Deconstructing Data Science

David Bamman, UC Berkeley

Info 290
Lecture 20: Networks

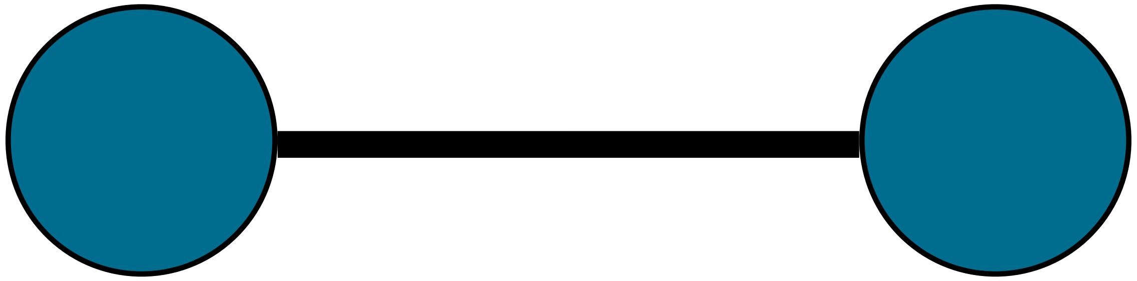
Apr 6, 2016

Nodes

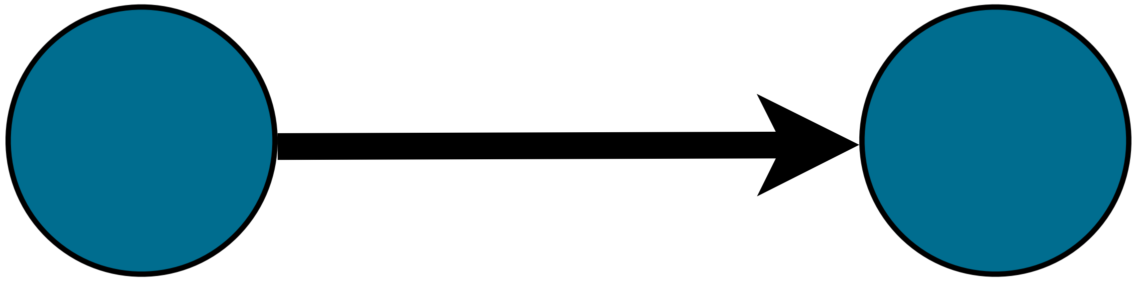
- People
- Web pages
- Servers
- Articles

Edges

Undirected



Directed

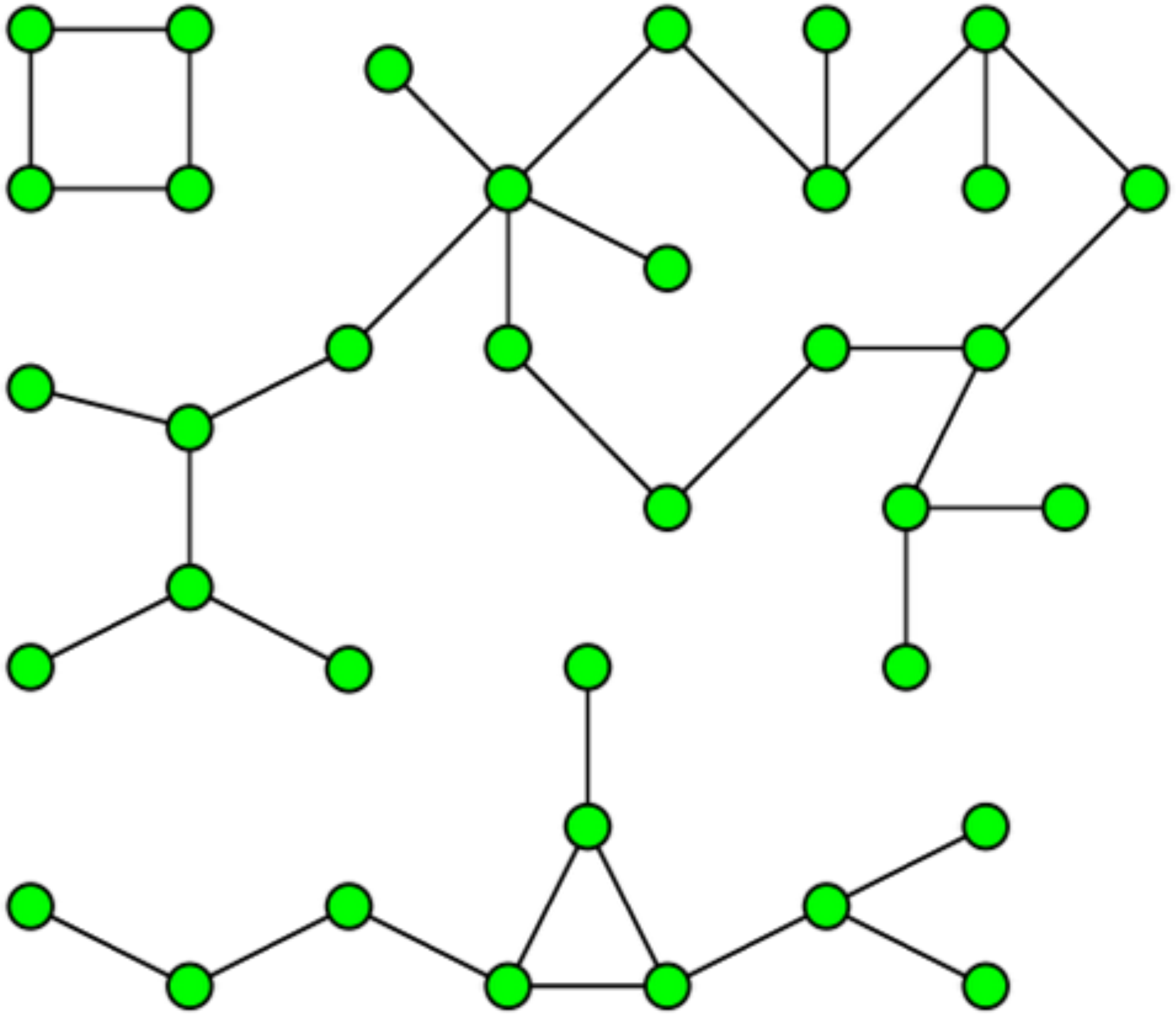


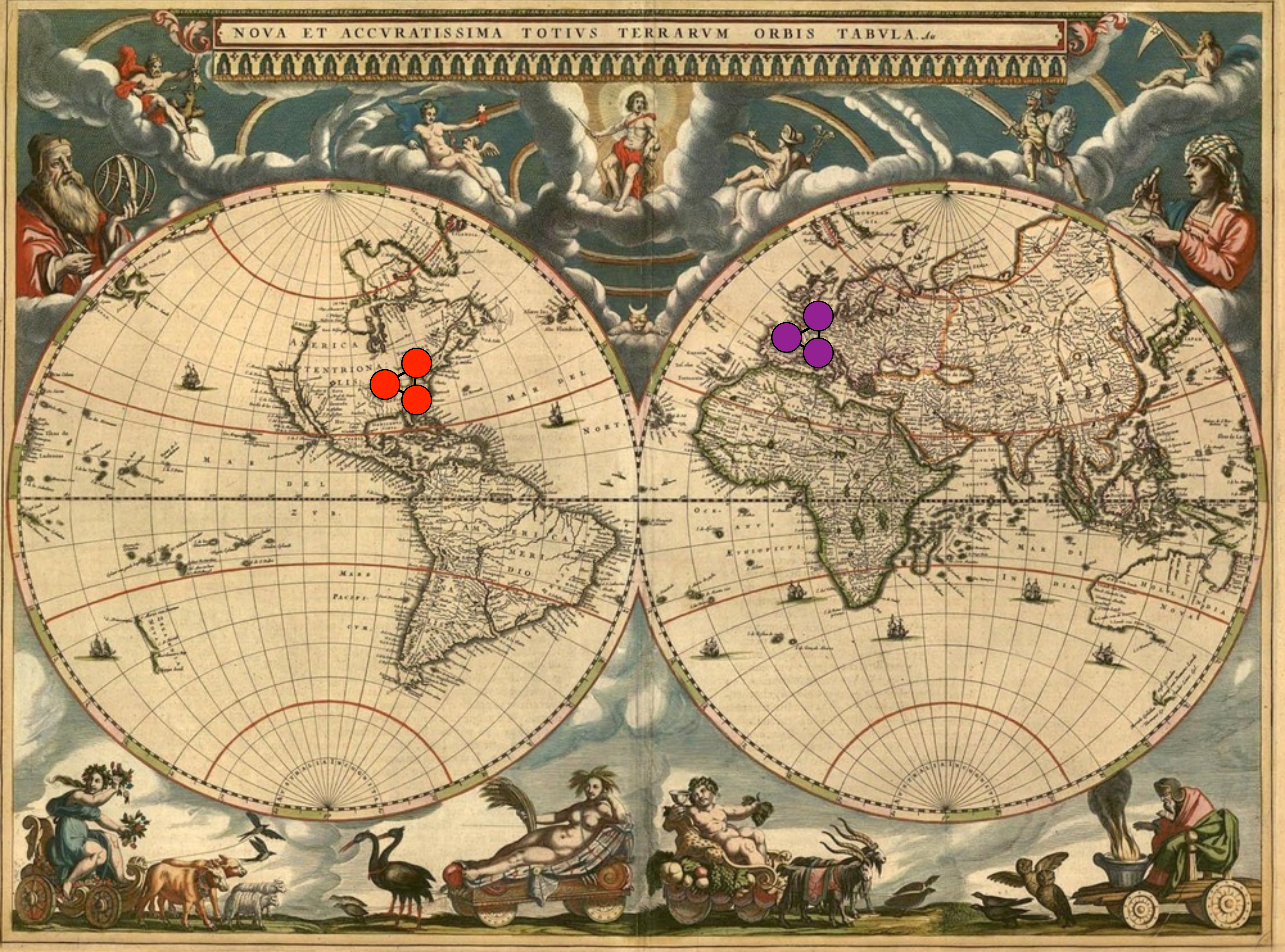
Connectivity

Connected component: subset of nodes where

— every node in the subset has a path to every other node

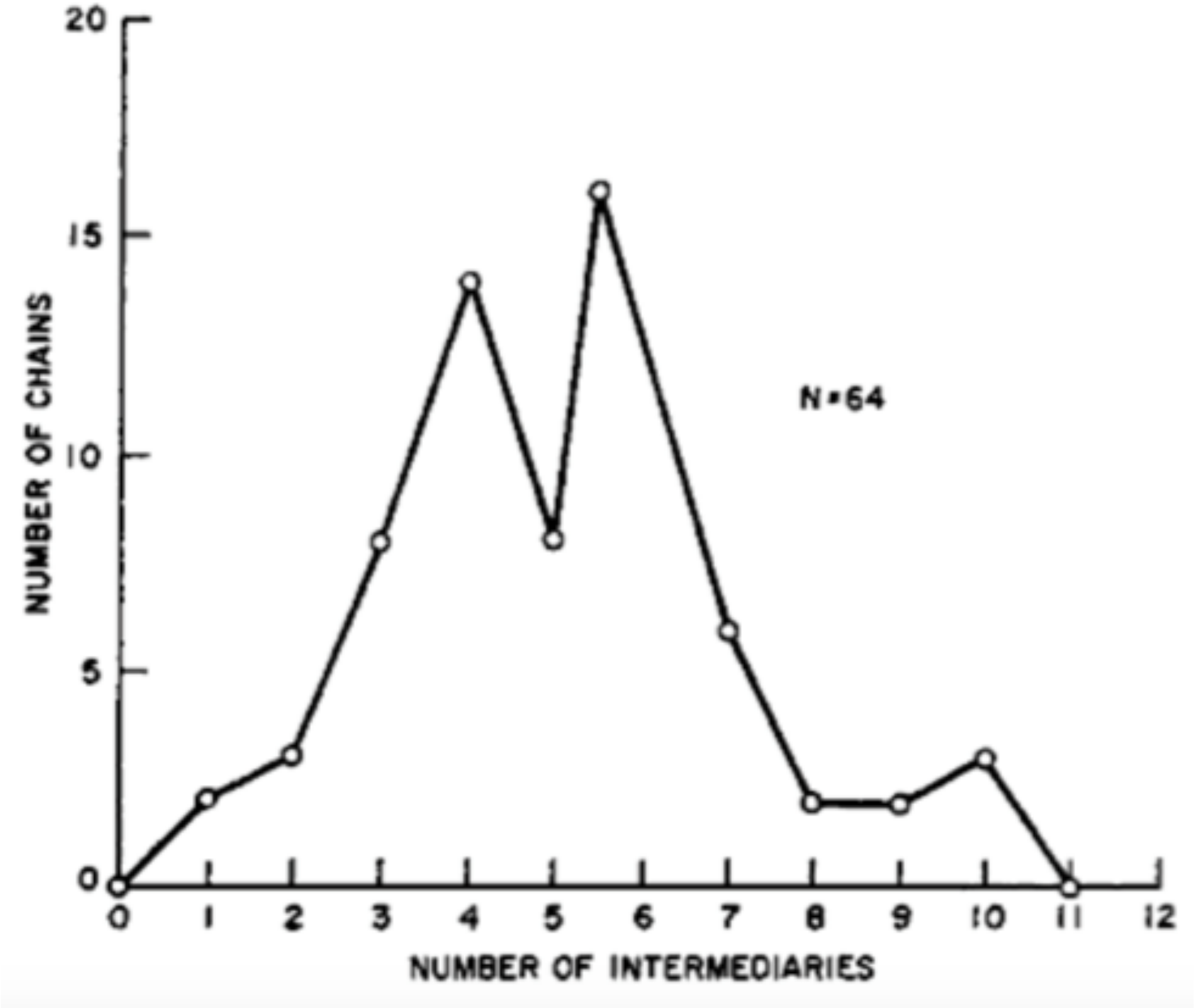
— that subset is not part of a larger set with that property




NOVA ET ACCVRATISSIMA TOTIVS TERRARVM ORBIS TABVLA. Au
AMERICA SEPTENTRIONALIS
MAR DEL NORT.
MAR DEL ZVR.
AMERICA MERIDIONALIS
MARE PACIFICVM
OCEANVS AETHIOPICVS
MAR DI INDIA
NOVA HOLLANDIA

Small-world phenomenon

- Stanley Milgram, “The Small World Problem,” *Psych. Today* (1967)
- 296 people asked to get a letter to a target near Boston by sending it to someone they knew on a first-name basis

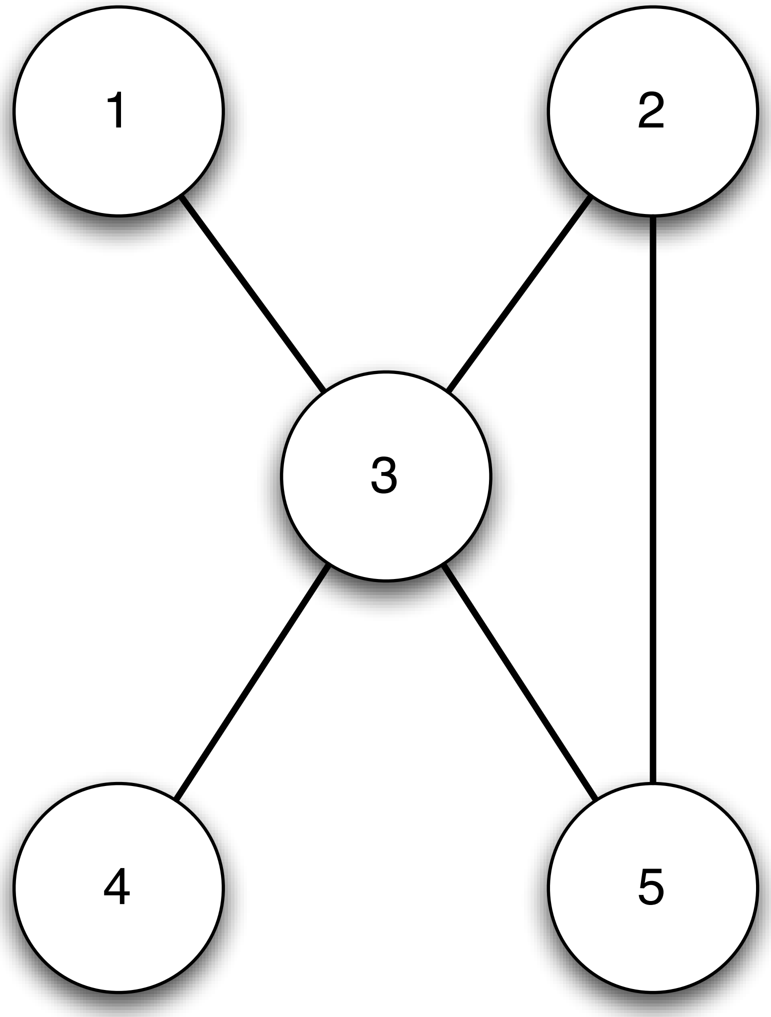


Data

- Co-authorship networks
- Citation networks
- Social networks
- Hyperlink networks

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

Adjacency Matrix



From:

To:

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

Adjacency Matrix

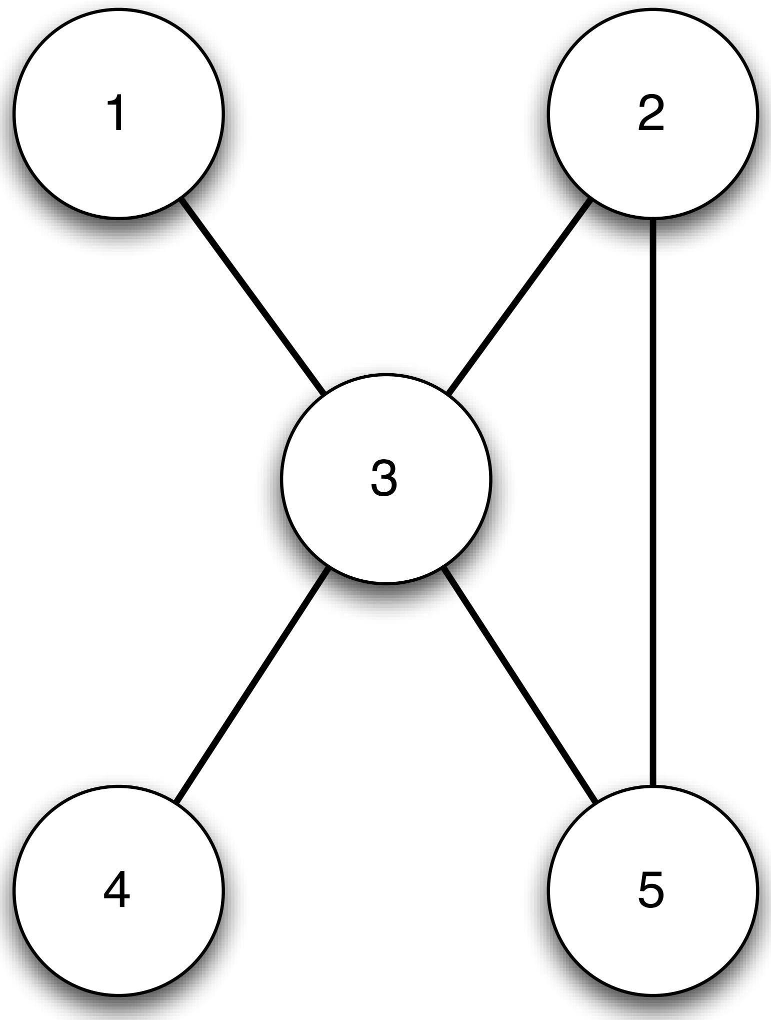
$$A_{3,1} = 1$$

From:

To:

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

Degree (centrality)



From:

To:

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

Degree (centrality)

$$\text{Degree}(3) = \sum_{i=1}^{5} A_{3,i}$$

$$= A_{3,1} + A_{3,2} +$$
$$A_{3,3} + A_{3,4} + A_{3,5}$$

From:

To:

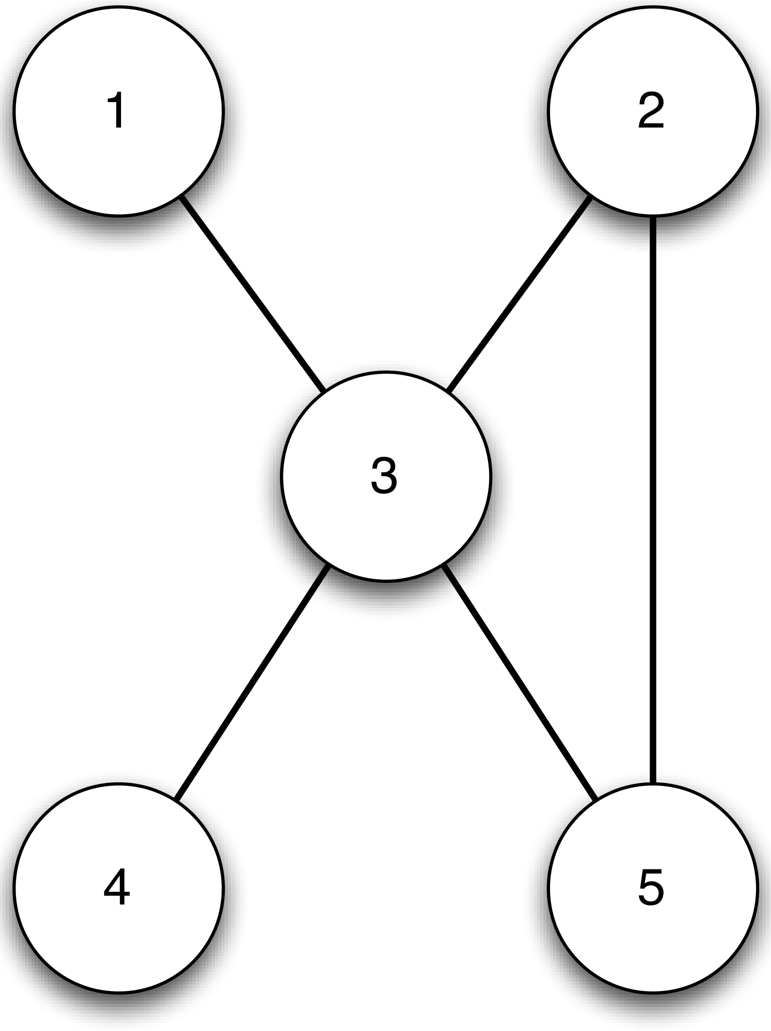
	1	2	3	4	5
1			1		
2			1		1
3	1	1		1	1
4			1		
5		1	1		

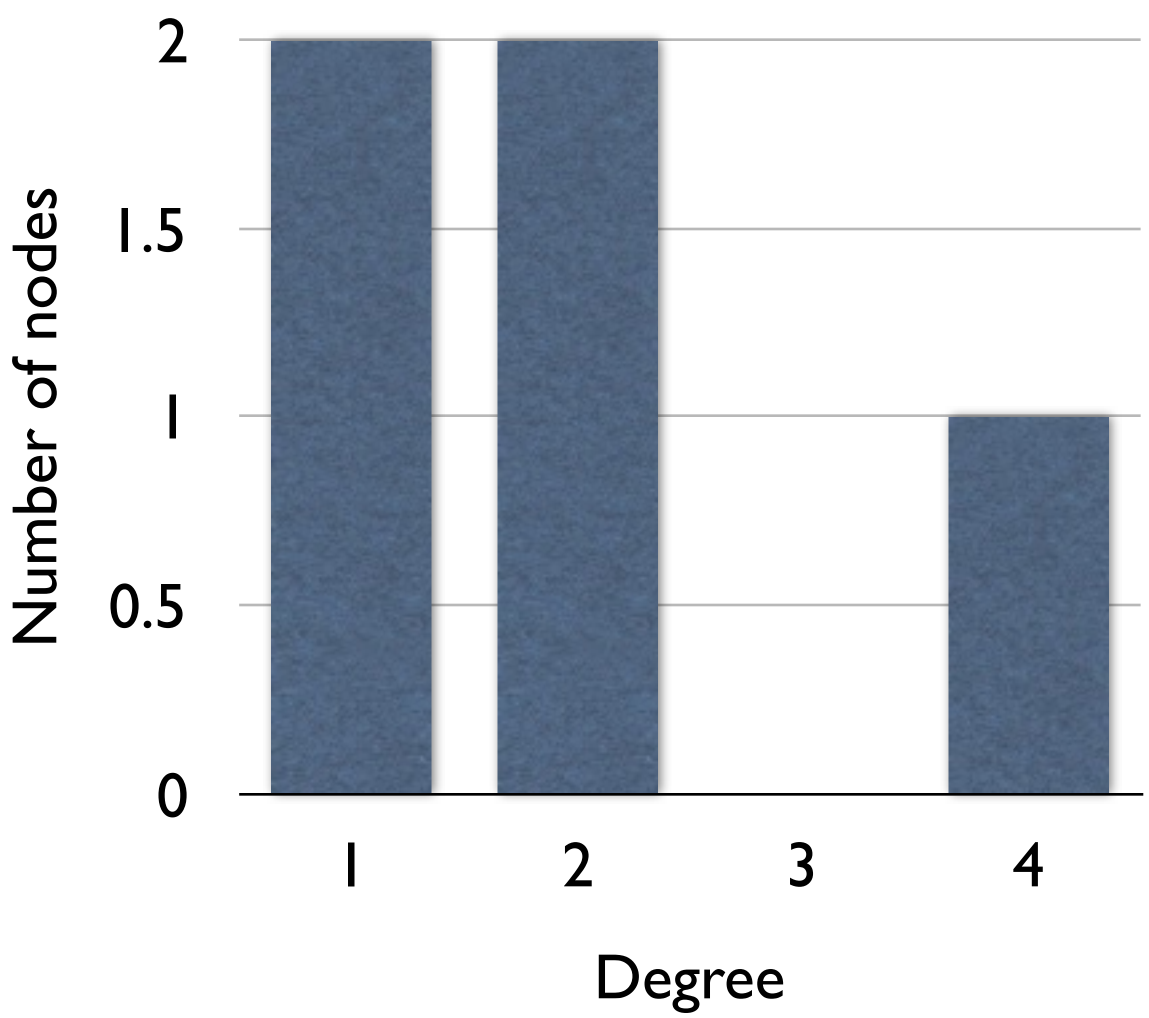
$$\text{Degree(i)} = \sum_{j} A_{i,j}$$

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

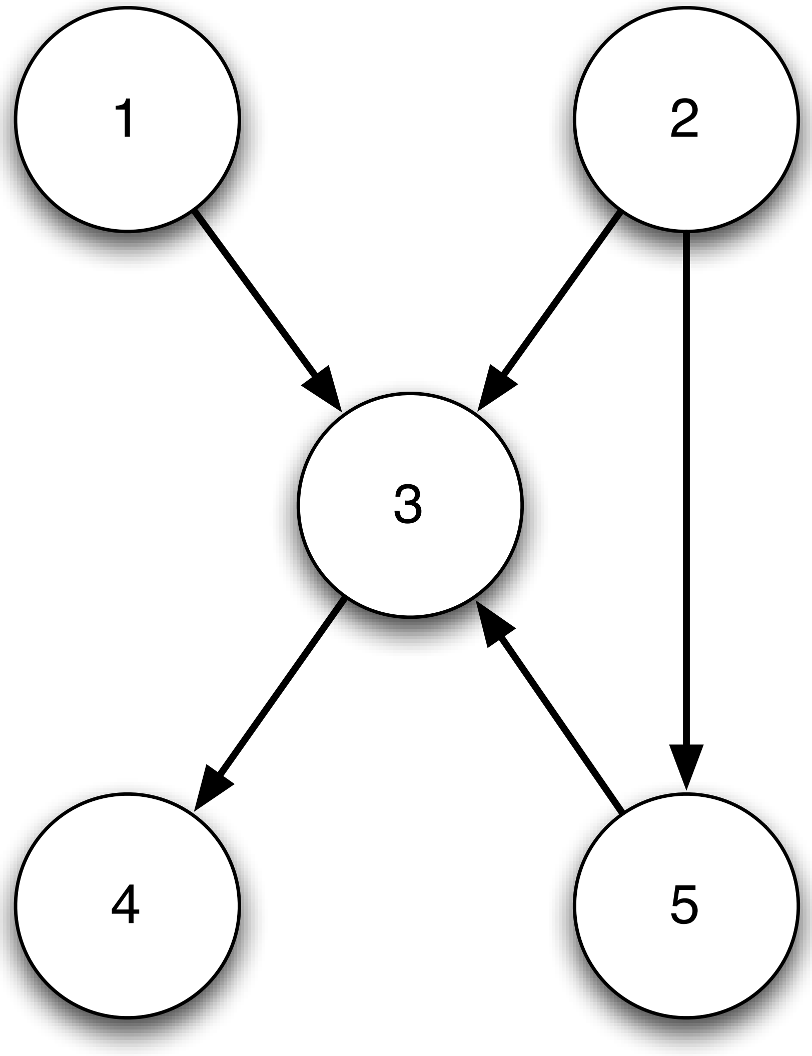
Degree
1
2
4
1
2

Degree distribution





(Directed) Adjacency Matrix



From:

To:

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

Under what circumstances is degree important?

Centrality

- Eigenvector centrality

$$\mathit{centrality}(i) = \sum_j \left[A_{i,j} \times \mathit{centrality}(j)\right]$$

- Katz centrality

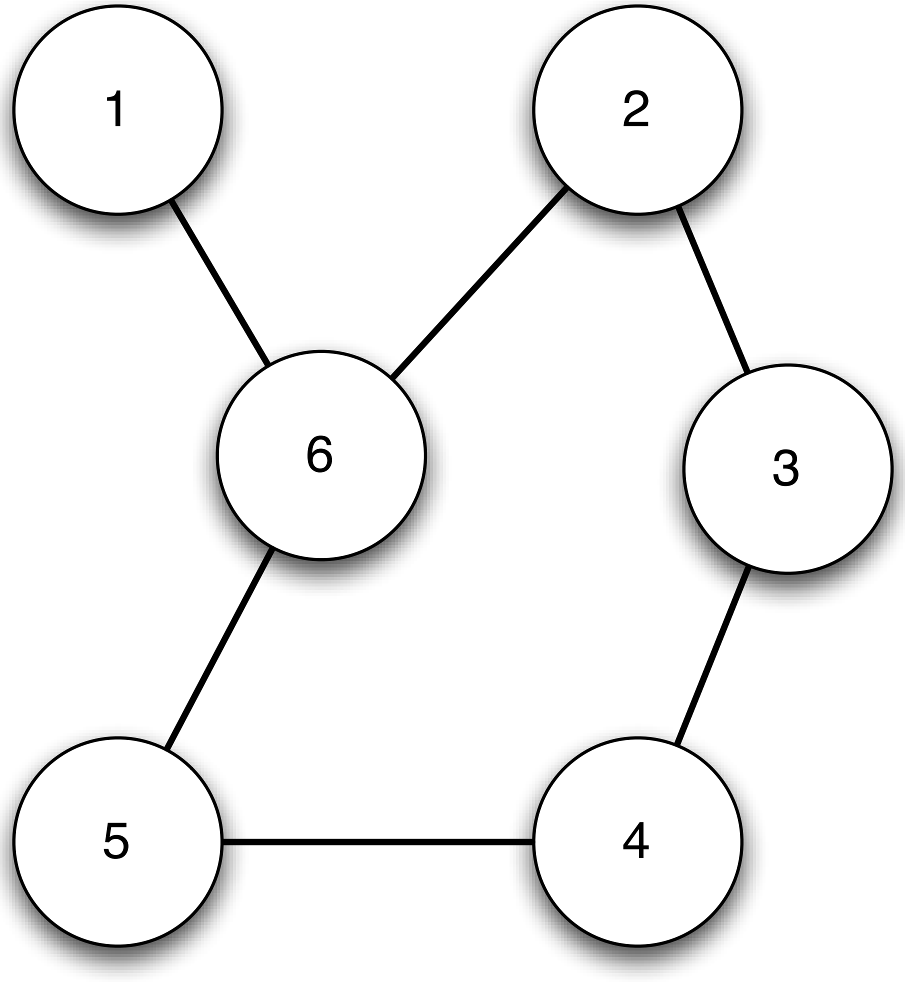
$$\mathit{centrality}(i) = \alpha \times \sum_j \left[A_{i,j} \times \mathit{centrality}(j)\right] + \beta$$

- PageRank

$$\mathit{centrality}(i) = \alpha \times \sum_j \left[A_{i,j} \times \frac{\mathit{centrality}(j)}{\mathit{outdegree}(j)}\right] + \beta$$

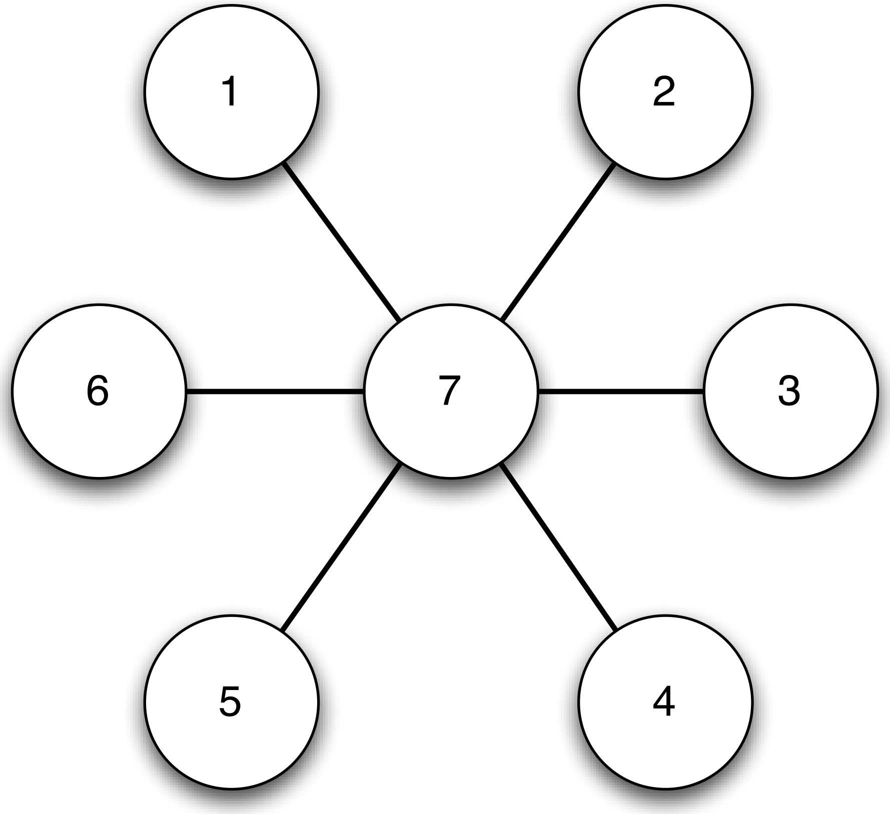
Geodesic path

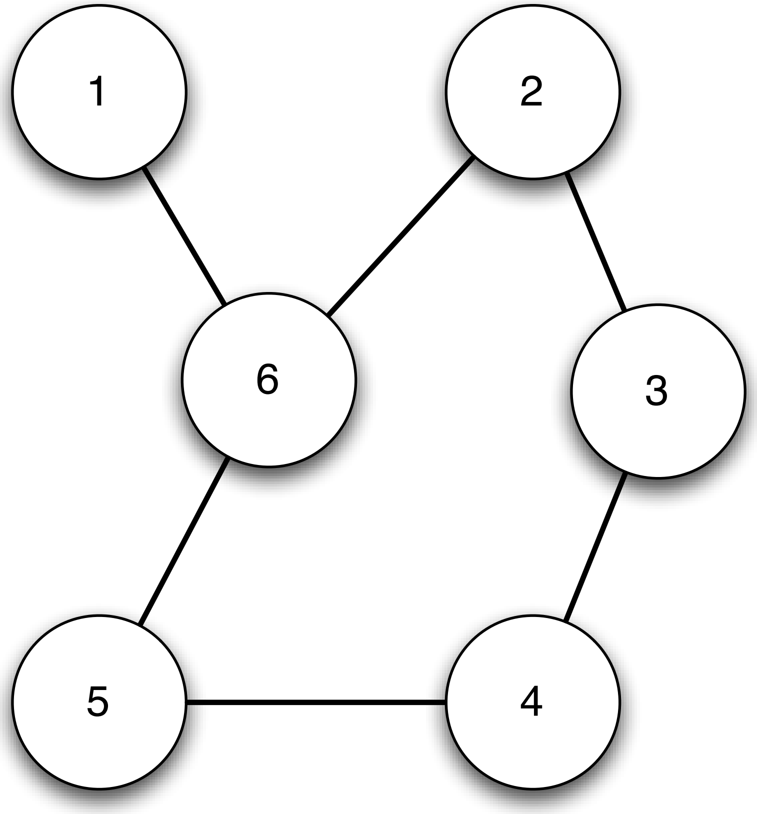
Shortest path between two nodes



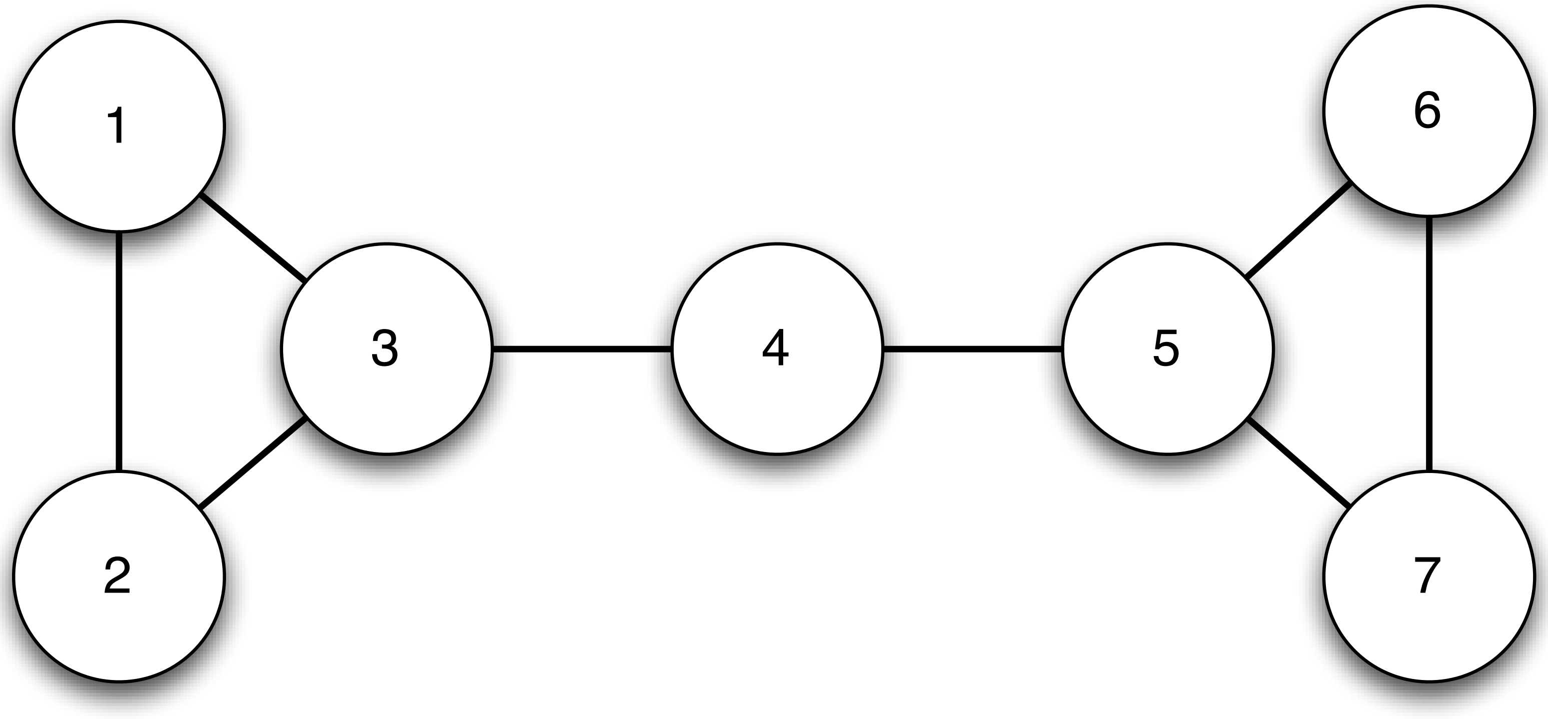
Closeness centrality

$$\mathit{centrality}(i) = \frac{\sum_j \mathit{shortest\_path}(i,j)}{n}$$





Betweenness centrality



$$betweenness(i) = \sum_{s,t} \mathbb{I}\{i \text{ is on the path from s to t}\}$$

Summary: centrality

What’s important?	Measure
Number of friends	Degree centrality
Number or importance of friends	Eigenvector, Katz centrality; PageRank
Distance from others	Closeness centrality
Middleman	Betweenness centrality

Tie strength

- “Strong” ties vs. “weak” ties

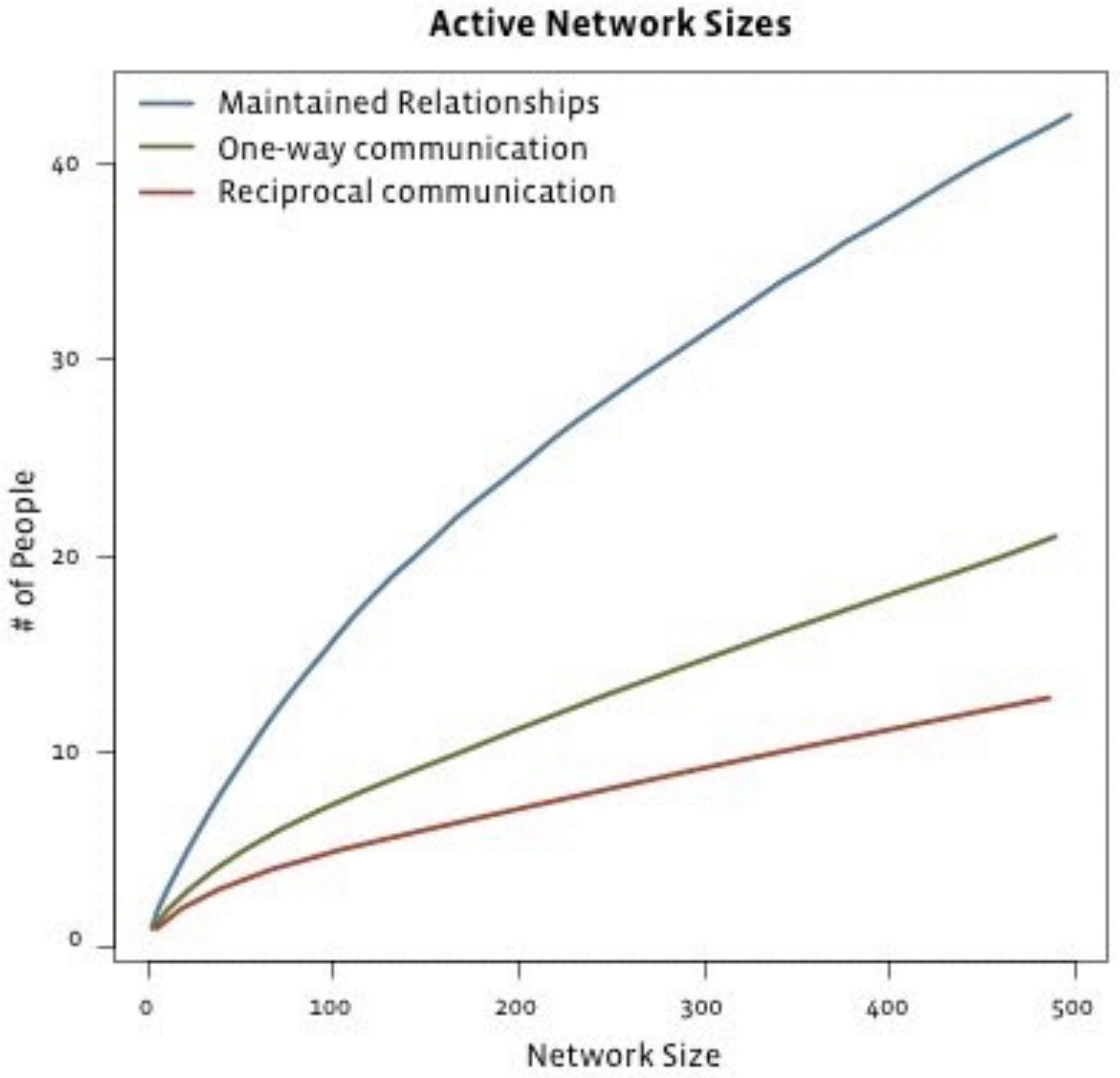
Tie strength

Marlow et al. (2009). Random sample of users over 30 days in 2009.

Maintained: click on news feed story/visit profile 3+ times

One-way: any directed message

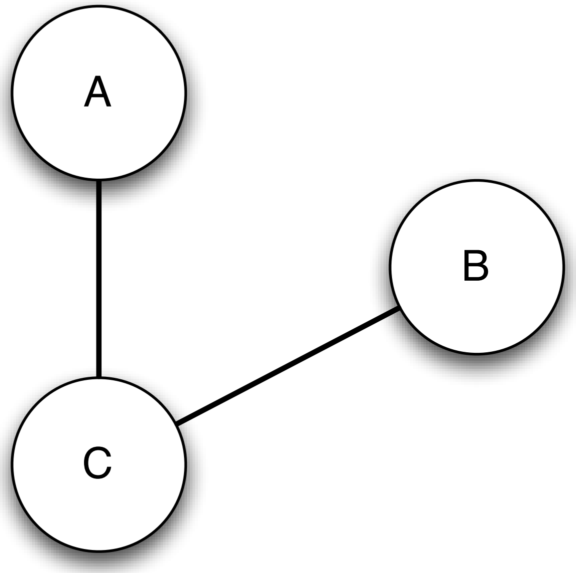
Reciprocal: reciprocated message

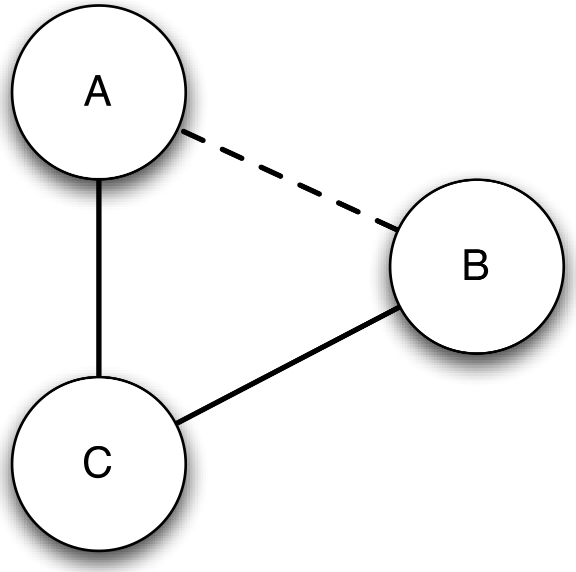


Triadic closure

Two people (A and B) have a friend (C) in common; A and B are likely to become friends.

More likely the stronger the tie is between A-C and B-C.

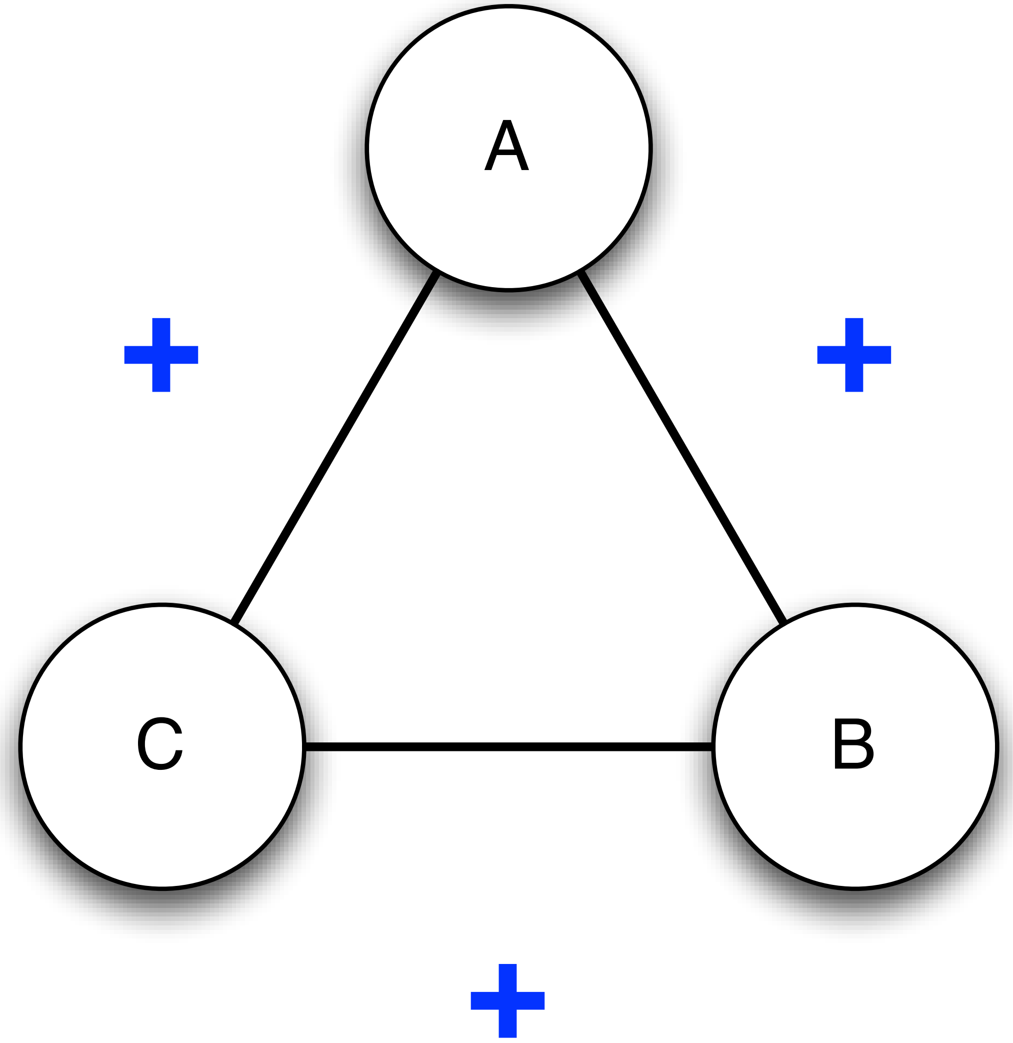




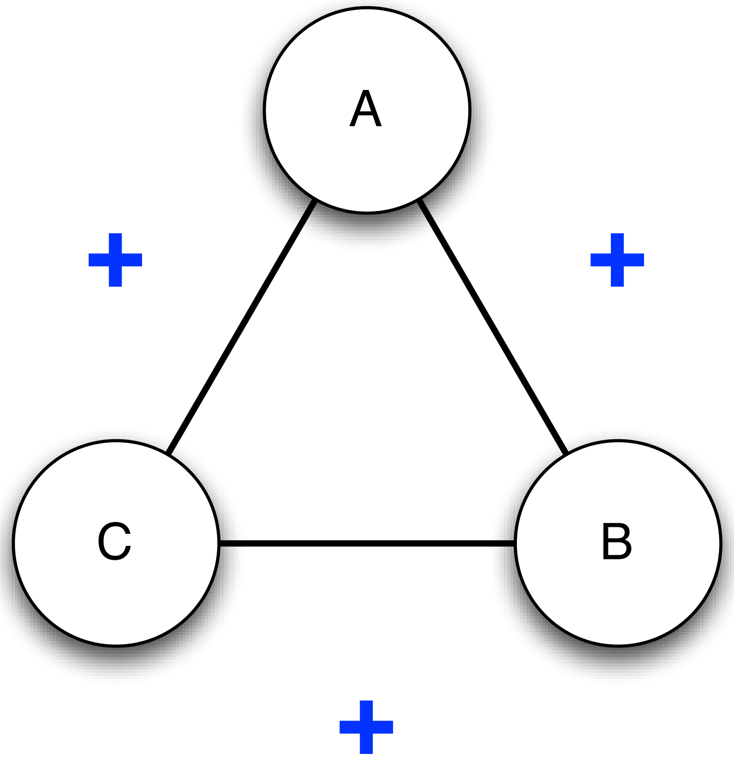
Triadic closure

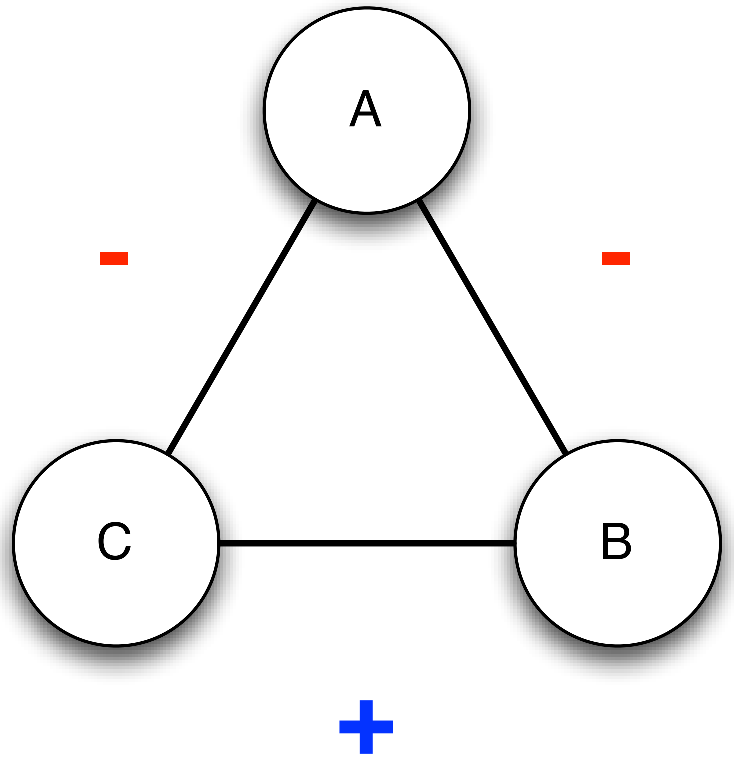
- Why?
    - A and B have more opportunity to interact if both are friends with the same person
    - A and B may trust each other if they’re both friends with the same person
    - C has a matchmaking incentive

Structural balance

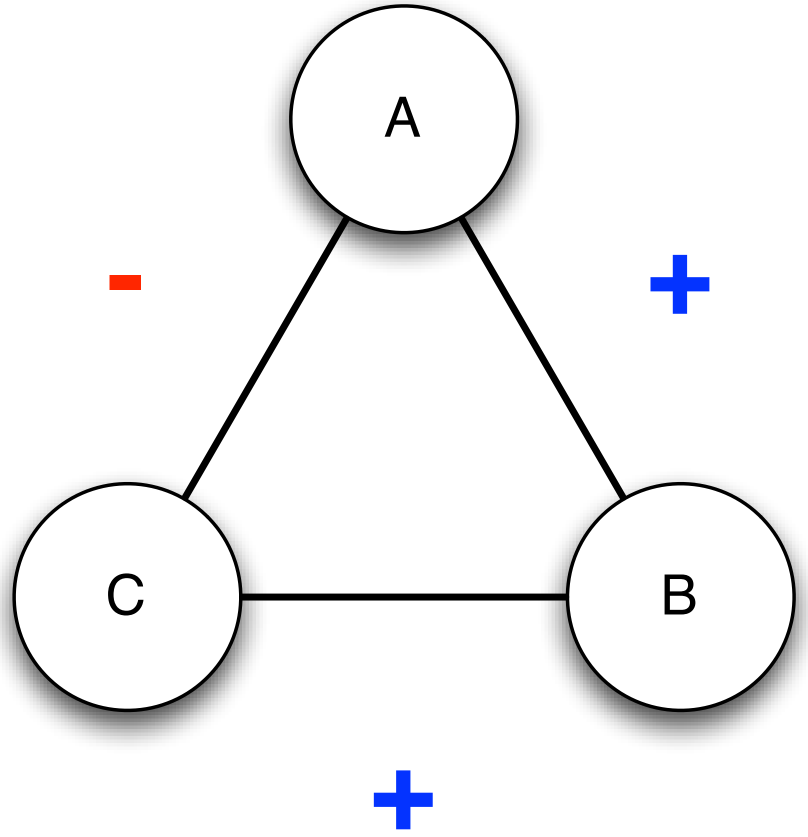


Structural balance





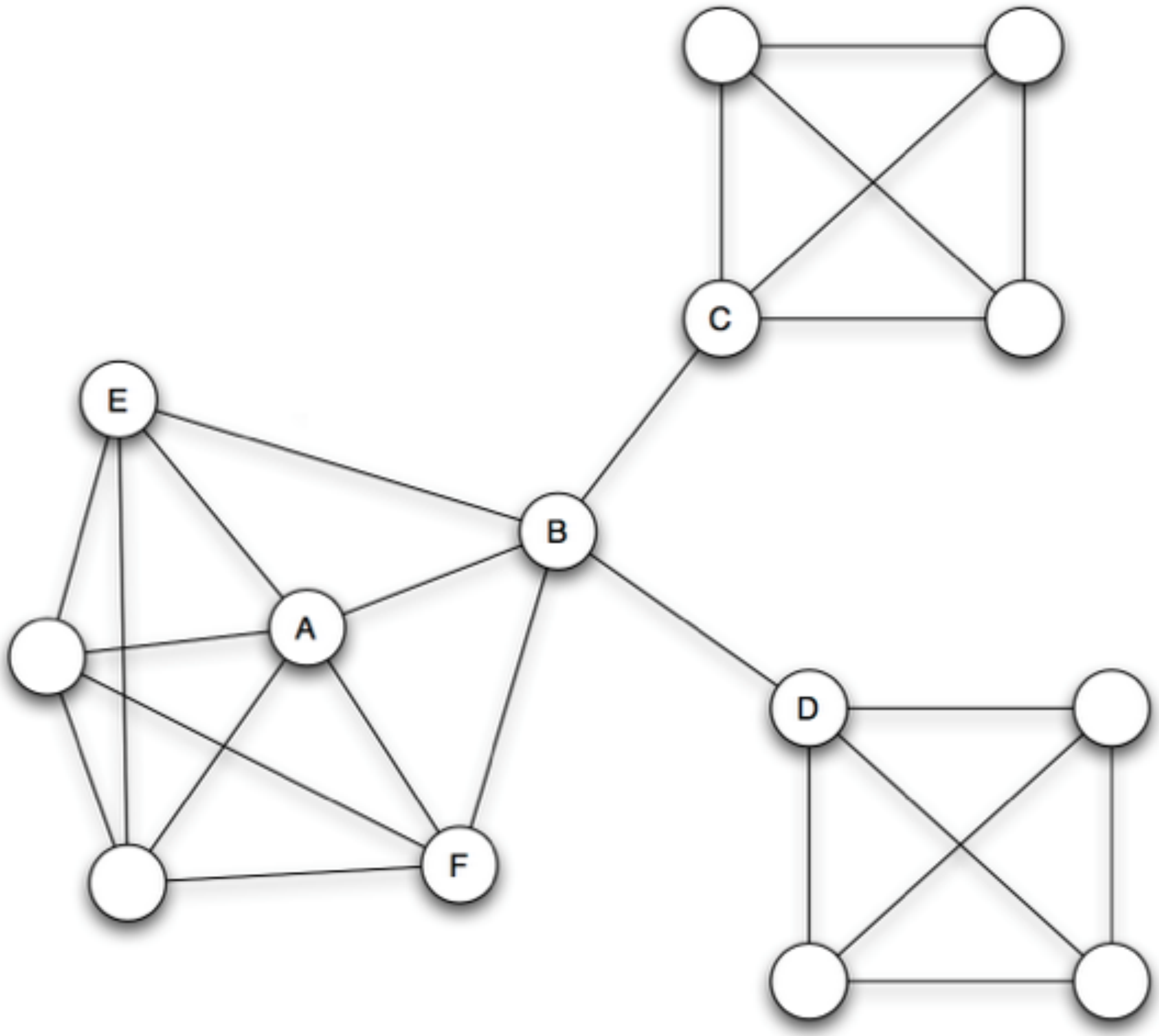
Structural balance



Clustering coefficient

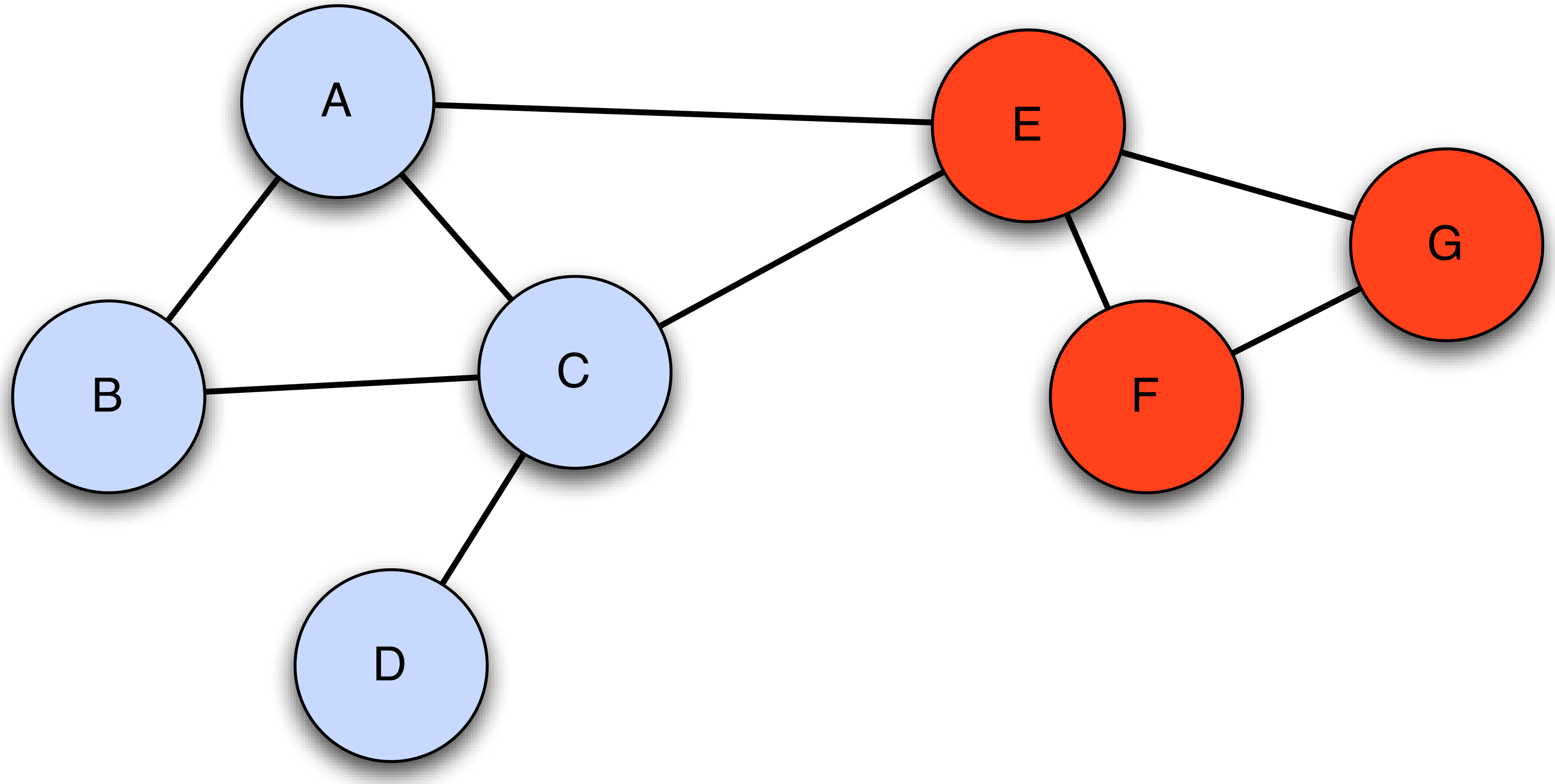
- Probability that two randomly selected friends of A will be friends with each other

structural bridges

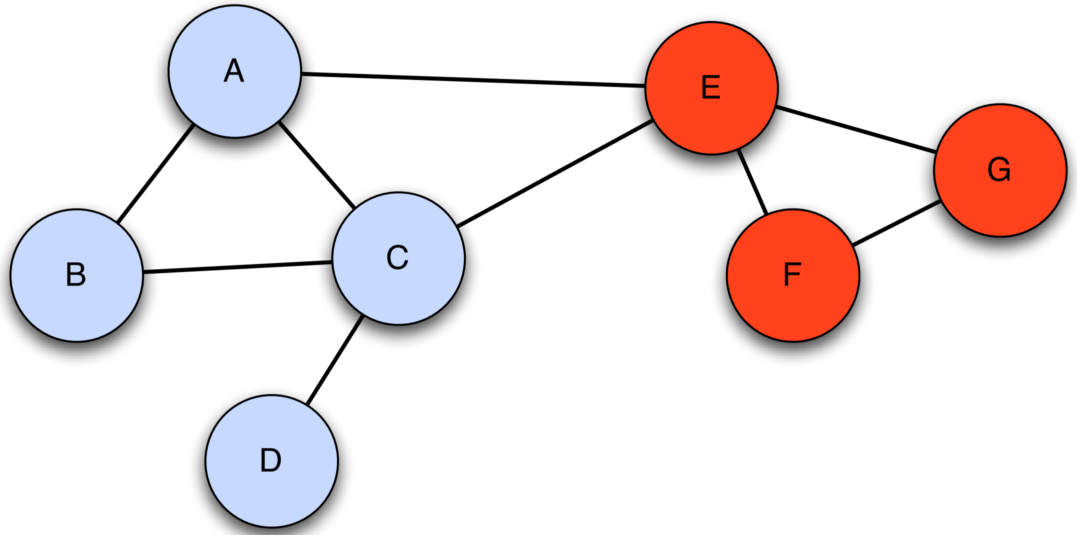


- early access to information
- ability to combine different sources of information
- gatekeeper between components

Assortativity



Assortativity



$$\frac{1}{2}\sum_{i,j}\left[A_{i,j} \times 1\{\text{if } node(i) = node(j)\}\right]$$

$$-\frac{1}{2}\sum_{i,j}\left[\frac{outdegree(i) \times outdegree(j)}{2m} \times 1\{\text{if } node(i) = node(j)\}\right]$$

m = total number of edges in network

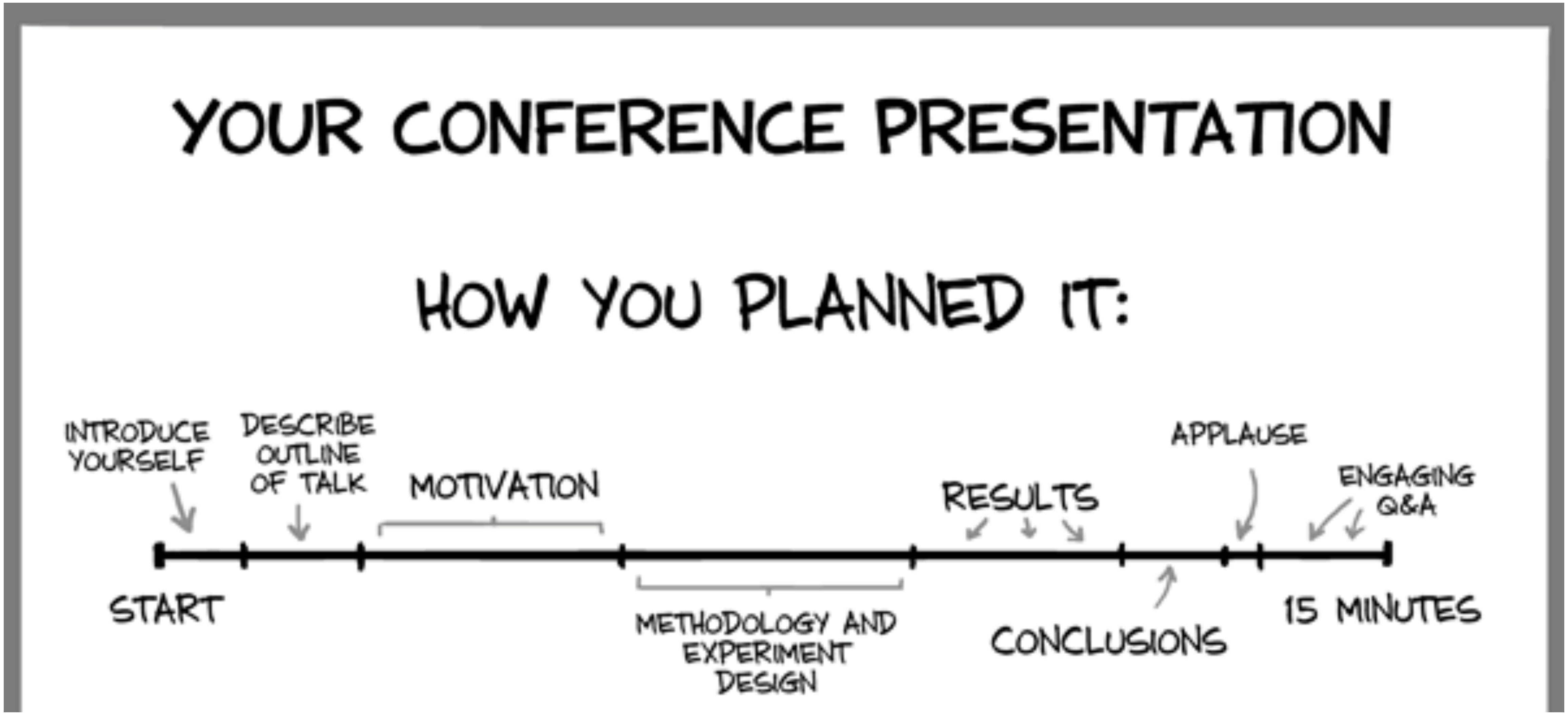
Assortativity

- Al Zamal et al. (2012), “Homophily and Latent Attribute Inference: Inferring Latent Attributes of Twitter Users from Neighbors”

Project presentation

Monday April 25 (6) + Wednesday April 27 (5)

10 min presentation +
3-5 min questions


YOUR CONFERENCE PRESENTATION
HOW YOU PLANNED IT:
INTRODUCE YOURSELF
DESCRIBE OUTLINE OF TALK
MOTIVATION
RESULTS
APPLAUSE
ENGAGING Q&A
START
METHODOLOGY AND EXPERIMENT DESIGN
CONCLUSIONS
15 MINUTES

Final report

- 8 pages, single spaced.

- Complete description of work undertaken
    - Data collection
    - Methods
    - Experimental details
    - Comparison with past work
    - Analysis

- See many of the papers we've read this semester for examples.

Final report

- Clarity.   For the reasonably well-prepared reader, is it clear what was done and why? Is the paper well-written and well-structured?

- Originality. How original is the approach or problem presented in this paper? Does this paper break new ground in topic, methodology, or content? How exciting and innovative is the research it describes?

- Soundness. Is the technical approach sound and well-chosen? Second, can one trust the claims of the paper -- are they supported by proper experiments, proofs, or other argumentation?

- Substance. Does this paper have enough substance, or would it benefit from more ideas or results? Do the authors identify potential limitations of their work?

- Evaluation. To what extent has the application or tool been tested and evaluated? Does this paper present a compelling argument for

- Meaningful comparison. Do the authors make clear where the presented system sits with respect to existing literature? Are the references adequate? Are the benefits of the system/application well-supported and are the limitations identified?

- Impact. How significant is the work described? Will novel aspects of the system result in other researchers adopting the approach in their own work?