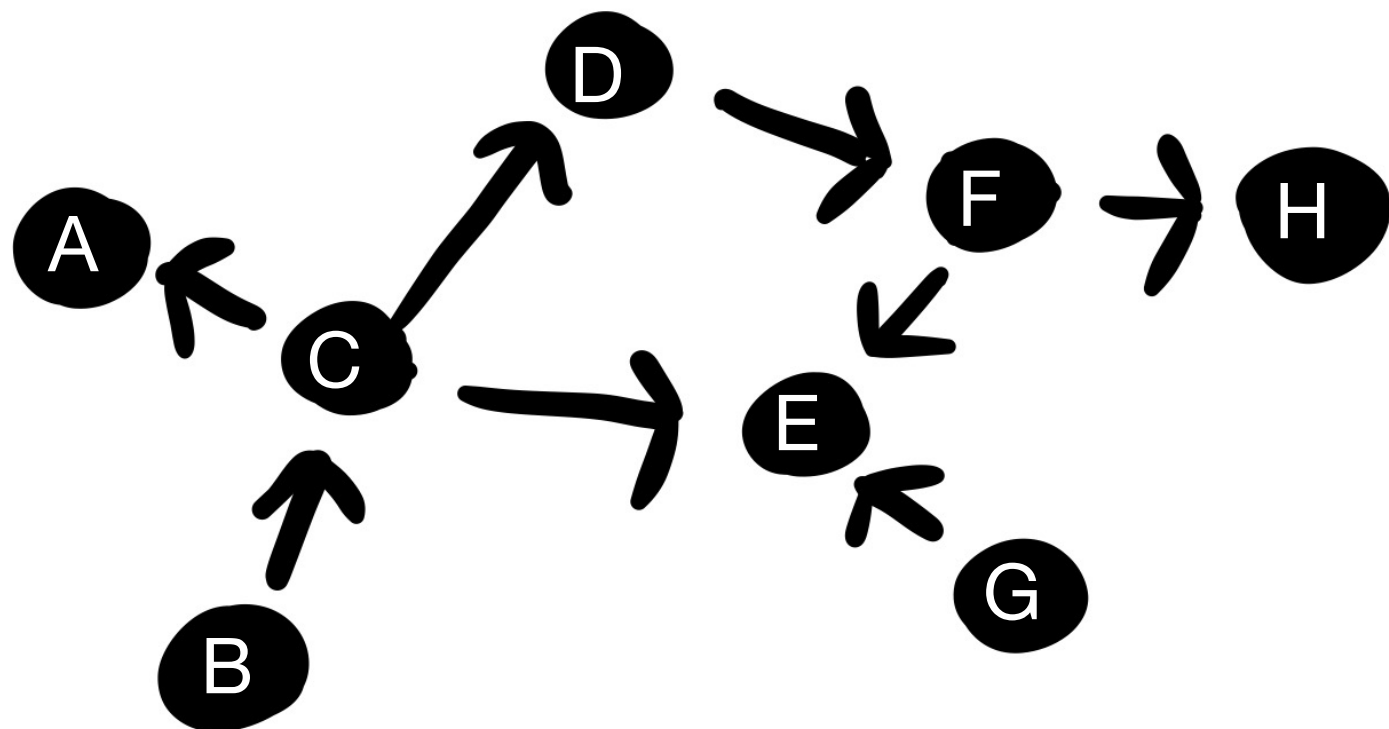
**How to
represent
graphs in
computers?**





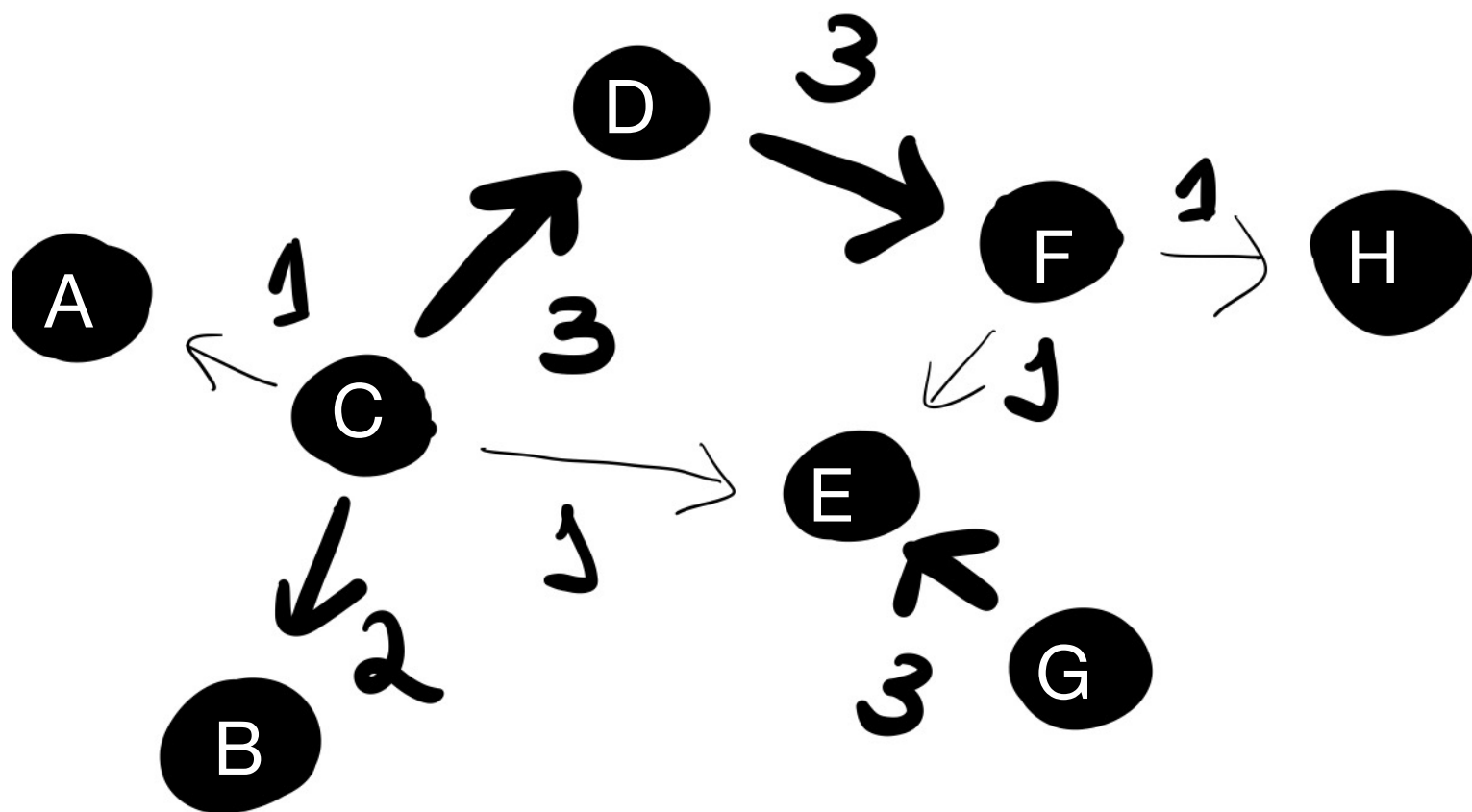
	A	B	C	D	E	F	G	H
A	0	0	1	0	0	0	0	0
B	0	0	1	0	0	0	0	0
C	1	1	0	1	1	0	0	0
D	0	0	1	0	0	1	0	0
E	0	0	1	0	0	1	1	0
F	0	0	0	1	1	0	0	1
G	0	0	0	0	1	0	0	0
H	0	0	0	0	0	1	0	0

[(A,C), (D,F),
 (B,C), (E,F),
 (C,D), (E,G),
 (C,E), (F,H)]



	A	B	C	D	E	F	G	H
A	0	0	0	0	0	0	0	0
B	0	0	0	0	0	0	0	0
C	1	1	0	1	1	0	0	0
D	0	0	0	0	0	1	0	0
E	0	0	0	0	0	0	0	0
F	0	0	0	0	1	0	0	1
G	0	0	0	0	1	0	0	0
H	0	0	0	0	0	0	0	0

[(**C,A**), (D,F),
 (B,C), (**F,E**),
 (C,D), (**G,E**),
 (C,E), (F,H)]



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

[(C,A,1), (D,F,3),
(C,B,2), (F,E,1),
 (C,D,3), (G,E,3),
 (C,E,1), (F,H,1)]

Adjacency Matrix

Edge List

Space

$$O(V^2)$$

$$O(V + E)$$

Edge lookup

$$O(1)$$

$$O(E)$$

Add vertex

$$O(V^2)$$

$$O(1)$$

Add edge

$$O(1)$$

$$O(1)$$

Graphs are a Modeling Tool

- **Wrong mindset:** Data is naturally represented as a graph of type X.
- **Correct mindset:** Modeling data as a graph of type X lets me do Y
 - *better understand the data*
 - *test a specific hypothesis*
 - *Make predictions*





MENU

APPETIZER

Graphs Refresher

MAIN COURSE

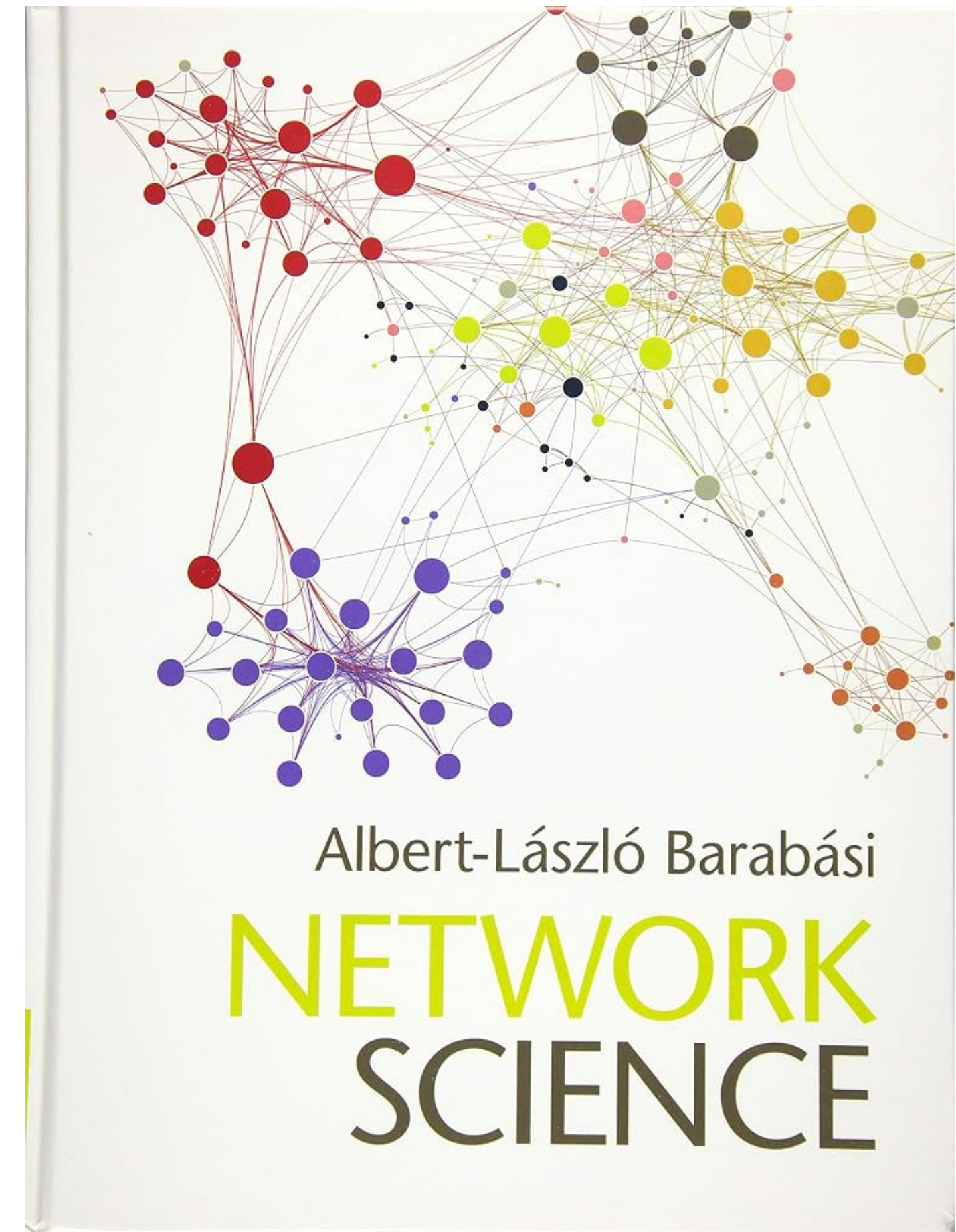
Network Properties

DESSERT

Node Importance

Real World Graphs

- Real-world networks tend to share certain properties!
- Remarkable, given the diversity of networks
- “Network Science”



Sparsity

- Each node is connected to only a small fraction of the other nodes!
 - *Why? Because connections often require continuous effort!*
- **Dunbar's number:** the cognitive limit to how many stable relationships we can maintain!
 - *Humans > Monkeys > Apes*

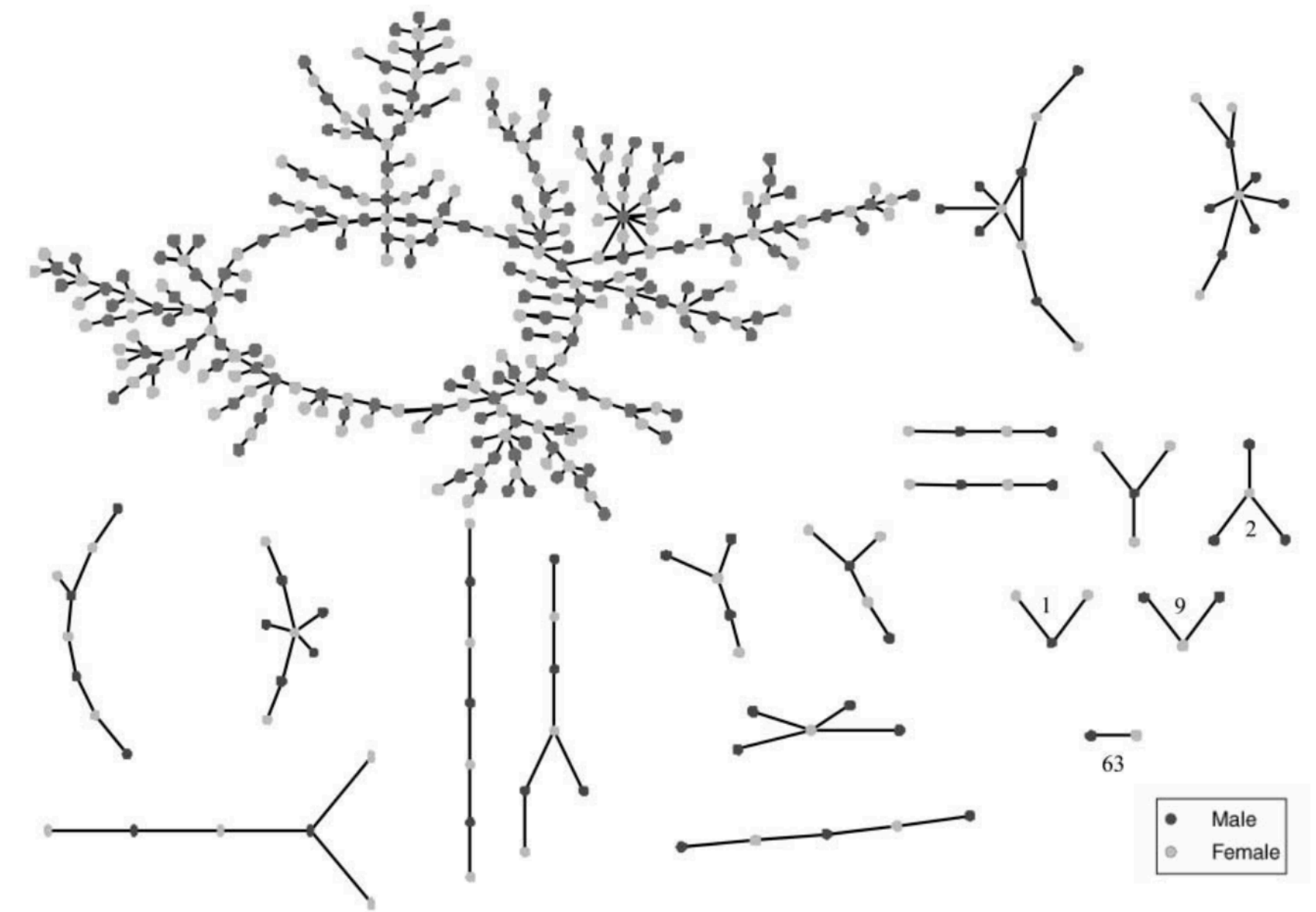
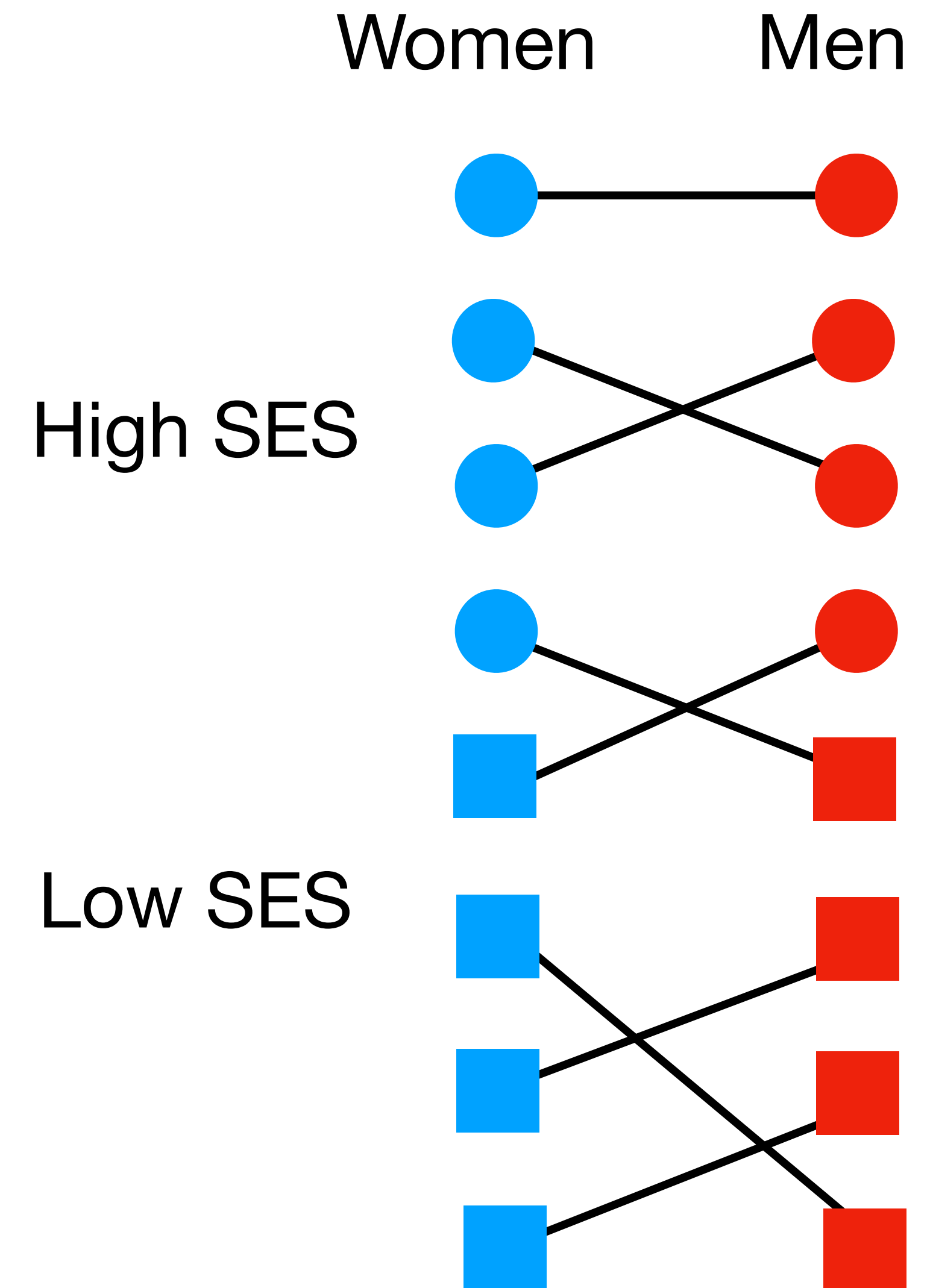


FIG. 2.—The direct relationship structure at Jefferson High

Bearman, Peter S., James Moody, and Katherine Stovel.
"Chains of affection: The structure of adolescent romantic
and sexual networks." *American journal of sociology* 110.1
(2004): 44-91.

Homophily

- Connections typically happen with nodes that are more similar!
- In the romantic networks described, dating was highly related to socioeconomic status
 - *Proportion of pairs with matching SES was 30% higher than for random pairs*



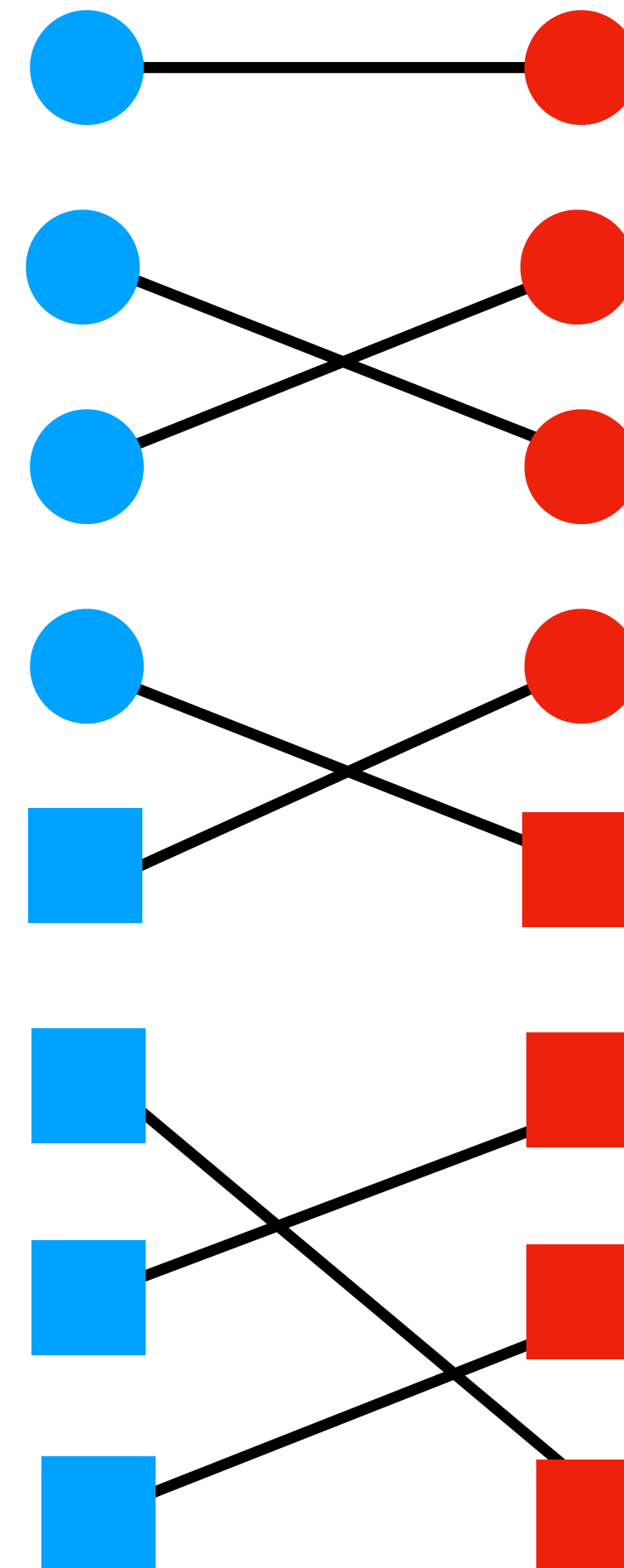
Measuring it!

- Expected edge type:
 - $A-A$: 2
 - $A-B$: 4
 - $B-A$: 2
- Alternative: randomize!

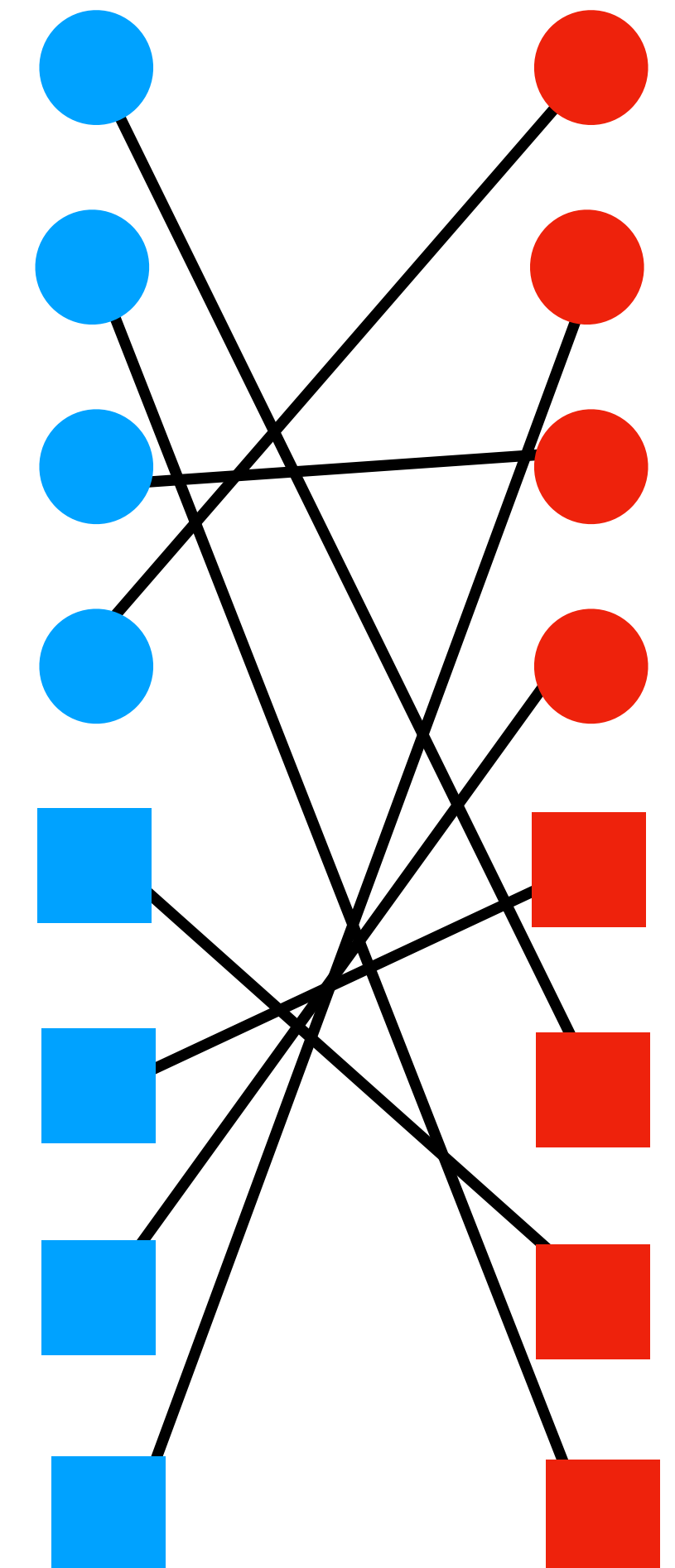
●● High SES

■ Low SES

Women Men



Women Men



Homophily vs. Influence

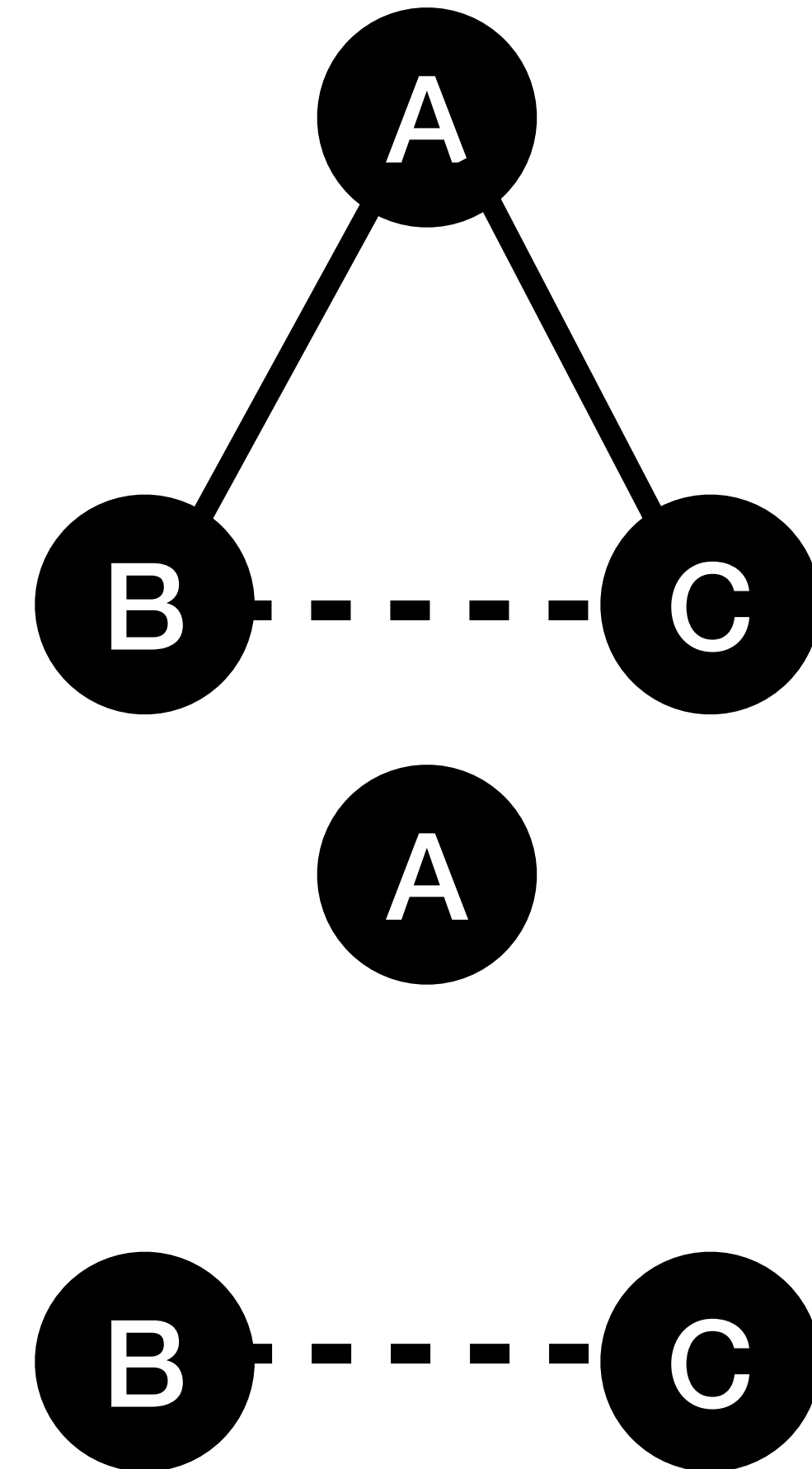
- Connections can also make nodes more similar!
- Are you a heavy metal fan because your friends are? Or did you befriend them because they like the same music as you?
 - *Probably a bit of both!*



Triadic Closure

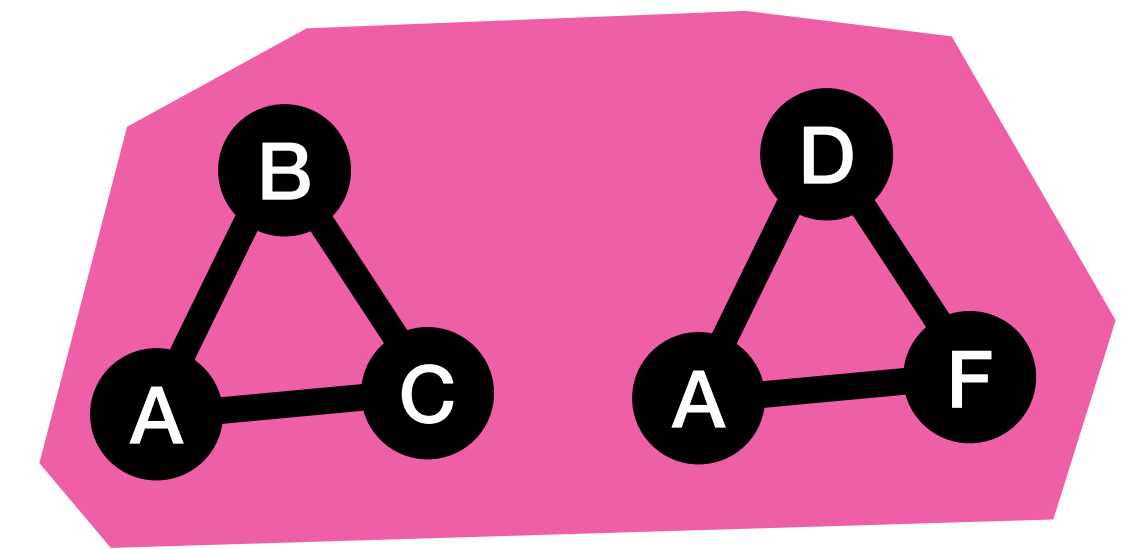
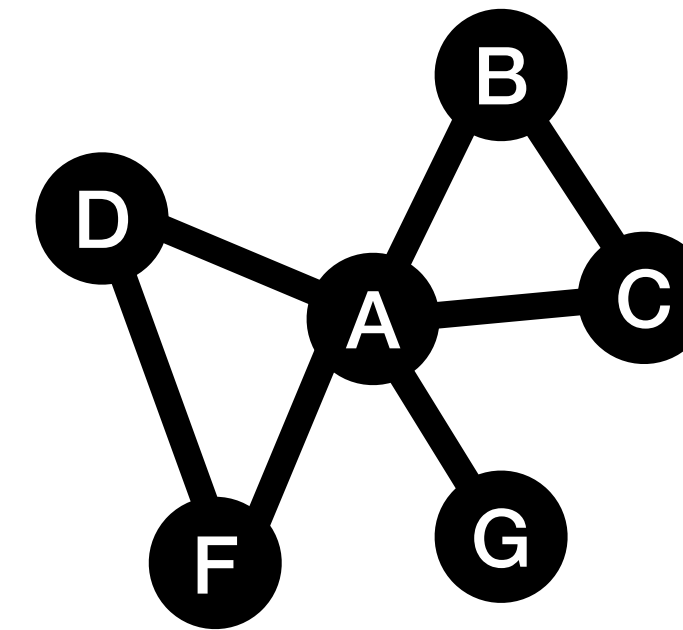
If nodes *B* and *C* share a mutual connection *A*, then they are likely to be connected!

- *Why? Because connections often require continuous effort!*

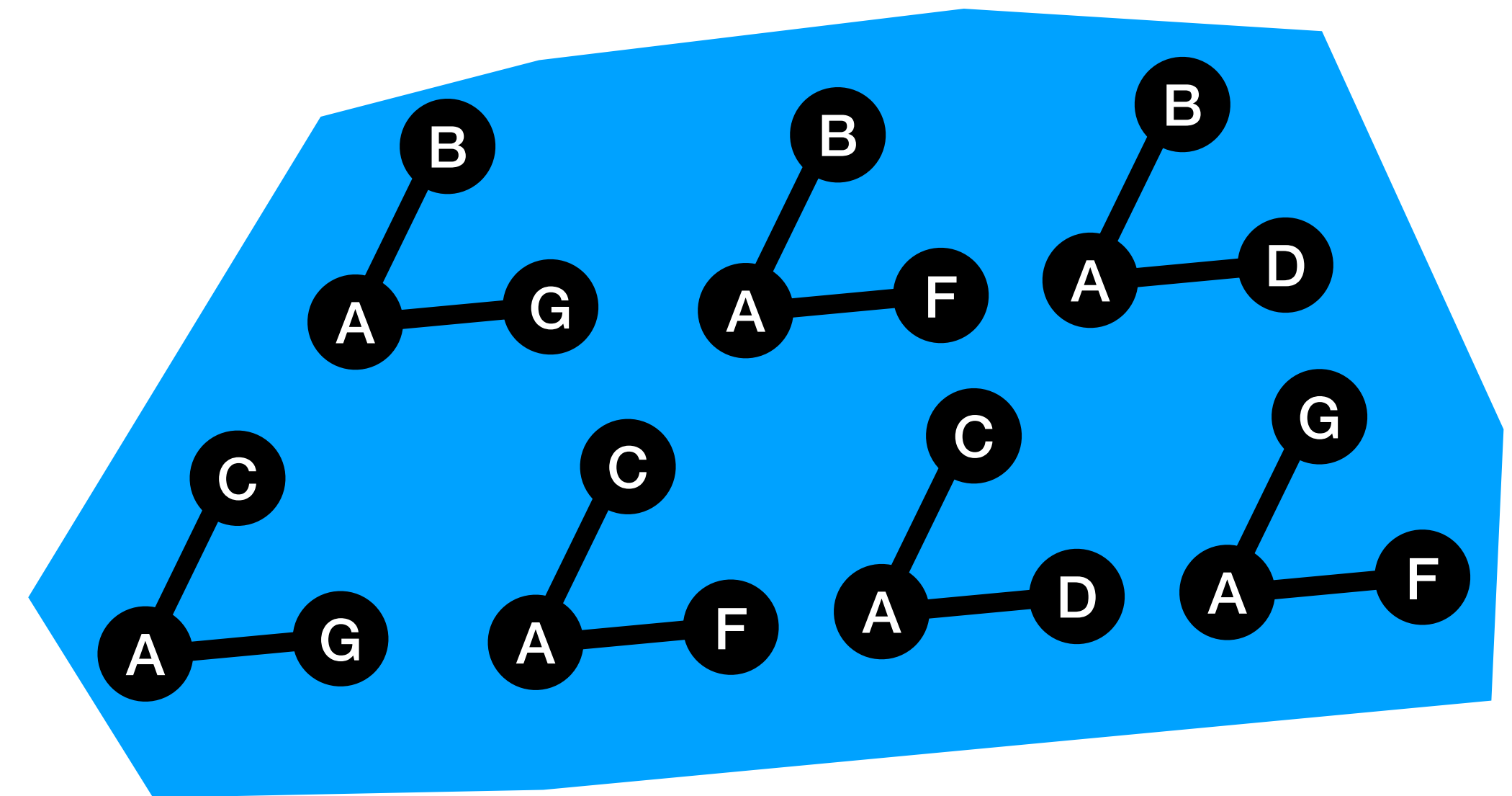


Clustering Coefficient

For a given node, what percentage of their neighbors are connected?

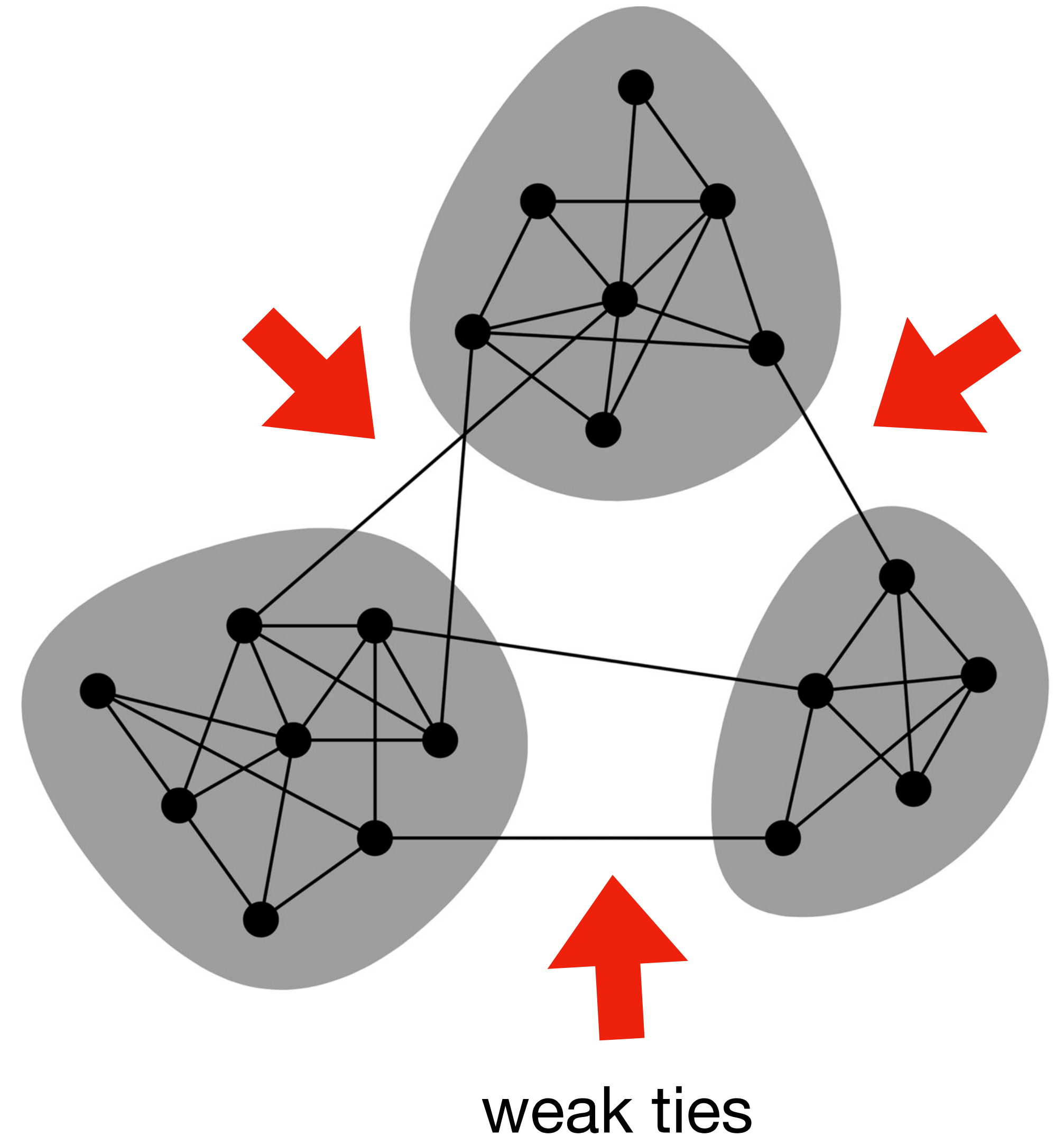


For a graph, the average clustering coefficient measures the extent to which



Community Structure

- Triadic closure makes networks cluster into locally dense communities
- Communities are connected via “weak ties”
- “The strength of weak ties” (Granovetter 1973), e.g, they disproportionately help job seekers!



Small worldness

- What is the typical shortest path length between two people?
- Small-world experiment (Milgram '67)
 - Found *300 people in Omaha and Wichita.*
 - *Goal: Get a letter to a stockbroker in Boston by mailing friends*
- Surprising finding: 6 steps!



Wikispeedia

This game is easy and fun:

- You are given two Wikipedia articles* (or you choose two yourself).
- Starting from the first article, your goal is to reach the second one, exclusively by following links in the articles you encounter. (For the articles you are given this is always possible.)
- Links you can take are colored like [this](#).
- Of course, it's more fun if you try to be as quick as possible...
- Next to wasting some precious time and learning interesting yet useless Wikipedia facts, you're also providing Bob (west@cs.mcgill.ca) with data for his [research project](#).

* The articles have been borrowed from the 4,600-article CD version of Wikipedia available at schools-wikipedia.org (version of 2007).

Let's see if I can find a path from **Banjul** to **Yellow River**

(Click on the target word if you don't know what it means...)

I'd rather find a path from to

Games played by others:

Mission	Clicks	Avg. hardness**	
Argyle diamond mine >> California Southern Railroad	avg. 9, rec. 9	5	Play!
Ununquadium >> Rainforest	avg. 6, rec. 6	-	Play!
Indian Peafowl >> Interpol	avg. 5, rec. 5	3	Play!
Liberia >> Reza Shah	avg. 5, rec. 5	-	Play!
Japanese war crimes >> Nile	avg. 4.3, rec. 4	1	Play!

** On a scale from 1 (easy) to 5 (brutal); if you want to make this more reliable, just hit the *rate* button after each game.



Bob West

<https://dlab.epfl.ch/wikispeedia/play/>



MENU

APPETIZER

Graphs Refresher

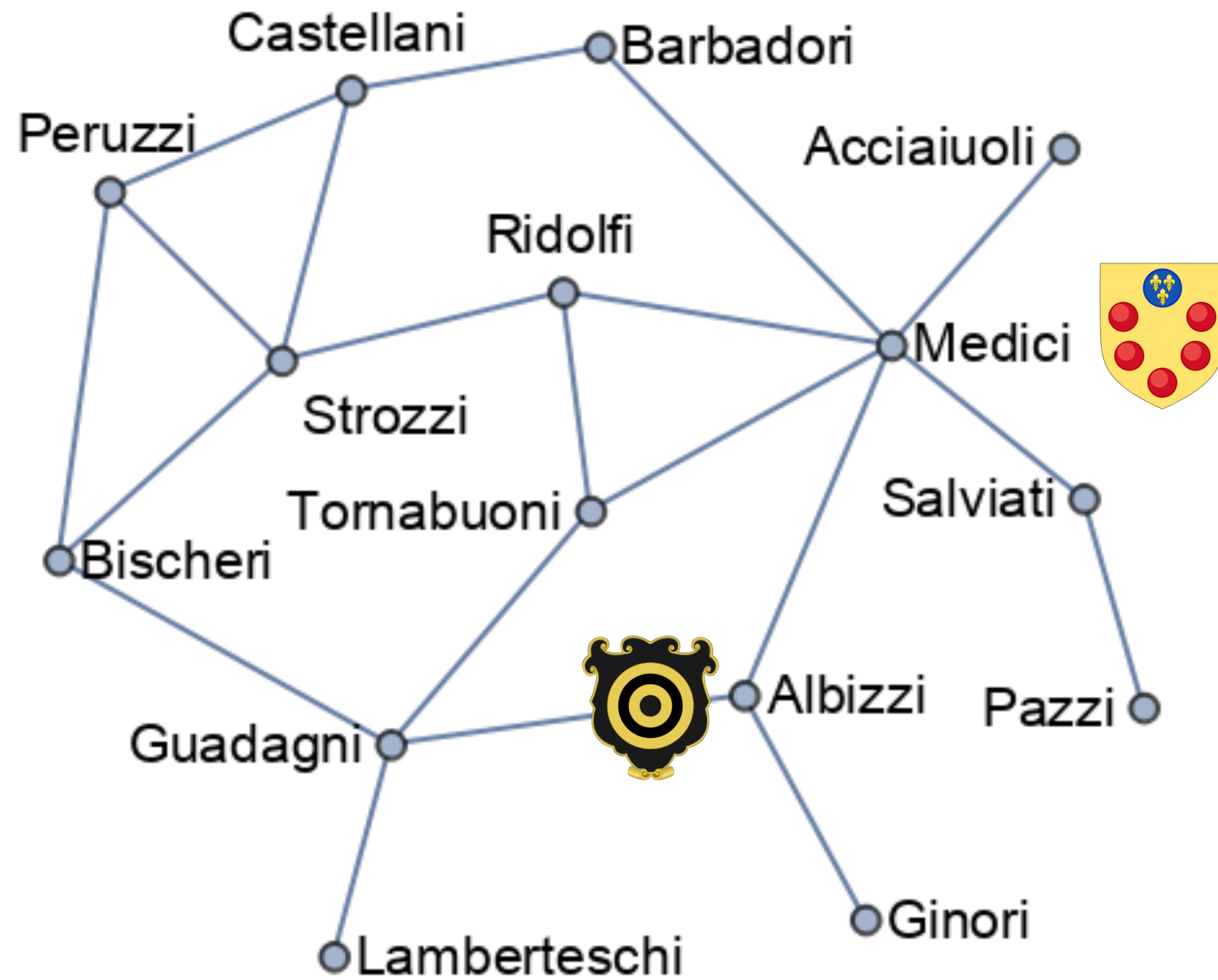
MAIN COURSE

Network Properties

DESSERT

Node Importance

Florentine Marriages Network



Robust Action and the Rise of the Medici, 1400–1434¹

John F. Padgett and Christopher K. Ansell
University of Chicago

We analyze the centralization of political parties and elite networks that underlay the birth of the Renaissance state in Florence. Class revolt and fiscal crisis were the ultimate causes of elite consolidation, but Medicean political control was produced by means of network disjunctures within the elite, which the Medici alone spanned. Cosimo de' Medici's multivocal identity as sphinx harnessed the power available in these network holes and resolved the contradiction between judge and boss inherent in all organizations. Methodologically, we argue that to understand state formation one must penetrate beneath the veneer of formal institutions, groups, and goals down to the relational substrata of peoples' actual lives. Ambiguity and heterogeneity, not planning and self-interest, are the raw materials of which powerful states and persons are constructed.

INTRODUCTION

Regardless of time or place, political centralization lies at the heart of state building. Less widely appreciated is the fact that the process of centralization is contradictory: its agents are forced to seek both reproduction and control. Centralization occurs, often abruptly, when founders emerge out of the soup of contending actors to establish (perhaps unintentionally) new rules for others' interaction. Reproduction ensues when rules induce roles, which induce interests, which induce strategic exchanges, which lock in patterns of collective action that depend on the

¹ Our colleague Paul McLean is a full joint participant in the larger project out of which this paper has been drawn. His help has been invaluable. We would also like to thank Wayne Baker, Ronald Breiger, Gene Brucker, Michael Cohen, Samuel Cohn, Walter Fontana, Mark Granovetter, the late David Greenstone, Wendy Griswold, the late David Herlihy, Alex Hicks, Ian Lustick, Charles Perrow, Tony Tam, Charles Tilly, and participants in the University of Chicago's Organizations and State-Building Workshop, the New School's "think and drink" seminar, and the Santa Fe Institute's Adaptive Organizations Conference for their many helpful comments. This article is dedicated to the memory of David Herlihy, whose quantitative research on Renaissance Florence made work like this possible.

© by The University of Chicago. All rights reserved.
0002-9602/90/9806-0001\$01.50

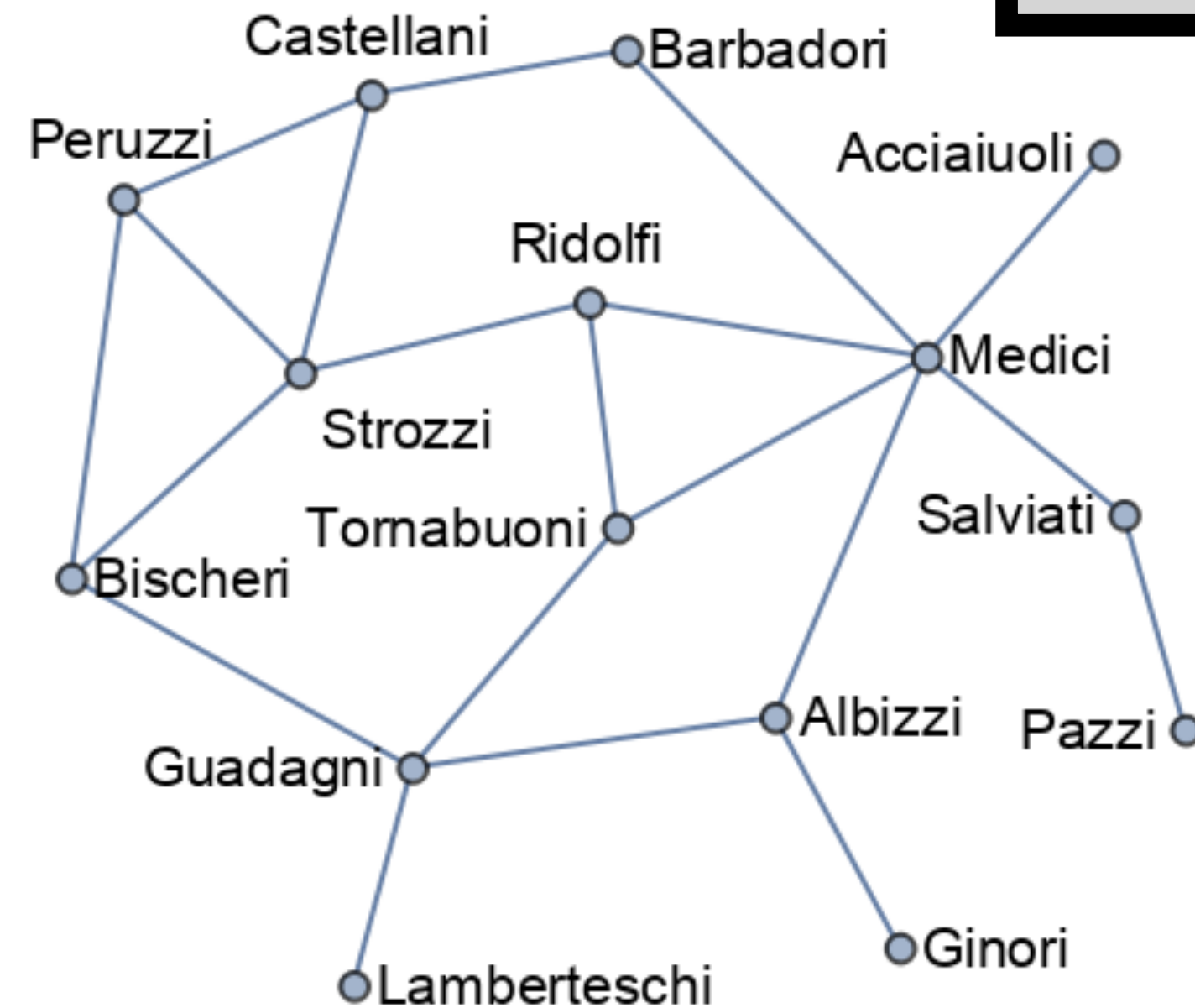
AJS Volume 98 Number 6 (May 1993): 1259-1319 1259

(Link)

Degree Centrality

- **Simple idea:** important nodes have more edges!

$$C_D(v) = \frac{\deg(v)}{N - 1}$$



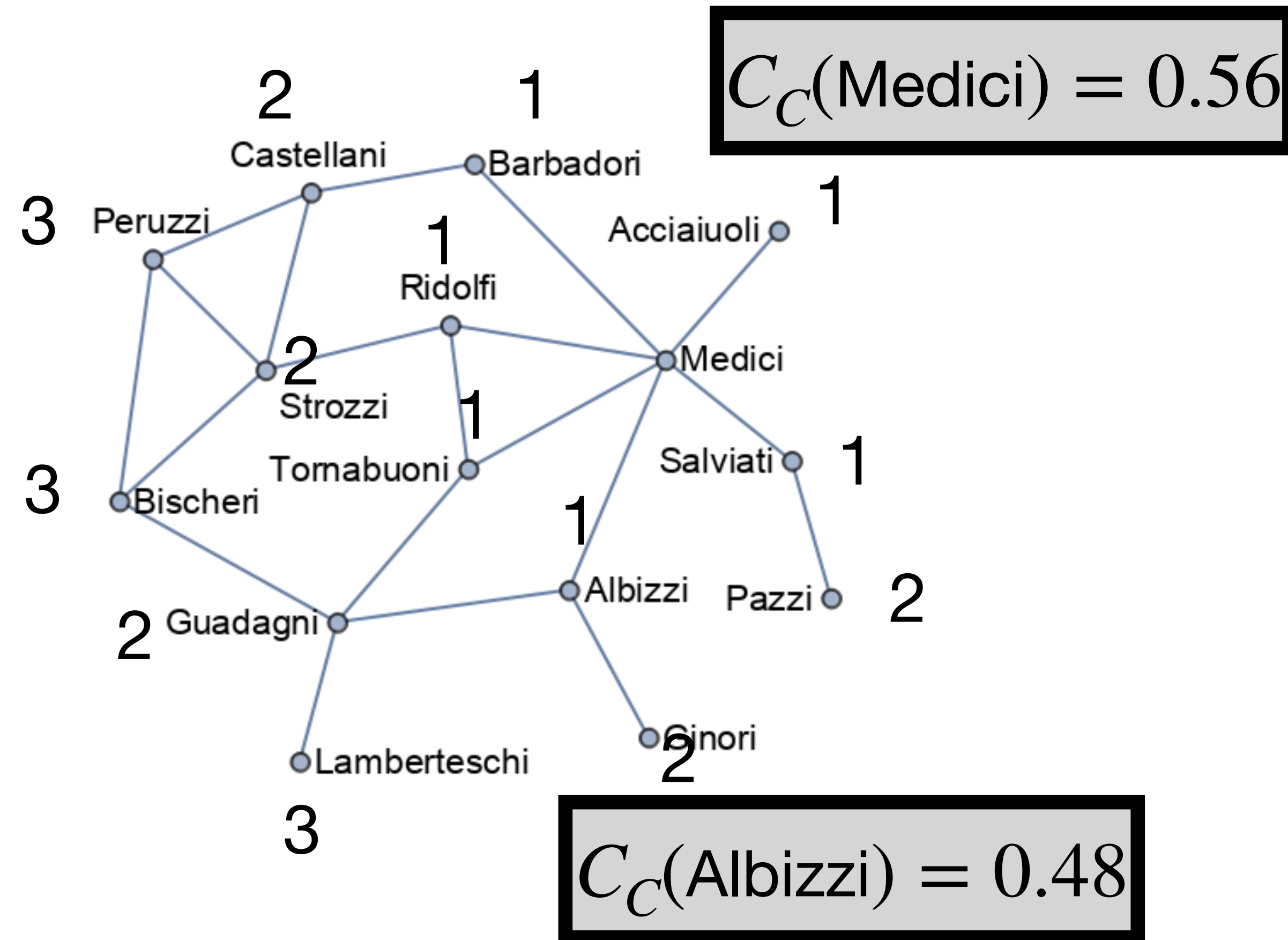
$$C_d(\text{Medici}) = 0.42$$

$$C_d(\text{Albizzi}) = 0.21$$

Closeness Centrality

- Sum of the shortest paths between a node and every other node!

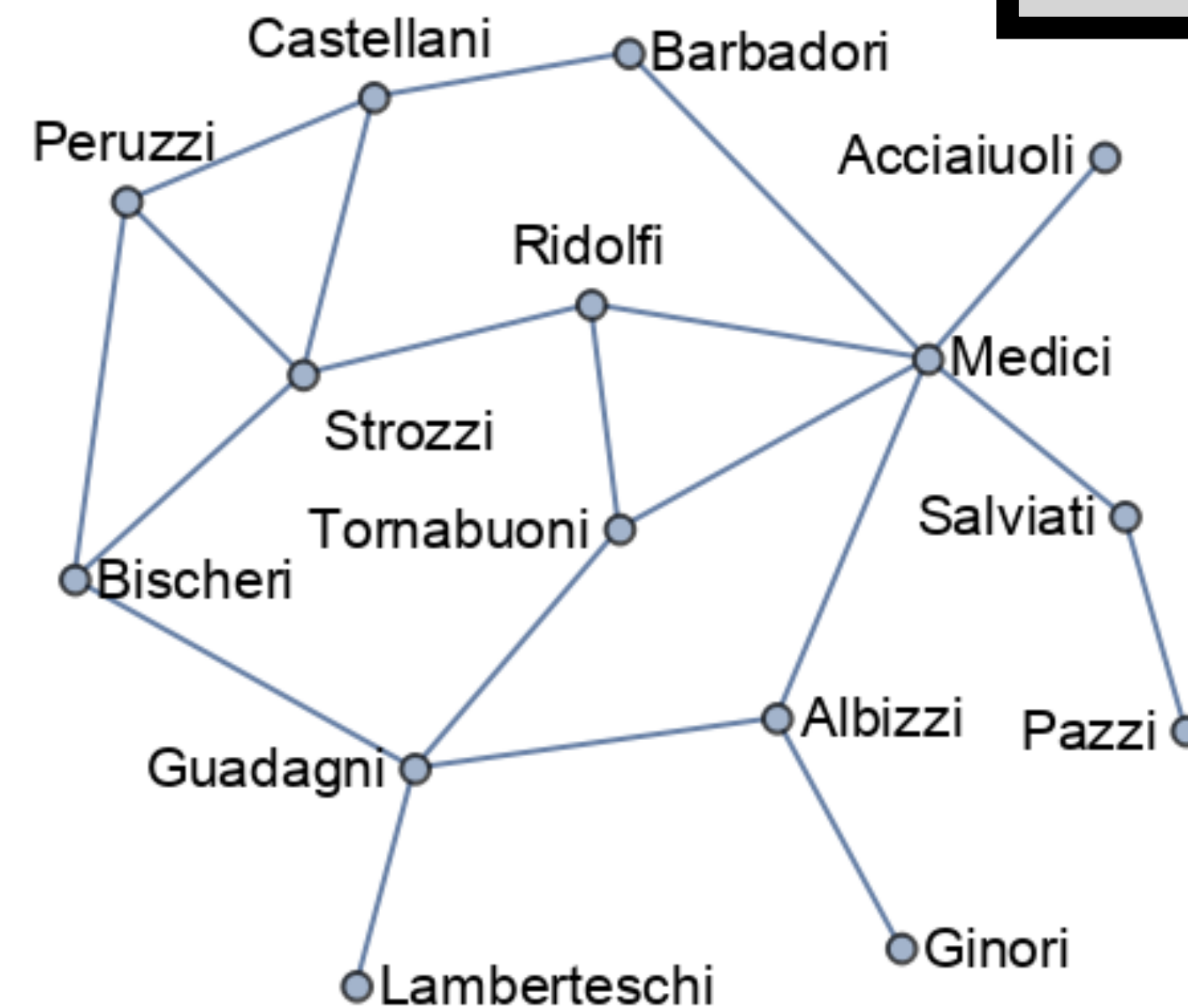
$$C_C(v) = \frac{N - 1}{\sum_u d(v, u)}$$



Betweenness Centrality

- Let σ_{st} be the number of shortest paths between pairs of nodes (s, t)
- Let $\sigma_{st}(v)$ be the numbers shortest paths that pass through v

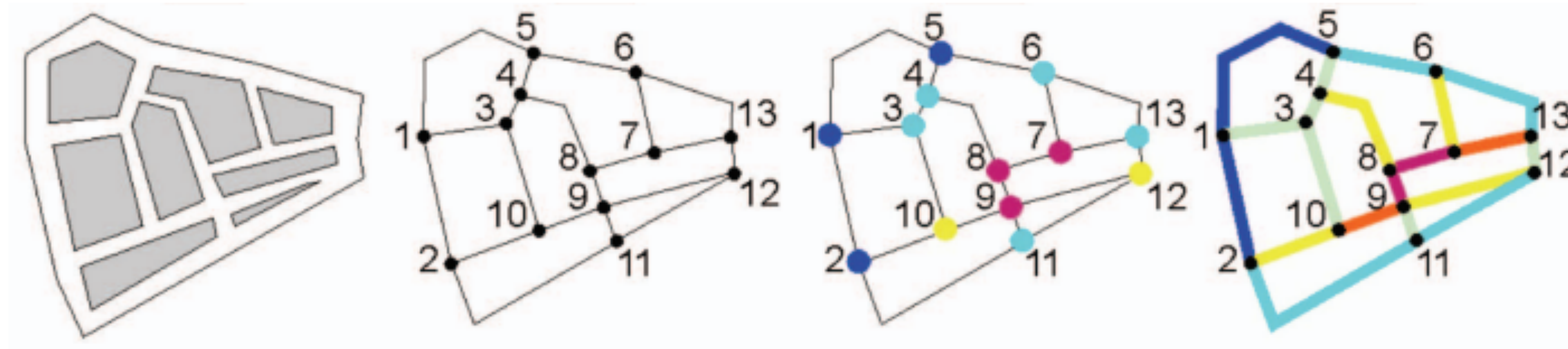
$$C_B(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}}$$



$$C_B(\text{Medici}) = 0.52$$

$$C_B(\text{Albizzi}) = 0.21$$

Example: Urban Planning



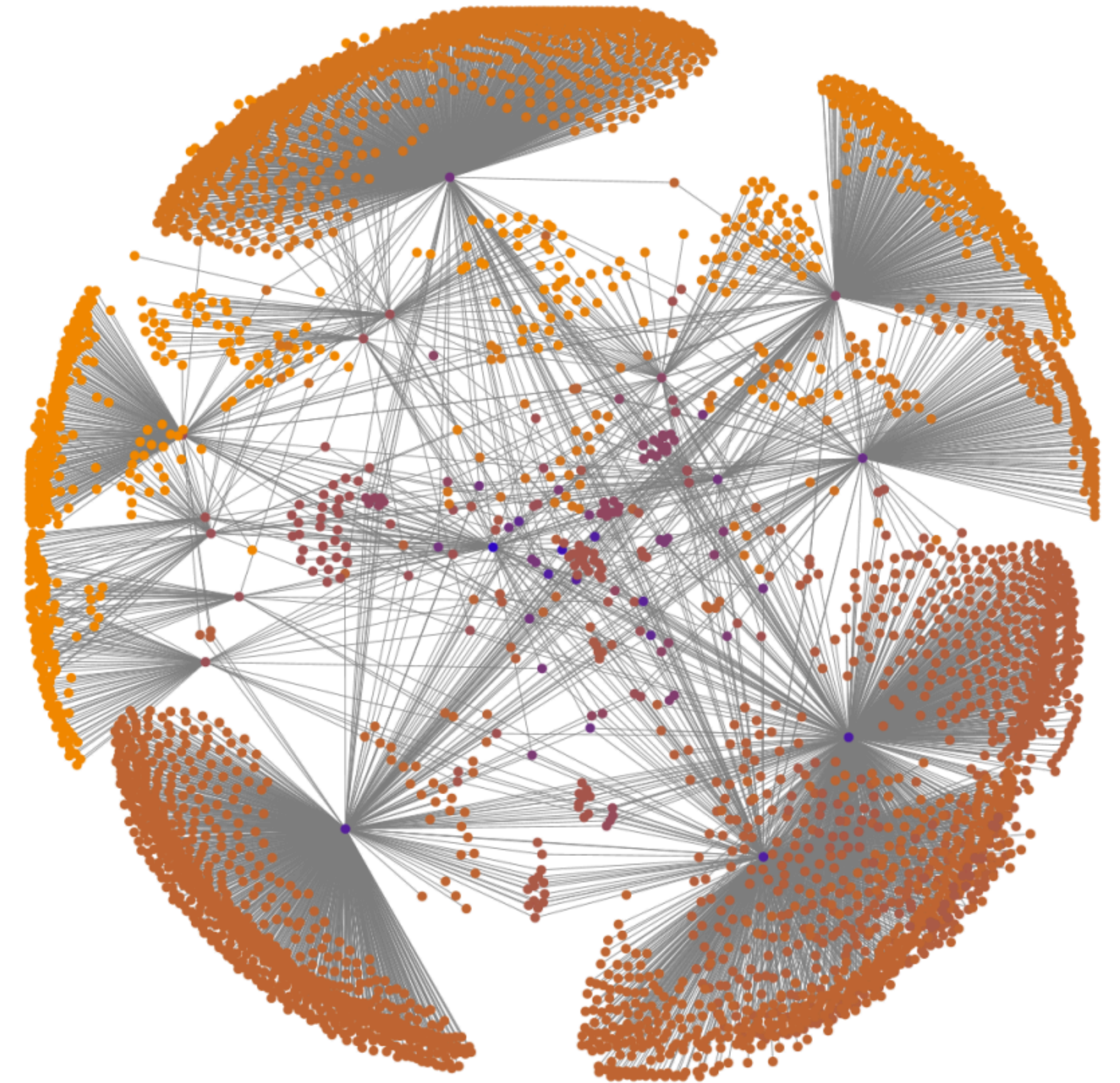
Crucitti, Paolo, Vito Latora, and Sergio Porta. "Centrality in networks of urban streets." *Chaos: an interdisciplinary journal of nonlinear science* 16.1 (2006).

Closeness: intersections easily accessible from the rest of the city.

Betweenness: intersections through which many shortest routes pass

Application: Search Engines

- How to find relevant pages?
- Hyperlinks are “votes.” A page is highly ranked if it has many votes!
- Should all votes be equal?



Dunne, Cody, and Ben Shneiderman. "Motif simplification: Improving network visualization readability with fan and parallel glyphs." *Human-Computer Interaction Lab Tech Report HCIL-2012-11*, University of Maryland (2012).

Page-Rank Centrality

- **Idea:** randomly walk through the graph n times. Jump to a random node with probability α .

$$C_{PR}(v) = \frac{\text{Visits to } v}{n}$$

