Structure between data points

David Bamman
Info 202: Information Organization and Retrieval

October 12, 2016

Perspectives

- Semantic
- Lexical
- Structural
- Architectural
- Implementational

Structural perspective

- Focuses on the patterns that emerge among individual relationships
- Network analysis, social network analysis

Questions about individuals

- Who are the most popular individuals in a network?
- Which individuals have the most influence?
- Who bridges different subgroups of users?
- If one is trying to disrupt a network, who should be removed?
- Are there different types of social actors that can be identified by unique network patterns?

Hansen and Smith, “Social Network Analysis in HCI”

Questions about overall structure

- How interconnected are a group of social actors?
- What is the distribution of individual network properties or social roles? For example, are there only a small percentage of “hubs” with a majority of “isolates”?
- Are there subgroups of highly connected users?
- What network properties or motifs (i.e., recurring network patterns) are related to social outcomes of interest?

Hansen and Smith, “Social Network Analysis in HCI”

Questions about flow

- How do the structures of social relationship vary over time?
- How does the importance of specific individuals, social roles, or clusters change over time?
- How does information spread through a network (e.g., Twitter)? How can information propagation be catalyzed or minimized?
- How does the use of new technologies spread through social networks? Who influences adoption of technology the most?

Hansen and Smith, “Social Network Analysis in HCI”

Nodes

- People
- Web pages
- Servers
- Articles

Edges

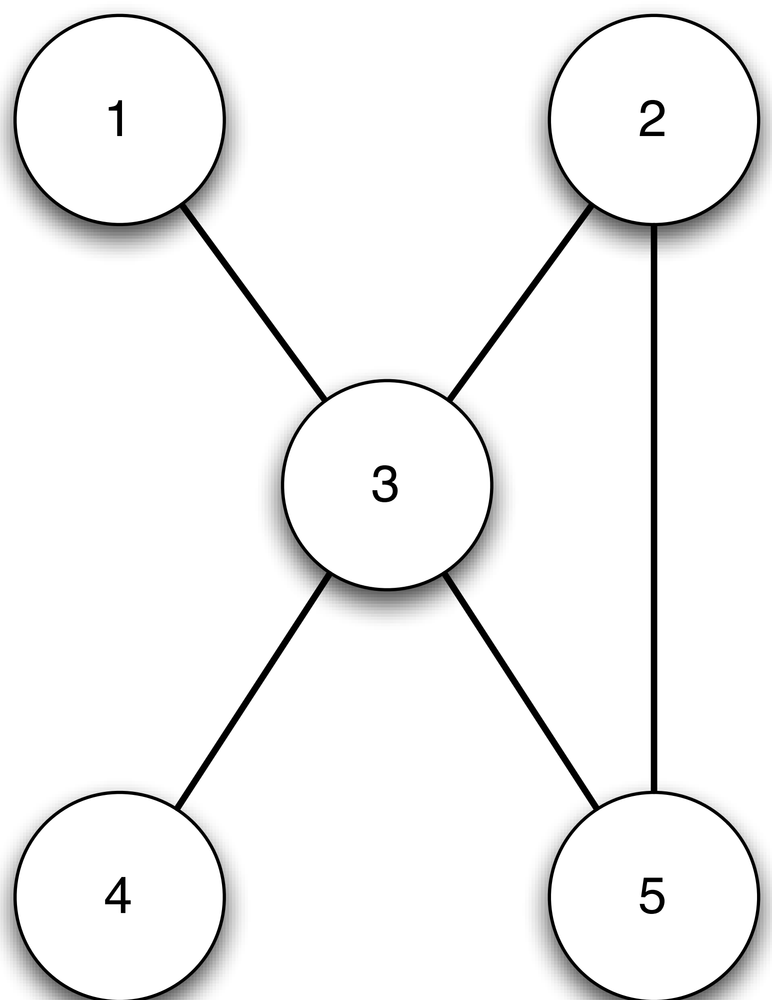
Undirected

Directed

Metrics for individuals

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

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

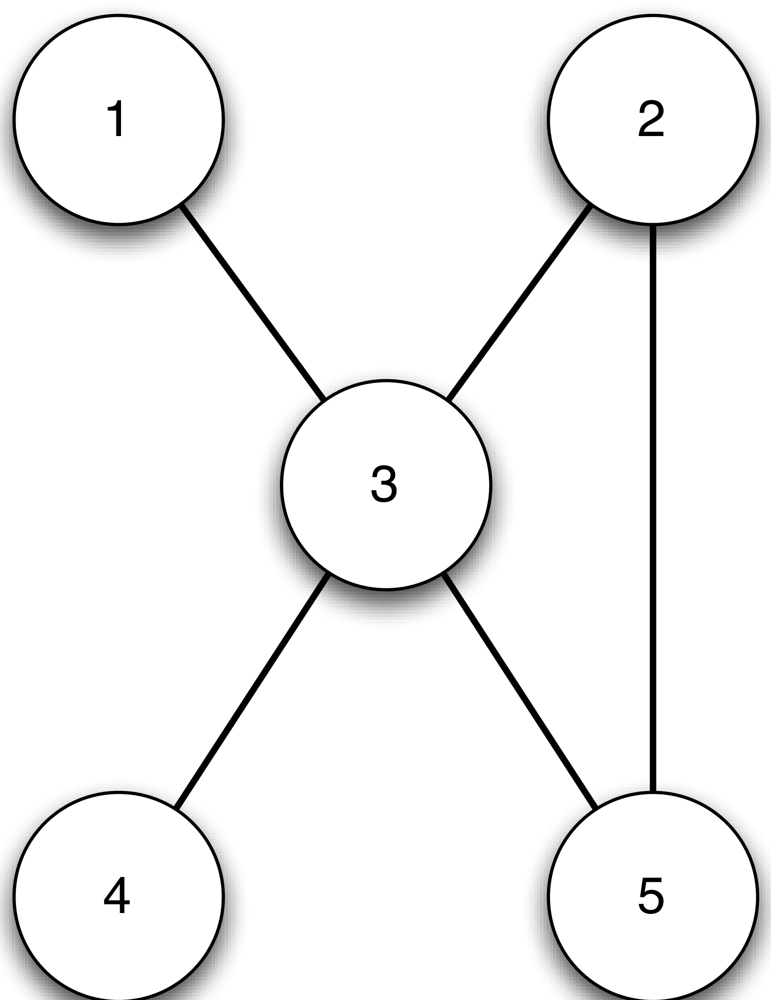
$$A_{3,1} = 1$$

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:

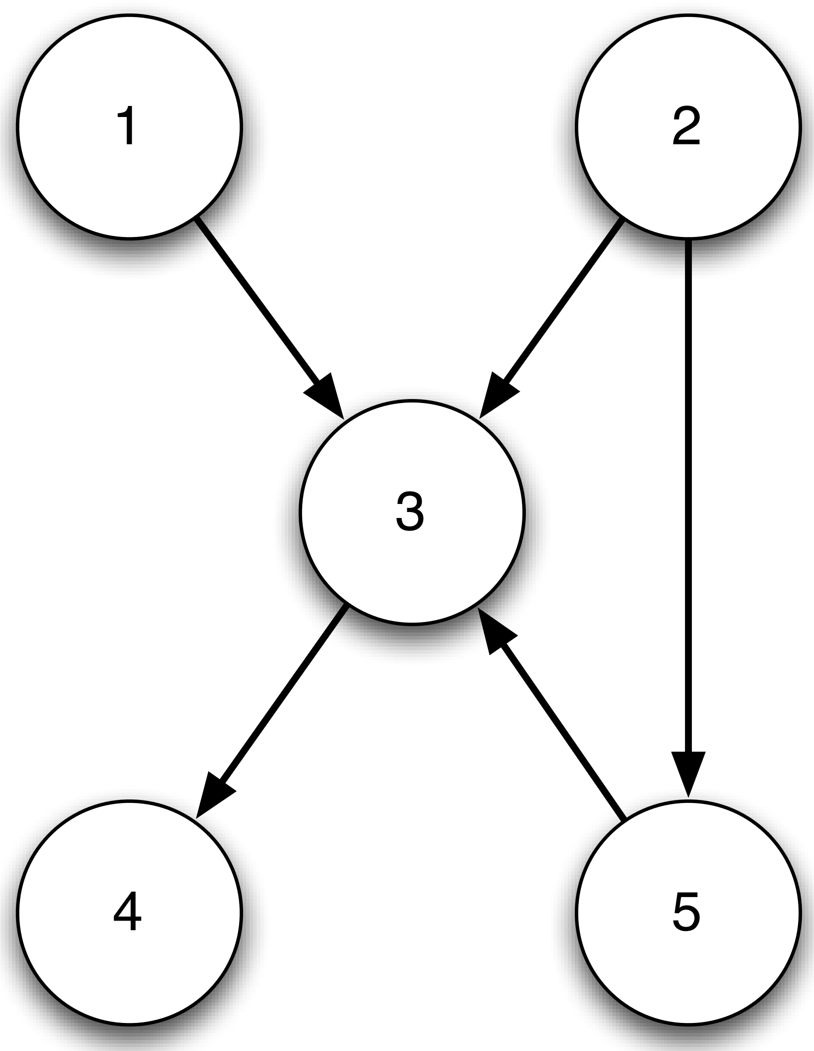
	1	2	3	4	5
1			1		
2			1		1
3	1	1		1	1
4			1		
5		1	1		

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

	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

(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

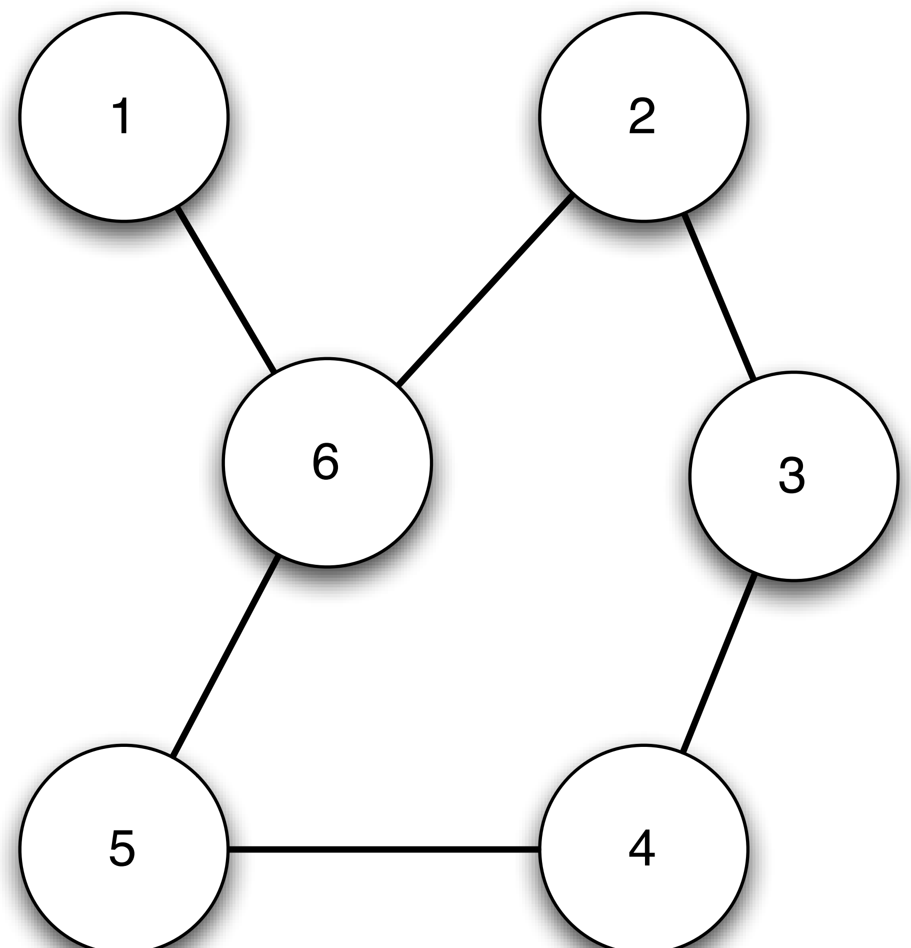
$$\mathit{centrality}(i) = \alpha \times \sum_j \left[A_{i,j} \times \mathit{centrality}(j)\right] + \beta$$

- PageRank

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

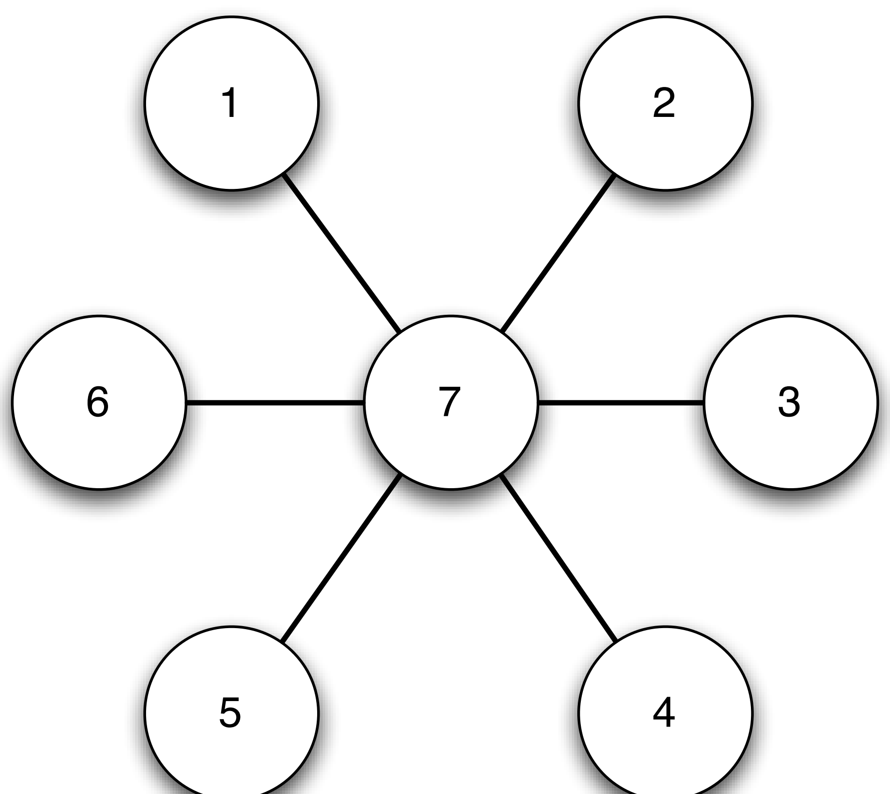
Geodesic path

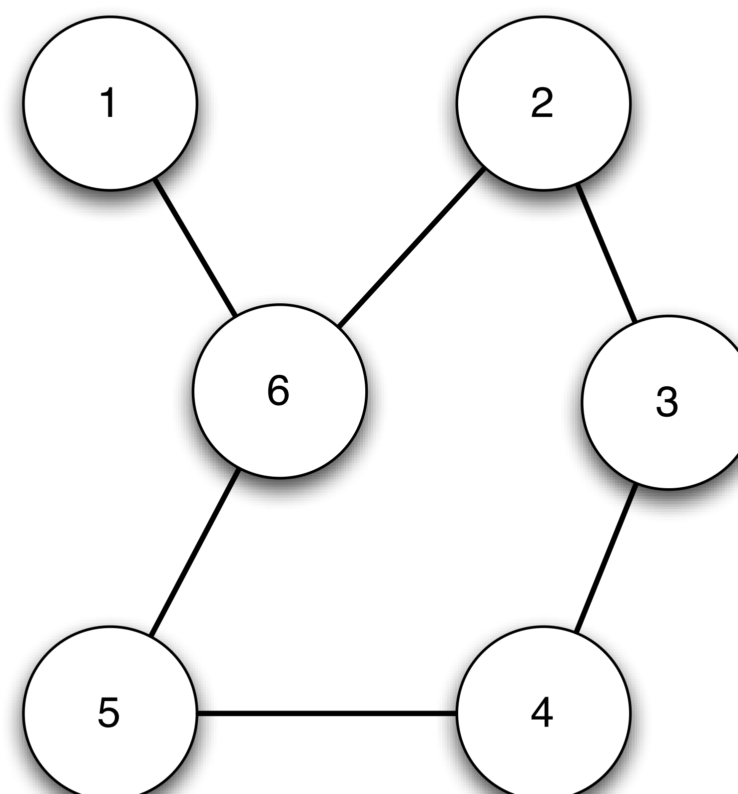
Shortest path between two nodes



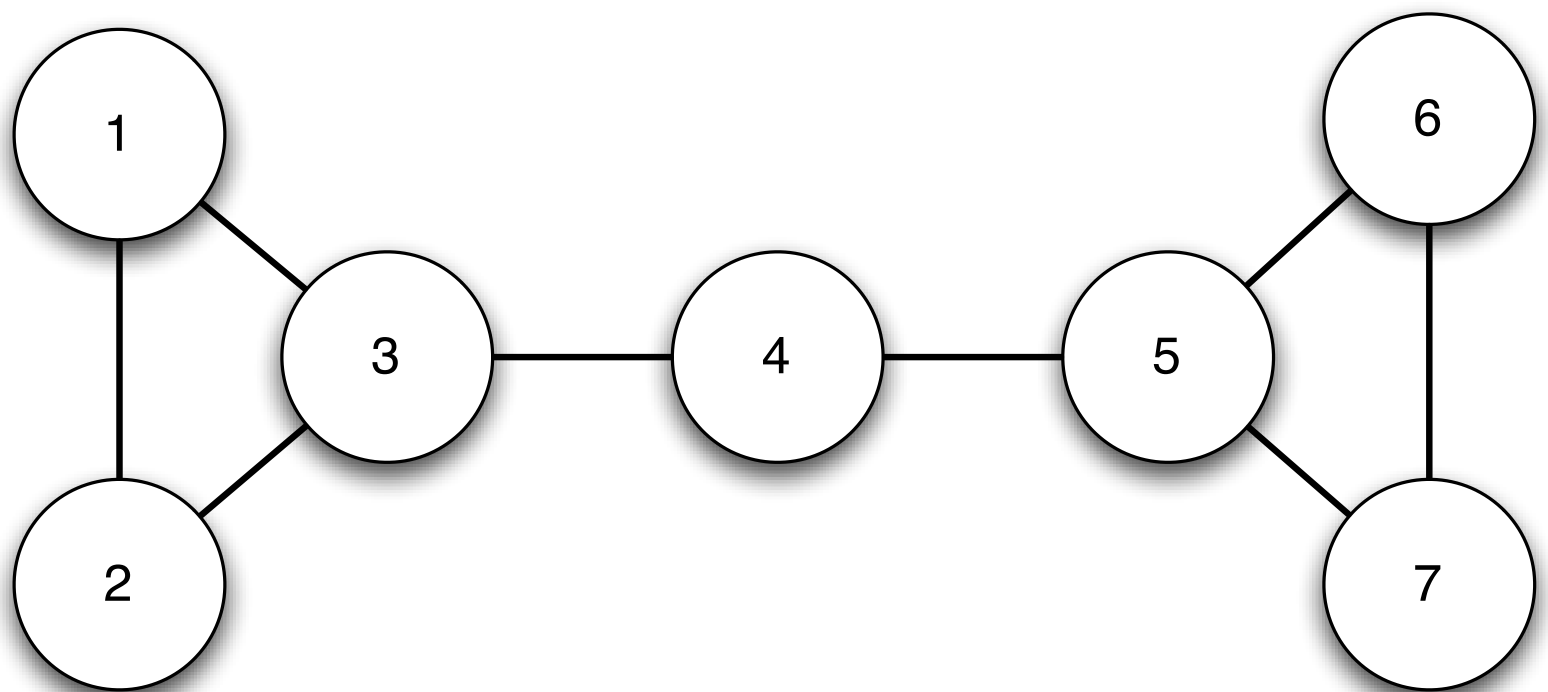
Closeness centrality

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





Betweenness centrality



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

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

Summary statistics

- Density
- Clustering coefficient
- Degree distribution
- Assortativity

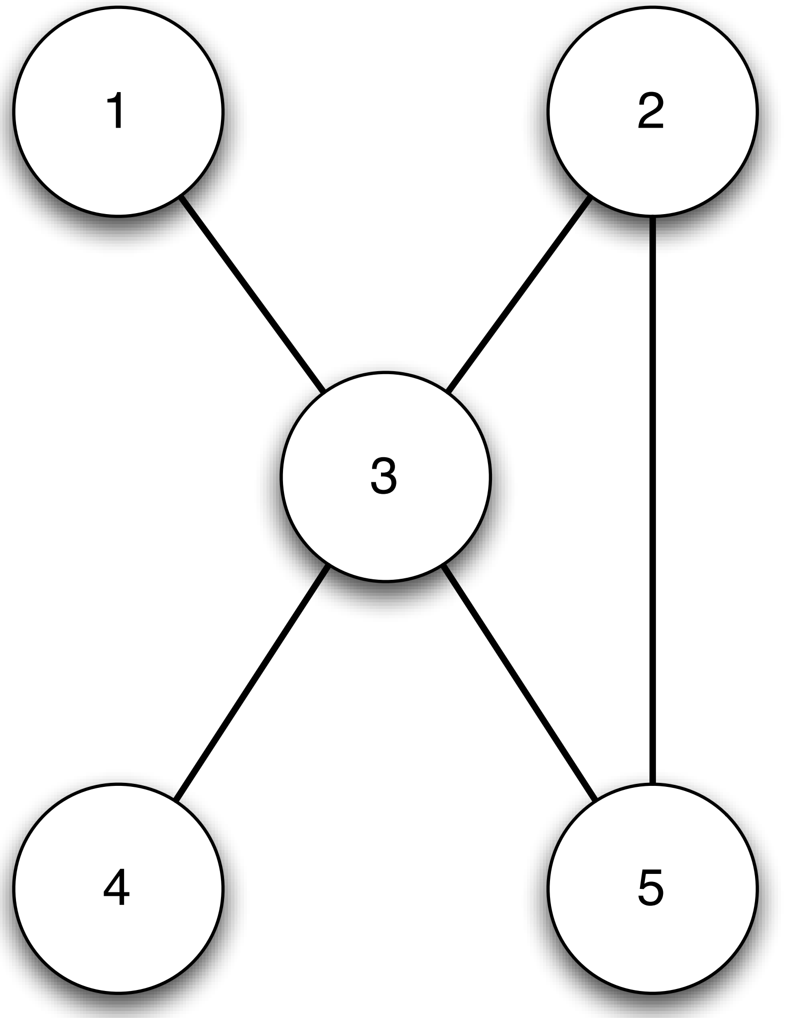
Density

How interconnected is the network?

Fraction of edges to total possible edges

$$\frac{2E}{N(N-1)}$$

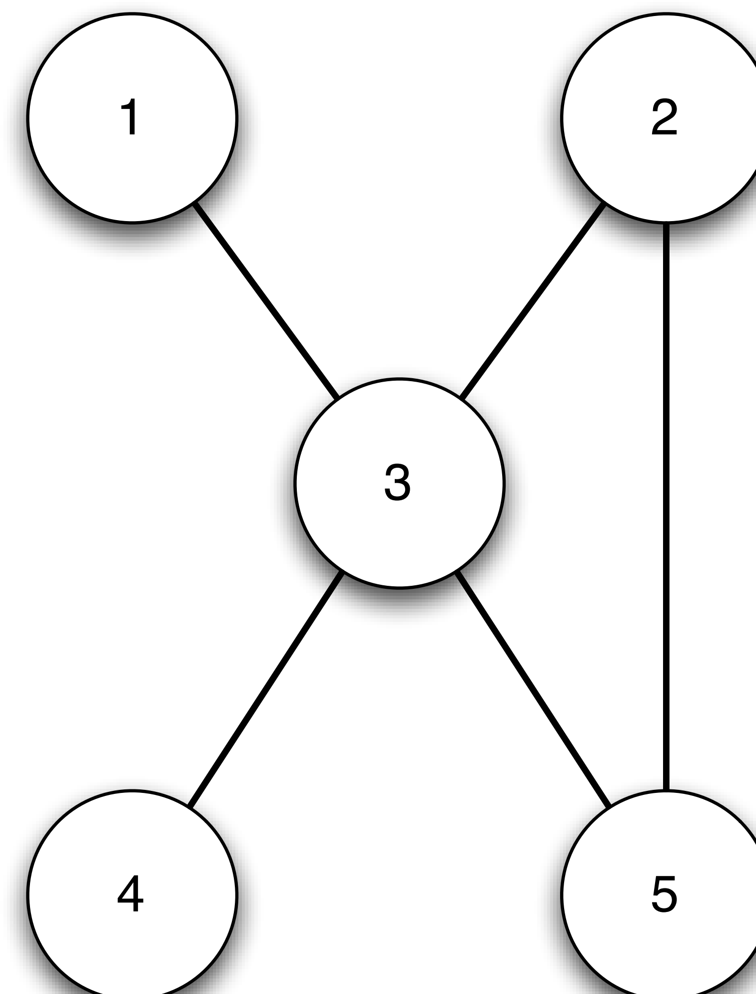
E = number of edges in network
N = number of nodes in network



Clustering coefficient

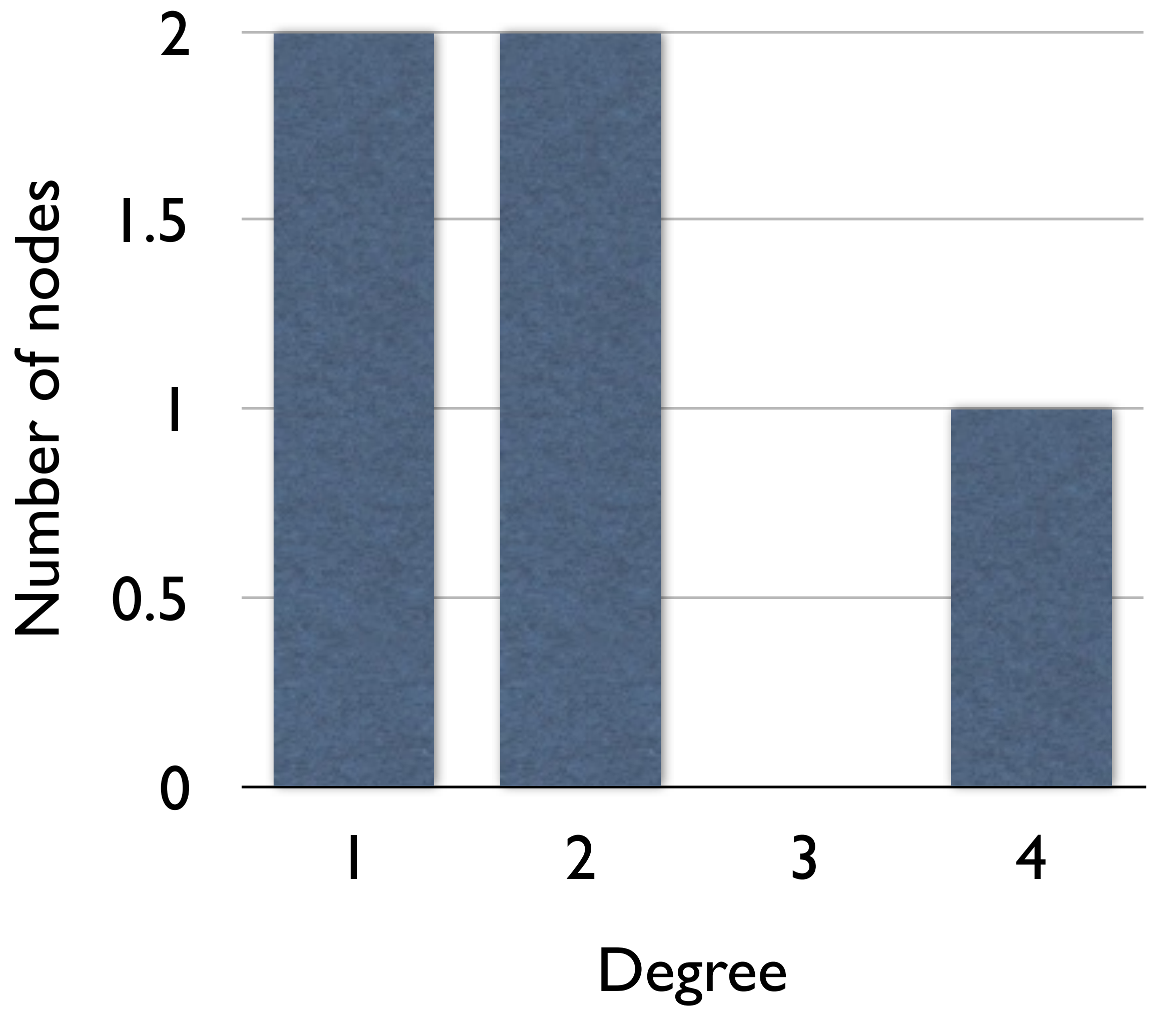
- Probability that two randomly selected friends of A will be friends with each other

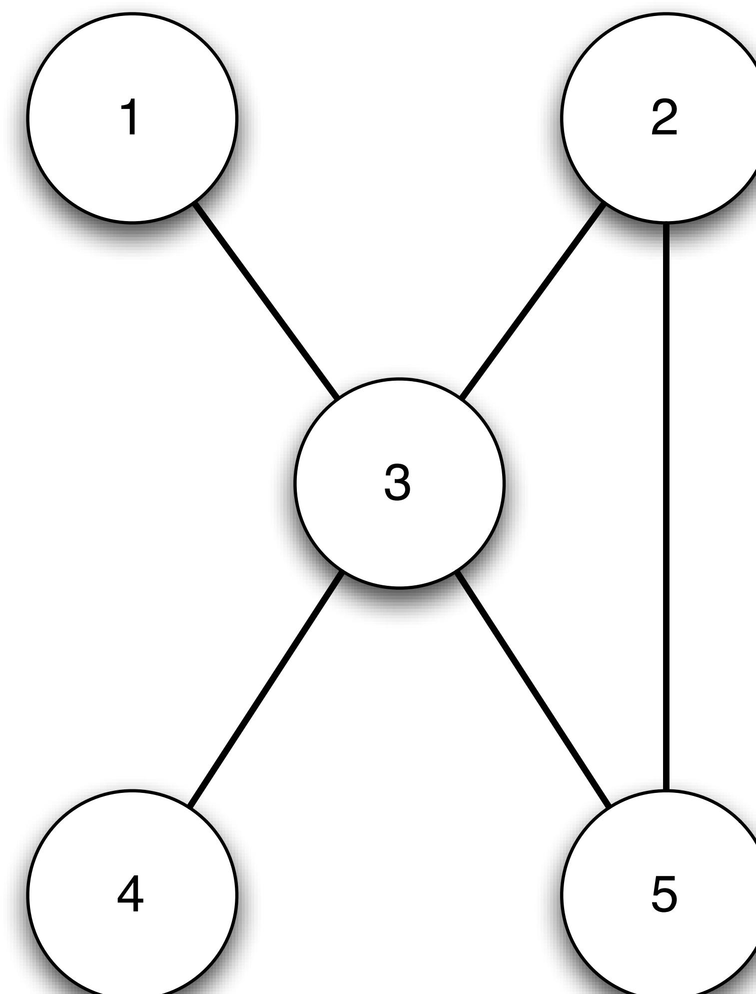
$$\frac{2e_i}{k_i(k_i - 1)}$$



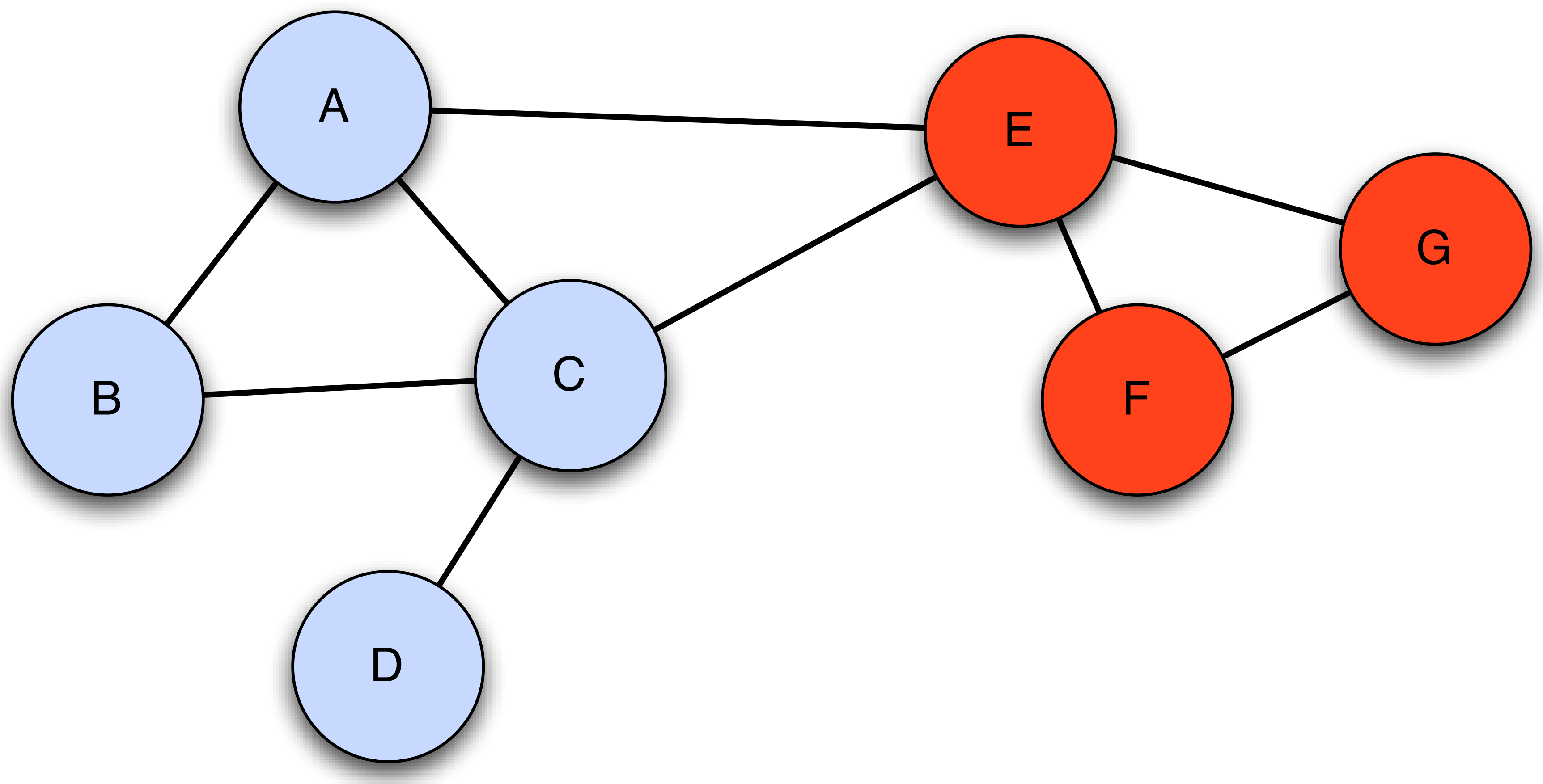
e_i = number of edges in network centered at node i
k_i = number of neighbors of node i

Degree distribution

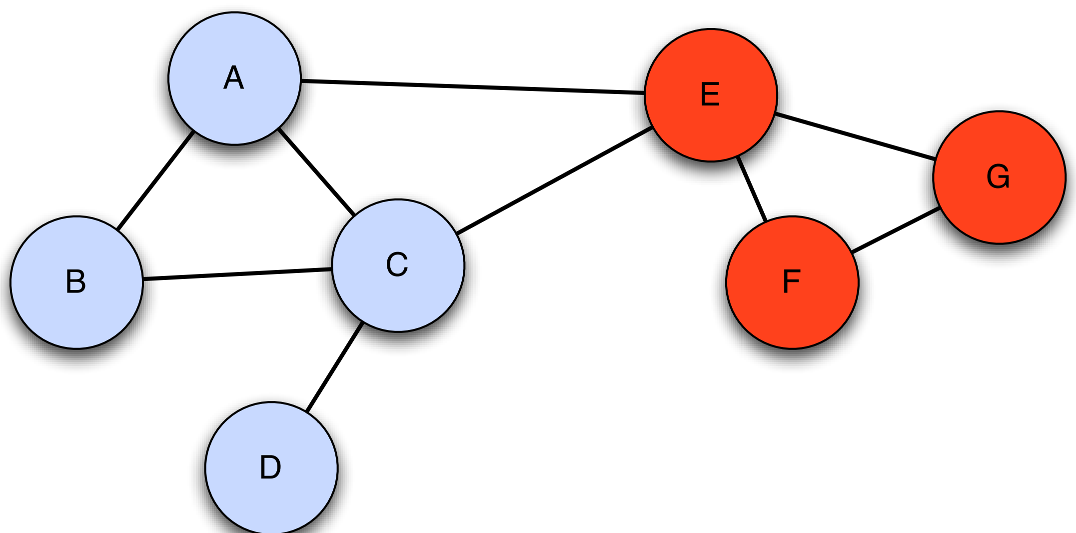




Assortativity



Assortativity



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

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

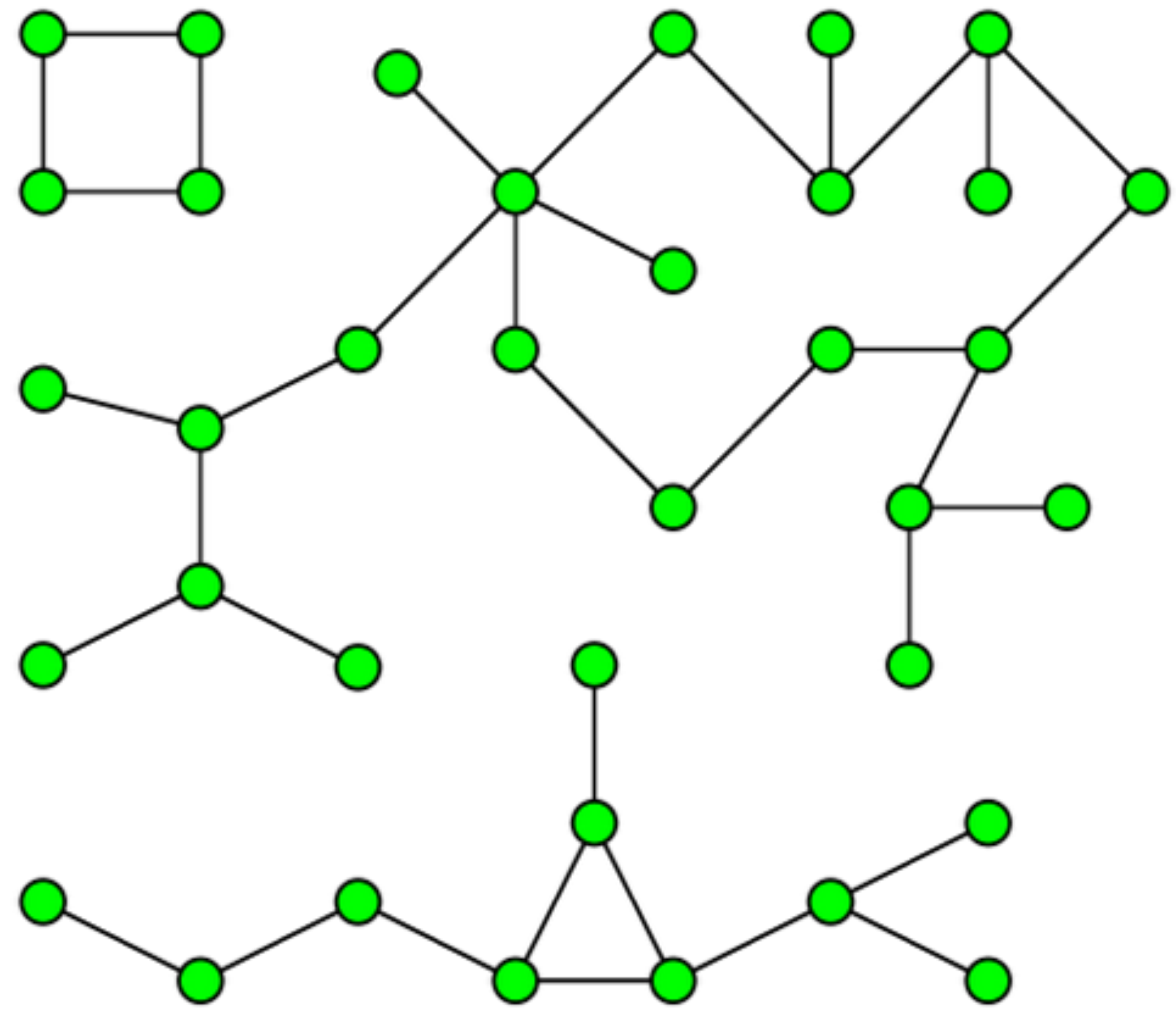
m = total number of edges in network

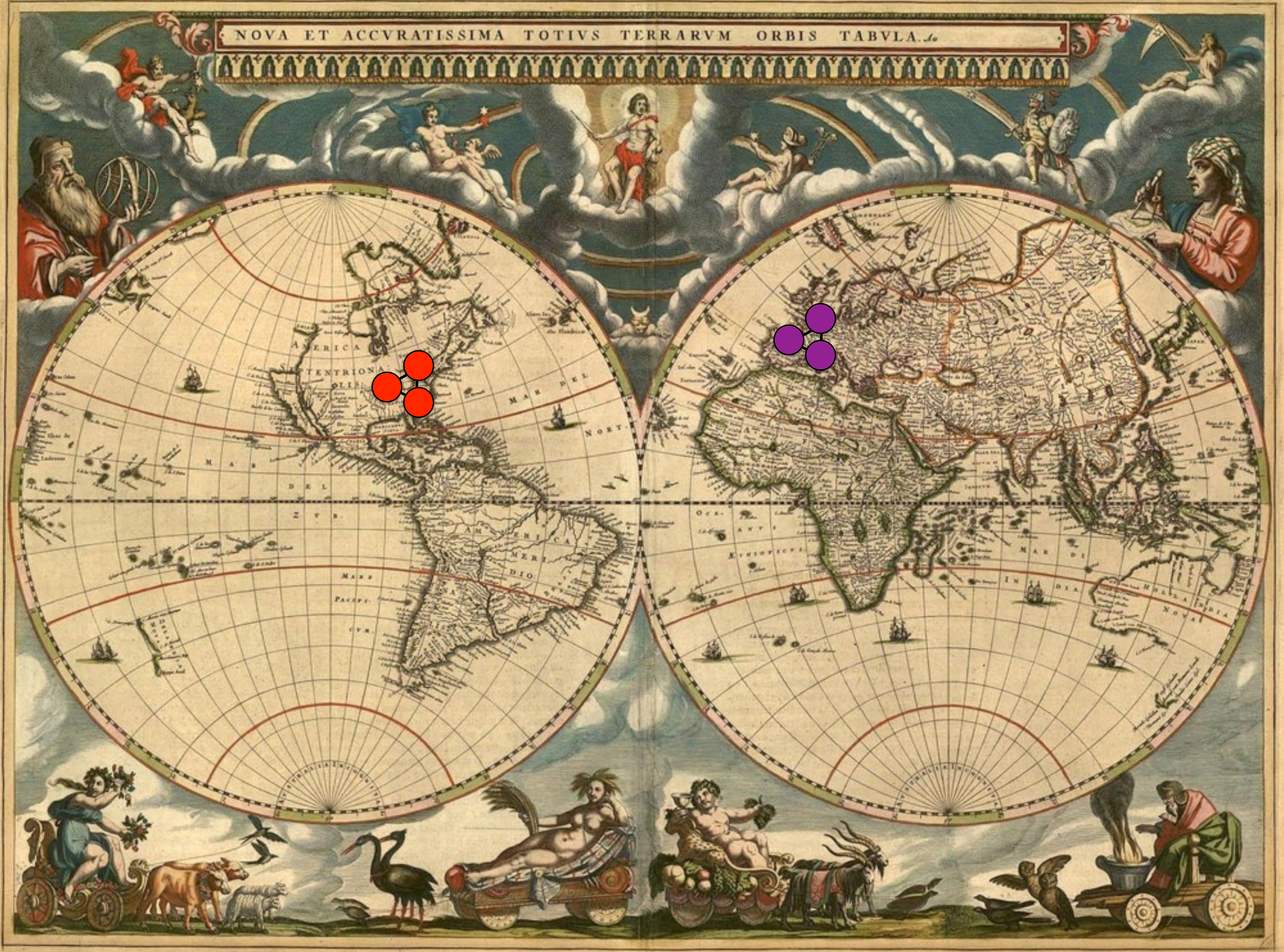
Connectivity

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

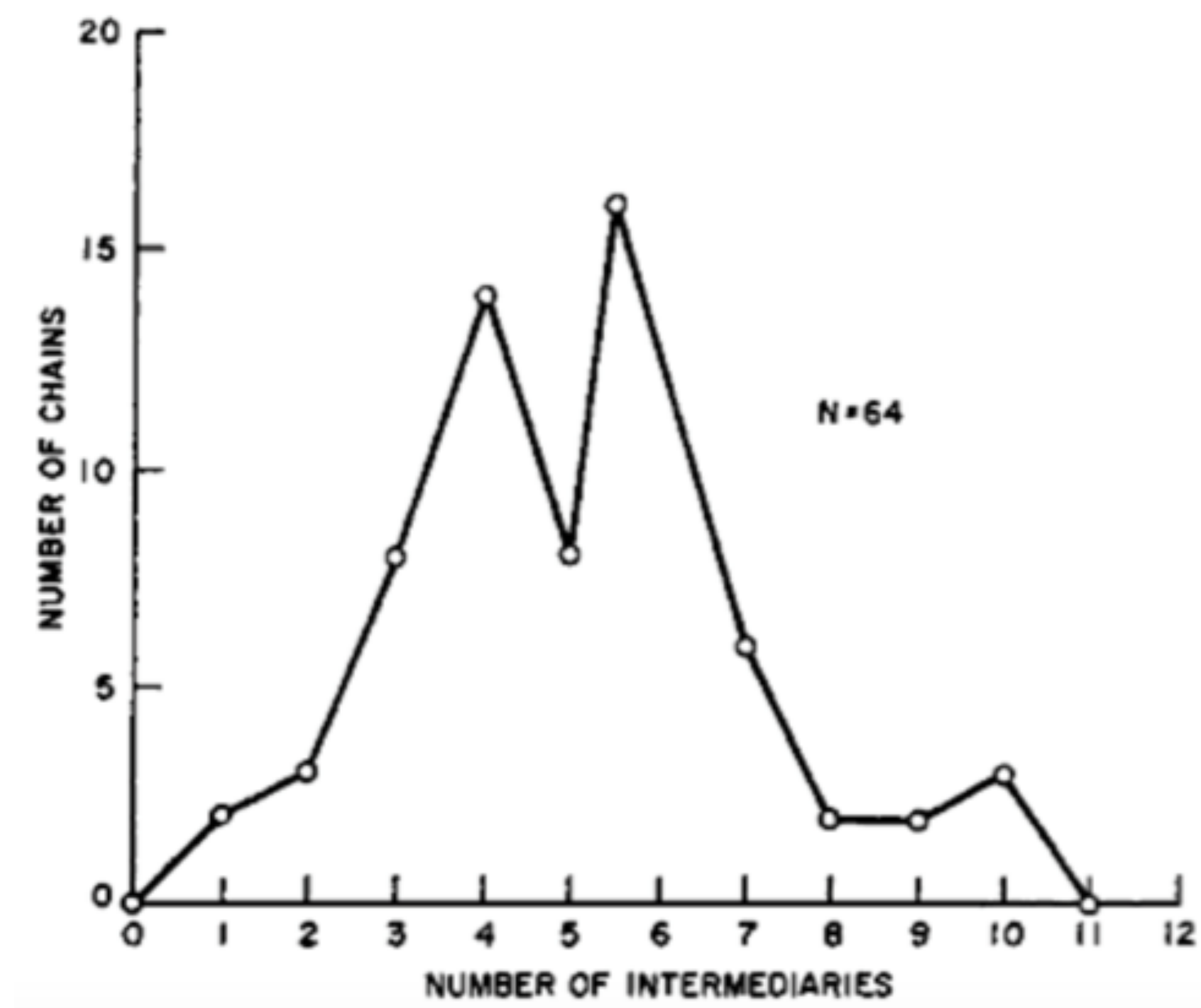



NOVA ET ACCVRATISSIMA TOTIVS TERRARVM ORBIS TABVLA.

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



Tie strength

- “Strong” ties vs. “weak” ties

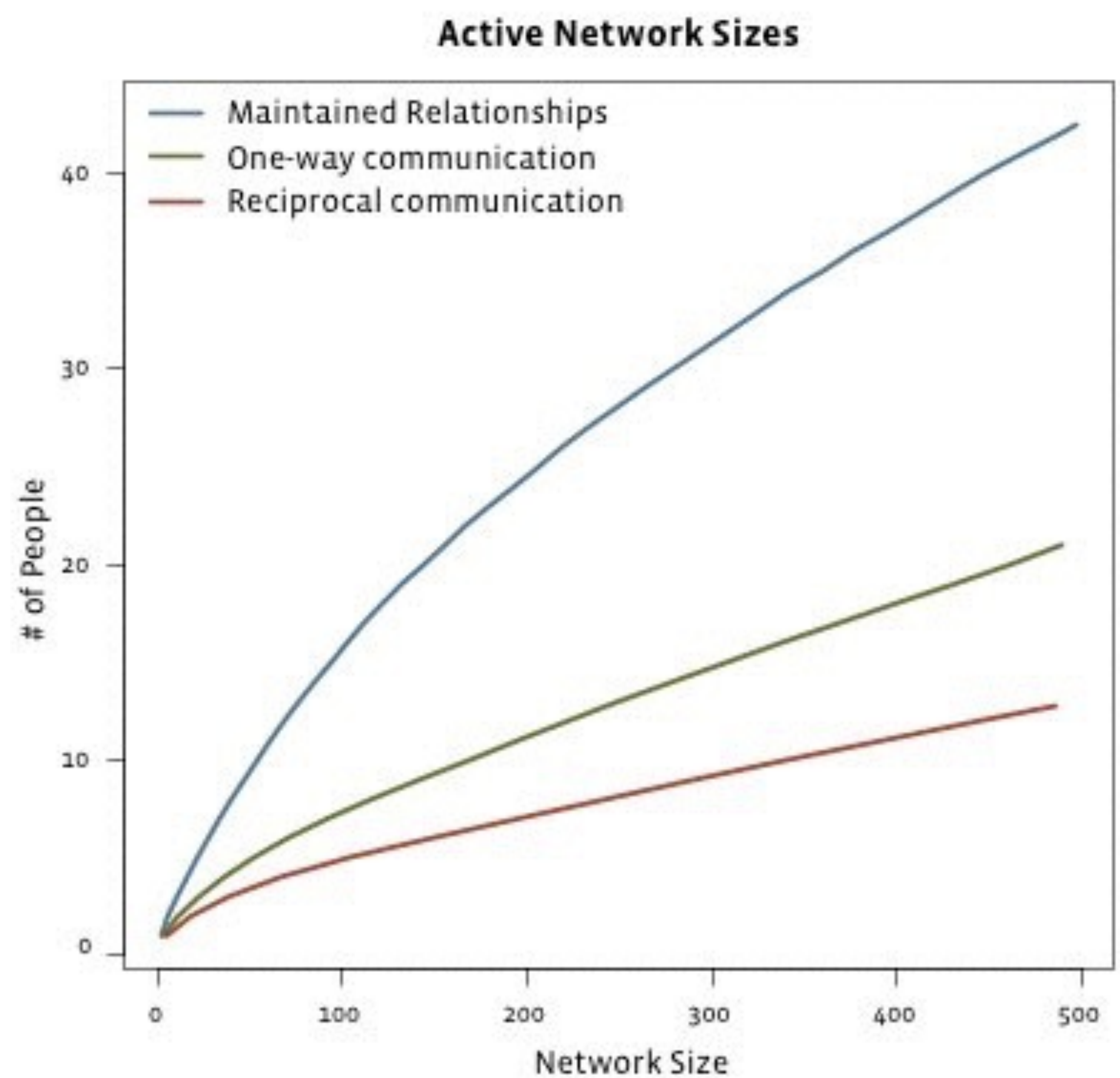
Tie strength

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

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

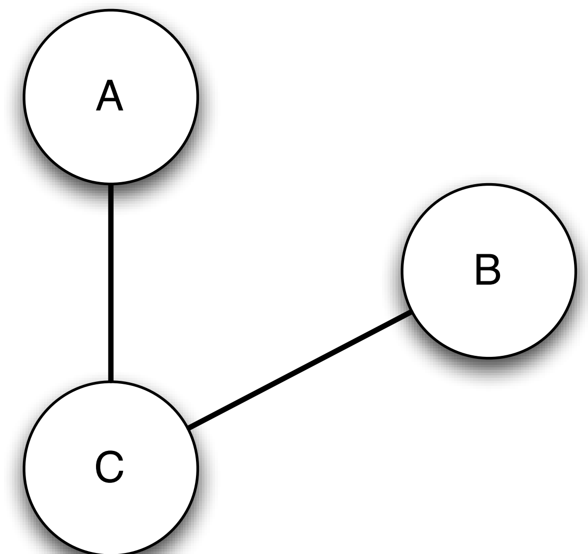
One-way: any directed message

Reciprocal: reciprocated message

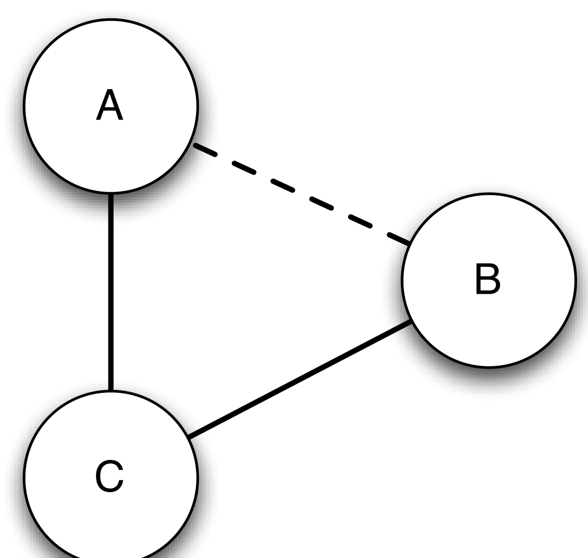


Triadic closure

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



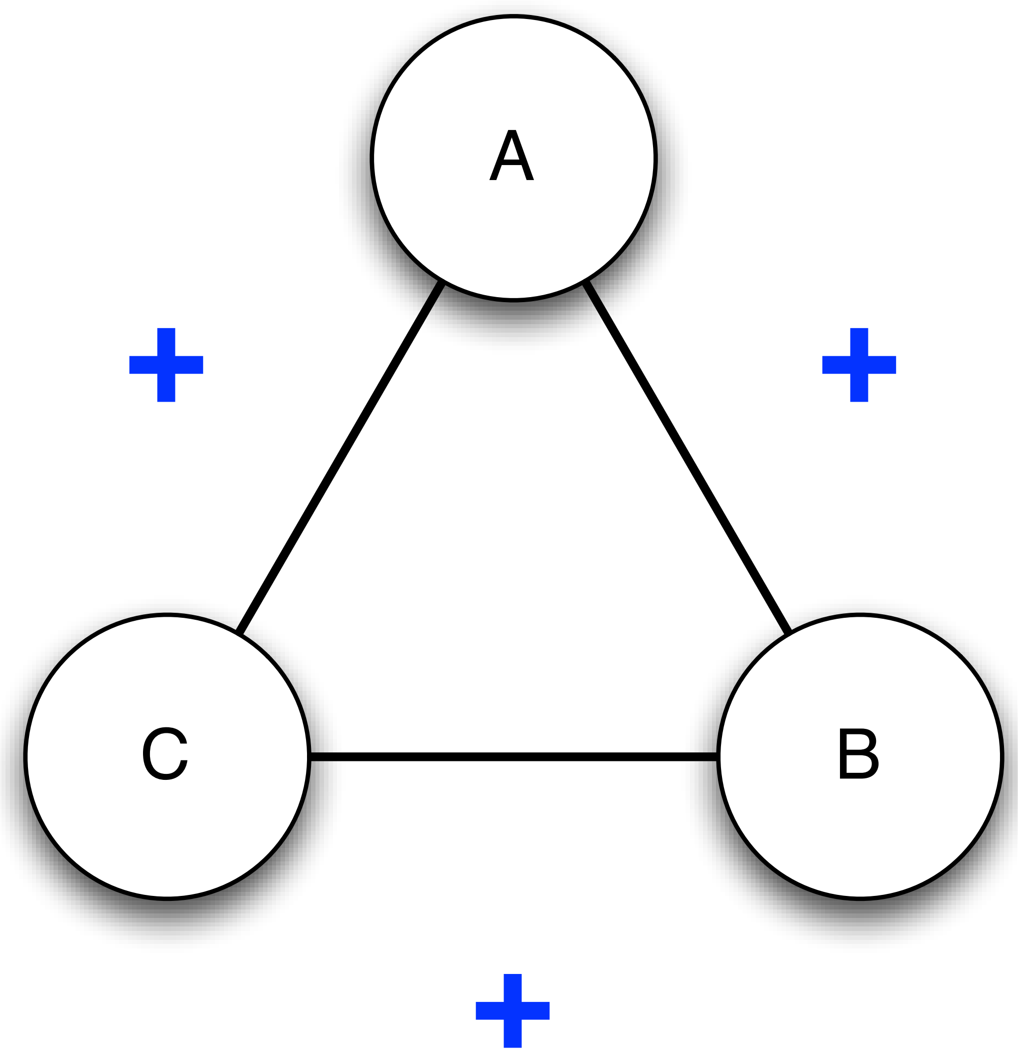
More likely the stronger the tie is between A-C and B-C.



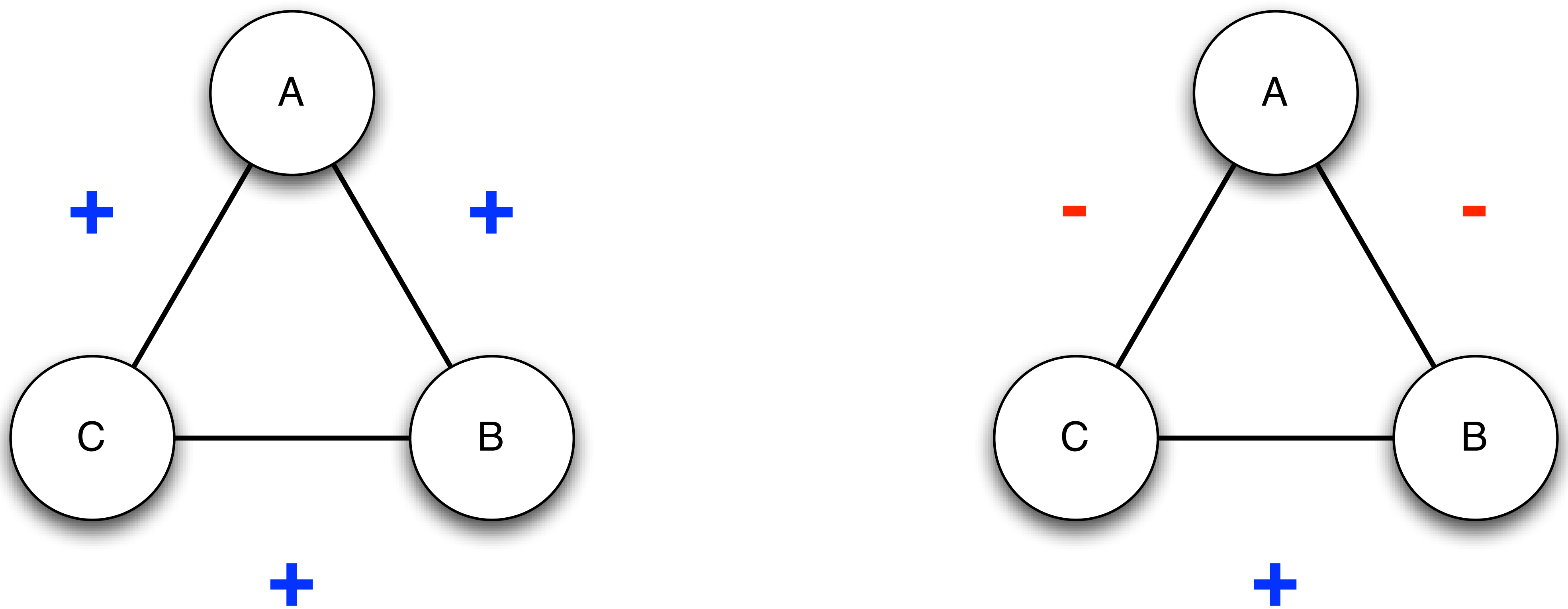
Triadic closure

- 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

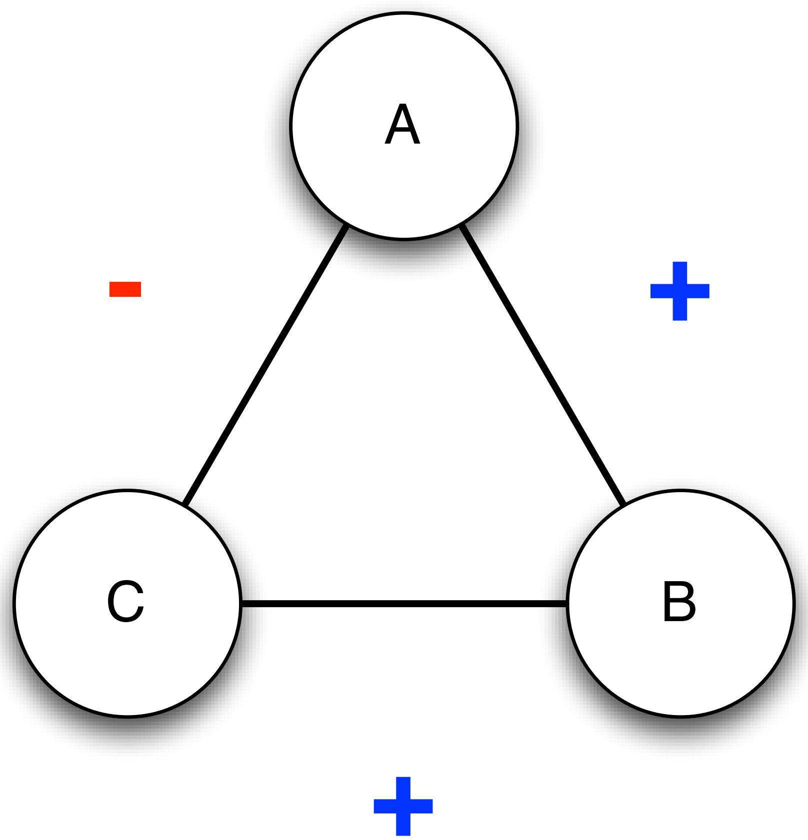
Structural balance



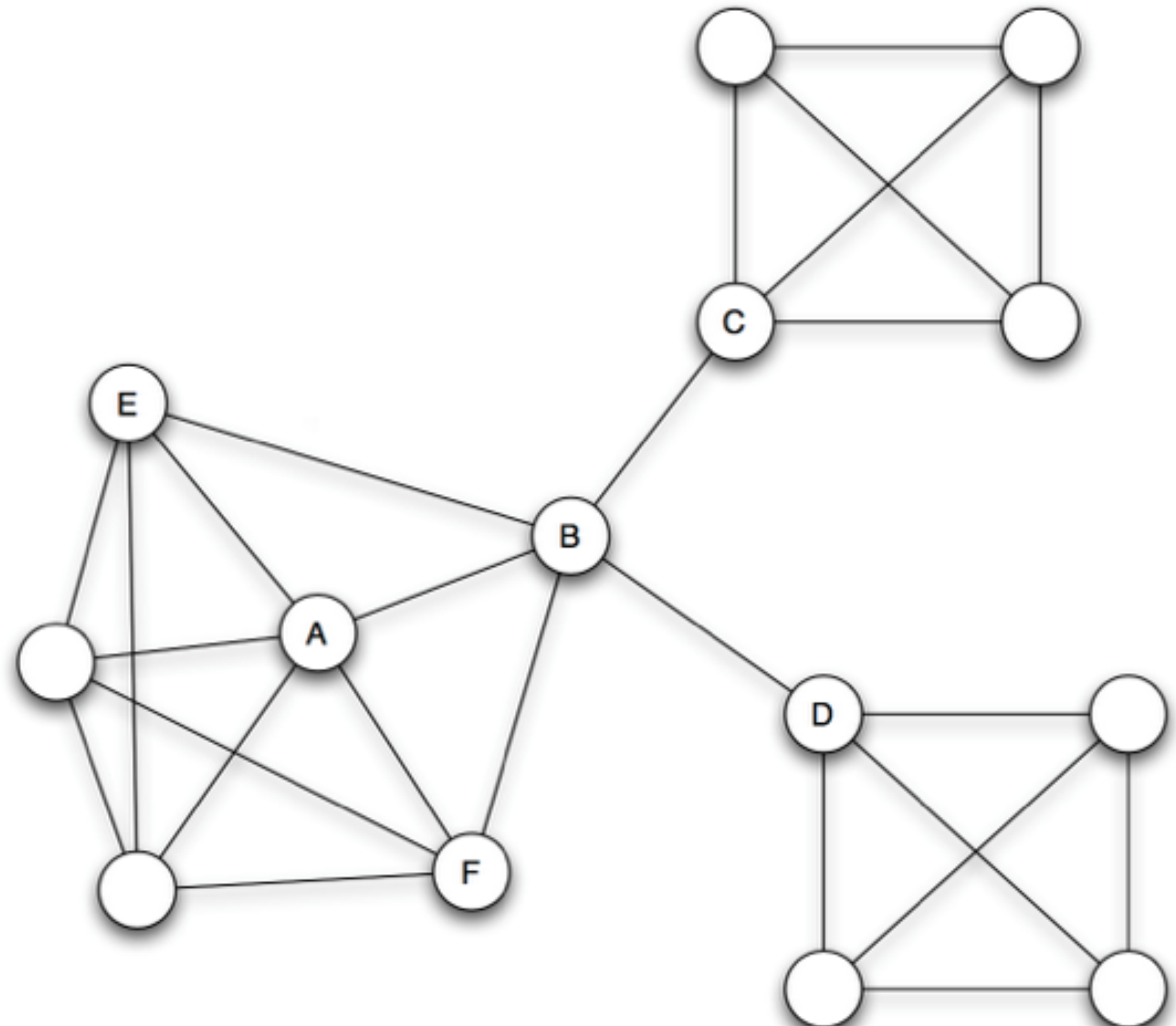
Structural balance



Structural balance



structural bridges



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

Networks

Network	Nodes	Edges	Information
Social	People		
Internet	Servers		
Citation network	Articles		
Web	Web pages		

Information flows

- Information effects (herding behavior)
- Direct-benefit effects
- Epidemics

Herding behavior

- Lines outside restaurants/clubs
- Crowd of people looking up (Milgram et al. 1969)

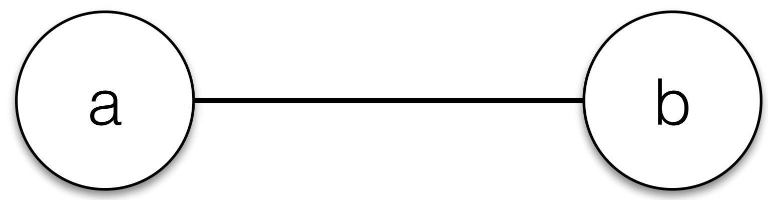
- Inference that observed choices are more powerful than own private information

Direct benefit effects

- Direct payoffs for making the same decisions others make

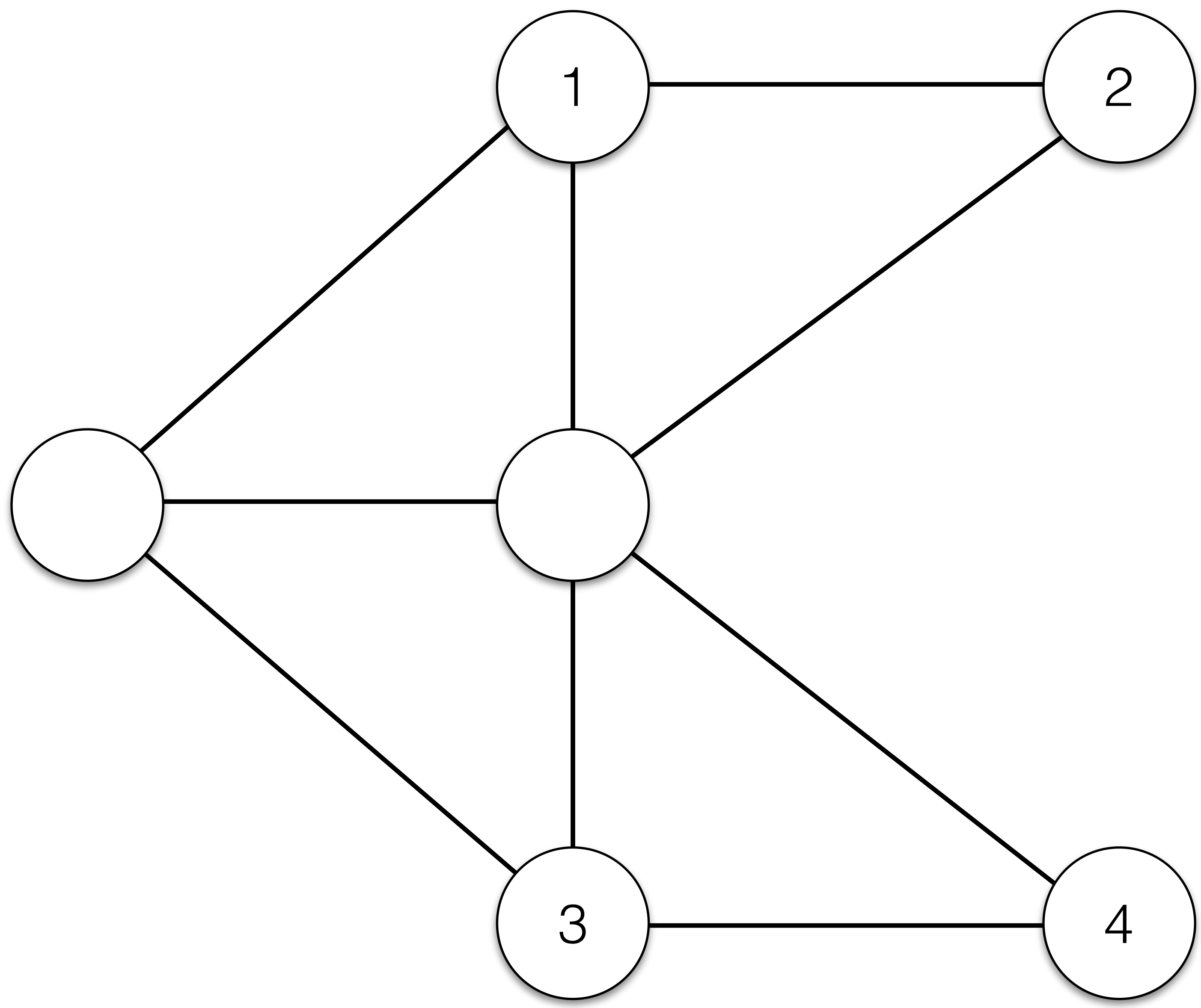
- Social networking sites
- Cell phone providers
- Mac/PC

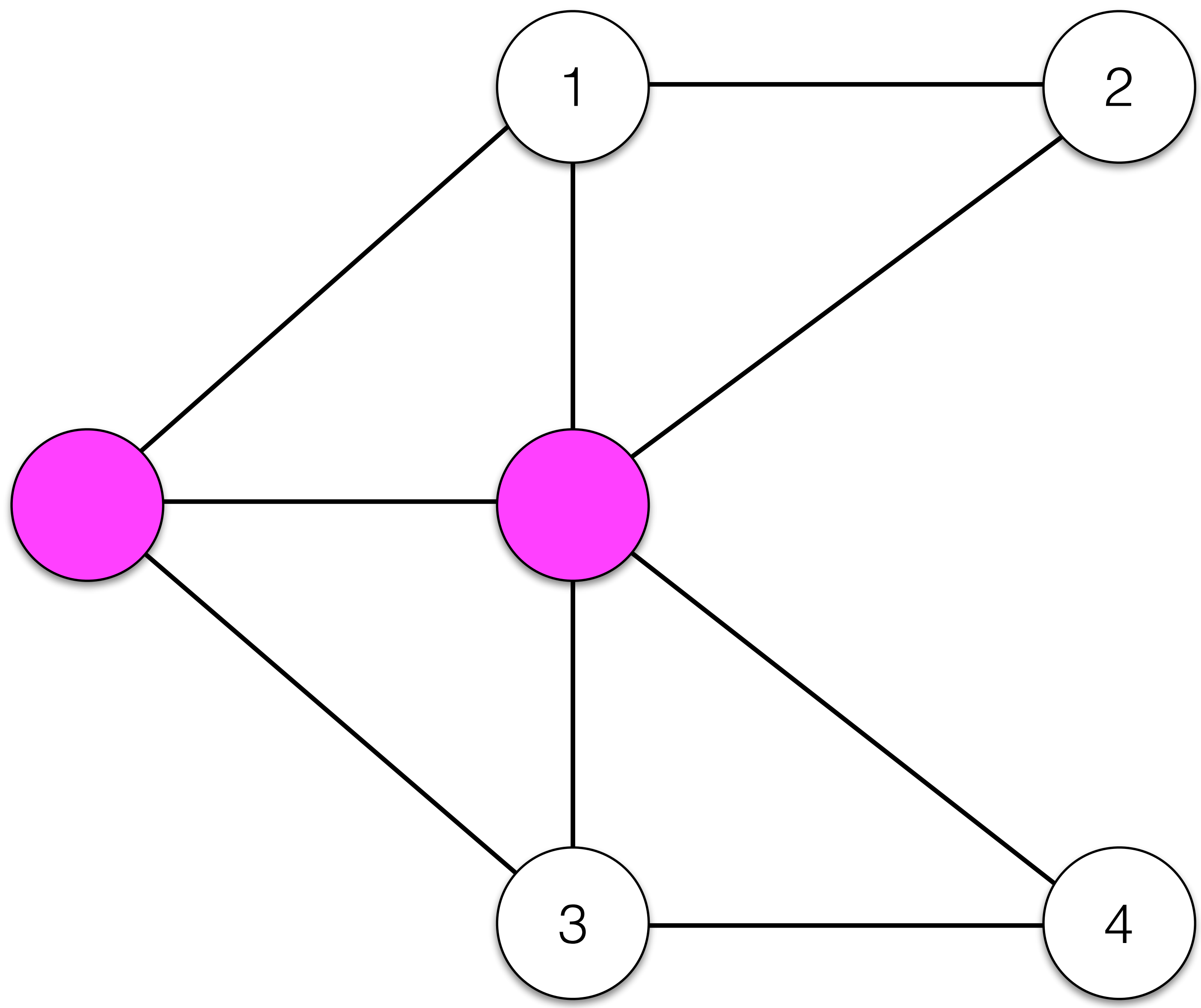
Direct benefit effects



- a and b adopt A, they get a payout of x
- a and b adopt B, they get a payout of y
- otherwise they get a payout of 0

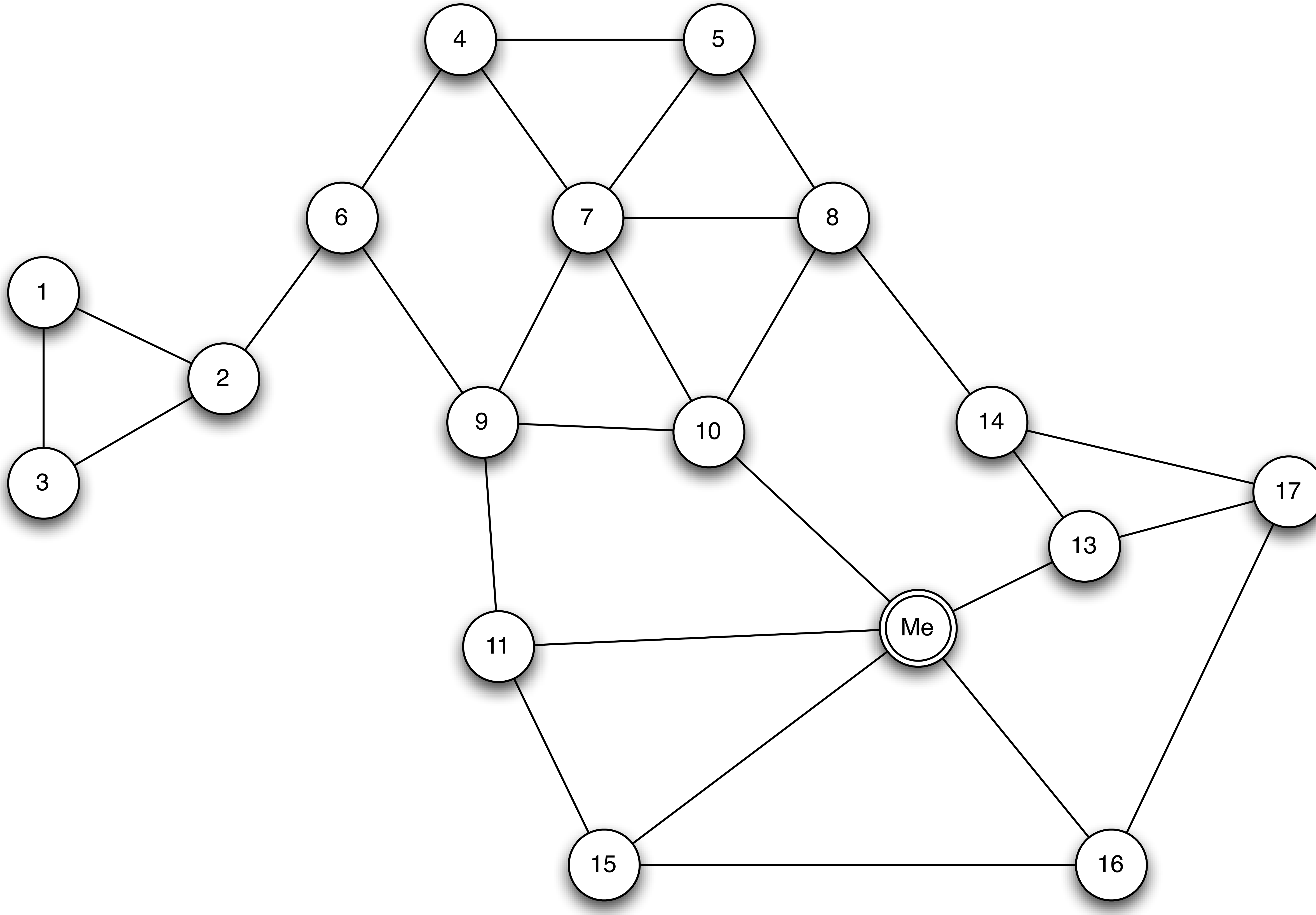
Easley and Kleinberg 2010

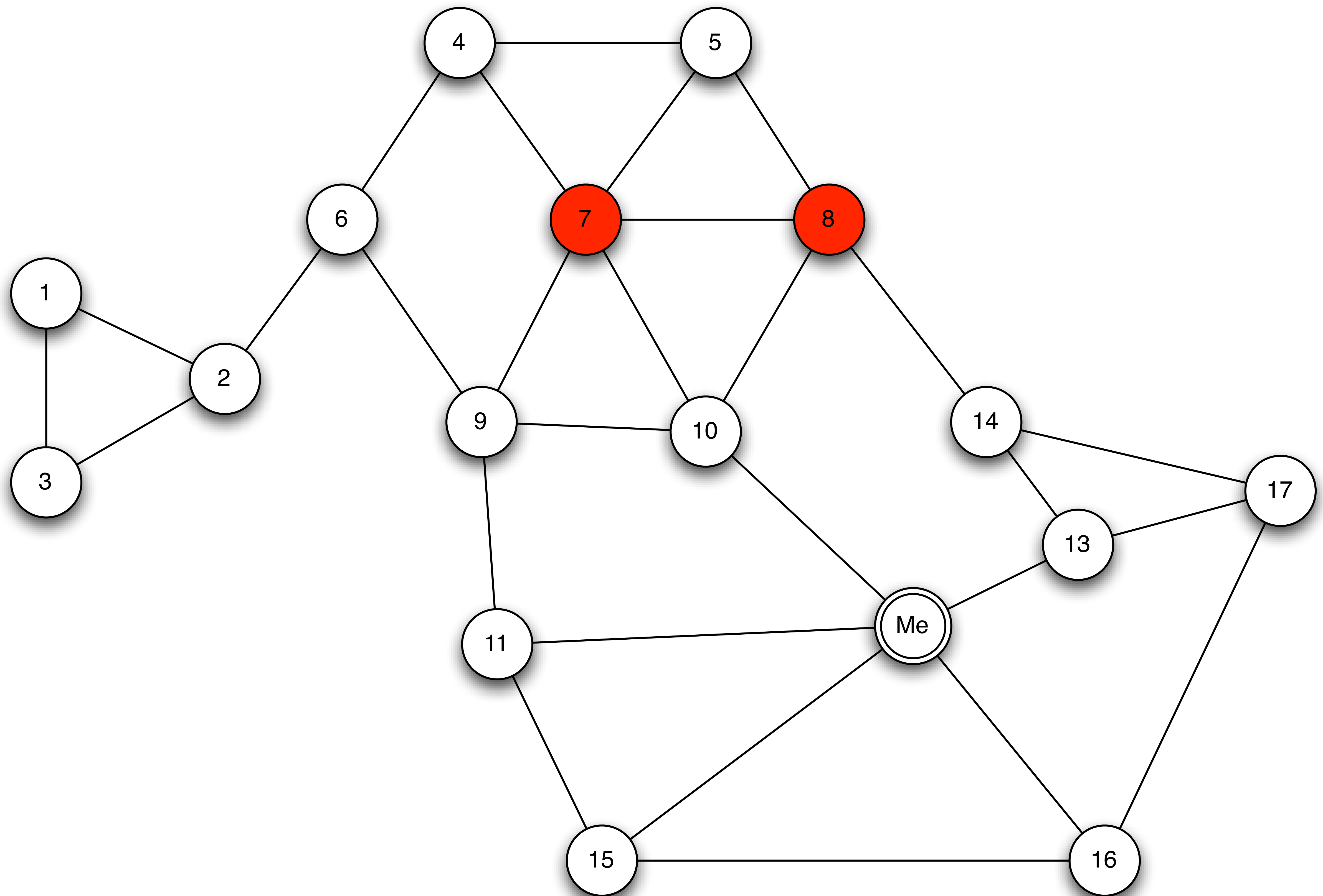

1
2
3
4

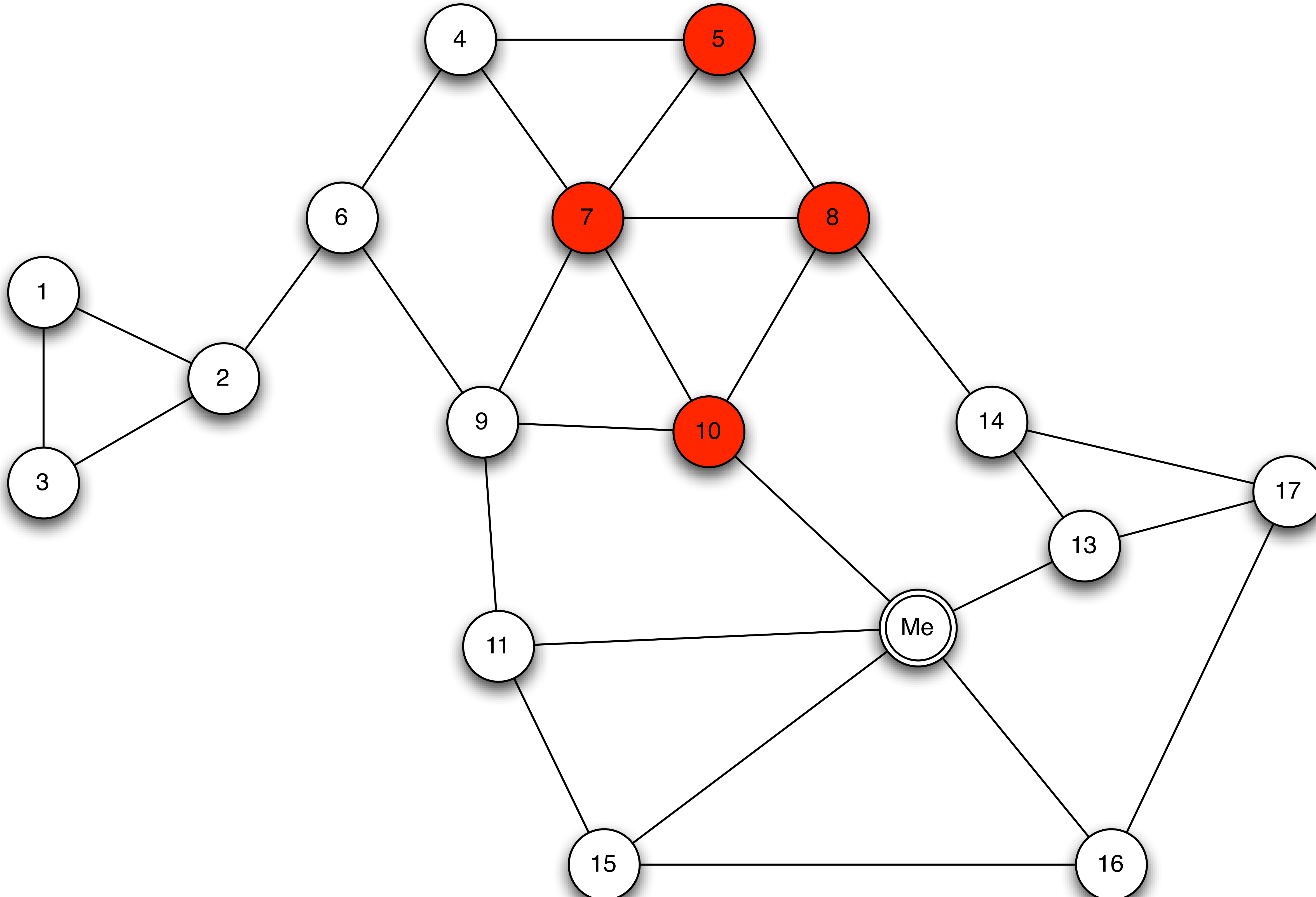


a=3
b=2

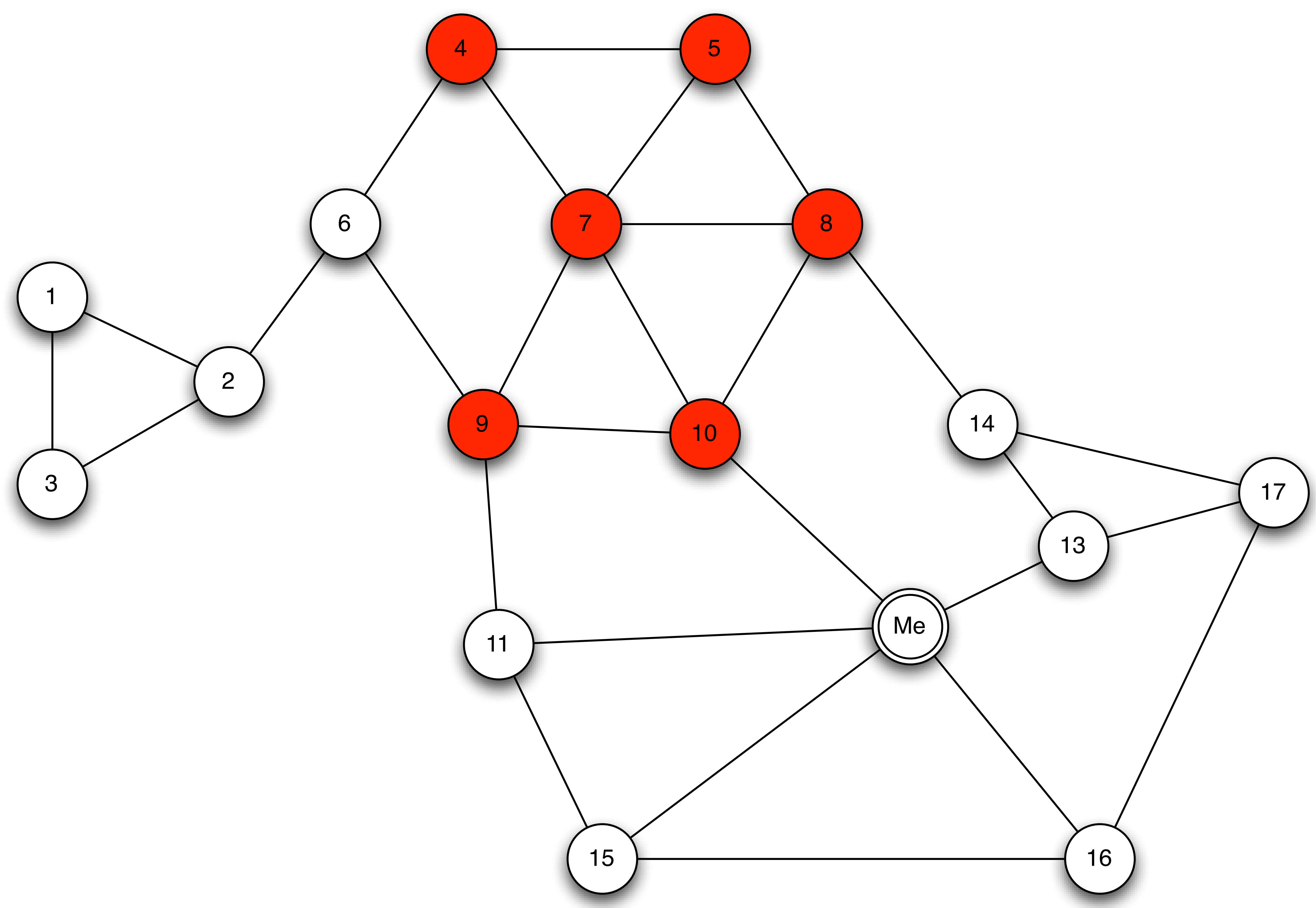
- The topology of the network has consequences for diffusion

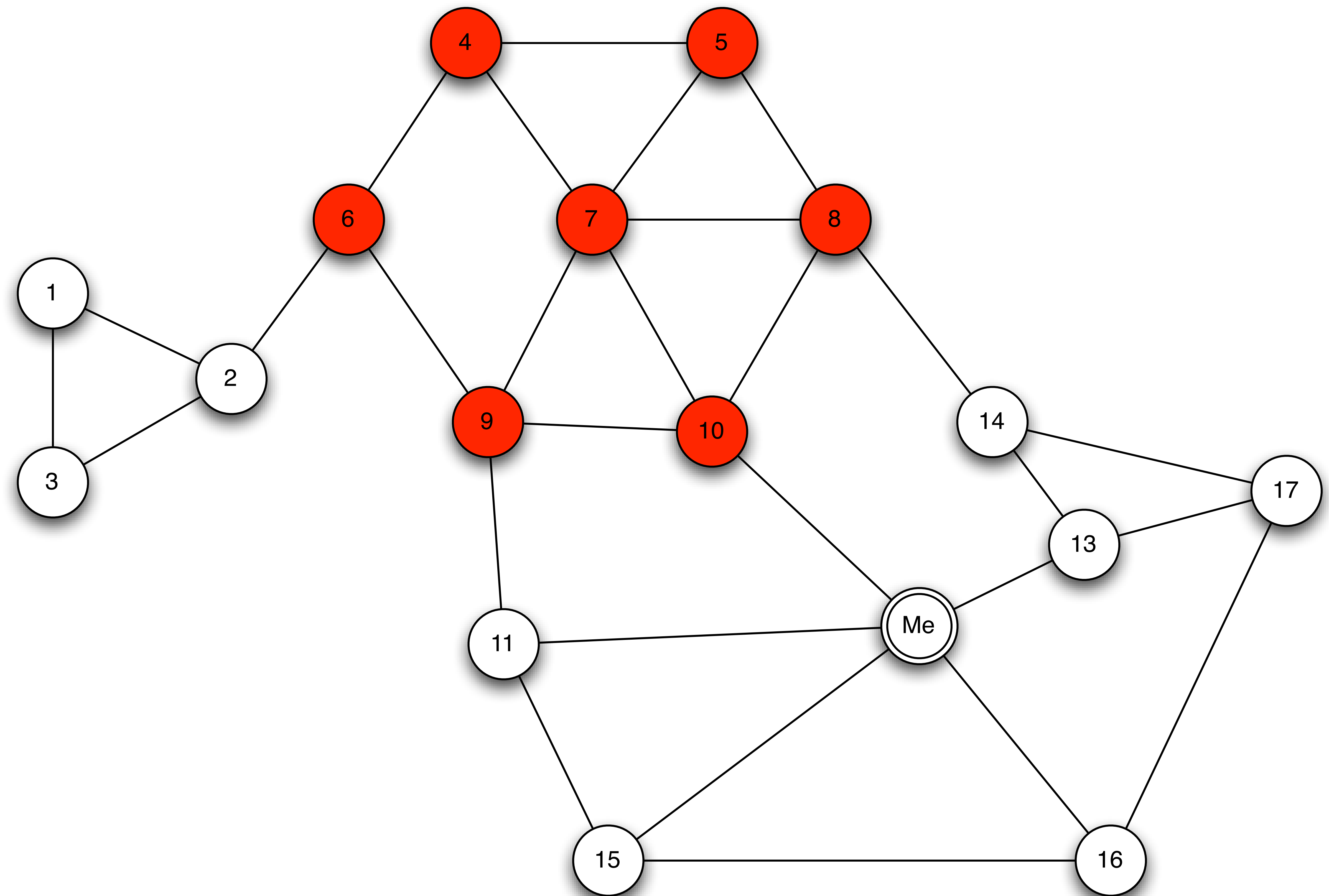

1
2
3
4
5
6
7
8
9
10
11
Me
13
14
15
16
17


1
2
3
4
5
6
7
8
9
10
11
Me
13
14
15
16
17



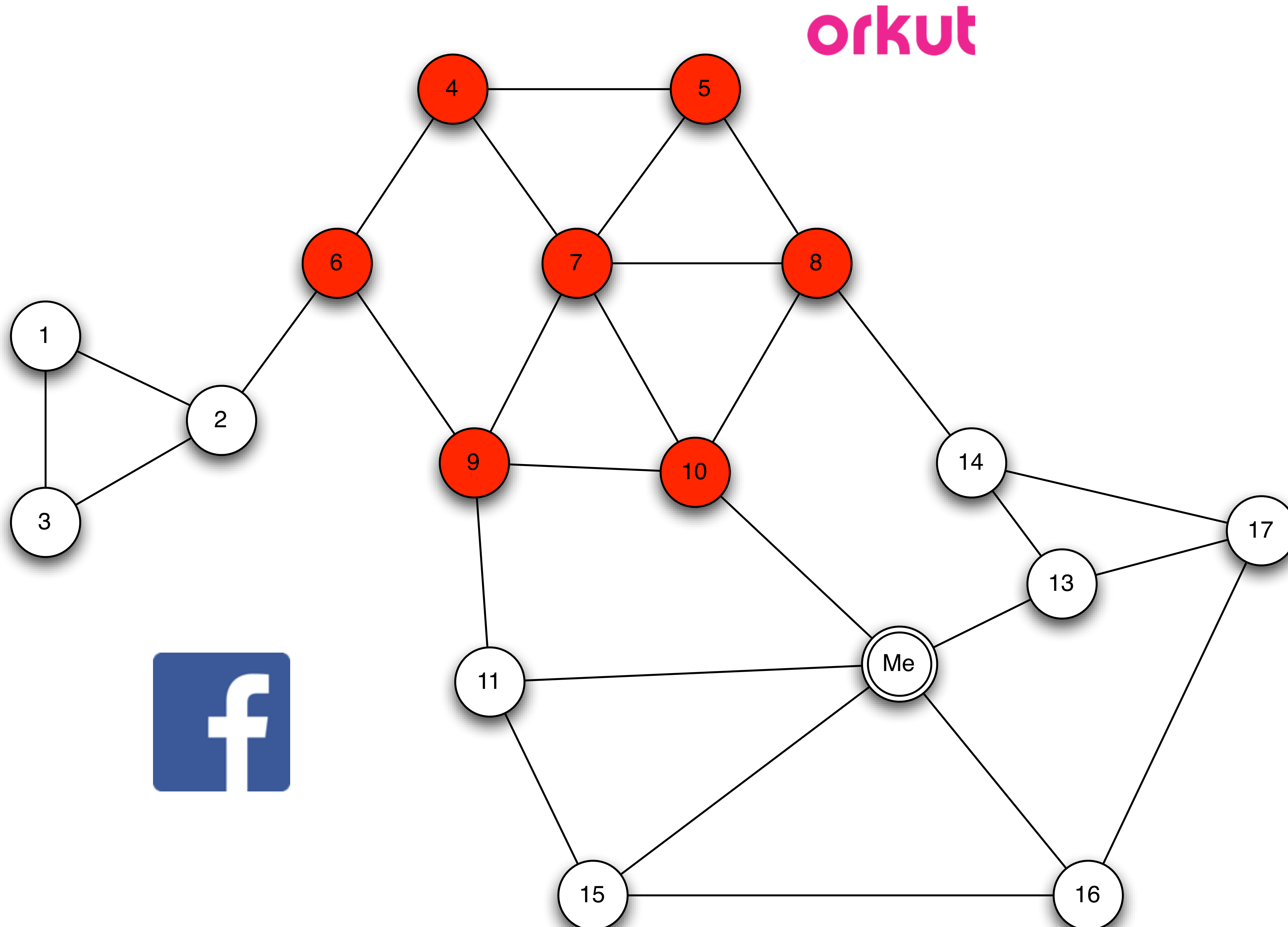
4
5
6
7
8
1
2
9
10
14
3
17
13
Me
11
15
16


4
5
6
7
8
1
2
9
10
14
3
17
13
Me
11
15
16

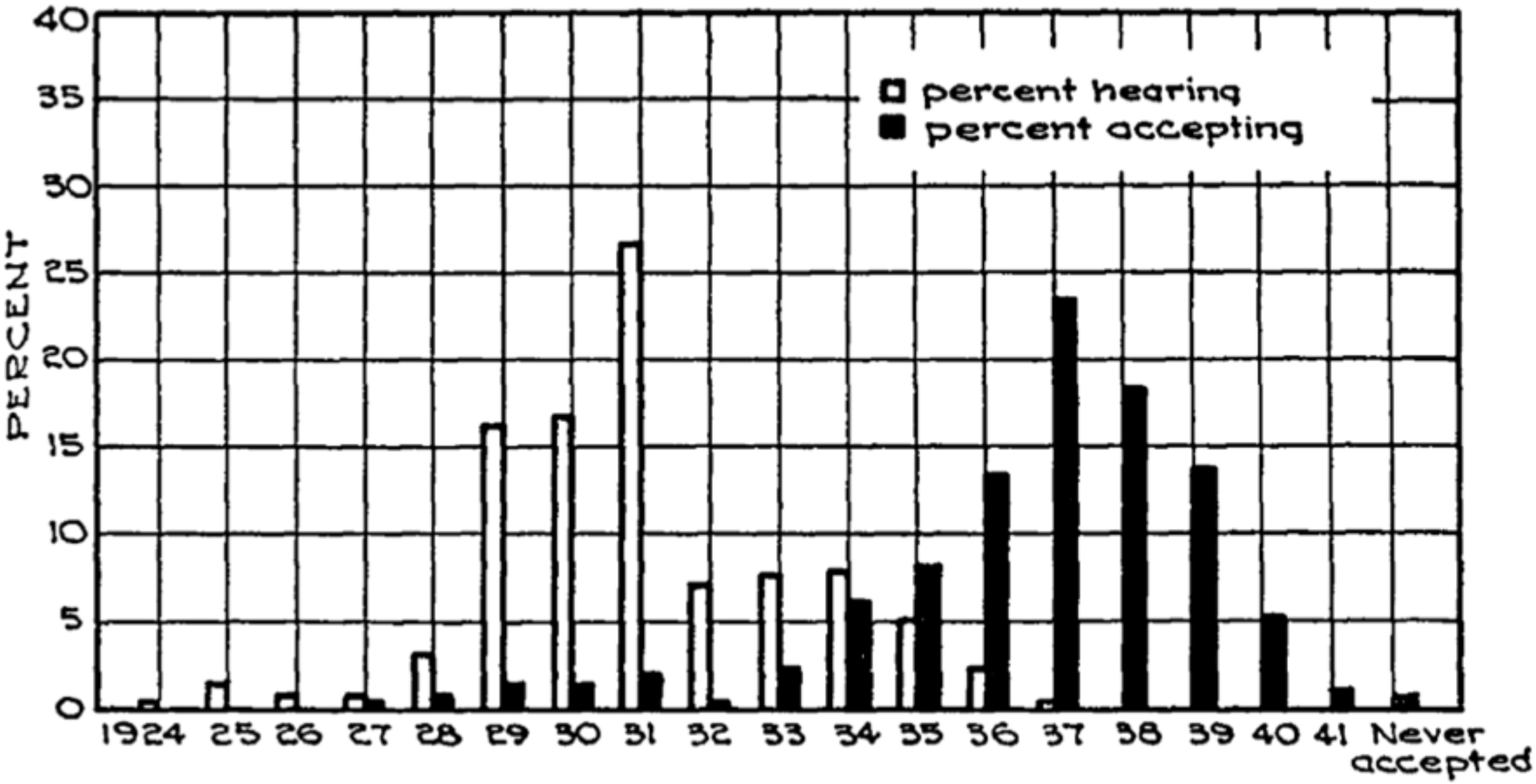

1
2
3
4
5
6
7
8
9
10
11
13
14
15
16
17
Me

- Tightly connected communities can hinder the spread of innovation

- Viral marketing: how do you choose the nodes where you can maximize adoption in the network?


orkut
4
5
6
7
8
1
2
9
10
14
3
17
13
11
Me
15
16

Information vs. adoption



Ryan & Gross (1943), “The Diffusion of Hybrid Seed Corn in Two Iowa Communities,” Rural Sociology

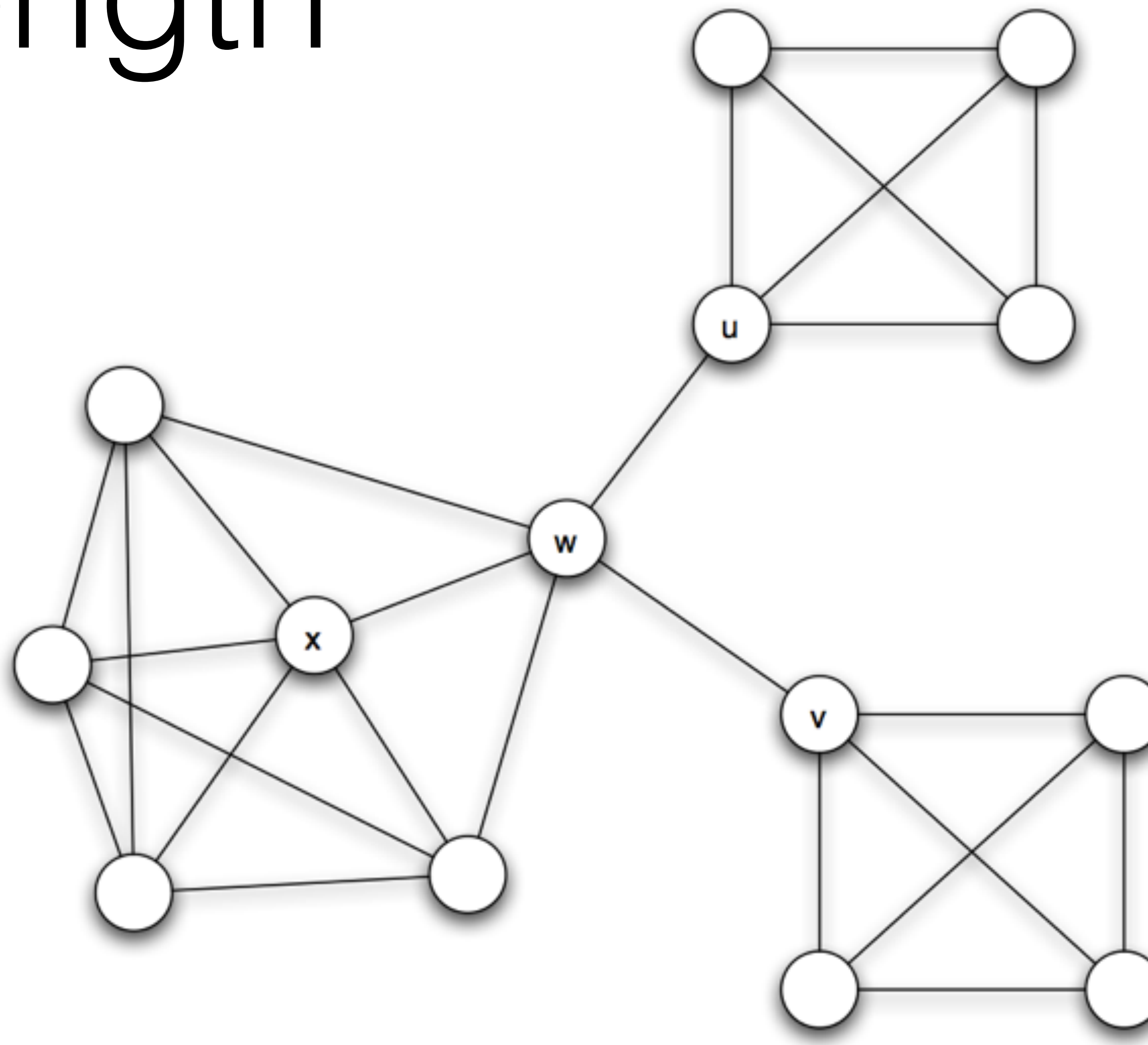
Diffusion of innovations

- Spread of a new technology/idea through a social network
- Common principles (Rogers 1995):
    - complexity. How easy can people understand it?
    - observability. How transparent is it when others are using it?
    - trialability. Can it be adopted incrementally?
    - compatibility. How comparable is it with existing practices?

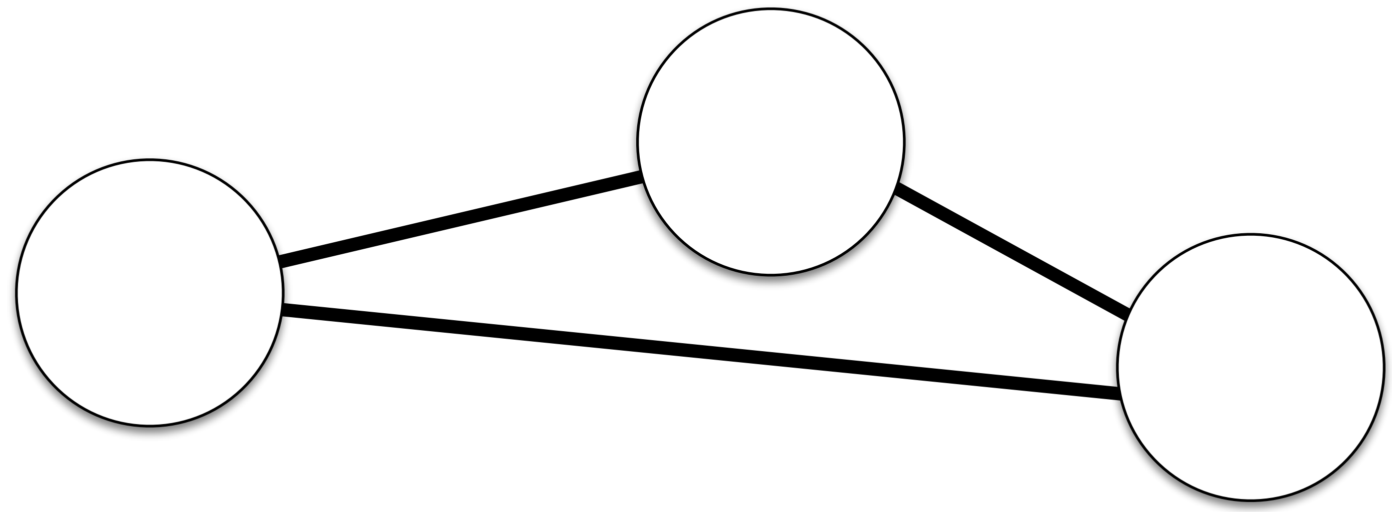
Tie strength

- Hearing about vs. adopting innovation

- Bridges are powerful for conveying awareness, but not uptake



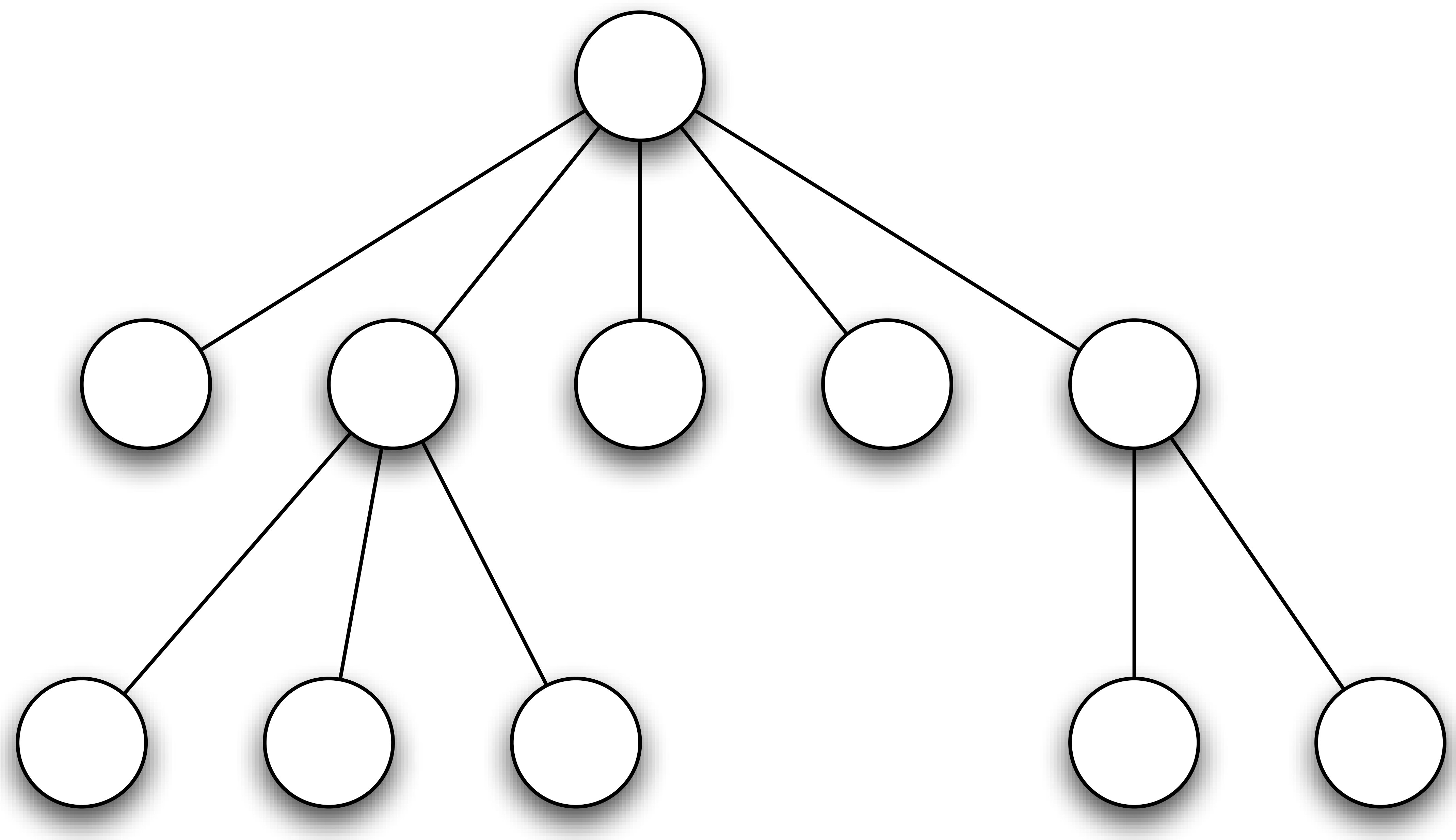
Diffusion as Epidemic



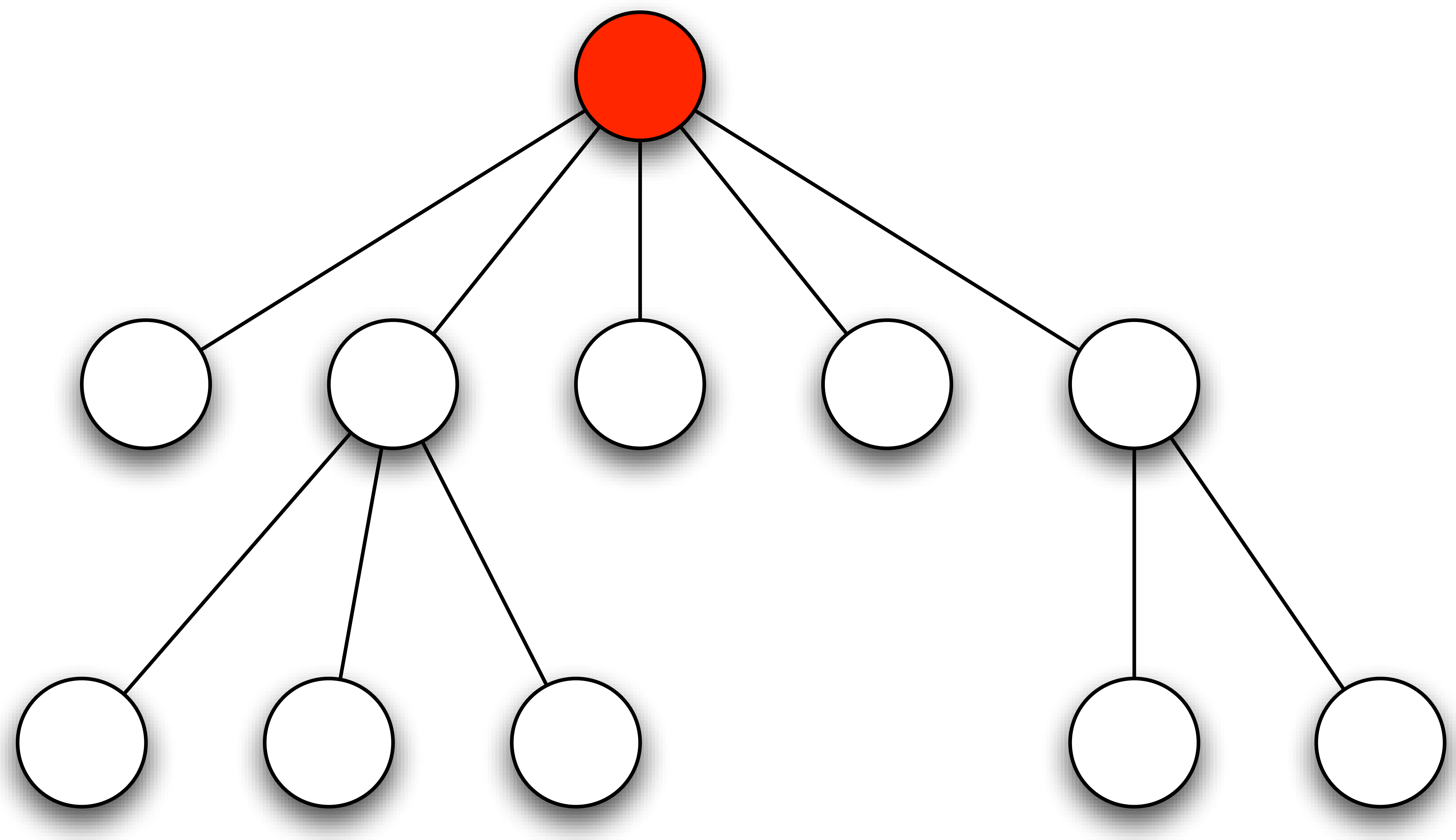
Network	Nodes	Edges	Information
	People		Disease

How does the network change as a function of the disease?

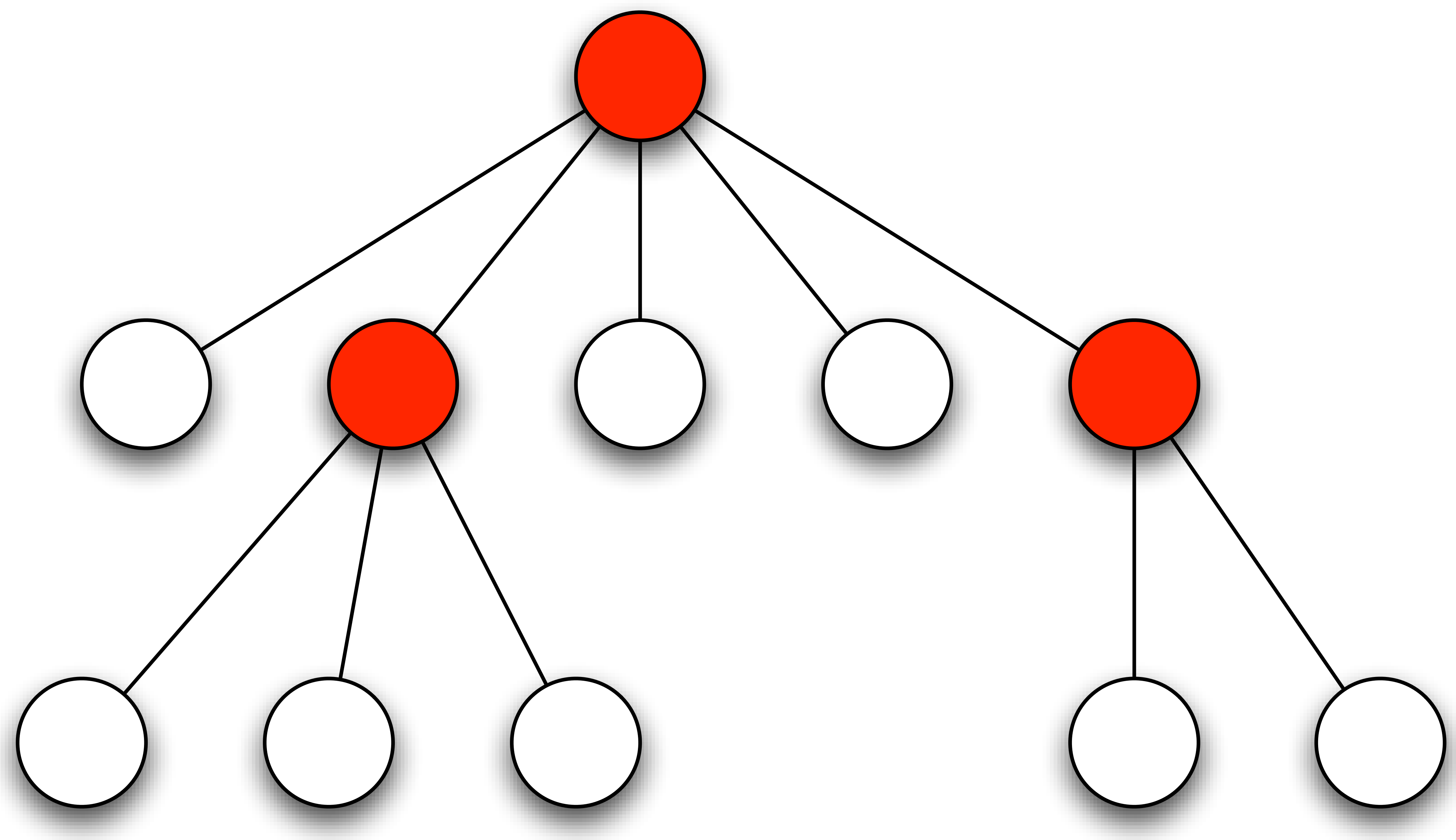
Diffusion as Epidemic



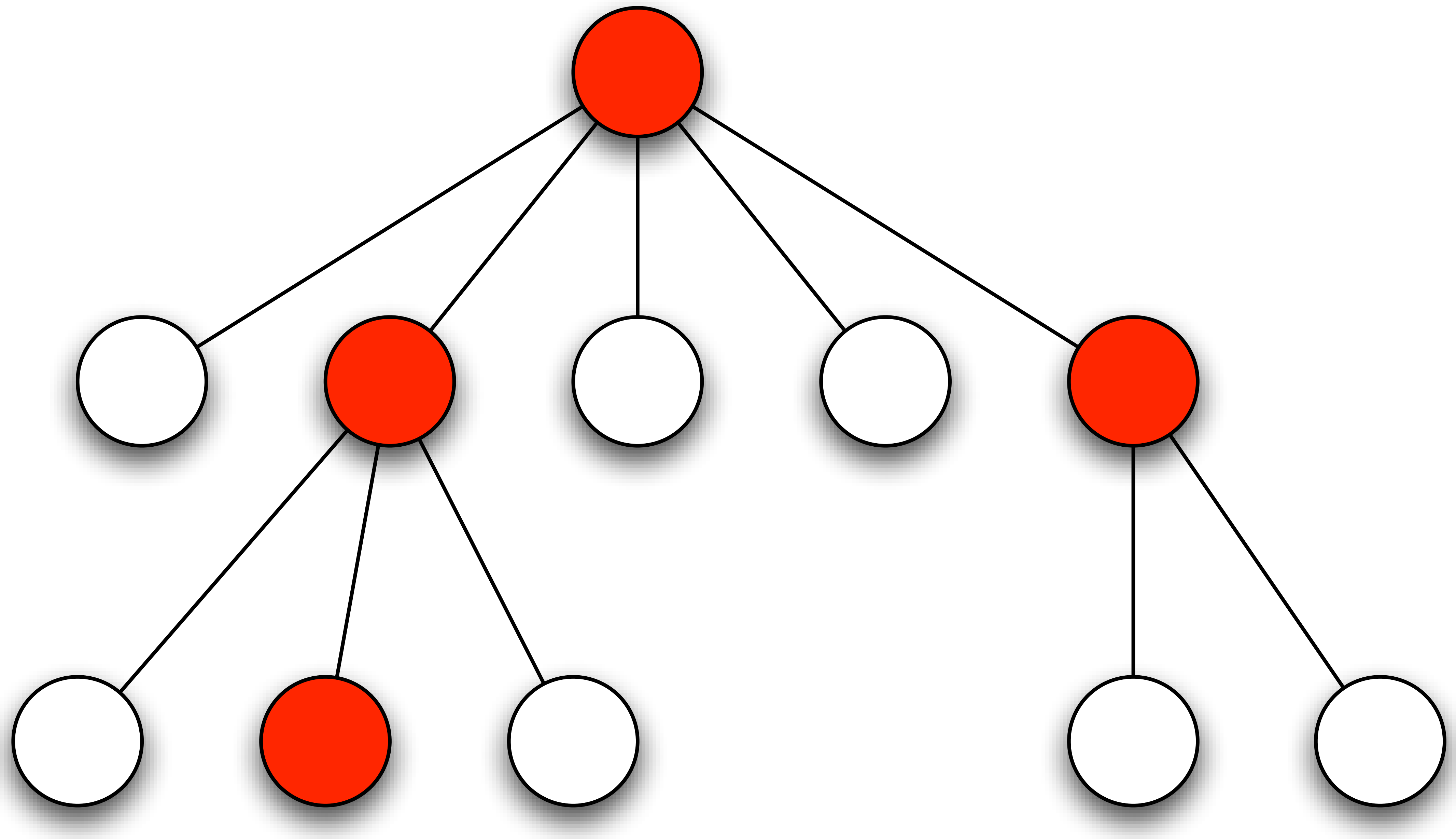
Diffusion as Epidemic



Diffusion as Epidemic



Diffusion as Epidemic



Basic Reproductive Number (R_0)

- Expected number of new infections caused by a randomly selected person in the population

Disease	R_0
1918 Flu	2-3
SARS	2-5
HIV	2-5
Polio	5-7
Smallpox	5-7
Measles	12-18

Basic Reproductive Number (R_0)

- In tree models, $R_0 = p \times k$
- p = probability of infecting 1 person
- k = number of people in contact with

decrease p by preventing spread of disease

decrease k by quarantine

Data

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

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