Social Networks Topics

Lecture 6

CSCI 4974/6971

19 Sep 2016

Today's Biz

1. Quick Review
2. Reminders
3. Social Networks Topics
4. Parallel Triangle Counting
5. Homework 1 solutions

Today's Biz

1. **Quick Review**
2. Reminders
3. Social Networks Topics
4. Parallel Triangle Counting
5. Homework 1 solutions

Quick Review

- Parallel SCC detection
  - Forward-Backward Algorithm: two searches from a given pivot, following out links and in links, overlap of two vertex sets discovered forms an SCC
- Centrality Measures:
  - PageRank: How important globally
  - Degree: How important locally
  - Betweenness: How much information might flow through
  - Closeness: How close to the rest of the graph

Today's Biz

1. Quick Review
2. **Reminders**
3. Social Networks Topics
4. Parallel Triangle Counting
5. Homework 1 solutions

Reminders

- Assignment 2: Thursday 29 Sept 16:00
- Project Proposal: Thursday 22 Sept 16:00
- Office hours: Tuesday & Wednesday 14:00-16:00 Lally 317
  - Or email me for other availability
- Class schedule:
  - Social net analysis methods
  - Bio net analysis methods
  - Random networks and usage

Today's Biz

1. Quick Review
2. Reminders
3. **Social Networks Topics**
4. Parallel Triangle Counting
5. Homework 1 solutions

Slides from Alexandros Nanopoulos, Stiftung Universität Hildesheim

Strong and Weak Ties

Objectives

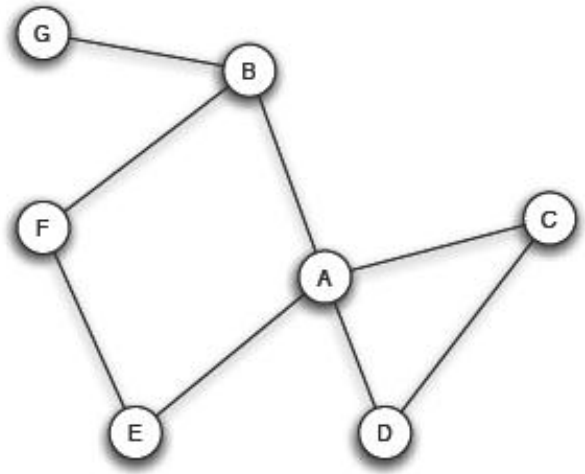
- How information flows through a social network
- How different nodes can play structurally distinct roles in this process
- How these structural considerations shape the evolution of the network itself over time

“Finding a job”

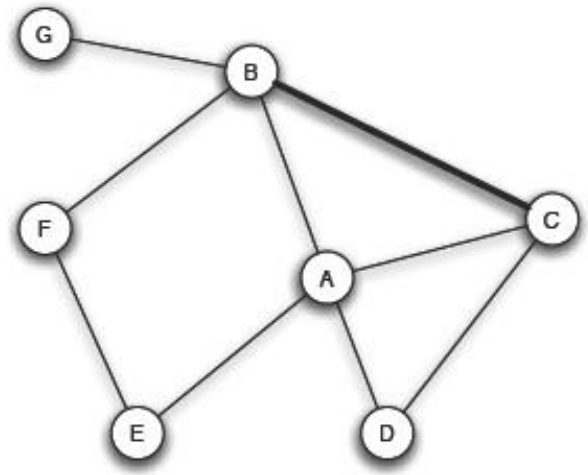
- Mark Granovetter (1960s)
  - interviewed people who had recently changed employers
  - how they discovered their new jobs?
  - many learned information through personal contacts
  - these contacts often described as acquaintances (weak ties) rather than close friends (strong ties)
- A bit surprising:
  - your close friends have the most motivation to help
  - why more distant acquaintances who are to thank?

Triadic Closure

- If B and C have a friend A in common, then edge between B and C tends to be produced
  - a triangle



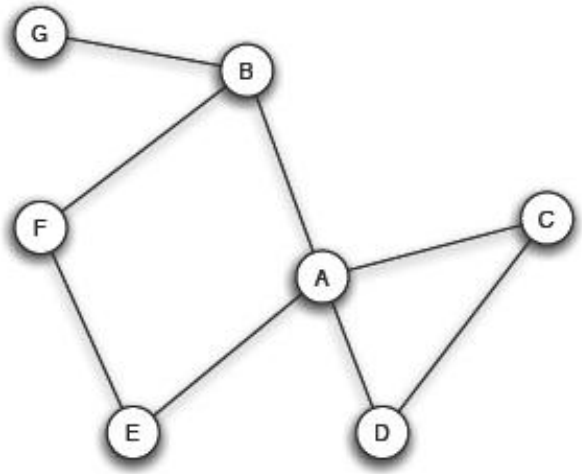
(a) *Before B-C edge forms.*



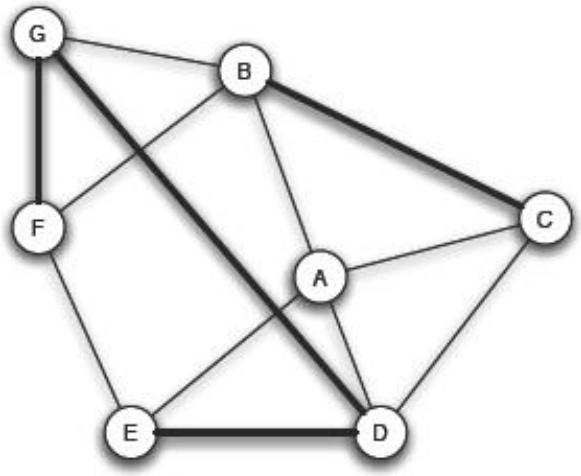
(b) *After B-C edge forms.*

Triadic Closure

- Observe snapshots of a social network at two distinct points in time
- Significant number of new edges form through this triangle-closing operation



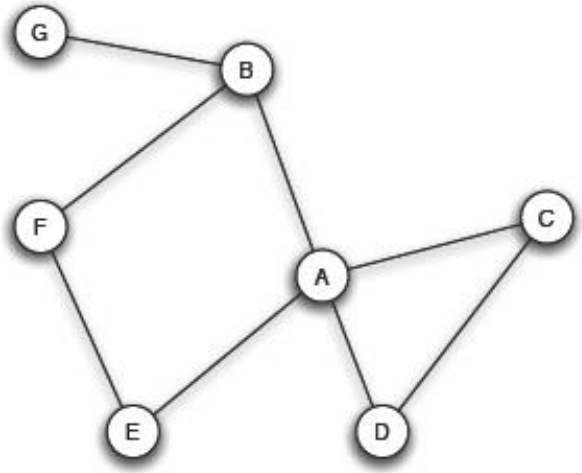
(a) *Before new edges form.*



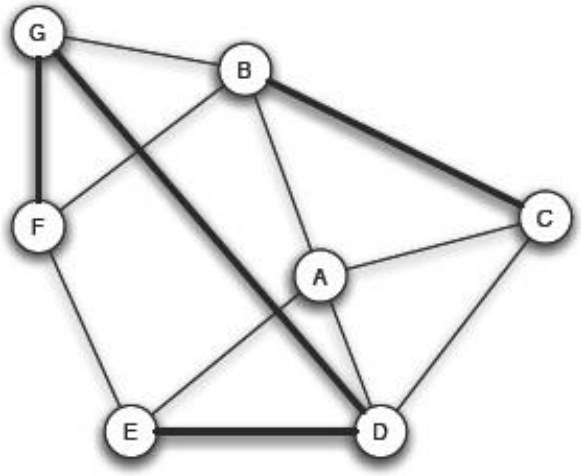
(b) *After new edges form.*

The Clustering Coefficient

- Measure to capture triadic closure
- The clustering coefficient of a node A, CC(A), is the fraction of pairs of A's friends that are connected to each other by edges
- Ex:
  - Figure(a): CC(A) = 1/6
  - Figure(b): CC(A) = ½
- CC ranges from 0 (when none of the
- node's friends are friends with each other) to 1 (when all of the node's friends are friends with each other)



(a) *Before new edges form.*



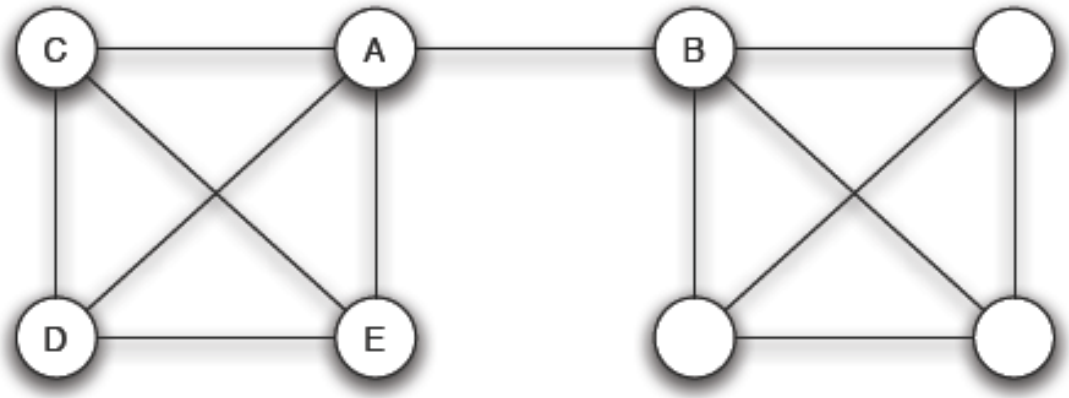
(b) *After new edges form.*

Reasons for Triadic Closure

- Why B and C more likely to become friends, when they have a common friend?
  - increased opportunity for B and C to meet
  - B and C trust each other
  - it becomes a source of latent stress in these relationships if B and C are not friends
    - teenage girls who have a low clustering coefficient in their network of friends are significantly more likely to contemplate suicide

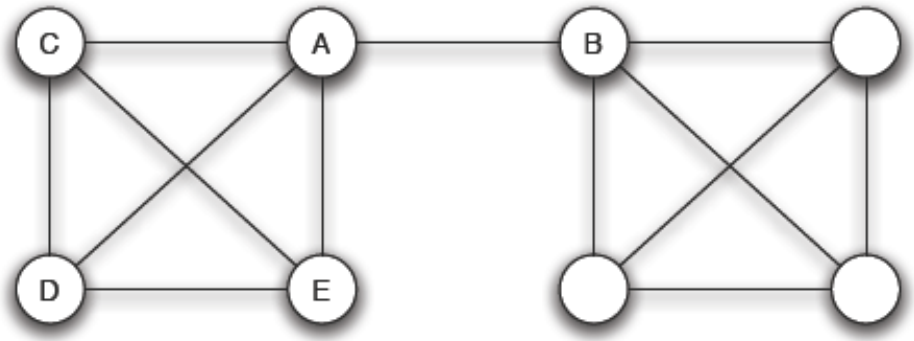
Bridges and Local Bridges

- A has 4 friends:
  - C, D, and E connected to a tightly-knit group
  - B reaches into a different part of the network
  - B offers access to new things



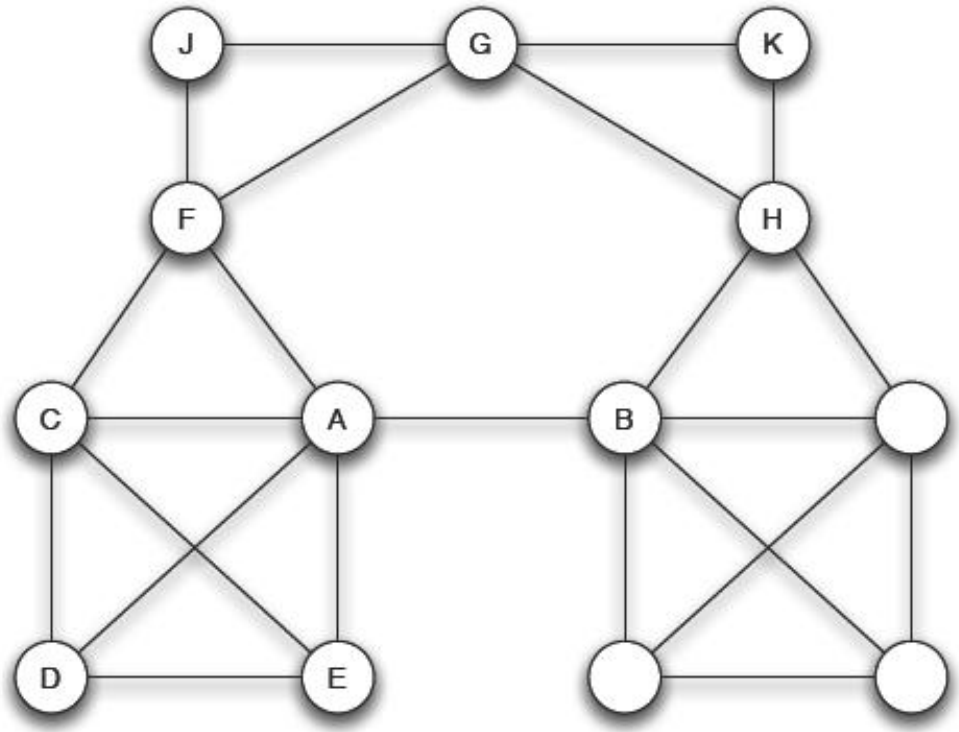
Bridges and Local Bridges

- Edge A-B is a bridge if deleting it causes A and B to be in different components
- Bridges are extremely rare in real social networks
  - giant component, many short paths



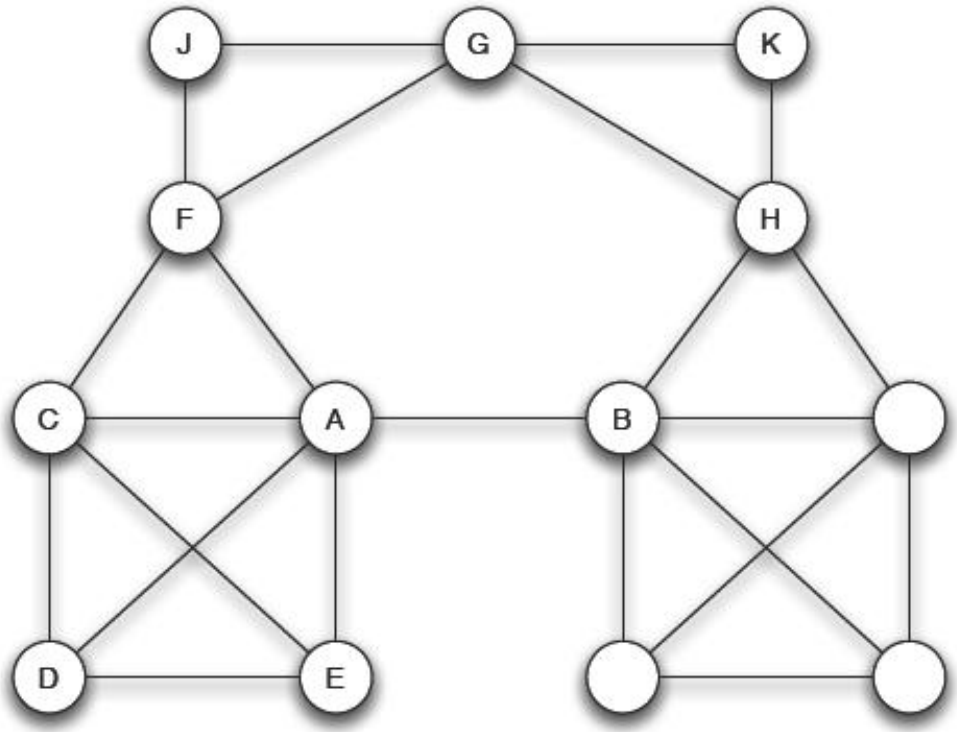
Bridges and Local Bridges

- Edge A-B is a local bridge if its endpoints A and B have nc friends in common
  - deleting A-B => d(A,B) increases more than 2
- Relation with triadic closure:
  - a local bridge does not belong to any triangle
- Local bridges provide their endpoints with access to parts of the network that they would otherwise be far away from



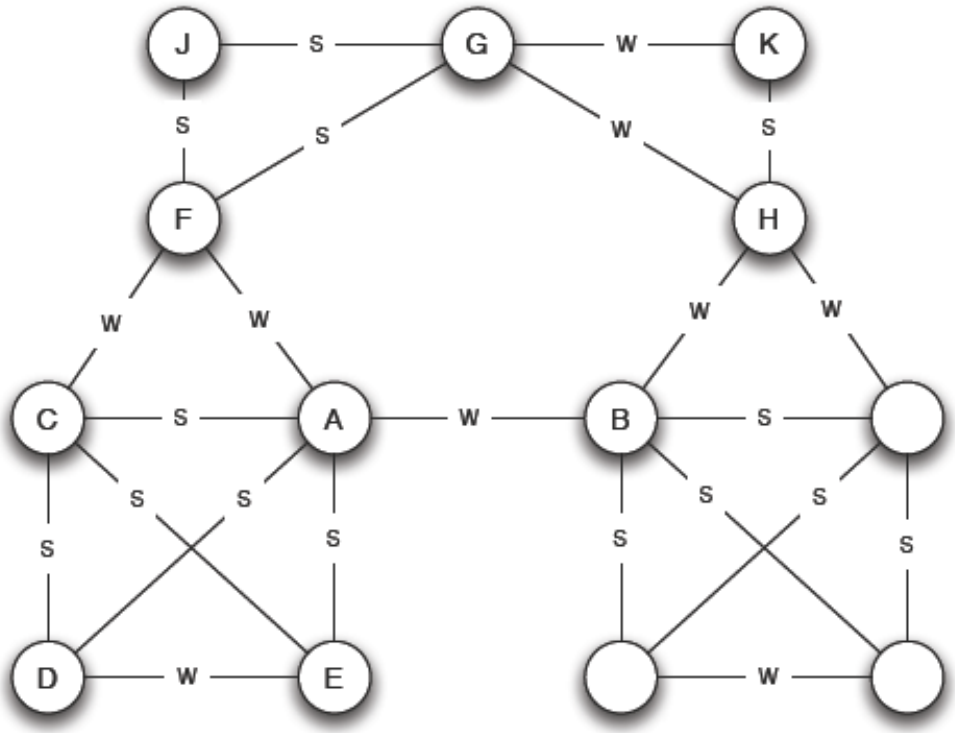
“Finding a job”

- if a node like A is going to get new information about a job, it might come often (not always) from a friend connected by a local bridge
- The closely-knit groups of close friends are eager to help, but they know roughly the same things with A
- **How to connect local bridges to acquaintances?**



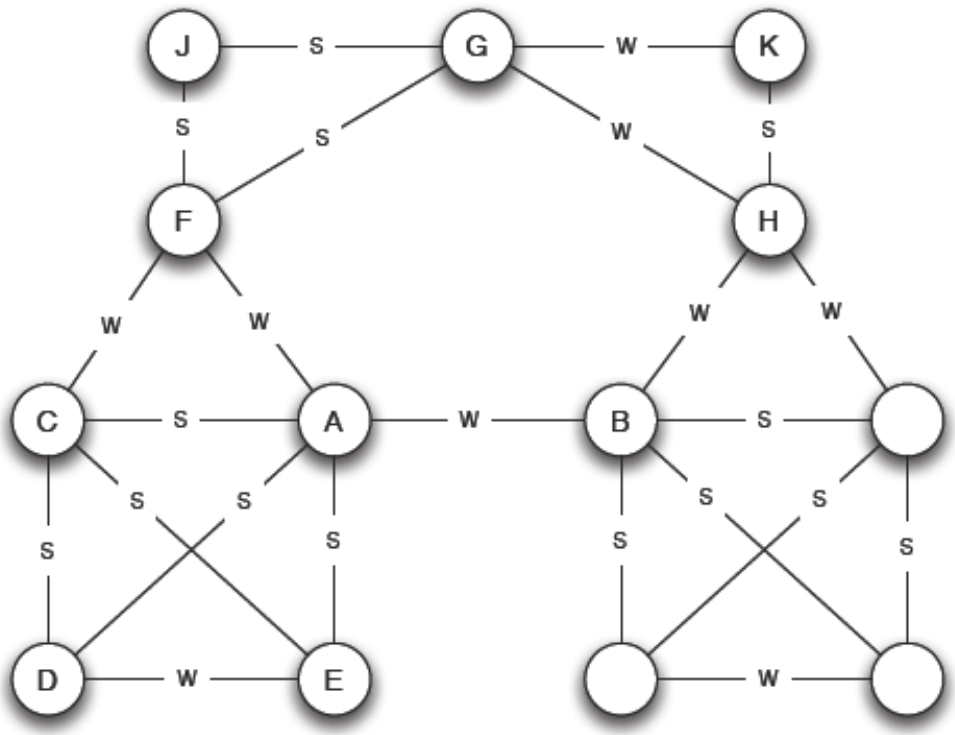
Strong vs. Weak Ties

- Classification of links in a social network:
  - strong ties (friends) vs. weak ties (acquaintances)
- Connection to triadic closure:
  - if A has edges to B and C, then edge B-C is especially likely to form if A's edges to B and C are both strong ties



The Strong Triadic Closure Property

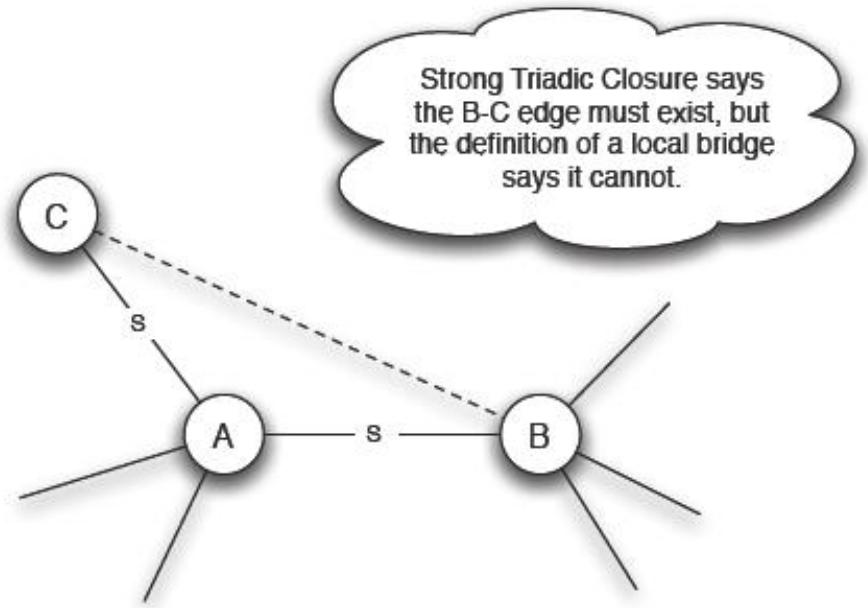
- A violates the Strong Triadic Closure Property if:
  - has strong ties to two other nodes B and C, and
  - there is no edge at all (either a strong or weak tie) between B and C
- A satisfies the Strong Triadic Closure Property if it does not violate it
- Ex (figure):
  - all nodes satisfy the Property
  - if edge A-F were strong tie, then A and F would both violate the Property (A-G is missing)
- Strong Triadic Closure Property is too extreme to hold across all nodes of a large social network
  - useful step as an abstraction to reality



Local Bridges and Weak Ties

- *If A satisfies the Strong Triadic Closure Property and is involved in at least two strong ties, then any local bridge it is involved in must be a weak tie.*

- *Proof.* Consider A that satisfies Strong Triadic Closure Property and is involved in at least two strong ties. Suppose A is involved in a local bridge to B that is a strong tie. Contradiction:
  - A-C the other strong tie
  - A-B local bridge => A and B must have no friends in common => B-C edge must not exist
  - A satisfies Strong Triadic Closure: A-B and A-C strong => B-C must exist (as weak or strong tie)



“Finding a job”

- The previous argument completes the connection between the weak ties (acquaintances) and local bridge (access to other parts of the network)
- But it is based on the assumptions of Strong Triadic Closure and is a simplification that:
  - holds approximately even when the assumption is relaxed
  - need to test on real-world data

Weak Ties and Local Bridges in Real Data

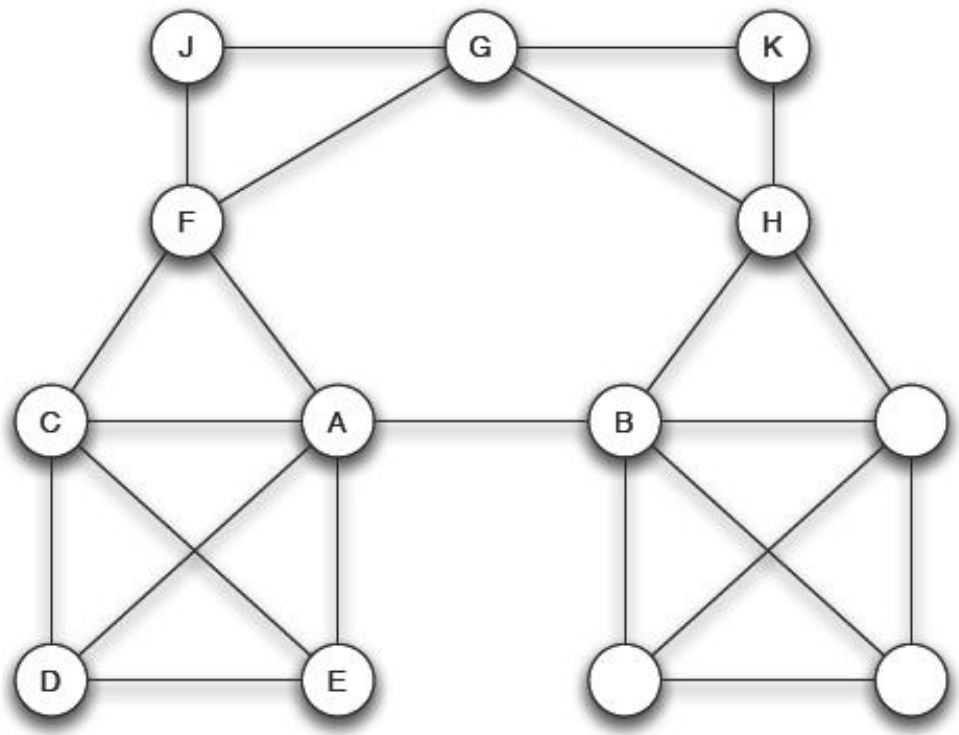
- Onnela et al.: traces of digital communication ("who-talks-to-whom" data)
  - cell phone records
  - 20% of a national population
  - 18-week observation period
  - a giant component (84%)
- How to measure weak ties and local bridges?
  - use the speaking time as strength
  - generalize definition of local bridge

Generalizing Weak Ties and Local Bridges

- So far sharp dichotomies:
  - an edge is either a strong tie or a weak tie, and
  - it is either a local bridge or it isn't
- For real data we need smoother gradations:
  - strength of an edge the total number of minutes between the two ends of the edge
  - neighborhood overlap of edge A-B:
    - N(A), N(B) are neighbors of A and B, resp.
    - $O(A\text{-}B) = |N(A) \cap N(B)| \; / \; |N(A) \cup N(B)|$
    - We don't count A or B themselves

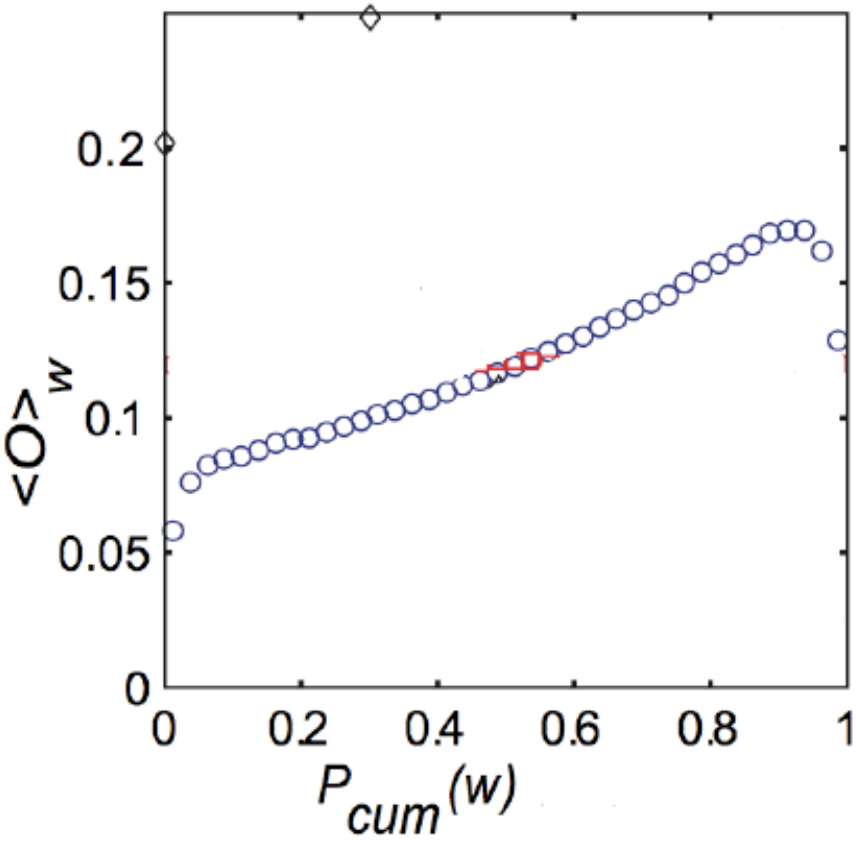
Generalizing Weak Ties and Local Bridges

- Ex(figure):
  - O(A-F) = 1/6
- Overlap(A-B) = 0 => A-B a local bridge
- Allows for “almost” local bridges
  - A-F vs. A-E
    - O(A-E) = 2/4



Weak Ties and Local Bridges in Real Data

- How the overlap of an edge depends on its strength?
- Lower overlap (almost local bridges) tend to have weaker strength
  - verifies theory
  - deviation at the end of the plot: people using cell-phones in unusual fashions



overlap as a function of strength (percentile)

Weak Ties and Local Bridges in Real Data

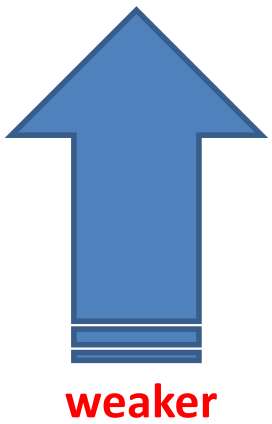
- How to test whether weak ties link together different tightly-knit communities that each contain a large number of stronger ties?
- Onnela et al. provided an indirect analysis:
  - deleted edges one at a time, starting with strongest ties => the giant component shrank steadily
  - deleted edges one at a time, starting with weakest ties => the giant component shrank more rapidly
- Verifies the theoretical expectation:
  - weak ties provide the more crucial connective structure for holding together communities

Tie Strength and Social Media

- Large lists of friends in social-networking tools
- How many of these correspond to strong and weak ties?
- Tie strengths can provide an important perspective on on-line social activity

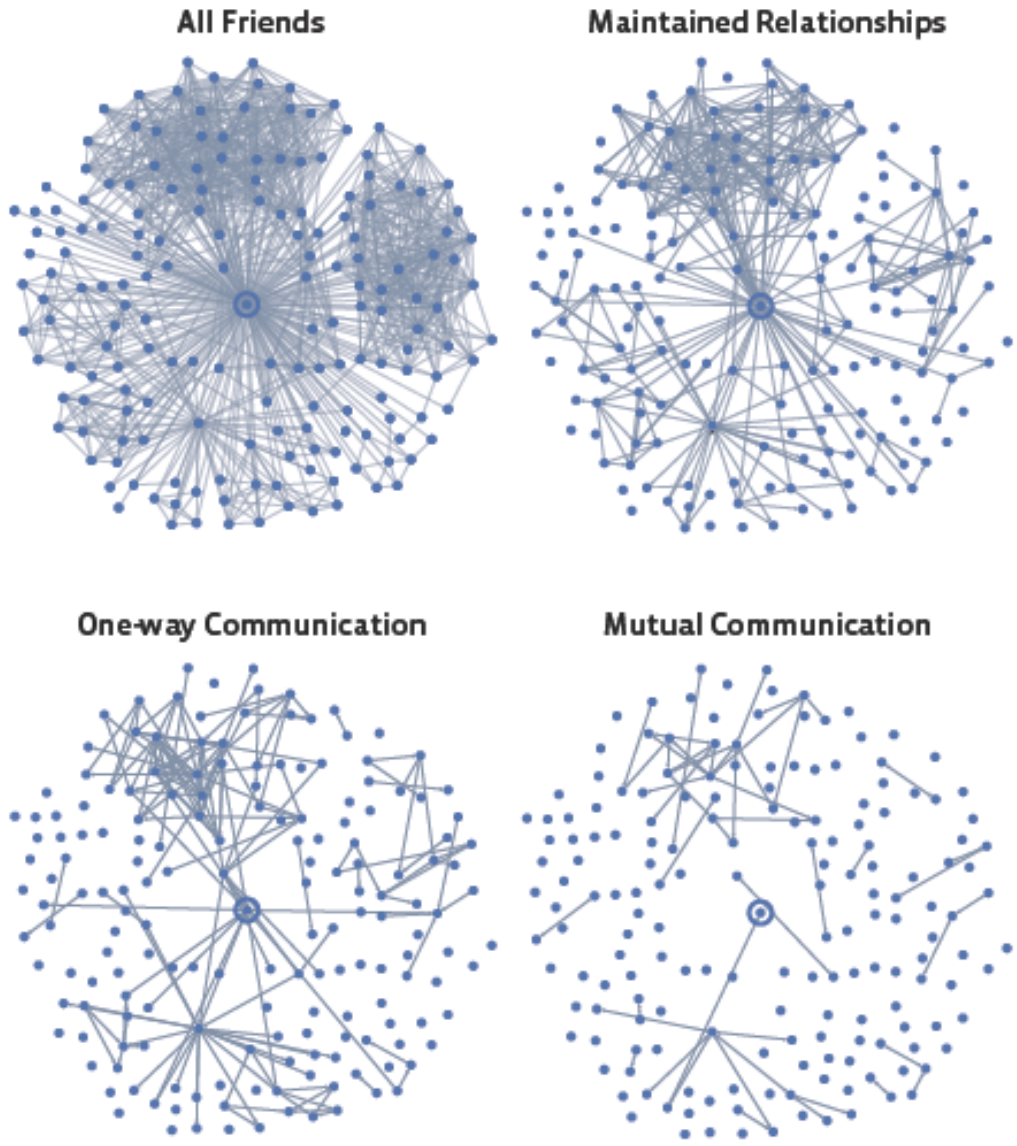
Tie Strength on Facebook

- Cameron and Marlow:
  - To what extent each social link is actually used for social interaction beyond being listed?
  - 3 categories of links (usage over a 1-month period)
    - mutual communication: user both sent and receive messages from the friend
    - one-way communication: user sent messages to the friend (regardless if replied)
    - maintained relationship: user followed information of the friend (regardless of messages)
      - "following information": clicking on content via Facebook's News Feed service or visiting the friend's profile
  - Categories not mutually exclusive:
    - mutual communication always belongs subset of one-way communication



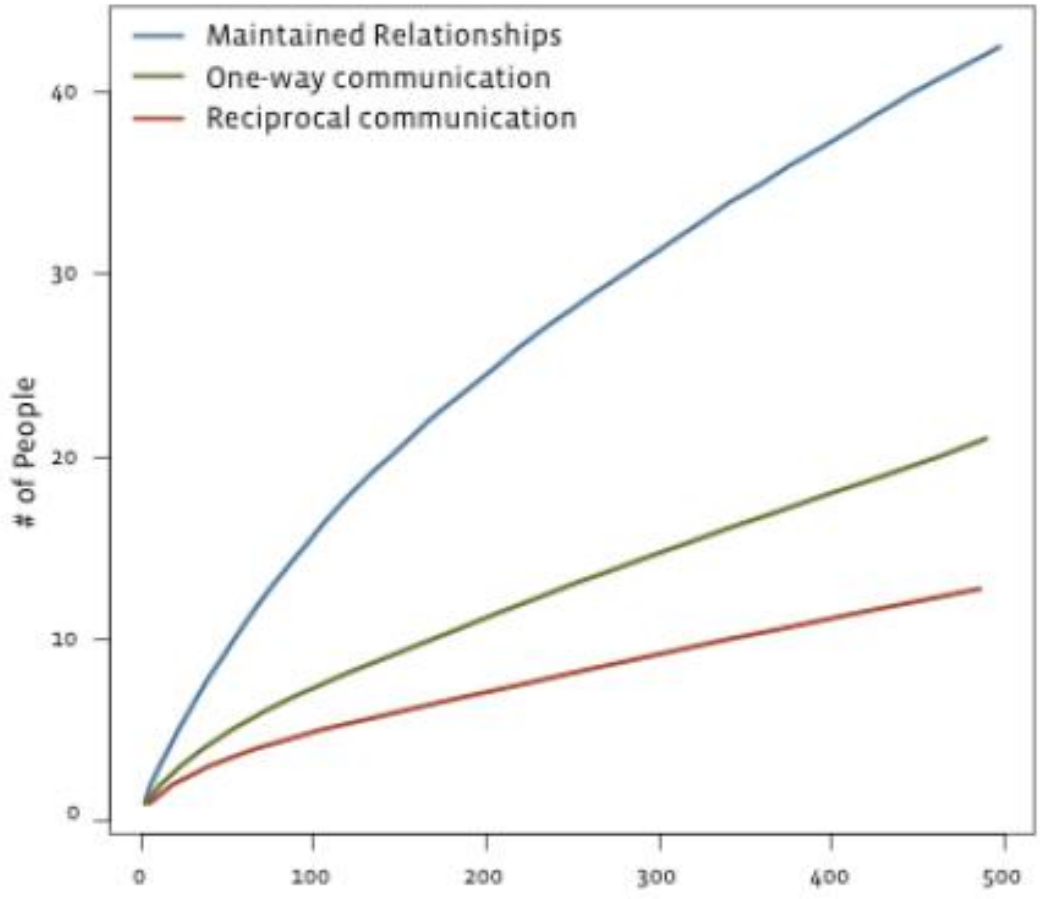
Example for a sample Facebook user

- Restricting to stronger ties thins out the network
- Triadic closure:
  - in upper and right part of “All Friends”
  - Maintained:
    - upper survives (current friends)
    - right hot (earlier friends, e.g., school)



Active Friendships in Facebook

- Users report large numbers of friends
  - up to 500
- Mutual communication (strong ties):
  - between 10 and 20
- Maintained (weak ties)
  - under 50



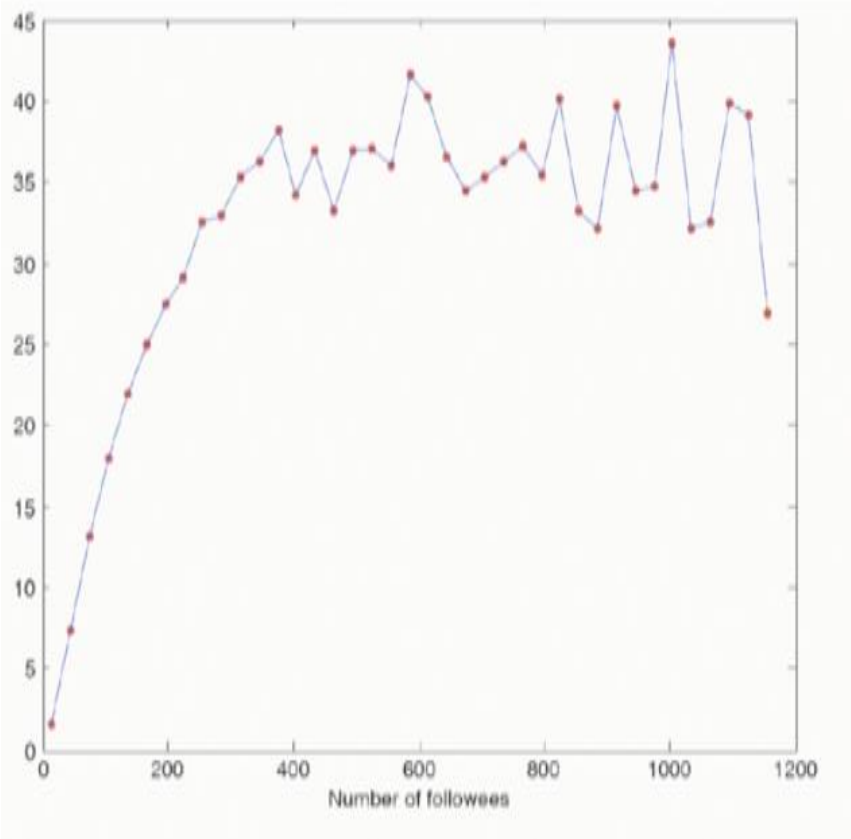
number of friends a user declares

Passive Engagement

- The power of media like Facebook:
  - maintained relationships (weak ties) enable passive engagement
    - keep up with friends by reading news about them (even) in the absence of communication
- Weak tie are middle ground between:
  - the strongest ties (mutual communication) and
  - inactive ties (friends only listed)
  - If only mutual communication allowed:
    - small list of friends (like those we call regularly)
- Weak ties maintain the social network highly connected:
  - everyone is passively engaged with each other and events/news propagate very quickly

Tie Strength on Twitter

- Huberman, Romero, and Wu:
  - Strong ties of a user A:
    - users that A directly communicates through tweets
  - Weak ties of a user A:
    - users that A follows without direct communication
- Below 50 strong ties even for over 1000 followees (weak ties)



number of a user's strong ties vs. weak ties

Reasons for weak ties

- Strong ties require investment of time and effort
- Both are constrained => we reach a limit
- "Dunbar's number" = 150
  - Strong ties limited by the size of the human brain
- Weak ties pose milder constraints
  - they need to be established but not necessarily maintained continuously
  - easier accumulate large numbers of weak ties

Understanding how on-line media affect social networks is a complex research problem (still open)

Networks in Their Surrounding Contexts

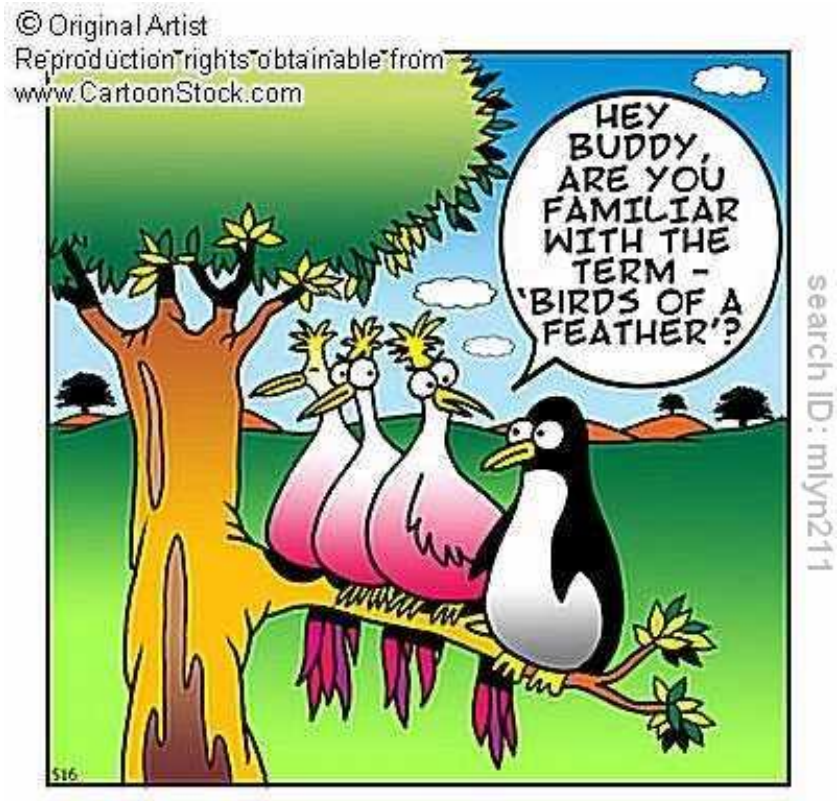
Objectives

- Examine additional processes (to triadic closure) that affect the formation of links in the network
- Surrounding contexts: factors that exist outside the nodes and edges of a network
- Represent the contexts together with the network in a common framework

Homophily

- Homophily principle: we tend have similar characteristics with our friends

“birds of a feather flock together”



- People of similar character, background, or taste tend to congregate or associate with one another (like likes like)
- expression appears in the 16th century, a literal translation of Plato's Republic

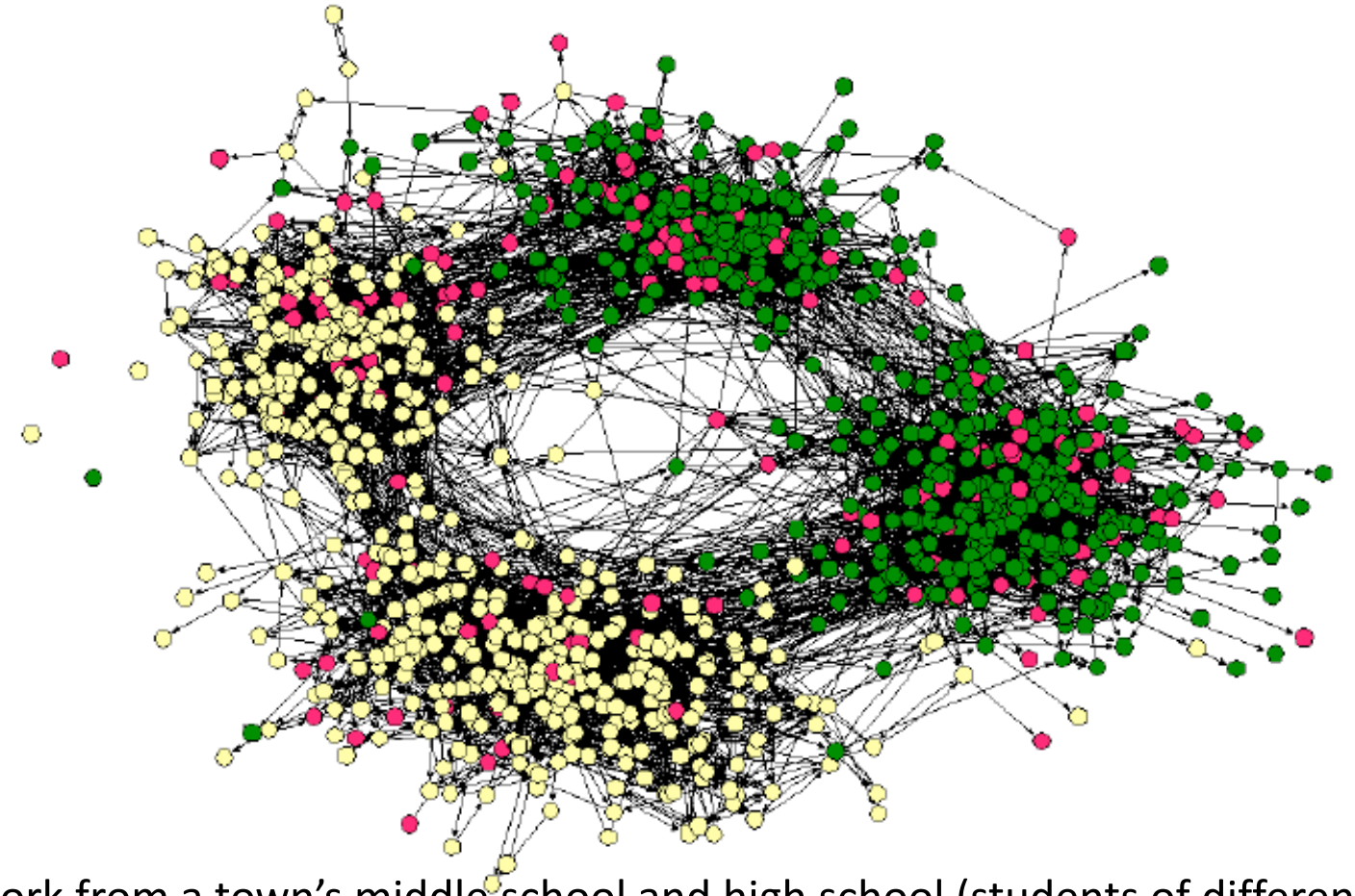
Homophily

- Links in a social network tend to connect people who are similar to one another
  - basic notions governing the structure of social networks
- Its role in modern sociology by influential work in the 1950s (Lazarsfeld and Merton)

Homophily vs. Triadic Closure for Link Formation

- With triadic closure:
  - a new link is added for reasons that are intrinsic to the network (need not look beyond the network)
  - Ex: a friendship that forms because two people are introduced through a common friend
- With homophily:
  - a new link is added for reasons that are beyond the network (at the contextual factors)
  - Ex: a friendship that forms because two people attend the same school or work for the same company

Example



Social network from a town's middle school and high school (students of different races drawn as differently colored circles)

2 divisions:

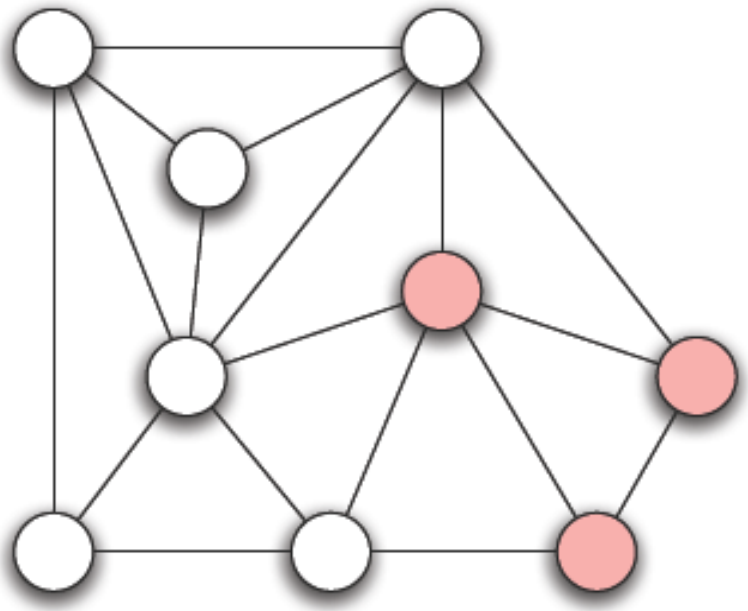
- one based on race and
- the other based on friendships in the middle and high schools

Homophily vs. Triadic Closure for Link Formation

- Strong interactions between intrinsic and contextual effects
- Both operating concurrently
- Triadic closure (intrinsic mechanism):
  - B and C have a common friend A
  - B and C have increased opportunities to meet
- Homophily (contextual mechanism):
  - B and C are each likely to be similar to A in a number of dimensions
  - also possibly similar to each other as well
- Most links arise from a combination of several mechanisms
  - difficult to attribute any individual link to a single mechanism

Measuring Homophily

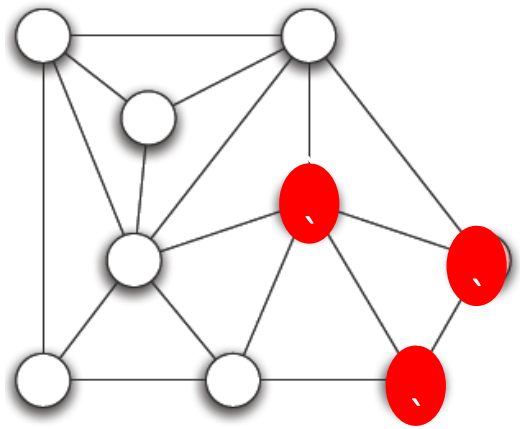
- Given a characteristic (like race, or age), how to test if a network exhibits homophily according to it?
- Ex friendship network:
  - Exhibits homophily by gender?
  - boys tend to be friends with boys, and girls tend to be friends with girls
  - cross-gender edges exist

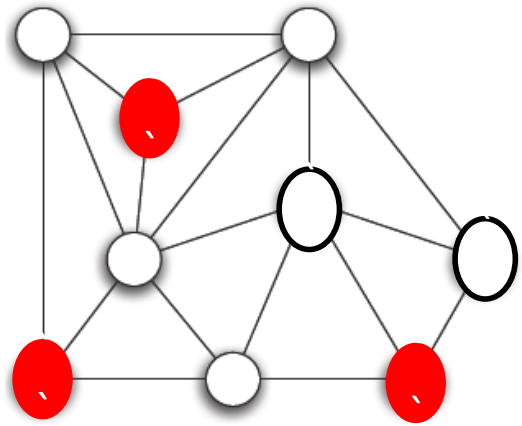


friendship network of a (hypothetical) classroom: shaded nodes are girls and the six unshaded nodes are boys

Measuring Homophily

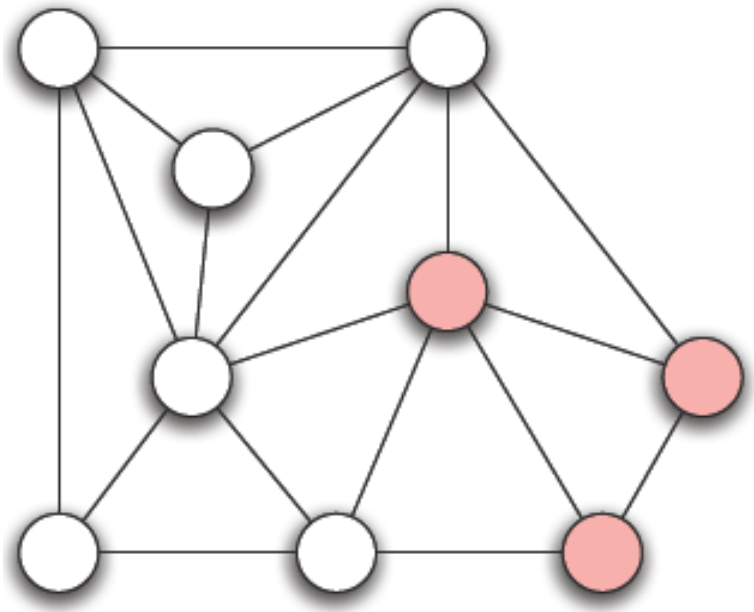
- Q: what would it mean for a network not to exhibit homophily by gender?
- A: number of cross-gender edges not **very different** from randomly assigning each node a gender
  - according to the gender balance in the original network





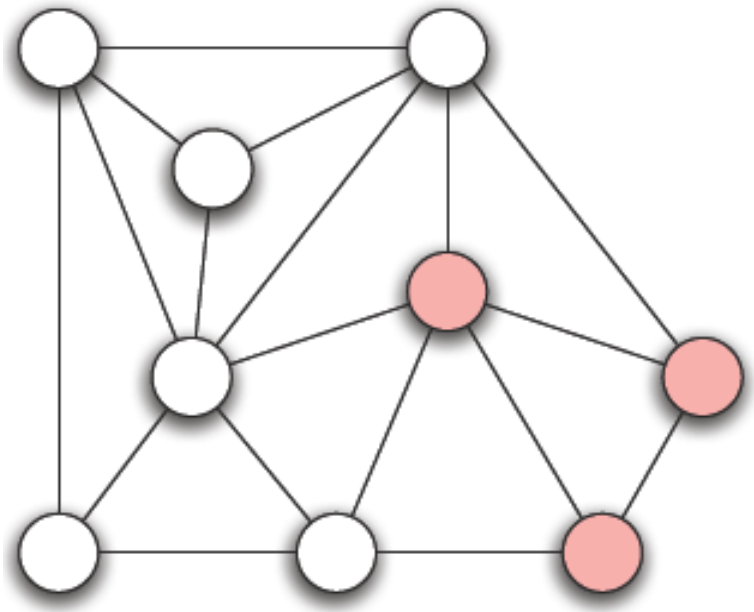
Measuring Homophily

- p the probability (fraction) of males
- q = 1-p the probability (fraction) of females
- For a given edge:
  - Homophily:
    - Prob(both ends male) = p*p
    - Prob(both ends female) = q*q
  - Cross gender:
    - Prob(ends male and female) = 2*p*q
- Homophily Test: If the fraction of cross-gender edges is **significantly** less than 2pq, then there is evidence for homophily



Measuring Homophily

- Ex:
  - p = 6/9 = 2/3
  - q = 1/3
  - 2pq = 4/9 = 8/18
  - 5/18 cross-gender edges
  - Test: 5/18 < 8/18 => some evidence of homophily
- Need definition of "significantly less than"
  - standard statistical significance
- What if cross-gender edges more than 2pq?
  - inverse homophily (Ex: network of romantic relationships)
- How to extend to characteristics with more than 2 states?



Mechanisms Underlying Homophily

- Homophily has 2 mechanisms for link creation
  - Selection: select friends with similar characteristics
    - individual characteristics drive the formation of links
    - involves immutable characteristics (determined at birth)
  - Social influence: modify behavior close to behaviors of friends
    - the reverse of selection
    - involves mutable characteristics (behaviors, activities, interests, beliefs, and opinions)

The Interplay of Selection and Social Influence

- Q: When homophily is observed, is it a result of selection or social influence?
  - Have people adapted their behaviors to become more like their friends, or have they selected friends who were already like them?
- A: Track the network and monitor the results of the two mechanisms (more details later)

The Interplay of Selection and Social Influence

- Most of the times, both mechanisms apply and interact with each other
- Studies show that teenage friends are similar to each other in their behaviors, and both selection and social influence apply:
  - teenagers seek social circles of people like them and peer pressure causes conform to behavioral patterns within these circles
- Q: how the two mechanisms interact and whether one is more strongly at work than the other? (more details later)

Affiliation

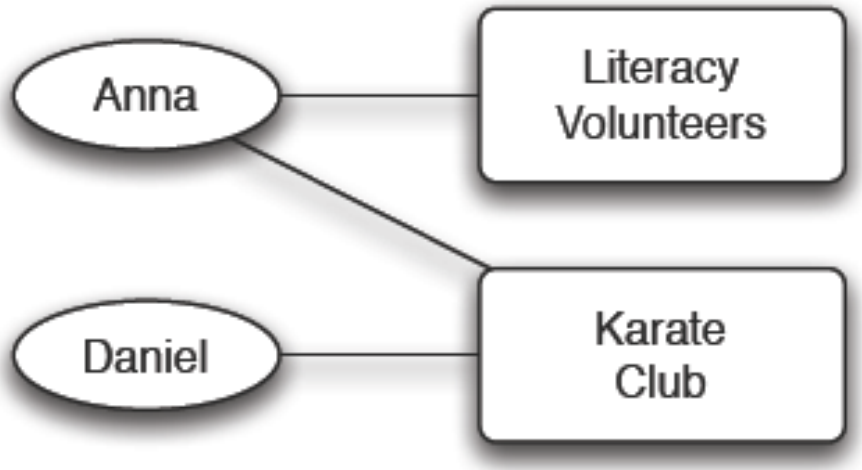
- Story so far:
  - Homophily groups together similar nodes
  - Selection and social influence determine the formation of links in a network
  - Similarity of nodes based on characteristics
- How to model these characteristics?
  - They represent surrounding contexts of networks
  - They exist “outside” the network
  - How to put these contexts into the network itself?

Affiliation

- Represent the set of activities a person takes part in (a general view of “activity”)
  - Ex: part of a particular company, organization, frequenting a particular place, hobby
- Refer to activities as foci: “focal points” of social interaction

Affiliation Networks

- Affiliation network:
  - bipartite graph
    - nodes divided into 2 sets
    - no edges joining a pair of nodes that belong to the same set
  - people affiliated with foci
- Ex:
  - Anna participates in both of the social foci on the right
  - Daniel participates in only one

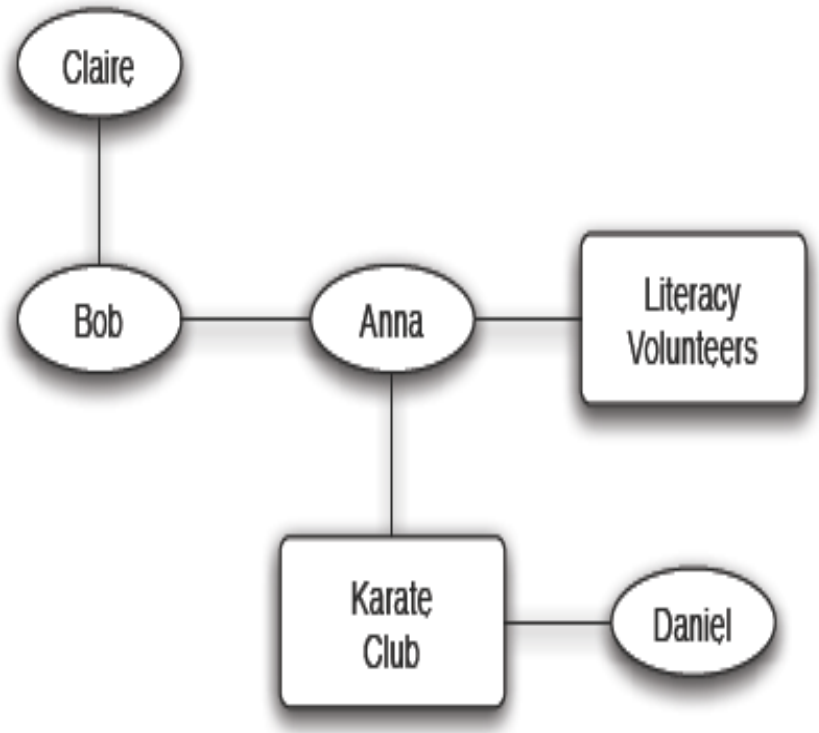


Co-Evolution of Social and Affiliation Networks

- Social networks change over time
  - new friendship links are formed
- Affiliation networks change over time
  - people become associated with new foci
- Co-evolution reflects interplay between selection and social influence
  - 2 people participate in a shared focus can become friends
  - if 2 people are friends, they can share their foci
- How to represent co-evolution with a single network?

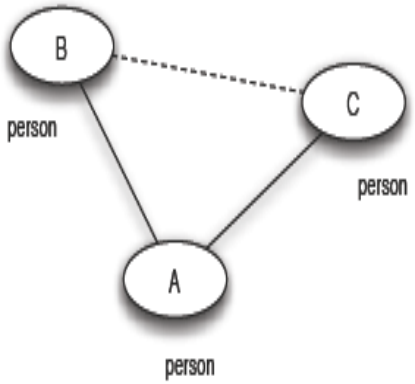
Social-affiliation networks

- Social-affiliation network contains:
  - a social network on the people and
  - an affiliation network on the people and foci

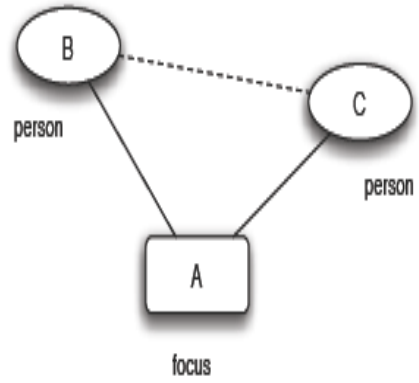


Social-affiliation networks

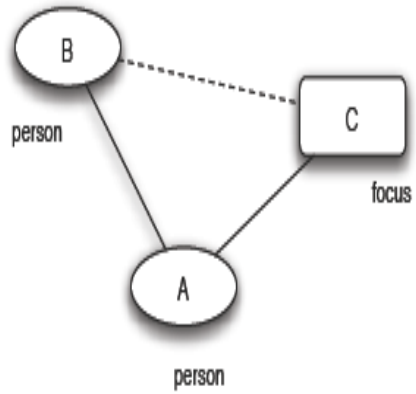
- In social-affiliation networks link formation as a closure process
- Several options for "closing" B-C
  - triadic closure: A, B, and C represent a person (already examined)
  - focal closure: B and C people, A focus
    - **selection**: B links to similar C (common focus)
  - membership closure: A and B people, C focus
    - **social influence**: B links to C influenced by A



(a) *Triadic closure*



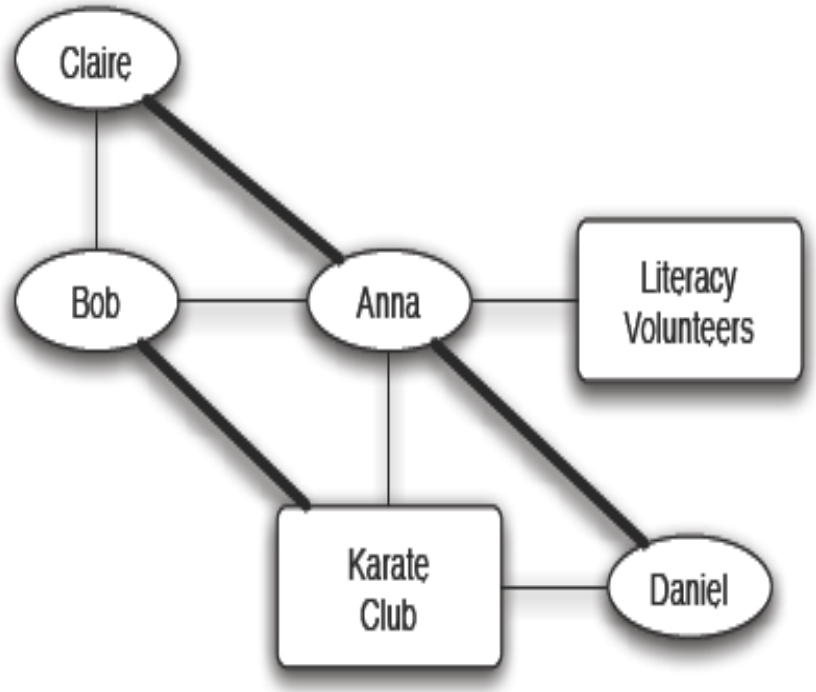
(b) *Focal closure*



(c) *Membership closure*

Example

- Bob introduces Anna to Claire
- Karate “introduces” Anna to Daniel
- Anna introduces Bob to Karate



Edges with **bold** are the newly formed

Tracking Link Formation in On-Line Data

- Story so far: a set of mechanisms that lead to the formation of links
  - triadic closure
  - focal closure
  - membership closure
- Tracking these mechanisms in large populations
  - their accumulation observable in the aggregate

Tracking triadic closure

- Likelihood of link as a function of common friends?

1. Two snapshots of the network
2. For each k, find all pairs of nodes with k common friends in the first snapshot, but not directly connected
3. T(k): fraction of these pairs connected in the second snapshot
   - empirical estimate of probability that a link will form between two people with k common friends
4. Plot T(k) as a function of k
   - T(0) is the rate of link formation when it does not close a triangle

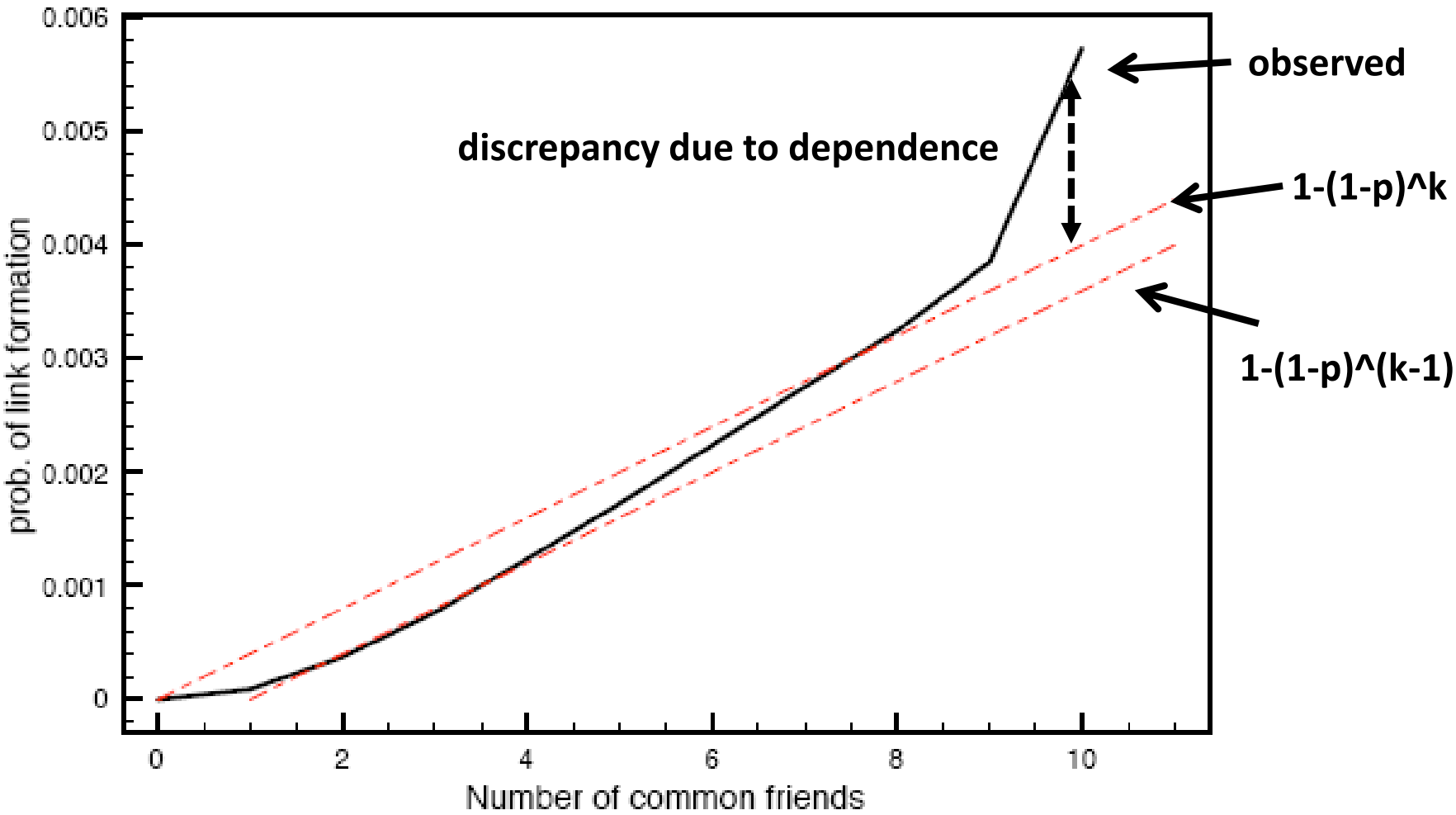
Tracking triadic closure

- Kossinets and Watts computed T(k)
  - full history of e-mail communication (“who-talks-to-whom”)
  - a one-year period
  - 22,000 students at a large U.S. university
  - observations in each snapshot were one day apart (average over multiple snapshots)

Tracking triadic closure

- Interpret the result compared to a baseline
- Assume that each common friend that 2 people have, gives them an independent probability **p** of forming a link
  - 2 people have k friends in common => the probability they fail to form a link is **(1-p)^k**
  - 2 people have k friends in common => probability that they form a link is **1-(1-p)^k**

Tracking triadic closure

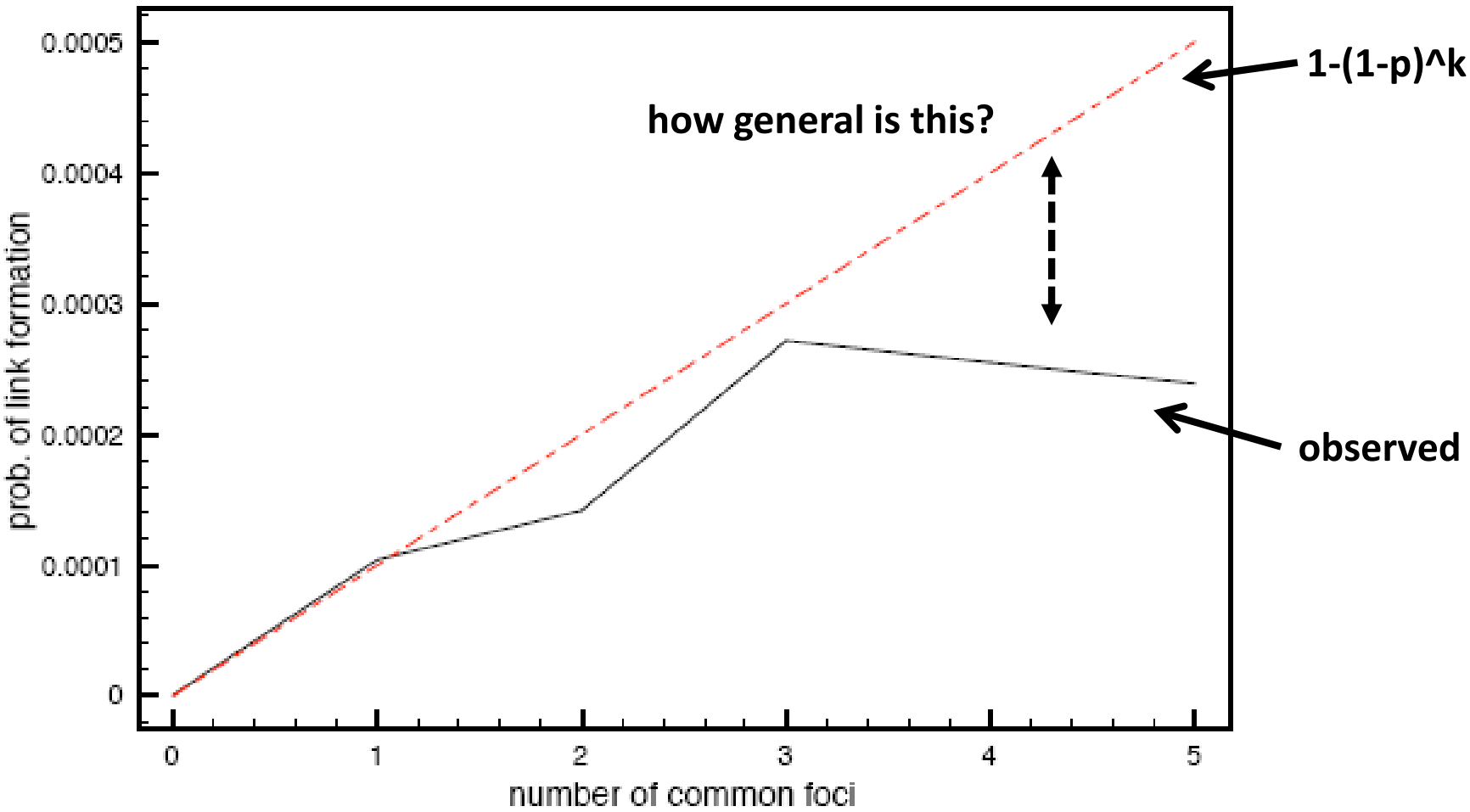


Tracking focal closure

- Likelihood of link formation as a function of the number of common foci?

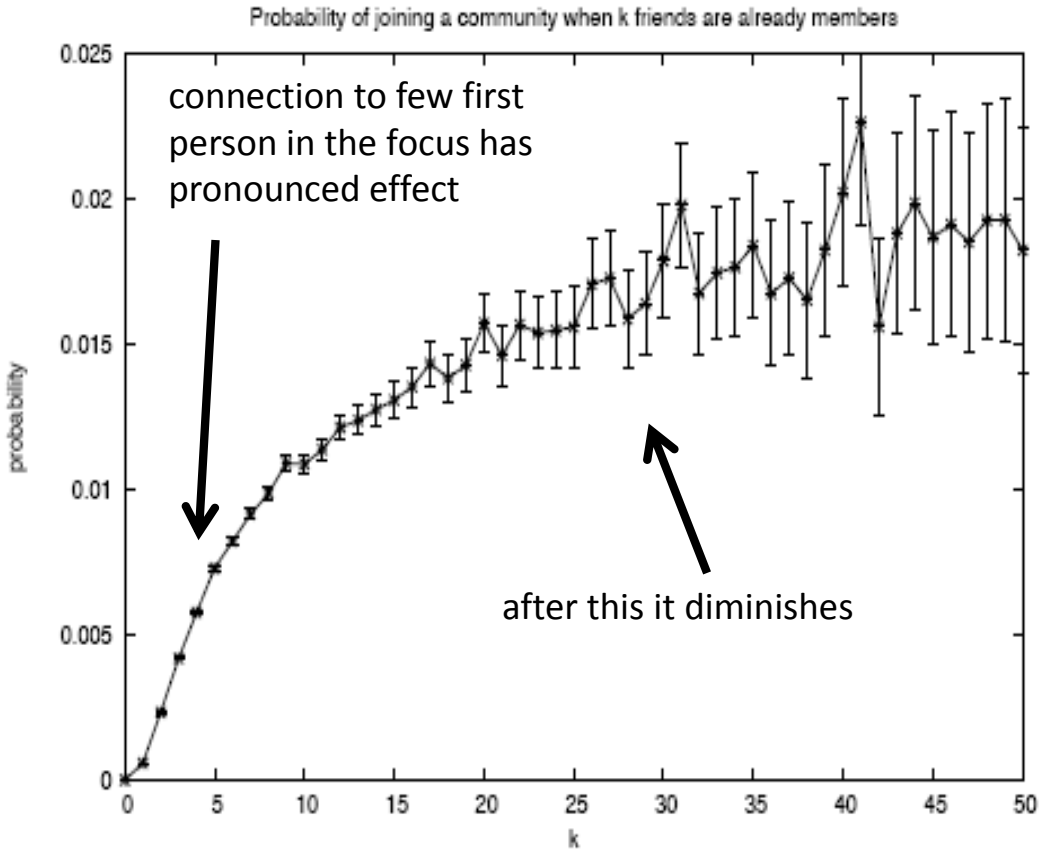
- Kossinets and Watts supplemented their university e-mail dataset with information about the class schedules
  - each class became a focus
  - students shared a focus if they had taken a class together

Tracking focal closure



Tracking membership closure

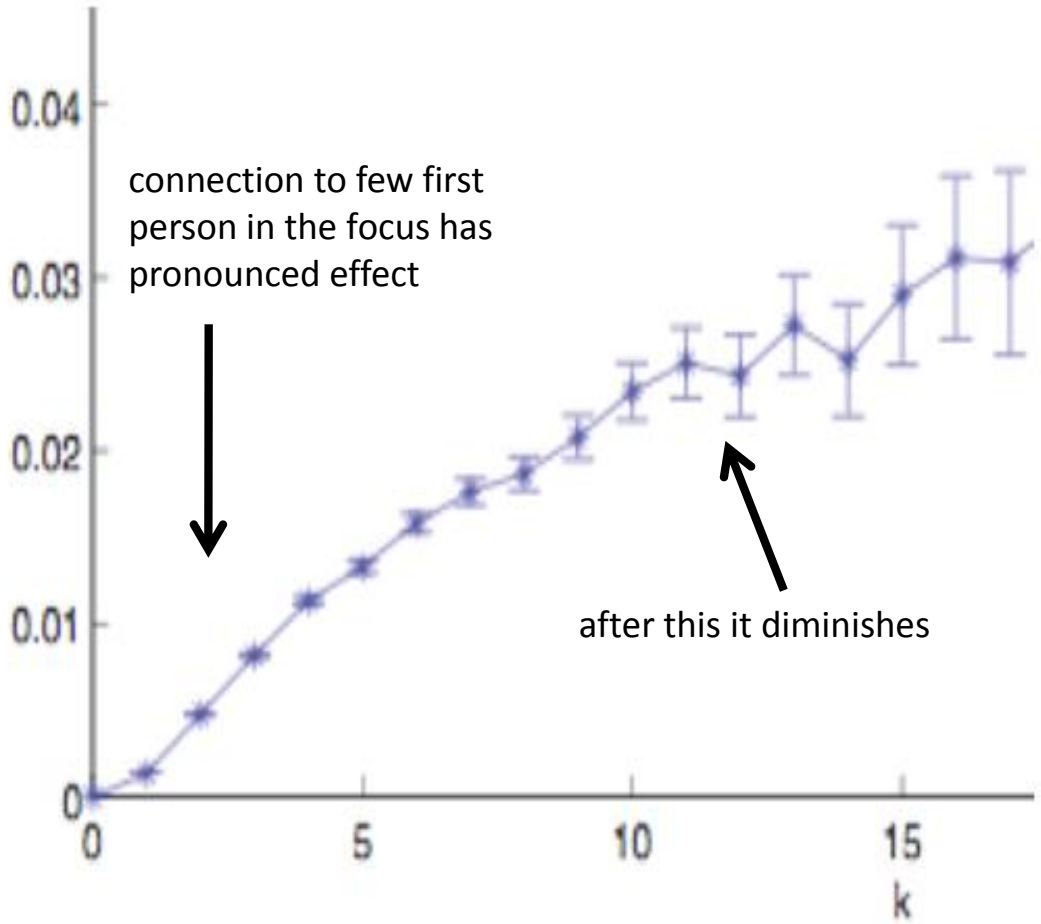
- Blogging site LiveJournal
  - social network (friendship links)
  - foci correspond to membership in user-defined communities



probability of joining a LiveJournal community as a function of the number of friends who are already members

Tracking membership closure

- Wikipedia editors
  - link editors when they communicated (user talk page)
  - each Wikipedia article defines a focus (editor associated with the articles he/she edited)



probability of editing a Wikipedia articles as a function of the number of friends who have already done so

Quantifying the Interplay Between Selection and Social Influence

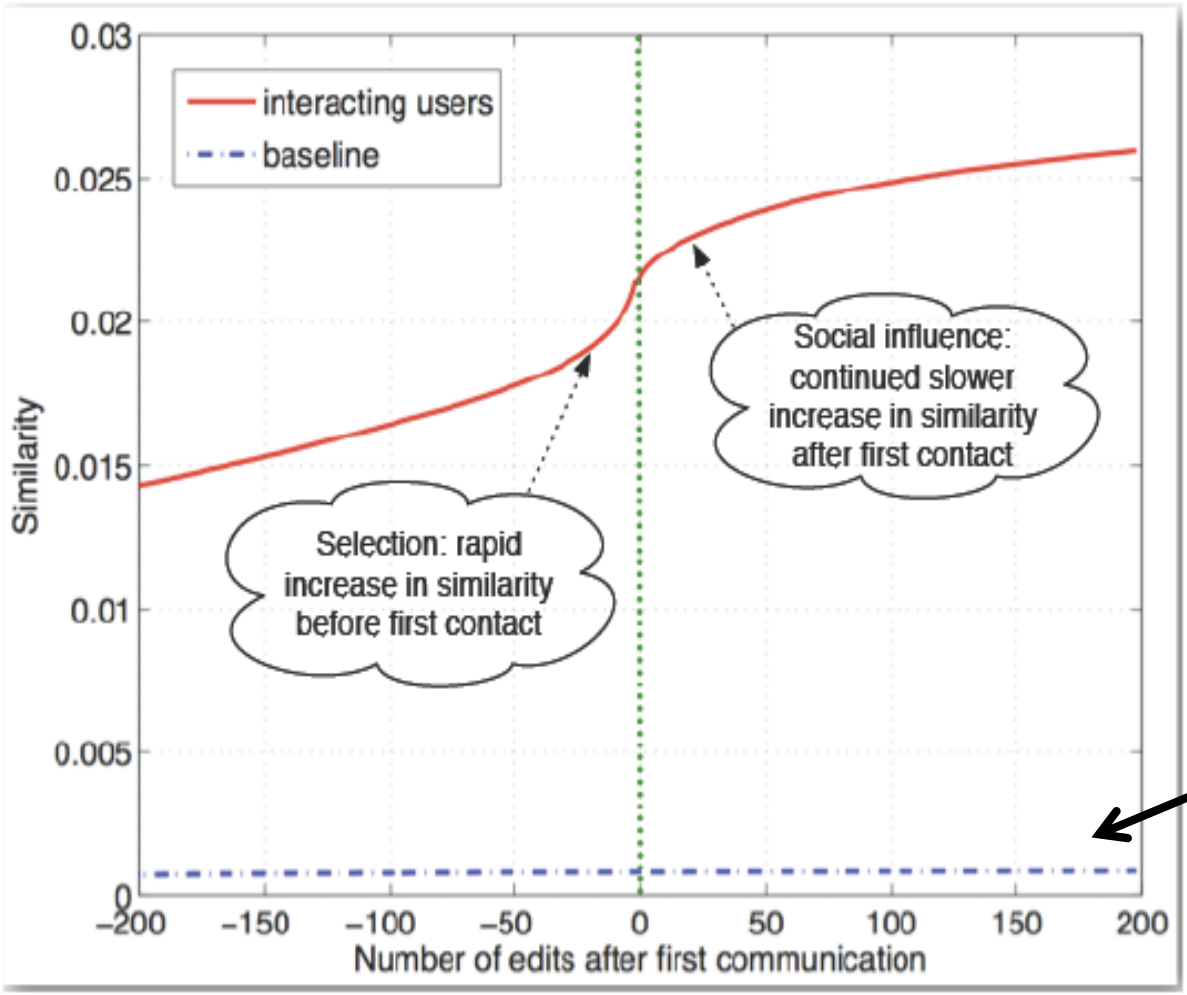
- How selection and social influence work together to produce homophily?
  - How do similarities in behavior between two Wikipedia editors relate to their pattern of social interaction over time?
  - Similarity between 2 Wikipedia editors A, B:

$$\frac{\text{number of articles edited by } \textit{both } A \text{ and } B}{\text{number of articles edited by } \textit{at least one of } A \text{ or } B}$$

- Is homophily (similarity) due to editors connected (talk) with those edited the same articles (selection), or because editors are led to edit articles by those they talk to (social influence)?

Quantifying the Interplay Between Selection and Social Influence

Record similarity over time for each pair of editors A and B who have ever talked



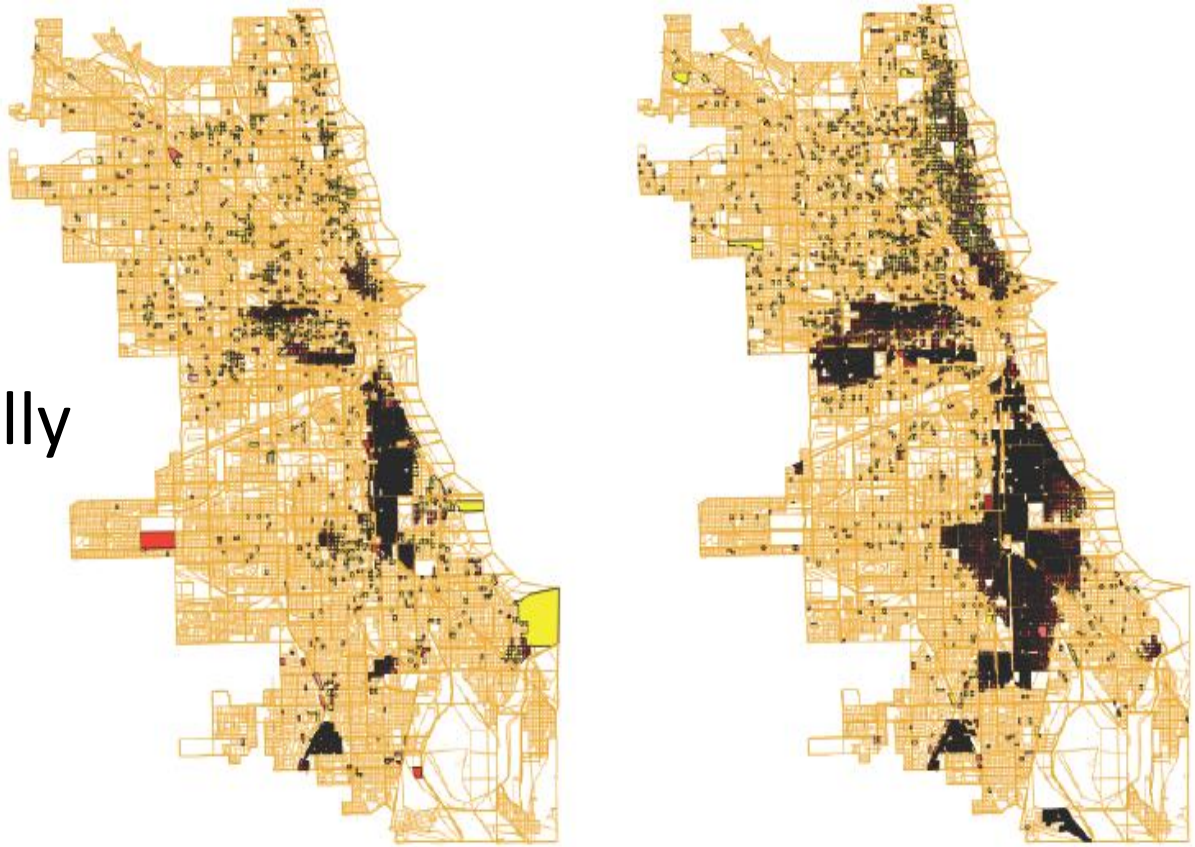
similarity of non-interacting pairs

"tick" in time whenever either A or B performs an action (editing or talking). Time 0 is the point at which they first talked

A SPATIAL MODEL OF SEGREGATION

Spatial patterns of segregation

- One of the most strong effects of homophily is in the formation of ethnically and racially homogeneous neighborhoods in cities
  - a process with a dynamic aspect
  - what mechanisms?



(a) *Chicago, 1940* (b) *Chicago, 1960*

In blocks colored yellow and orange the percentage of African-Americans is below 25, while in blocks colored brown and black the percentage is above 75

The Schelling Model

- How global patterns of spatial segregation can arise from the effect of homophily operating at a local level (Thomas Schelling)
  - an intentionally simplified mechanism
  - works even when no one individual explicitly wants a segregated outcome

The Schelling Model

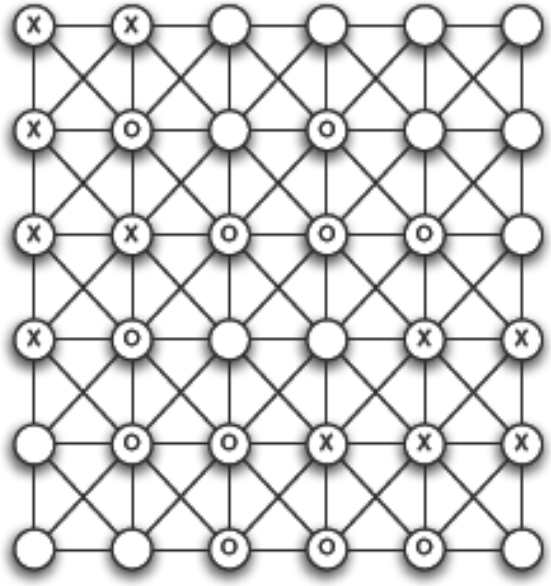
- Model assumptions:
  - Population of individuals called agents
  - Each agent of type X or type O
  - The two types represent some characteristic as basis for homophily (race, ethnicity, country of origin, or native language)
  - Agents reside in cells of a grid (simple model of a 2-D city map)
  - Some cells contain agents while others are unpopulated
  - Cell's neighbors: cells that touch it (including diagonal contact)

X	X				
X	O		O		
X	X	O	O	O	
X	O			X	X
	O	O	X	X	X
		O	O	O	

The Schelling Model

x	x				
x	o		o		
x	x	o	o	o	
x	o			x	x
	o	o	x	x	x
		o	o	o	

(a) *Agents occupying cells on a grid.*



(b) *Neighbor relations as a graph.*

Cells are the nodes and edges connect neighboring cells.
We will continue with the geometric grid rather than the graph.

The Schelling Model

- Local mechanism:
  - each agent wants to have at least some t other agents of its own type as neighbors (t the same for all)
  - unsatisfied agents have fewer than t neighbors of the same type as itself and move to a new cell
- Ex (figure):
  - agents with ID
  - t = 3

X1*	X2*				
X3	O1*		O2		
X4	X5	O3	O4	O5*	
X6*	O6			X7	X8
	O7	O8	X9*	X10	X11
		O9	O10	O11*	

(a) *An initial configuration.*

X3	X6	O1	O2		
X4	X5	O3	O4		
	O6	X2	X1	X7	X8
O11	O7	O8	X9	X10	X11
	O5	O9	O10*		

(b) *After one round of movement.*

The Dynamics of Movement

- Unsatisfied agents move in rounds
  - consider unsatisfied agents in some order
    - random or row-sweep
  - unsatisfied agents move to an unoccupied cell where will be satisfied
    - random or to nearest cell that satisfies them
  - may cause other agents to be unsatisfied
  - deadlocks may appear (no cell that satisfies)
    - stay or move randomly
- All variations have similar results
- Ex (figure):
  - t=3, one round, row-sweep, move to nearest cell, stay when deadlocks

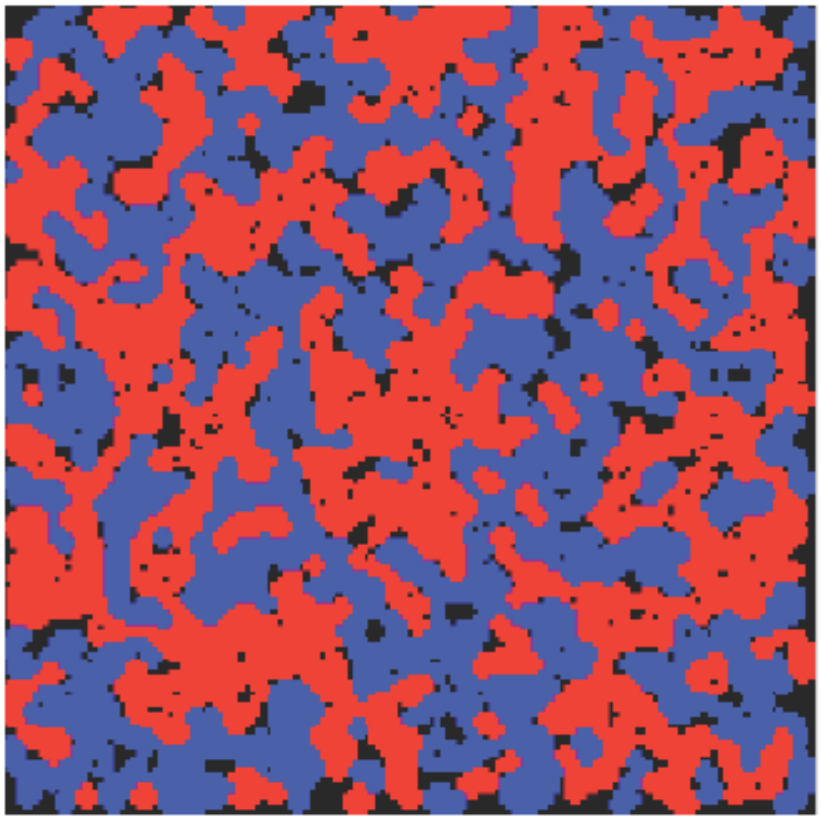
X1*	X2*				
X3	O1*		O2		
X4	X5	O3	O4	O5*	
X6*	O6			X7	X8
	O7	O8	X9*	X10	X11
		O9	O10	O11*	

(a) *An initial configuration.*

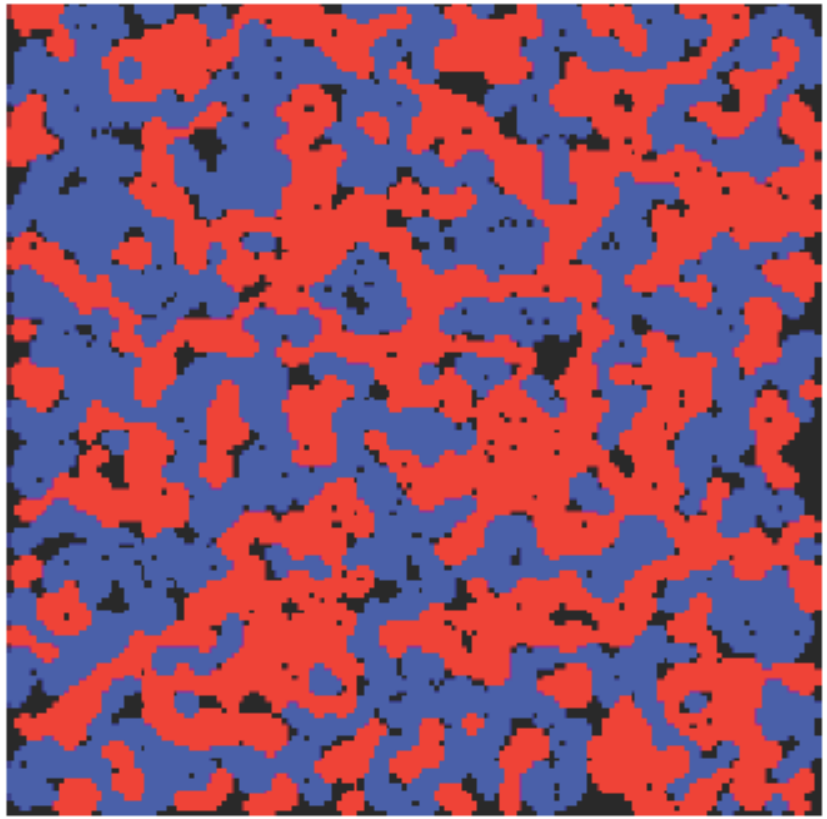
X3	X6	O1	O2		
X4	X5	O3	O4		
	O6	X2	X1	X7	X8
O11	O7	O8	X9	X10	X11
	O5	O9	O10*		

(b) *After one round of movement.*

Larger examples



(a) *A simulation with threshold 3.*



(b) *Another simulation with threshold 3.*

Two runs (50 rounds) of the Schelling model with unsatisfied agents moving to a random location. Threshold t=3, 150-by-150 grid with 10, 000 agents. Each cell of first type is red, of second type blue, or black if unoccupied.

Interpretations of the Model

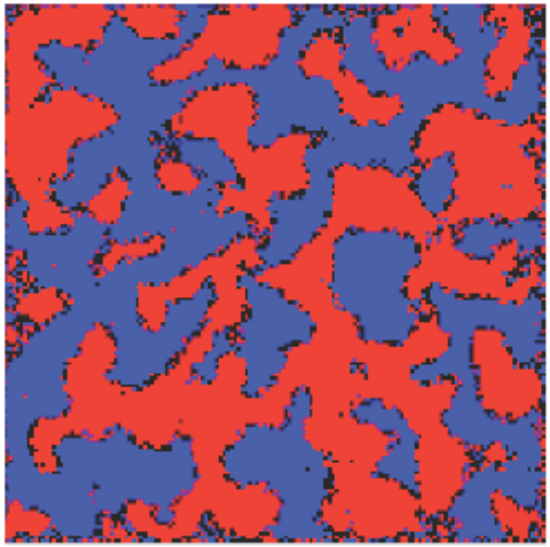
- Spatial segregation is taking place even though no individual agent is seeking it
  - agents just want to be near **t** others like them
  - when t=3, agents are satisfied being minority among its neighbors (5 neighbors of the opposite type)
- Ex (figure):
  - a checkerboard 4x4 pattern can make all agent satisfied (even for large grids)
  - we don't see this result in simulations

X	X	O	O	X	X
X	X	O	O	X	X
O	O	X	X	O	O
O	O	X	X	O	O
X	X	O	O	X	X
X	X	O	O	X	X

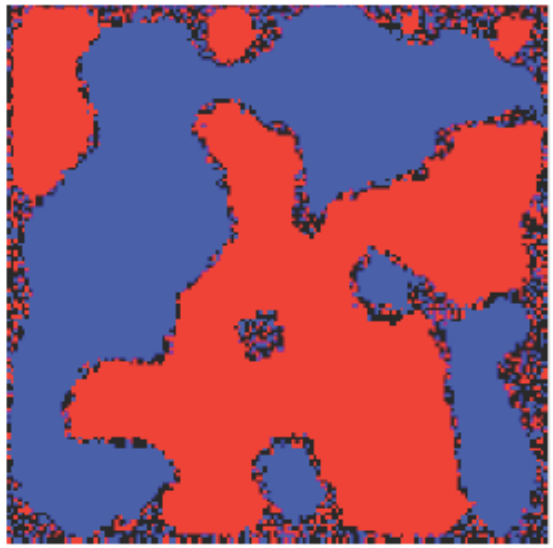
Interpretations of the Model

- More typically, agents form larger clusters
  - agents become unsatisfied and attach to larger clusters (where higher probability to be satisfied)
- The overall effect:
  - local preferences of individual agents have produced a global pattern that none of them necessarily intended

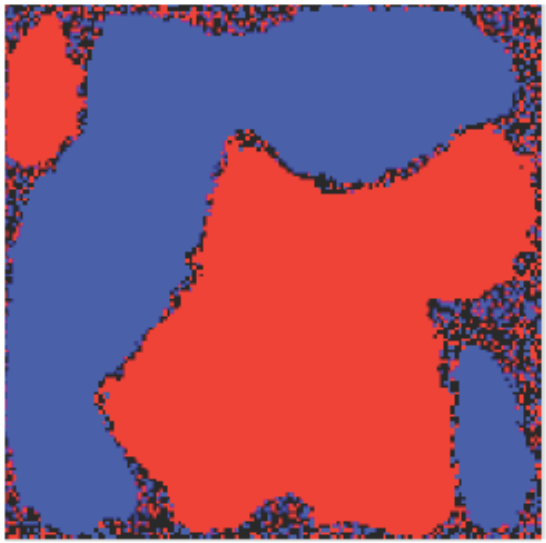
Interpretations of the Model



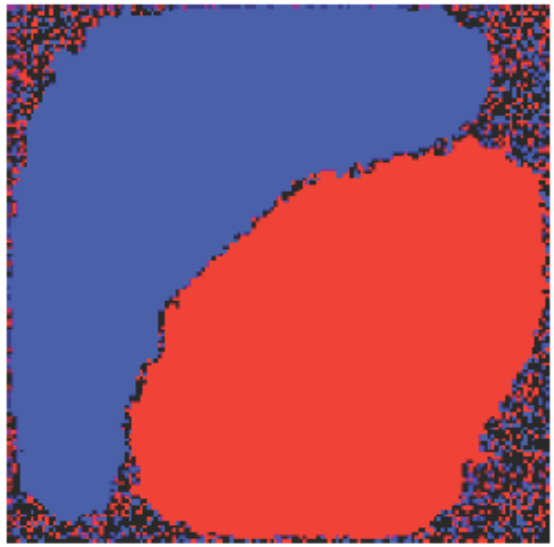
(a) *After 20 steps*



(b) *After 150 steps*



(c) *After 350 steps*



(d) *After 800 steps*

t=4, 150-by-150 grid, 10, 000 agents, varying number of rounds (steps), not shown until the end

Schelling model and Homophily

- The Schelling model is an example that, as homophily draws people together along immutable characteristics (race or ethnicity), it creates a natural tendency for mutable characteristics (decision about where to live) to change in accordance with the network structure

Today's Biz

1. Quick Review
2. Reminders
3. Social Networks Topics
4. **Parallel Triangle Counting**
5. Homework 1 solutions

Counting Triangles & The Curse of the Last Reducer

Slides from Siddharth Suri and Sergei Vassilvitskii, Yahoo! Research

Counting Triangles & The Curse of the Last Reducer

Siddharth Suri

Sergei Vassilvitskii

Yahoo! Research

Why Count Triangles?

Why Count Triangles?

Clustering Coefficient:

Given an undirected graph $G = (V, E)$

cc(v) = fraction of v's neighbors who are neighbors themselves

$$= \frac{|\{(u, w) \in E | u \in \Gamma(v) \wedge w \in \Gamma(v)\}|}{\binom{d_v}{2}}$$

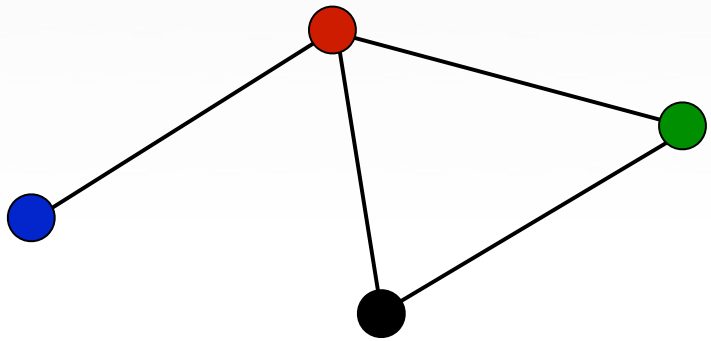
Why Count Triangles?

Clustering Coefficient:

Given an undirected graph $G = (V, E)$

cc(v) = fraction of v's neighbors who are neighbors themselves

$$= \frac{|\{(u,w) \in E | u \in \Gamma(v) \wedge w \in \Gamma(v)\}|}{\binom{d_v}{2}}$$



cc ( ) = N/A

cc ( ) = 1/3

cc ( ) = 1

cc ( ) = 1

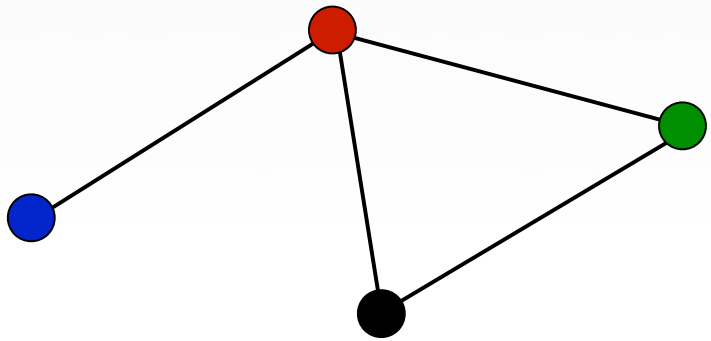
Why Count Triangles?

Clustering Coefficient:

Given an undirected graph $G = (V, E)$

cc(v) = fraction of v's neighbors who are neighbors themselves

$$= \frac{|\{(u,w) \in E | u \in \Gamma(v) \wedge w \in \Gamma(v)\}|}{\binom{d_v}{2}} = \frac{\#\Delta s \text{ incident on } v}{\binom{d_v}{2}}$$



cc ( ) = N/A

cc ( ) = 1/3

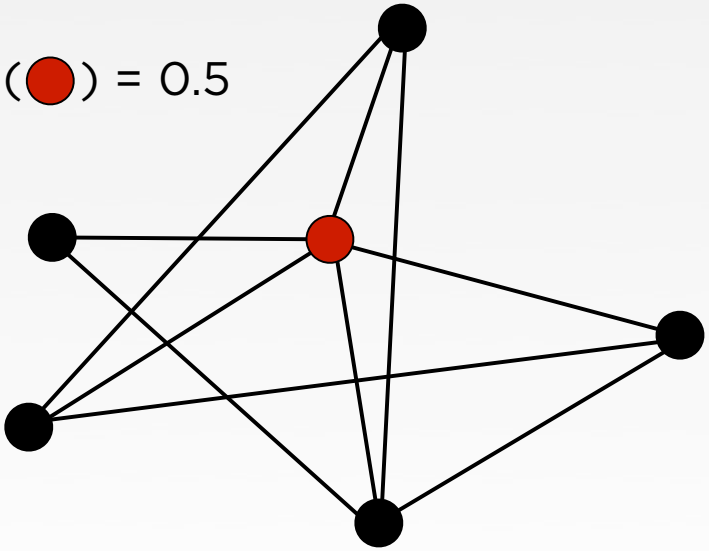
cc ( ) = 1

cc ( ) = 1

Why Clustering Coefficient?

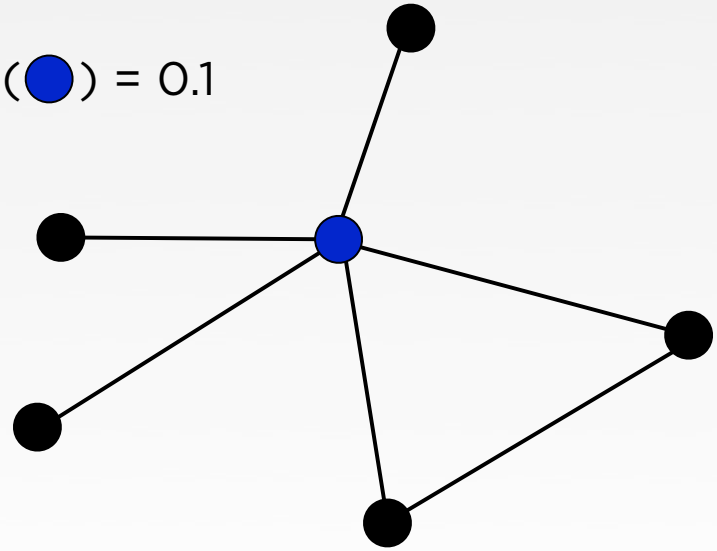
Captures how tight-knit the network is around a node.

cc ( ) = 0.5



vs.

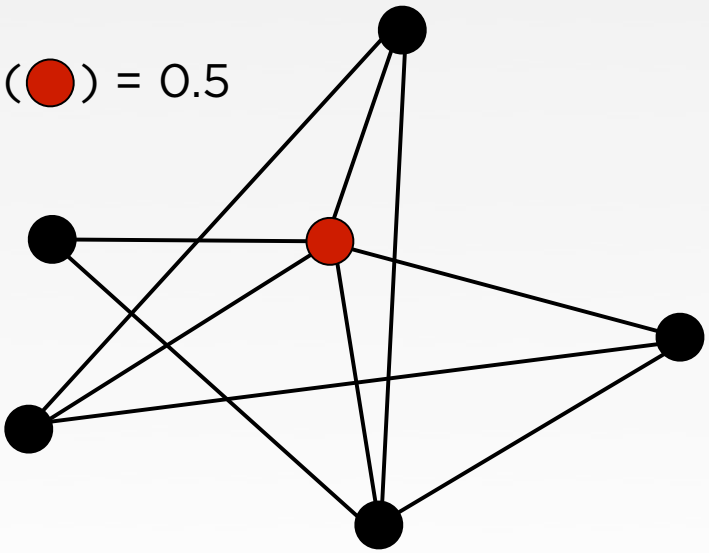
cc ( ) = 0.1



Why Clustering Coefficient?

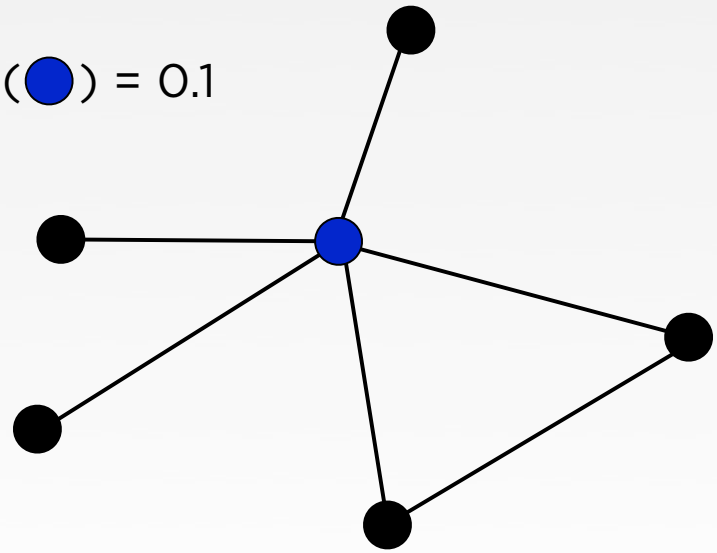
Captures how tight-knit the network is around a node.

cc ( ) = 0.5



vs.

cc ( ) = 0.1



Network Cohesion:

- Tightly knit communities foster more trust, social norms. [Coleman '88, Portes '88]

Structural Holes:

- Individuals benefit form bridging [Burt '04, '07]

Why MapReduce?

De facto standard for parallel computation on large data

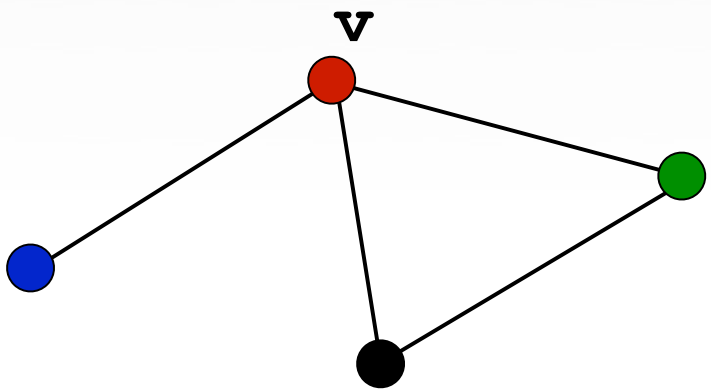
- Widely used at: Yahoo!, Google, Facebook,
- Also at: New York Times, Amazon.com, Match.com, ...
- Commodity hardware
- Reliable infrastructure

- Data continues to outpace available RAM !

How to Count Triangles

Sequential Version:

```
foreach v in V
    foreach u,w in Adjacency(v)
        if (u,w) in E
            Triangles[v]++
```

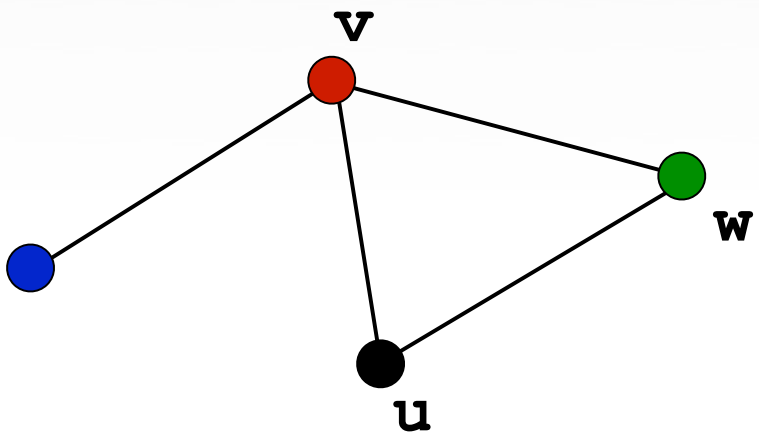



```
Triangles[v]=0
```

How to Count Triangles

Sequential Version:

```
foreach v in V
    foreach u,w in Adjacency(v)
        if (u,w) in E
            Triangles[v]++
```

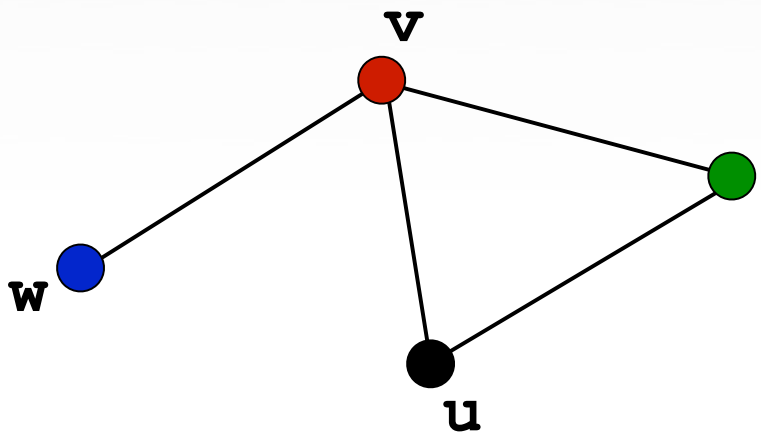


Triangles[v]=1

How to Count Triangles

Sequential Version:

```
foreach v in V
    foreach u,w in Adjacency(v)
        if (u,w) in E
            Triangles[v]++
```



Triangles[v]=1

How to Count Triangles

Sequential Version:

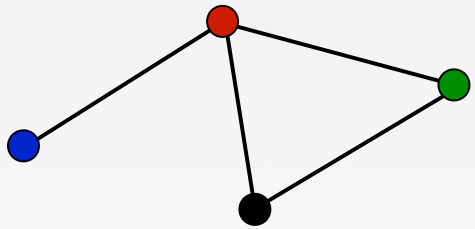
```
foreach v in V
    foreach u,w in Adjacency(v)
        if (u,w) in E
            Triangles[v]++
```

Running time: $\sum_{v \in V} d_v^2$

Even for sparse graphs can be quadratic if one vertex has high degree.

Parallel Version

Parallelize the edge checking phase



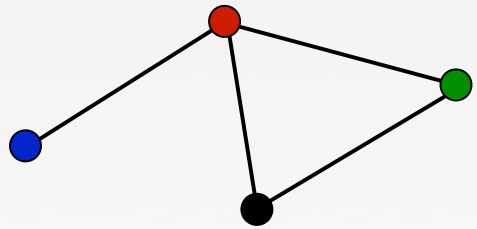
Parallel Version

Parallelize the edge checking phase

- Map 1: For each v send $(v, \Gamma(v))$ to single machine.
- Reduce 1: Input: $\langle v; \Gamma(v) \rangle$

  Output: all 2 paths $\langle (v_1, v_2); u \rangle$ where $v_1, v_2 \in \Gamma(u)$

  (●,●);● (●,●); ● (●,●);●



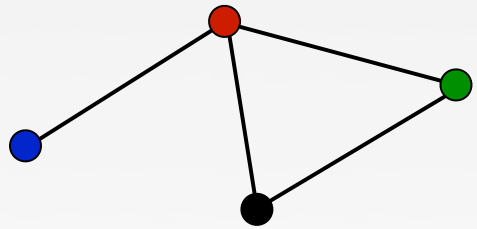
Parallel Version

Parallelize the edge checking phase

- Map 1: For each v send $(v, \Gamma(v))$ to single machine.
- Reduce 1: Input: $\langle v; \Gamma(v) \rangle$

  Output: all 2 paths $\langle (v_1, v_2); u \rangle$ where $v_1, v_2 \in \Gamma(u)$

  (●,●);● (●,●); ● (●,●);●



- Map 2: Send $\langle (v_1, v_2); u \rangle$ and $\langle (v_1, v_2); \$ \rangle$ for $(v_1, v_2) \in E$ to same machine.
- Reduce 2: input: $\langle (v, w); u_1, u_2, \ldots, u_k, \$? \rangle$

  Output: if $\$$ part of the input, then: $u_i = u_i + 1/3$

(●,●);●,$ ⟶ ● $+1/3$ ● $+1/3$ ● $+1/3$

(●,●);● ⟶

Data skew

How much parallelization can we achieve?

- Generate all the paths to check in parallel
- The running time becomes $\max_{v \in V} d_v^2$

Data skew

How much parallelization can we achieve?

- Generate all the paths to check in parallel
- The running time becomes $\max_{v \in V} d_v^2$

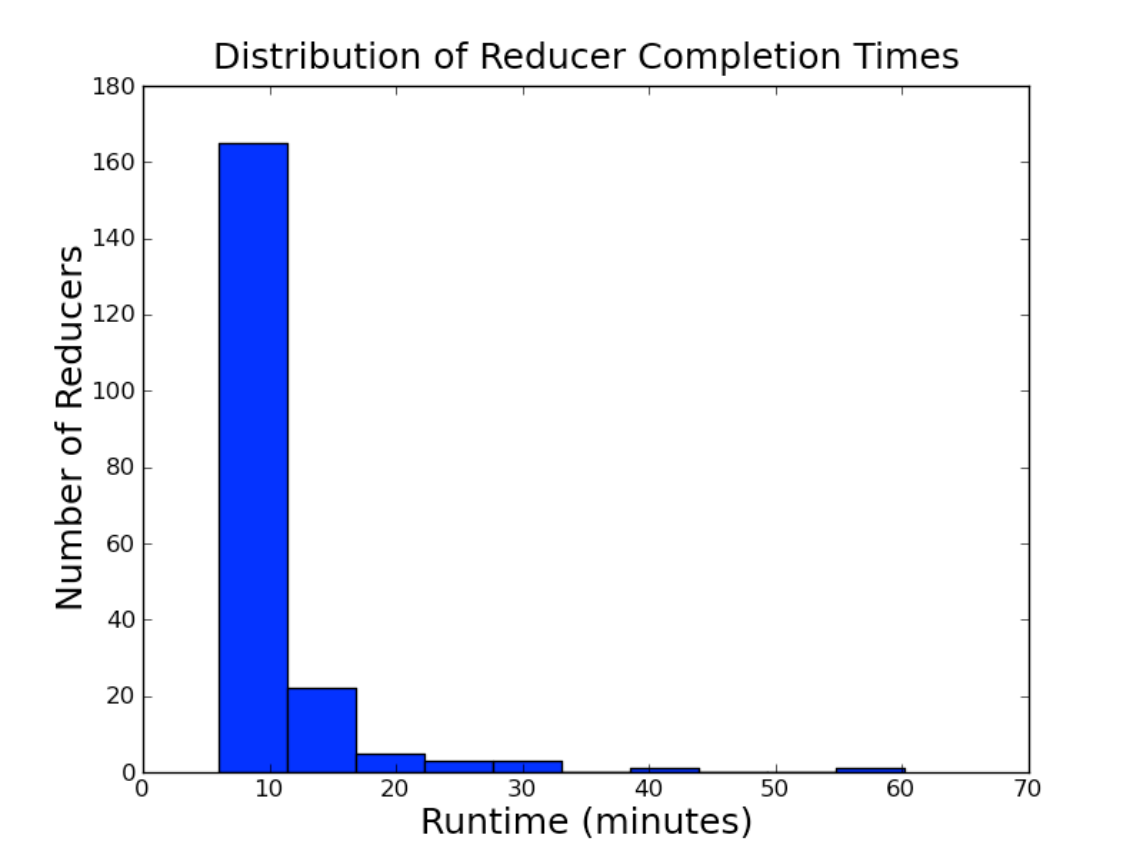
Naive parallelization does not help with data skew

- Some nodes will have very high degree
- Example. 3.2 Million followers, must generate 10 Trillion (10^13) potential edges to check.
- Even if generating 100M edges per second, 100K seconds ~ 27 hours.

“Just 5 more minutes”

Running the naive algorithm on LiveJournal Graph

- 80% of reducers done after 5 min
- 99% done after 35 min



Adapting the Algorithm

Approach 1: Dealing with skew directly

- currently every triangle counted 3 times (once per vertex)
- Running time quadratic in the degree of the vertex
- Idea: Count each once, from the perspective of lowest degree vertex
- Does this heuristic work?

Adapting the Algorithm

Approach 1: Dealing with skew directly

- currently every triangle counted 3 times (once per vertex)
- Running time quadratic in the degree of the vertex
- Idea: Count each once, from the perspective of lowest degree vertex
- Does this heuristic work?

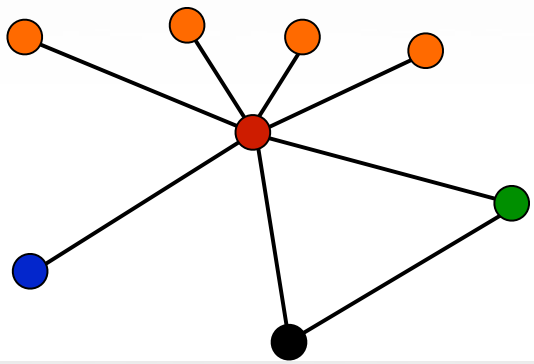
Approach 2: Divide & Conquer

- Equally divide the graph between machines
- But any edge partition will be bound to miss triangles
- Divide into overlapping subgraphs, account for the overlap

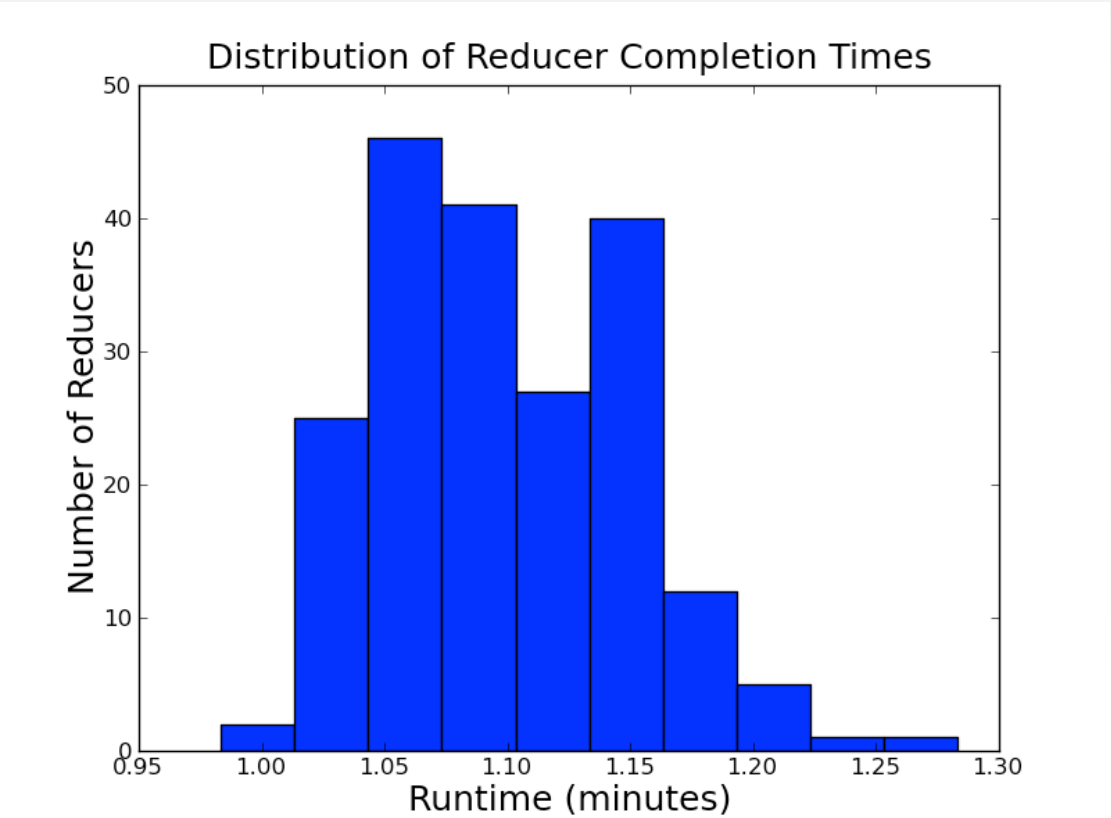
How to Count Triangles Better

Sequential Version [Schank '07]:

```
foreach v in V
   foreach u,w in Adjacency(v)
     if deg(u) > deg(v) && deg(w) > deg(v)
         if (u,w) in E
            Triangles[v]++
```



Does it make a difference?



Dealing with Skew

Why does it help?

- Partition nodes into two groups:
  - Low: $\mathcal{L} = \{v : d_v \leq \sqrt{m}\}$
  - High: $\mathcal{H} = \{v : d_v > \sqrt{m}\}$
- There are at most n low nodes; each produces at most $O(m)$ paths
- There are at most $2\sqrt{m}$ high nodes
  - Each produces paths to other high nodes: $O(m)$ paths per node

Dealing with Skew

Why does it help?

- Partition nodes into two groups:
  - Low: $\mathcal{L} = \{v : d_v \leq \sqrt{m}\}$
  - High: $\mathcal{H} = \{v : d_v > \sqrt{m}\}$
- There are at most n low nodes; each produces at most $O(m)$ paths
- There are at most $2\sqrt{m}$ high nodes
  - Each produces paths to other high nodes: $O(m)$ paths per node

- These two are identical !
- Therefore, no mapper can produce substantially more work than others.
- Total work is $O(m^{3/2})$, which is optimal

Approach 2: Graph Split

Partitioning the nodes:

- Previous algorithm shows one way to achieve better parallelization
- But what if even $O(m)$ is too much. Is it possible to divide input into smaller chunks?

Graph Split Algorithm:

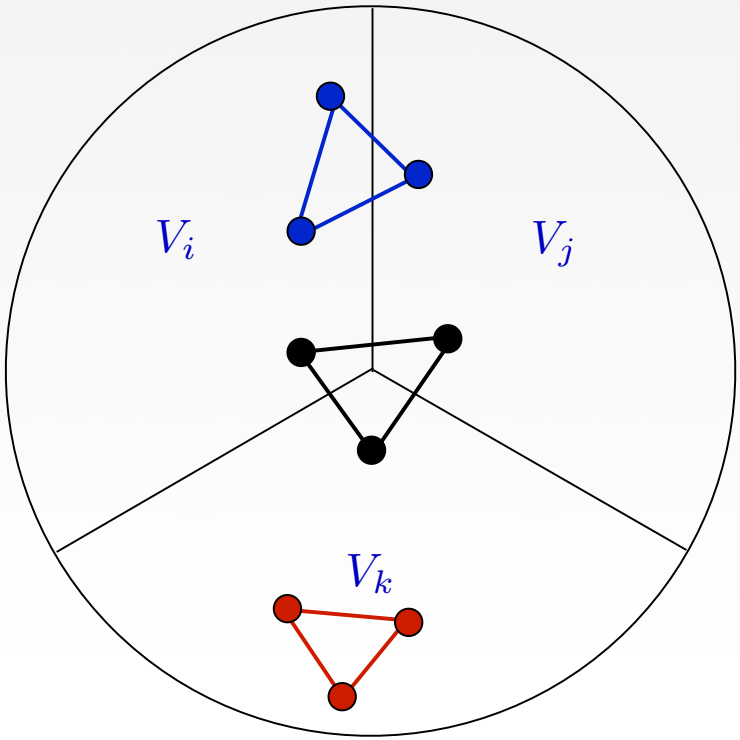
- Partition vertices into p equal sized groups $V_1, V_2, \ldots, V_p$.
- Consider all possible triples (V_i, V_j, V_k) and the induced subgraph:

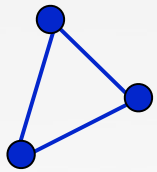
$$G_{ijk} = G\left[V_i \cup V_j \cup V_k\right]$$

- Compute the triangles on each G_{ijk} separately.

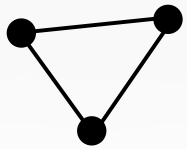
Approach 2: Graph Split

Some Triangles present in multiple subgraphs:

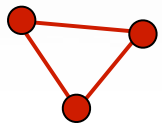




in p-2 subgraphs



in 1 subgraph



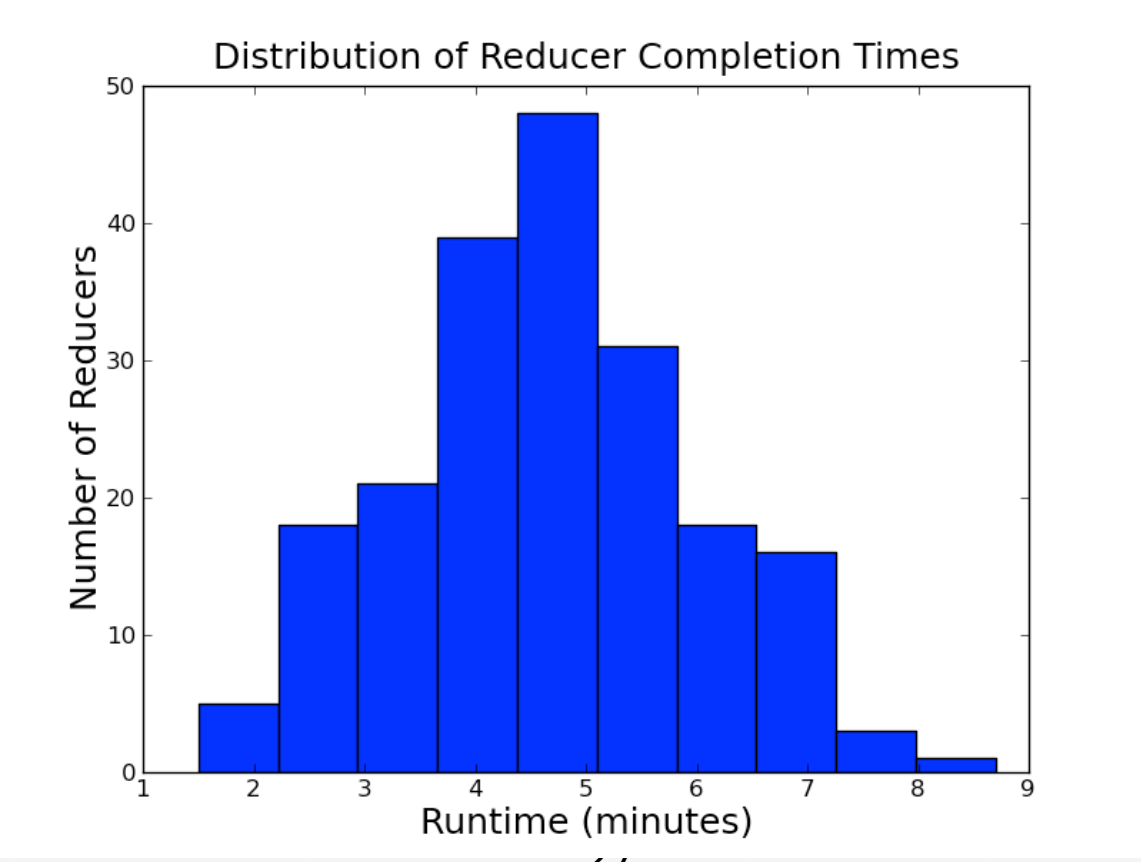
in ~p^2 subgraphs

Can count exactly how many subgraphs each triangle will be in

Approach 2: Graph Split

Analysis:

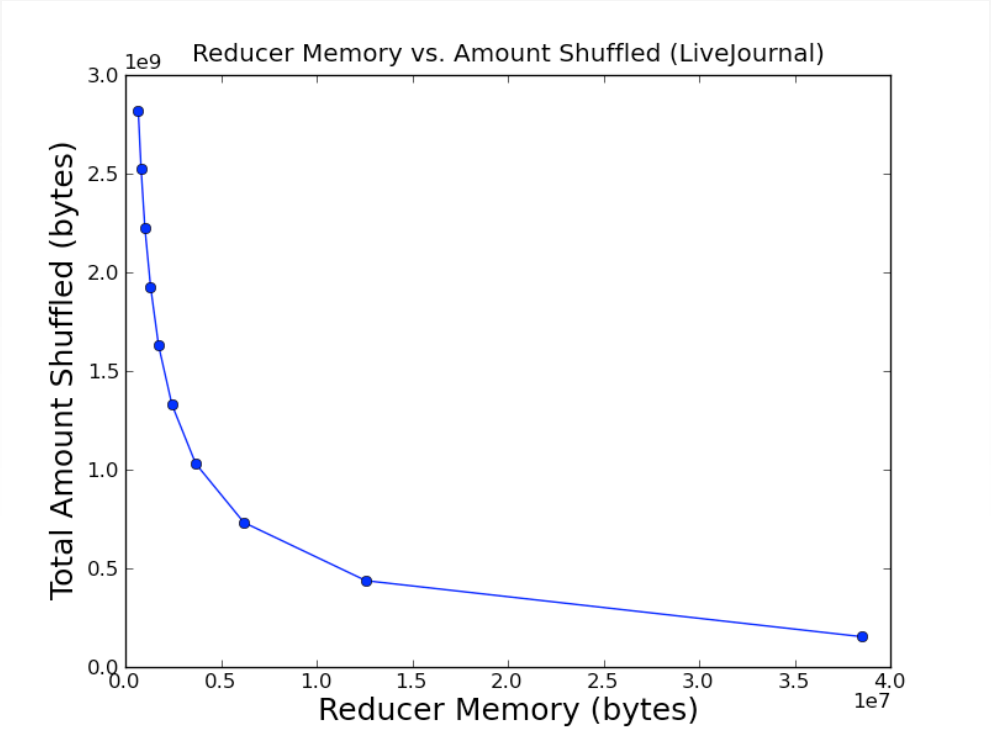
- Each subgraph has $O(m/p^2)$ edges in expectation.
- Very balanced running times



Approach 2: Graph Split

Analysis:

- Very balanced running times
- p controls memory needed per machine



Approach 2: Graph Split

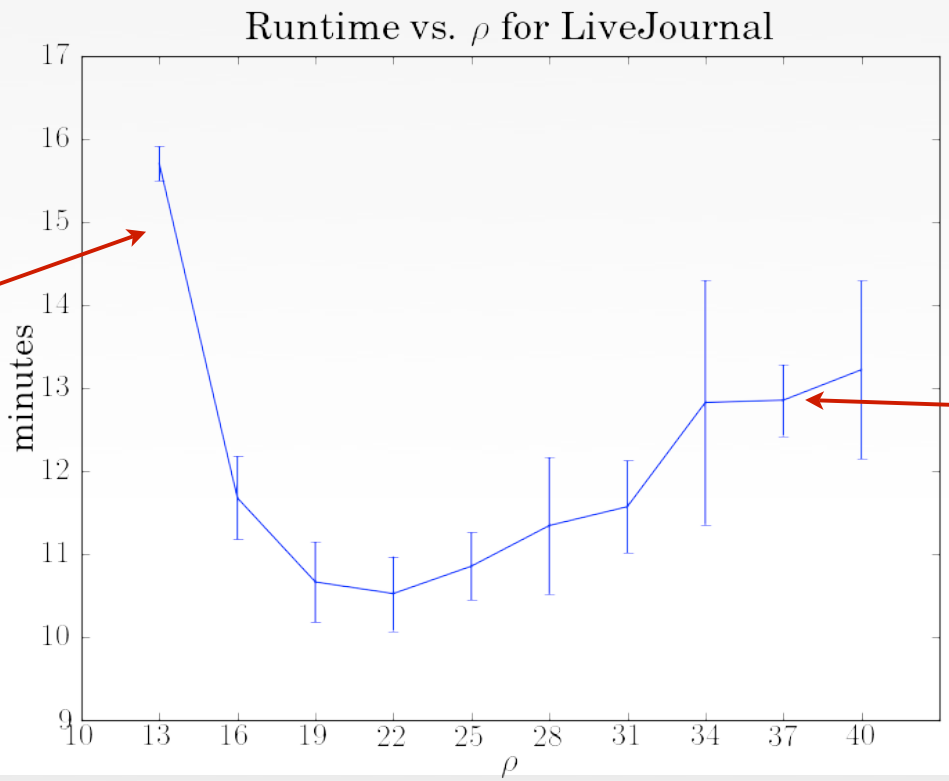
Analysis:

- Very balanced running times
- p controls memory needed per machine
- Total work: $p^3 \cdot O((m/p^2)^{3/2}) = O(m^{3/2})$, independent of p

Approach 2: Graph Split

Analysis:

- Very balanced running times
- p controls memory needed per machine
- Total work: $p^3 \cdot O((m/p^2)^{3/2}) = O(m^{3/2})$, independent of p



Input too big: paging

Shuffle time increases with duplication

Overall

Naive Parallelization Doesn't help with Data Skew

Related Work

- Tsourakakis et al. [09]:
  - Count global number of triangles by estimating the trace of the cube of the matrix
  - Don't specifically deal with skew, obtain high probability approximations.
- Becchetti et al. [08]
  - Approximate the number of triangles per node
  - Use multiple passes to obtain a better and better approximation

Conclusions

Conclusions

Think about data skew.... and avoid the curse

Conclusions

Think about data skew.... and avoid the curse

- Get programs to run faster

Conclusions

Think about data skew.... and avoid the curse

- Get programs to run faster
- Publish more papers

Conclusions

Think about data skew.... and avoid the curse

- Get programs to run faster
- Publish more papers
- Get more sleep

Conclusions

Think about data skew.... and avoid the curse

- Get programs to run faster
- Publish more papers
- Get more sleep
- ..

Conclusions

Think about data skew.... and avoid the curse

- Get programs to run faster
- Publish more papers
- Get more sleep
- ..
- The possibilities are endless!

Today's Biz

1. Quick Review
2. Reminders
3. Social Networks Topics
4. Parallel Triangle Counting
5. **Homework 1 Solutions**