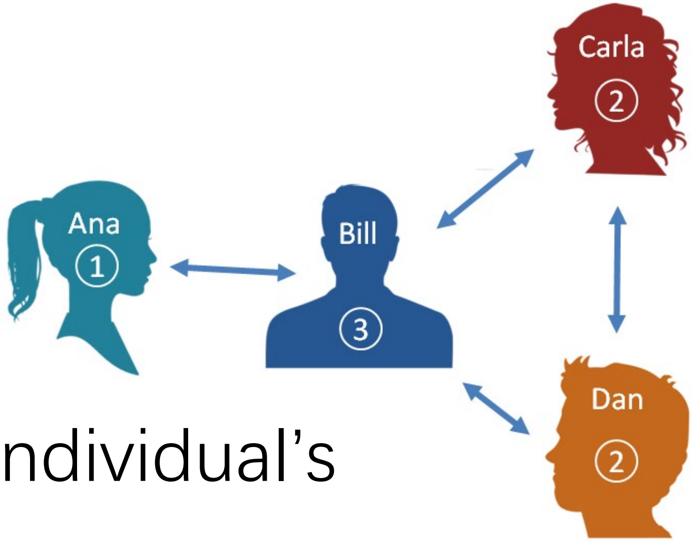
CS3319 Foundations of Data Science

6. Social Networks

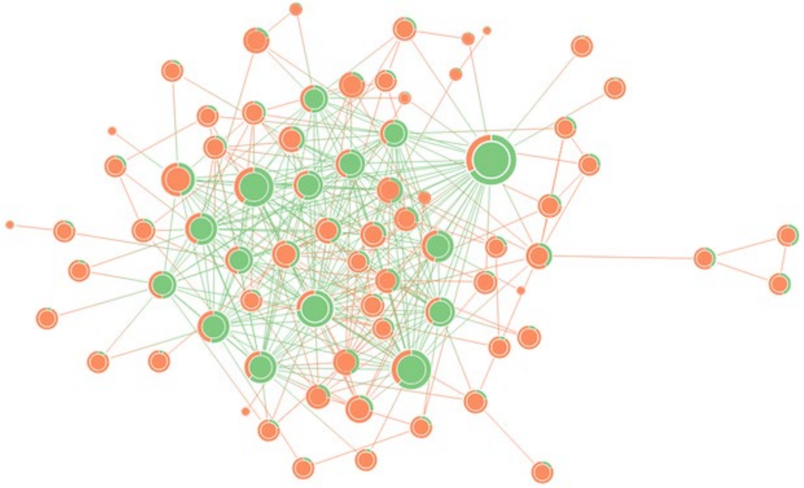
Jiaxin Ding
John Hopcroft Center

上海交通大学
约翰·霍普克罗夫特计算机科学中心
John Hopcroft Center for Computer Science

- Do you have more friends than your friends?



- What about the expectation of the number of an individual's friend vs the number of her/his friend's friends?
- Friendship paradox



- Average friend number of an individual:

$$\mu = \frac{\sum_{v \in V} d(v)}{|V|} = \frac{2|E|}{|V|}$$

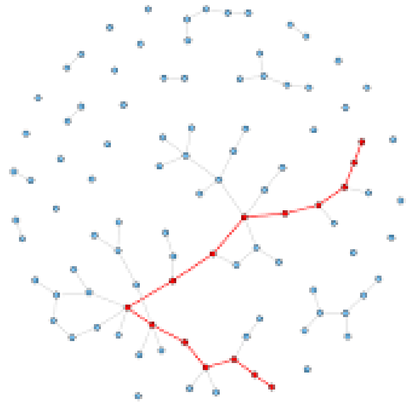
- Average friend number of his friend:

$$\sum_v \left(\frac{d(v)}{|E|} \frac{1}{2} \right) d(v) = \frac{\sum_v d(v)^2}{2|E|}$$

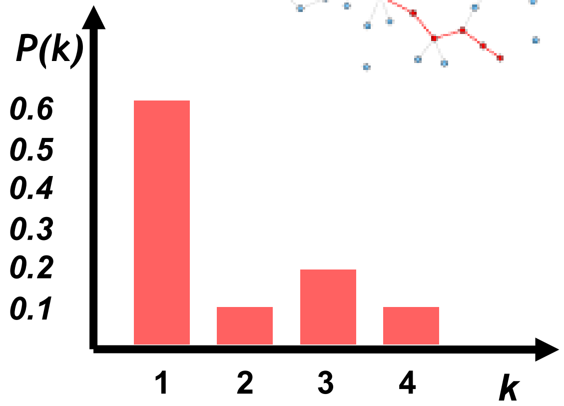
$$= \frac{|V|}{2|E|}(\mu^2 + \sigma^2) = \frac{\mu^2 + \sigma^2}{\mu} = \mu + \frac{\sigma^2}{\mu}$$

Key Network Properties

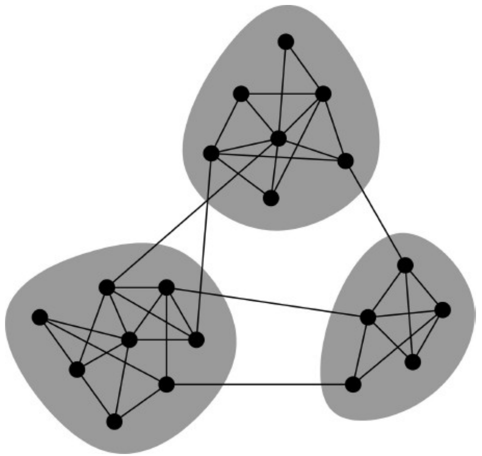
- **Network diameter:** The maximum distance between any pair of nodes in a graph



- **Degree distribution:** Probability that a randomly chosen node has degree $\boldsymbol{k}$

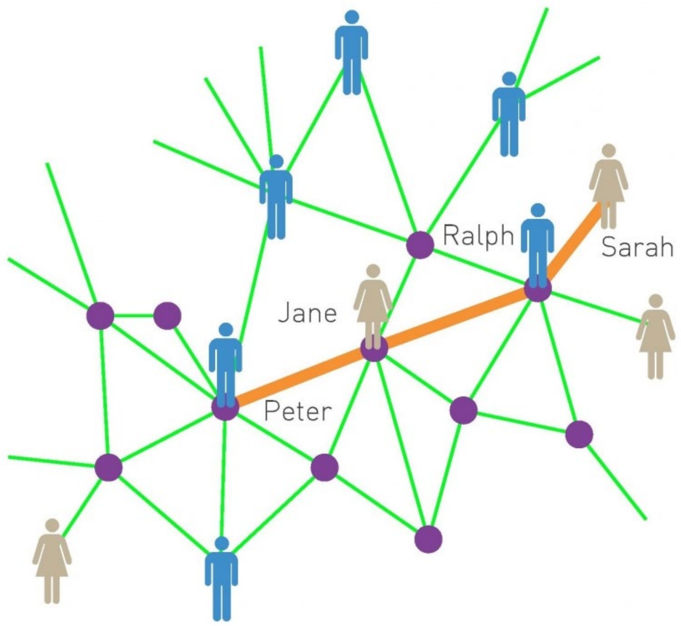


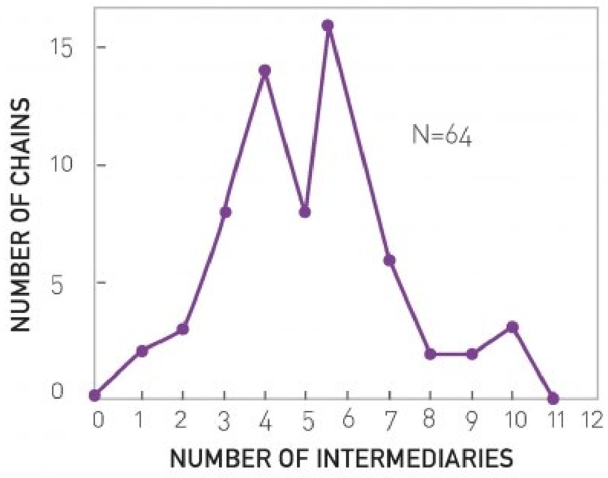
- **Community/cluster:** tightly connected component
  - Social networks usually have clusters (high clustering coefficient)



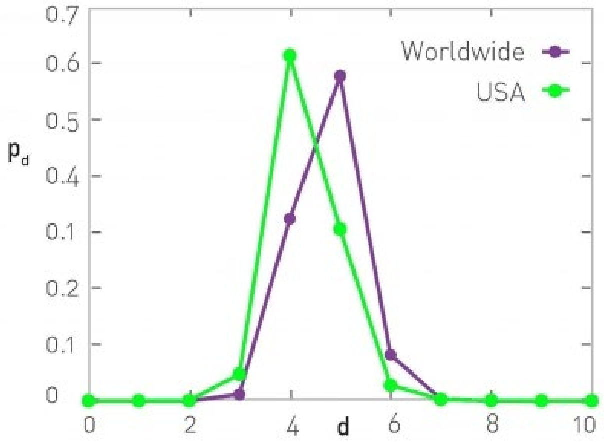
Network Diameter: Small World

- Milgram's six degrees of separation



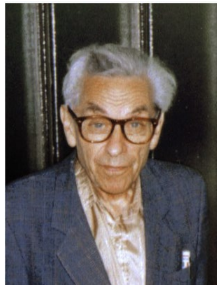


$Milgram's\ experiment$



$facebooks\ experiment$

Erdős number
in academic world



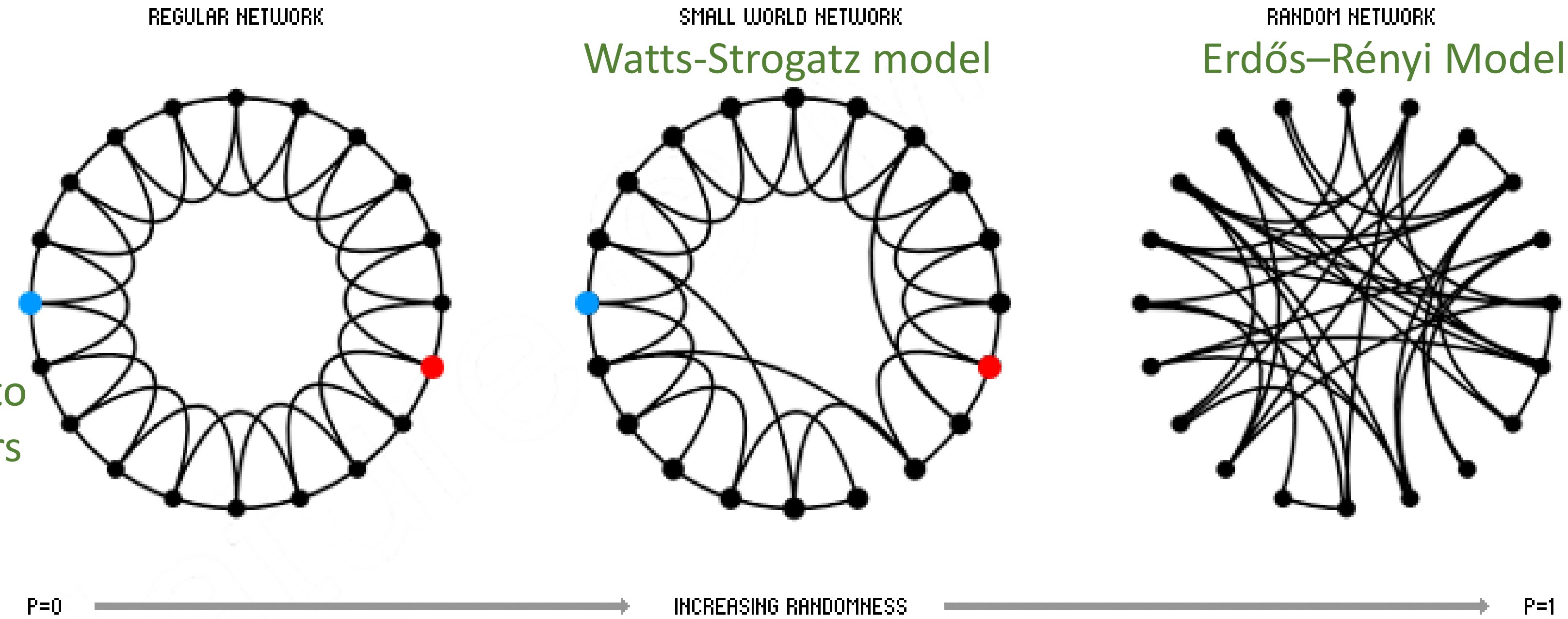
Network Diagram: Small World

Social network Observation:

- Low diameter
- High clustering

- Small world model: Watts-Strogatz model
  - Rewire randomly with probability p from the regular lattice network
  - Strength: model low diameter, high clustering coefficient in real world
  - Limitation: does not lead to the correct degree distribution

Lattice graph: Nodes locate on a ring and only connect to their neighbors



Watts-Strogatz model

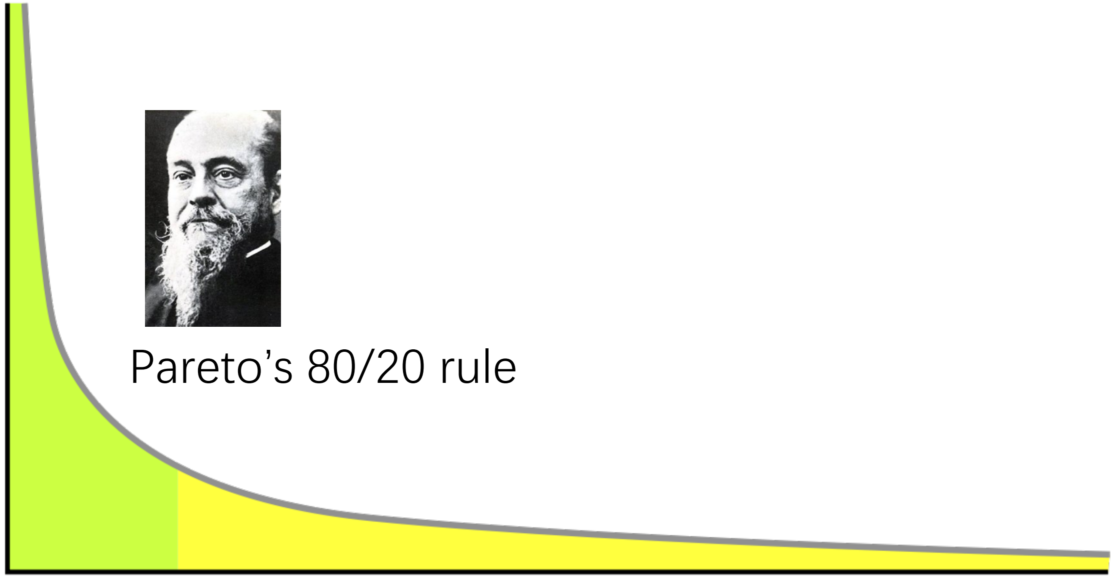
Erdős–Rényi Model

Random graph: every possible edge occurs independently with probability p

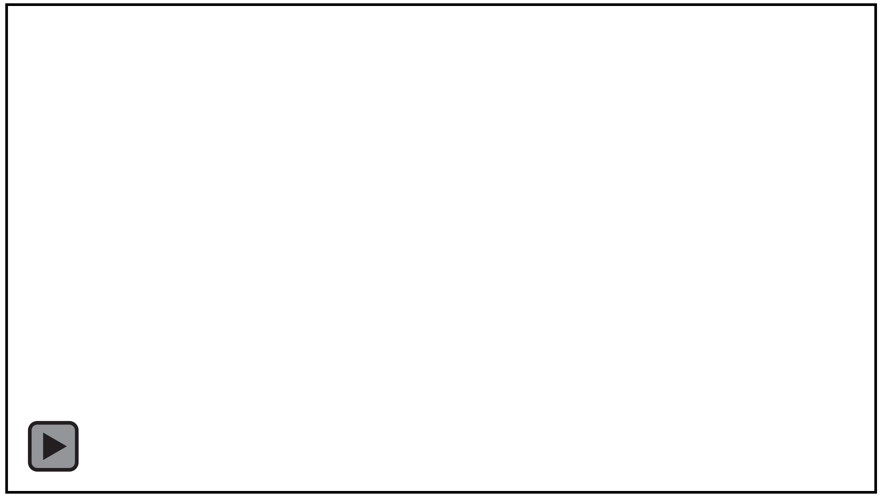
High clustering High diameter	High clustering Low diameter	Low clustering Low diameter

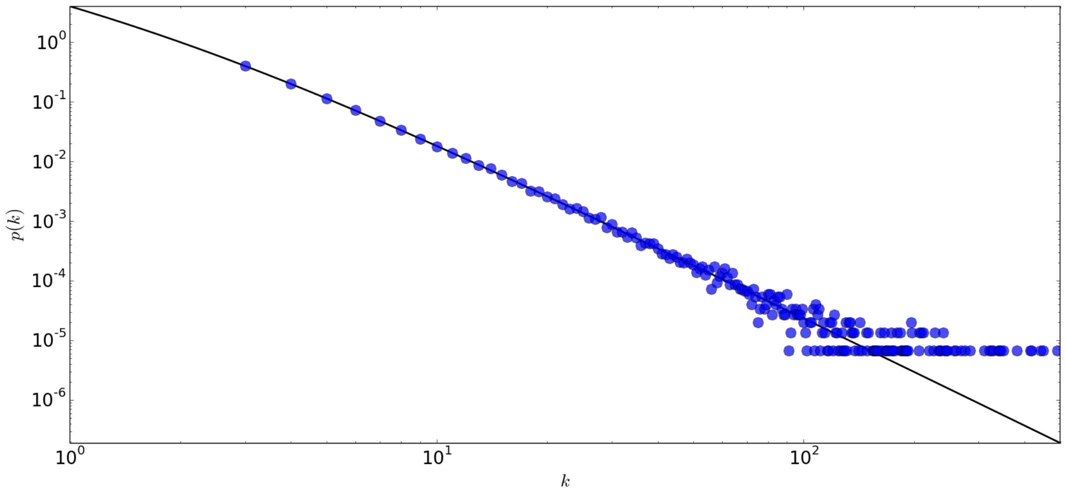
Degree distribution: Scale-free Network

- Degree distribution follows power law
  - $P(k) \sim k^{-\gamma}$
  - Scale-free



Pareto’s 80/20 rule

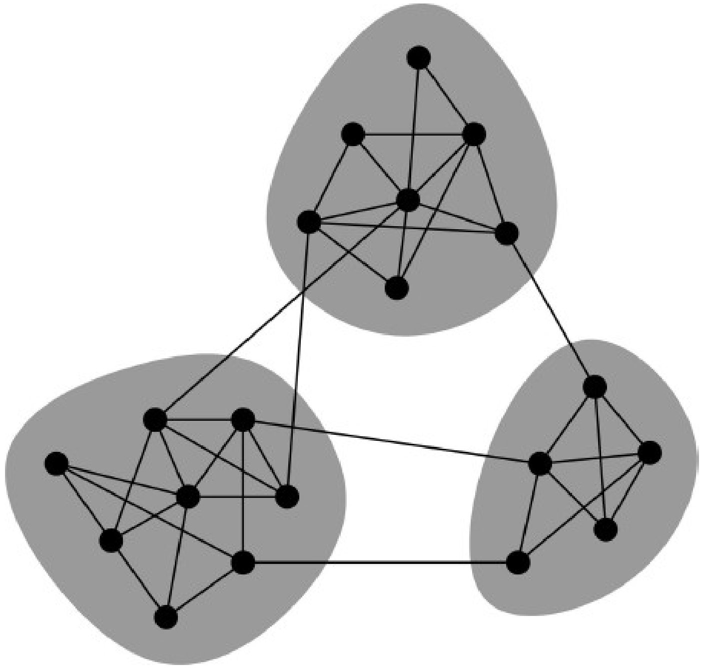




Video generated by Hawoong Jeong at University of Notre Dame

Networks & Communities

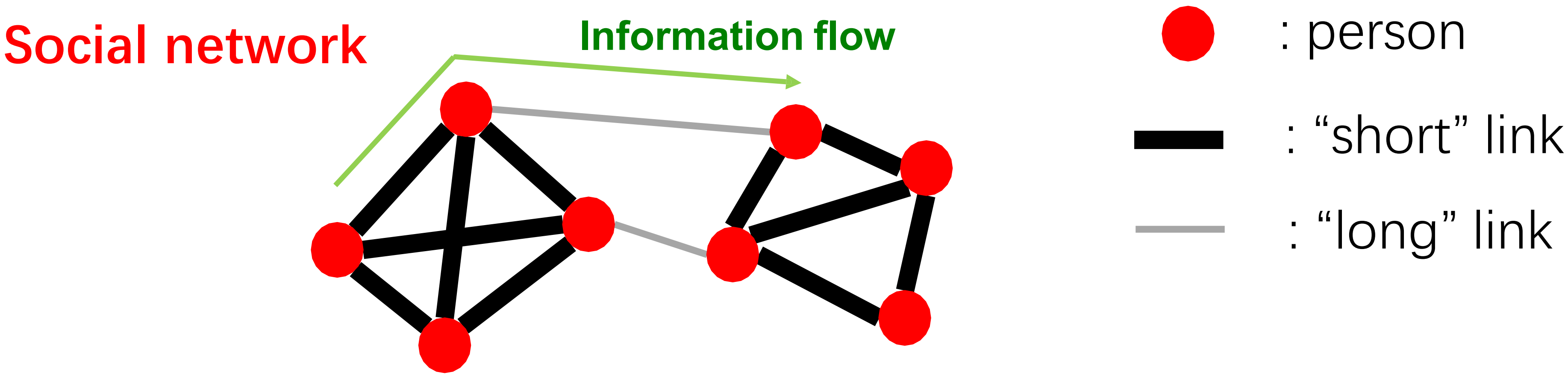
- We often think of networks "looking" like this:



- What led to such a conceptual picture?

Networks: Flow of Information

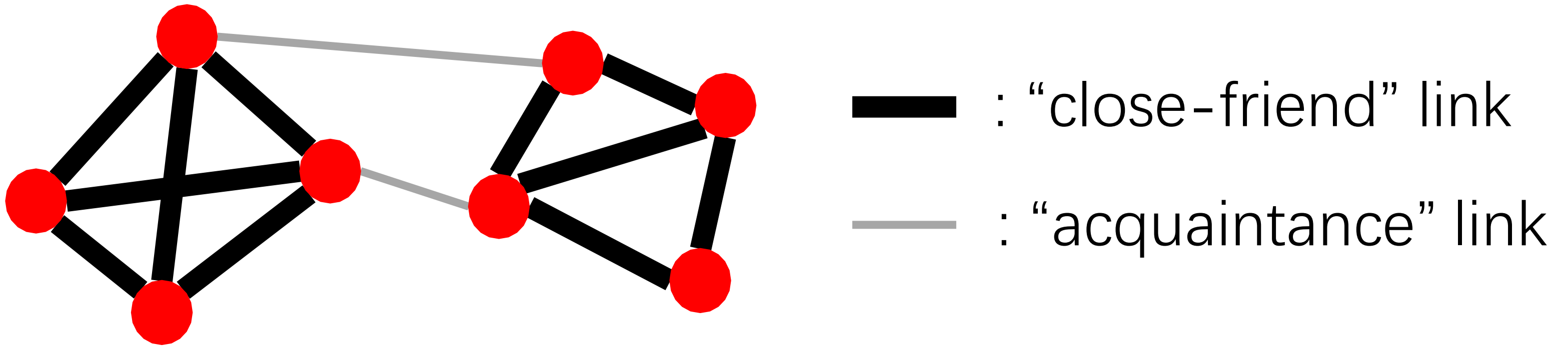
- **How does information flow through the network?**
  - People are "**embedded**" in a social network.
  - There are different links ("short" vs. "long") in the network, through which information flows.



Flow of Job Information

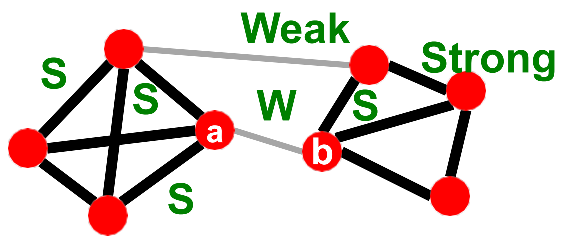
- **How do people find out about new jobs?**
  A. Close friend **B. Acquaintances**
  - Mark Granovetter, part of his PhD in 1960s
  - People find the information **through personal contacts**
- **But:** Contacts were often **acquaintances** rather than close friends
  - **This is surprising:** One would expect your friends to help you out more than casual acquaintances
- **Why is it that acquaintances are most helpful?**

Personal contact network



Granovetter's Explanation

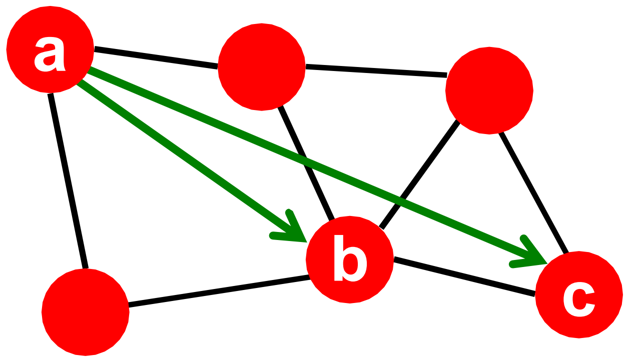
- **Granovetter makes a connection between the social and structural role of an edge**
- **First point: Structure**
  - tightly-connected edges are also socially **strong**
  - Long-range edges spanning different parts of the network are socially **weak**
- **Second point: Information**
  - Structurally embedded edges are heavily redundant in terms of information access
  - Long-range edges allow you to gather information from different parts of the network and get a job



Triadic Closure

- How community (tightly-connected cluster of nodes) forms?

Which edge is more likely, a-b or a-c?



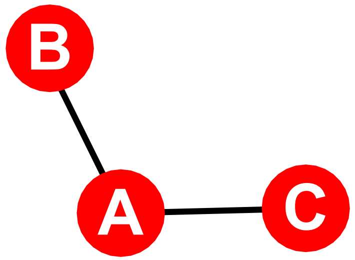
If two people in a network have a friend in common, then there is an increased likelihood they will become friends themselves.

Reasons for Triadic Closure

- **Triadic closure: strongly connected component**

Reasons for triadic closure:

- If $\boldsymbol{B}$ and $\boldsymbol{C}$ have a friend $\boldsymbol{A}$ in common, then:
  - $\boldsymbol{B}$ is **more likely to meet** $\boldsymbol{C}$
    - (since they both spend time with $\boldsymbol{A}$)
  - $\boldsymbol{B}$ and $\boldsymbol{C}$ **trust** each other
    - (since they have a friend in common)
  - $\boldsymbol{A}$ has **incentive** to bring $\boldsymbol{B}$ and $\boldsymbol{C}$ together
    - (since it is hard for $\boldsymbol{A}$ to maintain two disjoint relationships)



If you design an experiment to validate that stronger connection between two people results in stronger connected component around them, what measure and data will you use?

Edge Strength in Real Data

- **For many years Granovetter's theory was not tested**
- But, today we have large who-talks-to-whom graphs:
  - Email, Messenger, Cell phones, Facebook

- **Onnela et al. 2007:**
  - Cell-phone network of 20% of EU country's population
  - **Edge weight:** # phone calls

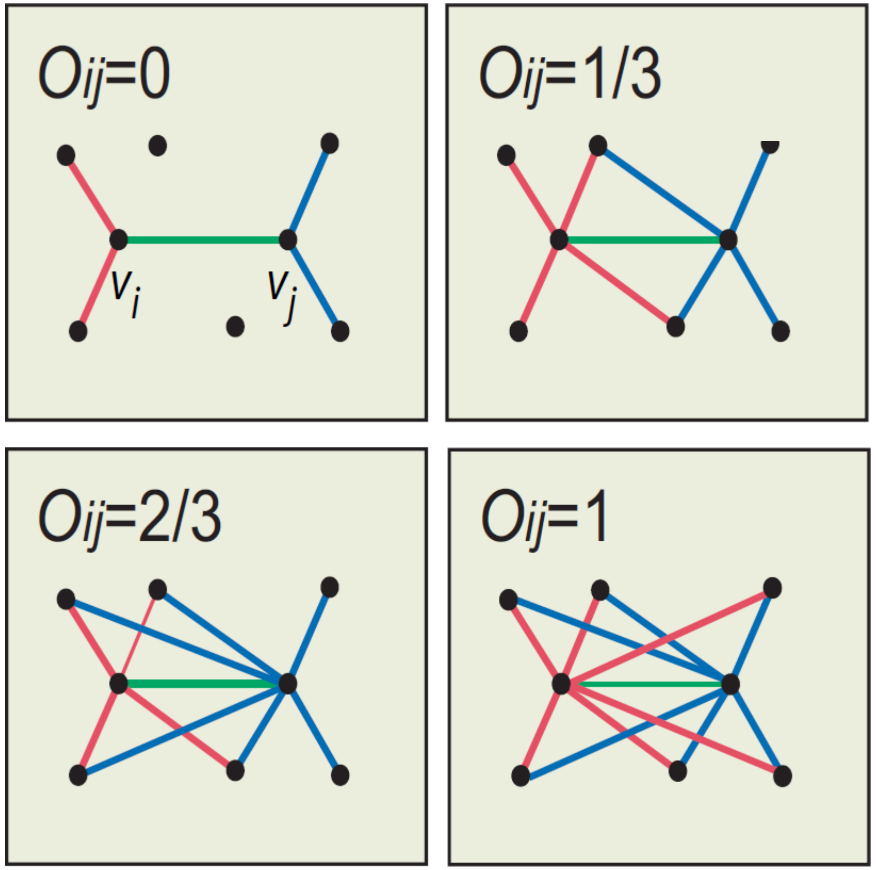
Edge Overlap

- **Edge overlap:**

$$O_{ij} = \frac{|(N(i) \cap N(j)) - \{i,j\}|}{|(N(i) \cup N(j)) - \{i,j\}|}$$

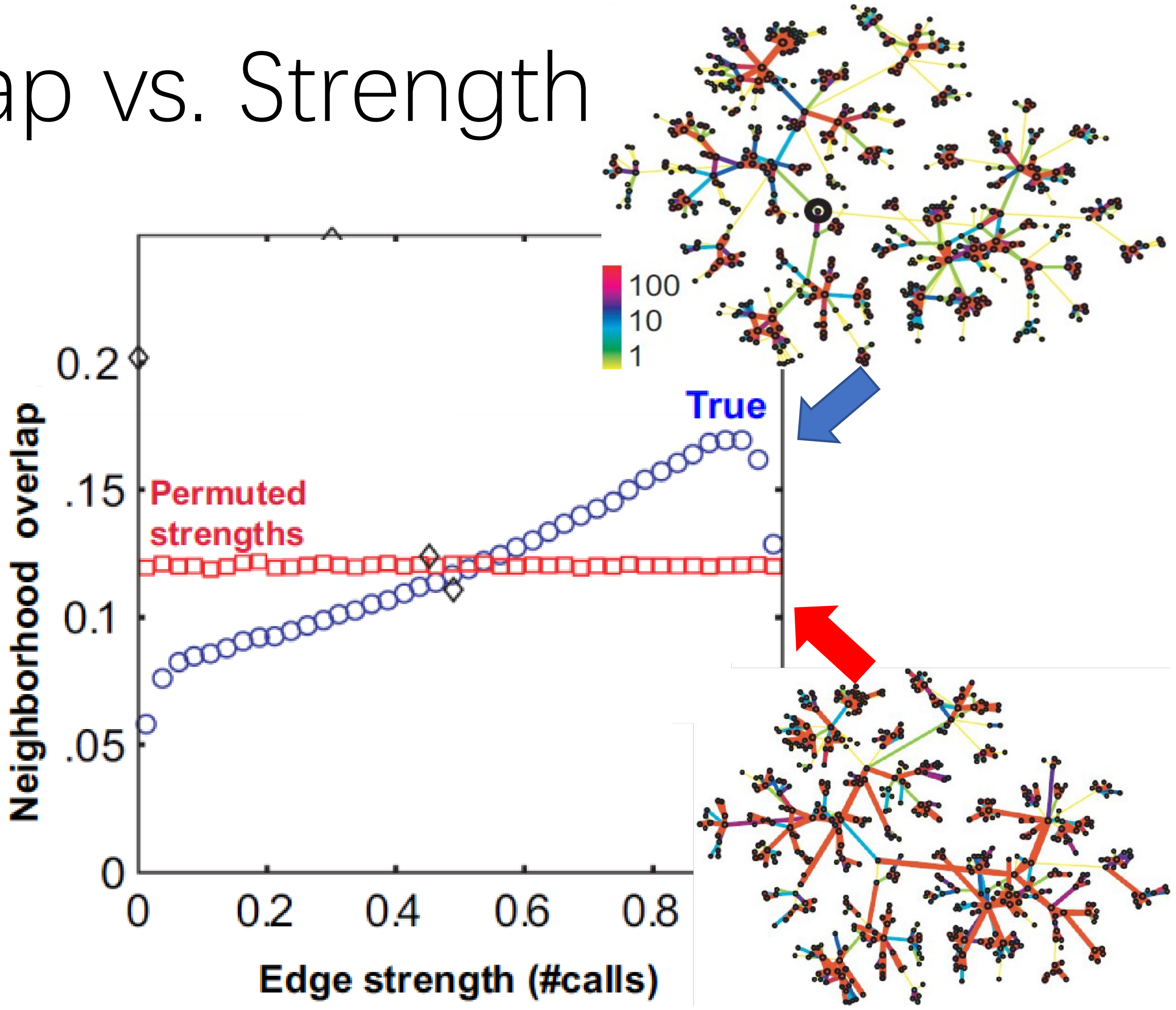
  - $N(i)$⋯ the set of neighbors of node i

- **Note: Overlap = 0** when an edge is a **local bridge**



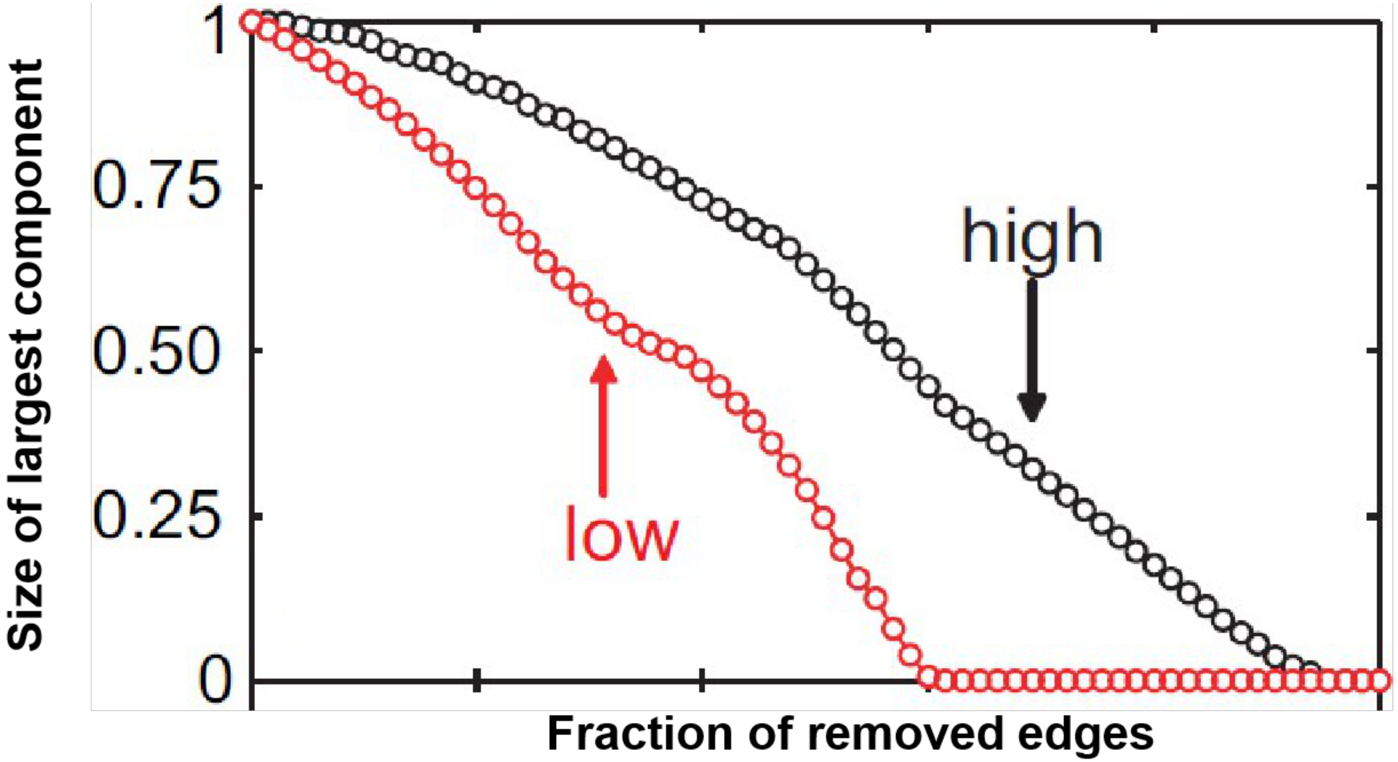
Phones: Edge Overlap vs. Strength

- **Cell-phone network Legend:**
  - **True:** The data
  - **Permuted strengths:** Keep the network structure but randomly reassign edge strengths
- **Observation:**
  - **Highly used links have high overlap!**

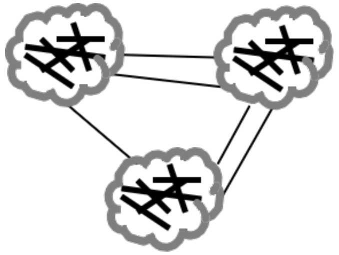


Edge Removal by Overlap

- Removing edges based on **edge overlap**
  - Low to high
  - High to low



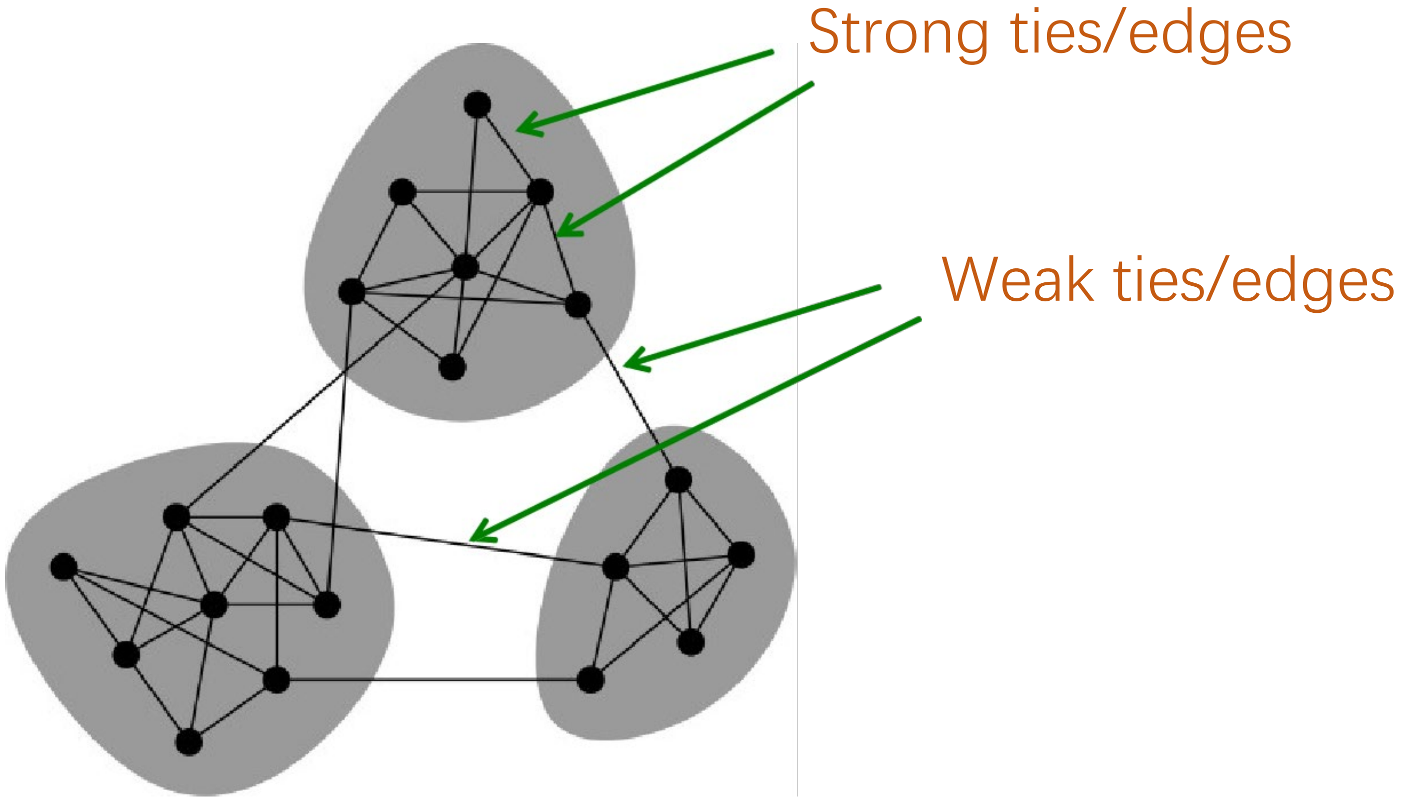
Low disconnects the network sooner



Conceptual picture of network structure

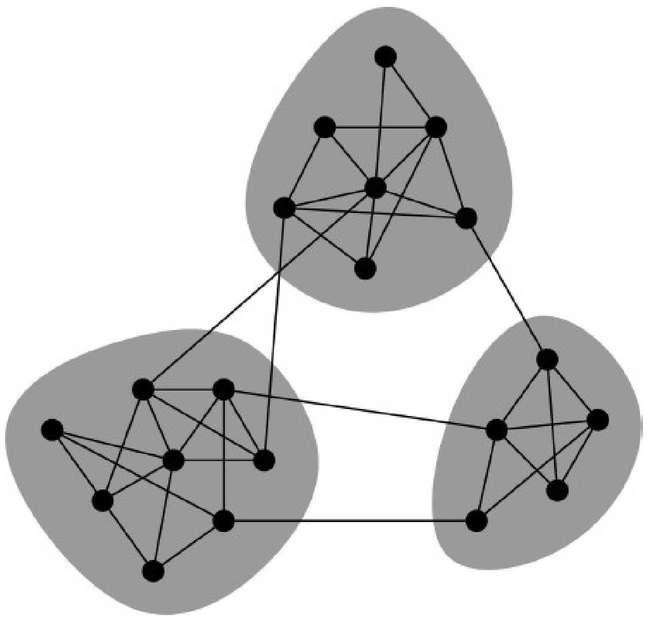
Conceptual Picture of Networks

- **Granovetter's theory leads to the following conceptual picture of networks**



Networks Communities

- Granovetter's theory suggests that networks are composed of **tightly connected sets of nodes**



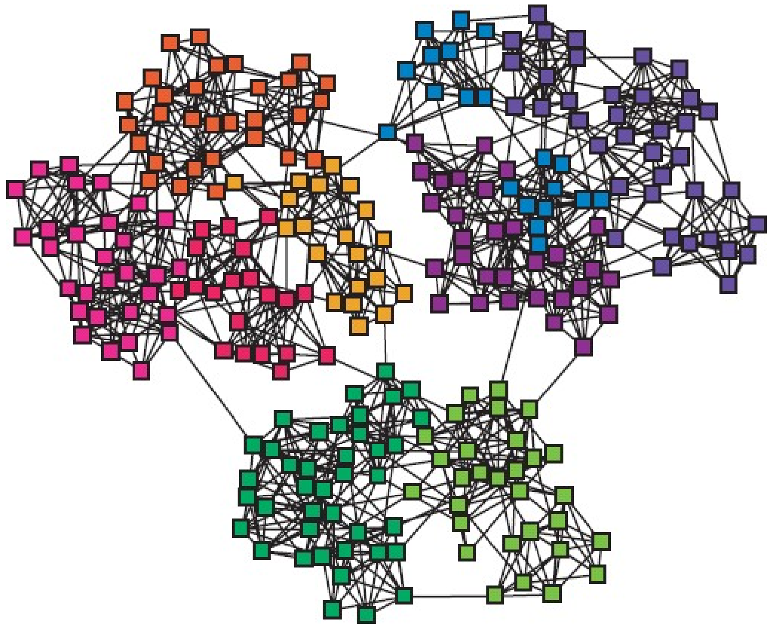
Communities, clusters, groups, modules

- **Network communities:**
  - Sets of nodes with **lots** of **internal** connections and **few external** ones (to the rest of the network).

Finding Network Communities

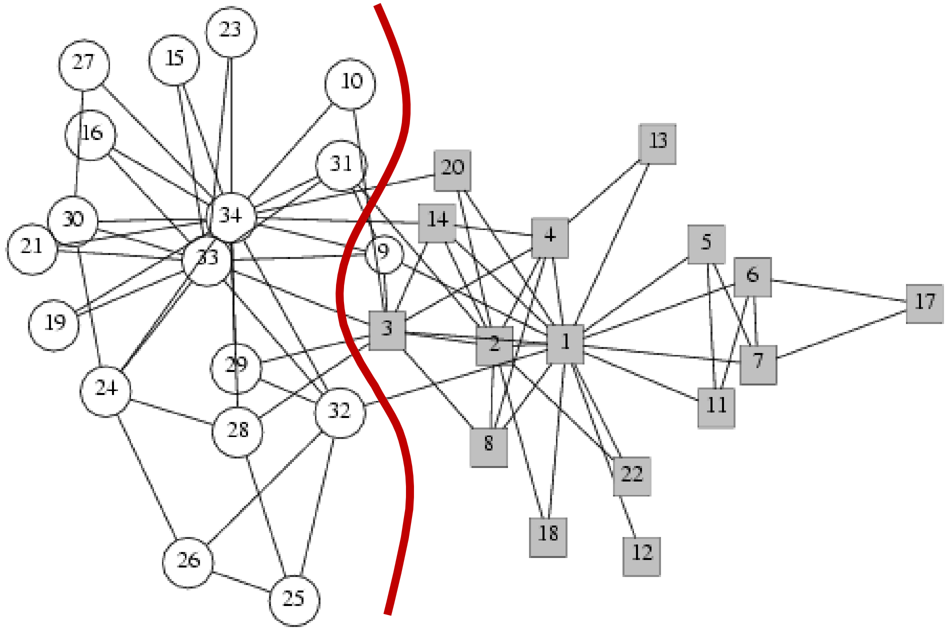
- **How do we automatically find such densely connected groups of nodes?**

- Ideally such automatically detected clusters would then correspond to real groups



Communities, clusters, groups, modules

Social Network Data



- **Zachary's Karate club network:**
  - Observed social ties & rivalries in a university karate club
  - During the study, conflicts led the group to split
  - Split could be explained by a minimum cut in the network