

Homophily & Network Formation

Advanced Social Computing

Department of Computer Science
University of Massachusetts, Lowell
Fall 2020

Hadi Amiri
hadi@cs.uml.edu



Announcements

- **HW2:**
 - Due date: 9/23, 3:30 PM
 - Submission link on Piazza

Lecture Topics

- Homophily
 - Selection
 - Social Influence
- Affiliation Networks
- Network Formation

Homophily

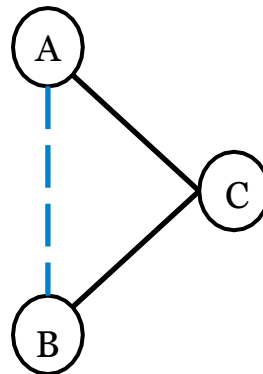
- The principle that we tend to be similar to our friends!
 - your friends are generally similar to you in terms of your characteristics!
 - **Immutable characteristics**
 - race, country of birth, etc., (determined at birth).
 - **Mutable characteristics**
 - location, occupations, affluence, interests, beliefs, opinions, etc (change through time).
- Factors that exist outside the nodes and edges of a network (**surrounding contexts**)

Homophily- Cnt.

- Links in social networks tend to connect people who are *similar* to one another
 - Formation of links in networks!

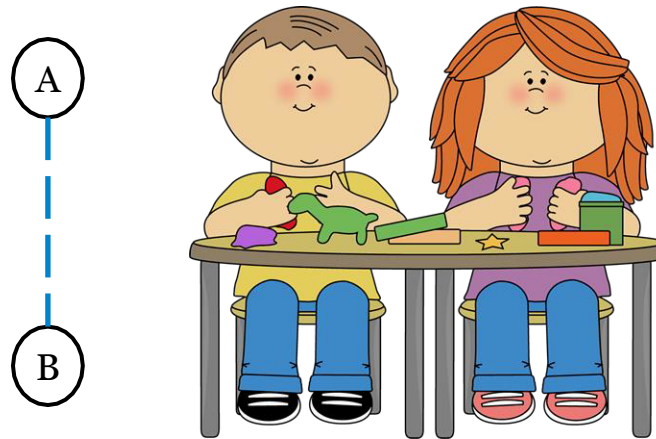
Homophily- Cnt.

- Formation of a new link (friendship):
 - **Case 1: Triadic Closure**
 - Two people will connect through a common friend!
 - Link is added for reasons that are **intrinsic** to the network itself.
 - We don't need to look beyond the network to understand where the links came from.



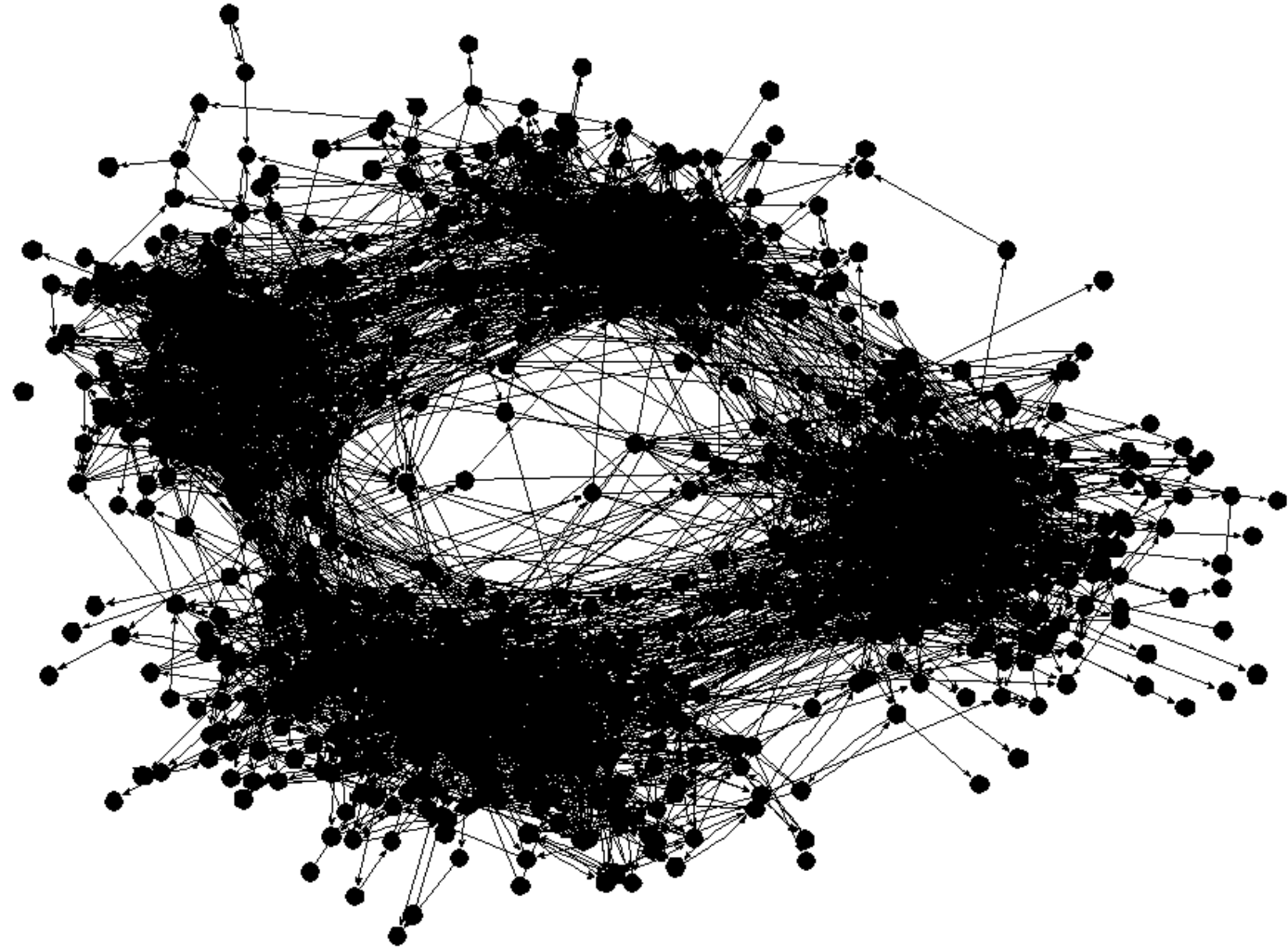
Homophily- Cnt.

- Formation of a new link (friendship):
 - **Case 2: Homophily**
 - Two people attend the same school / work for the same company!
 - The link is added for **contextual** reasons that are beyond the network.



Homophily- Cnt.

- Social net among students in a middle and high school.



Two context features:

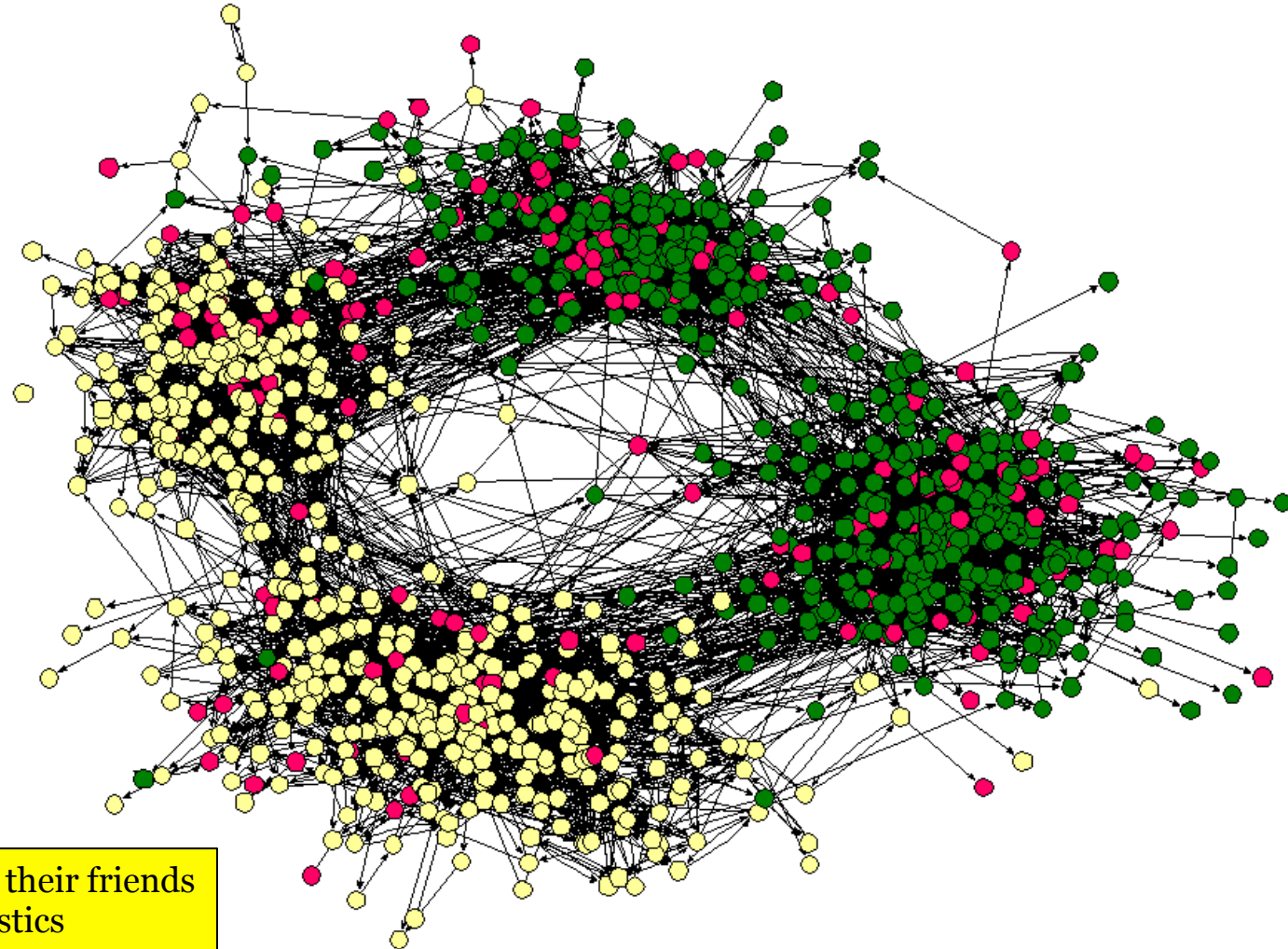
a. Race

b. School

Color the nodes based on race.

Homophily- Cnt.

- Social net among students in a middle and high school.



Two context features:

a. Race

b. School

Color the nodes based on race.

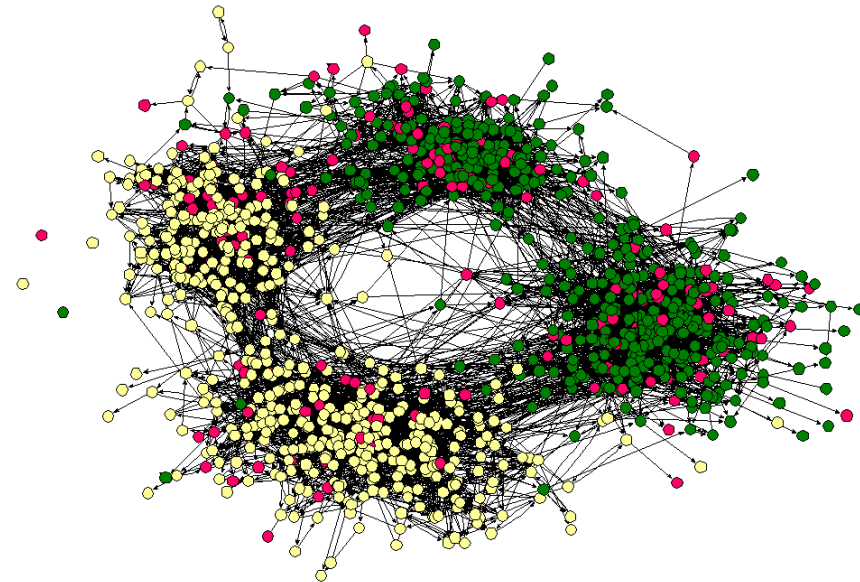
The network exhibits homophily with respect to:

Race (left to right), and School (top to bottom)!

Students tend to be similar to their friends in terms of specific characteristics

Homophily- Cnt.

- Which factors are more dominant for link formation?
 - **Hard to say!**
 - Most links arise from a combination of several factors
 - network intrinsic effects, and
 - contextual effects.



Homophily vs. Triadic Closure

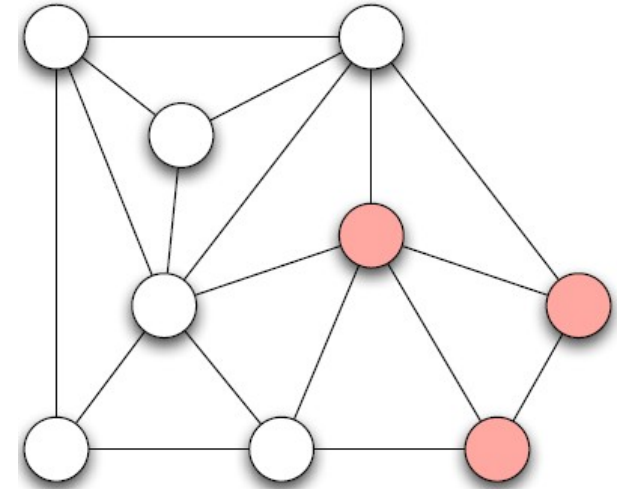
- Triadic closure
 - intrinsic factor:
 - A and B have a common friend C
 - A and B have increased opportunities to meet
- Homophily
 - contextual factor:
 - A and B are likely to be similar in a number of beyond network dimensions
- Both operate concurrently
 - Most links form due to a combination of several factors
 - Difficult to attribute any individual link to a single factor

Measuring Homophily

- Given a particular factor (like race, or age), how can we test if a network exhibits homophily according to this factor?

Measuring Homophily- Cnt.

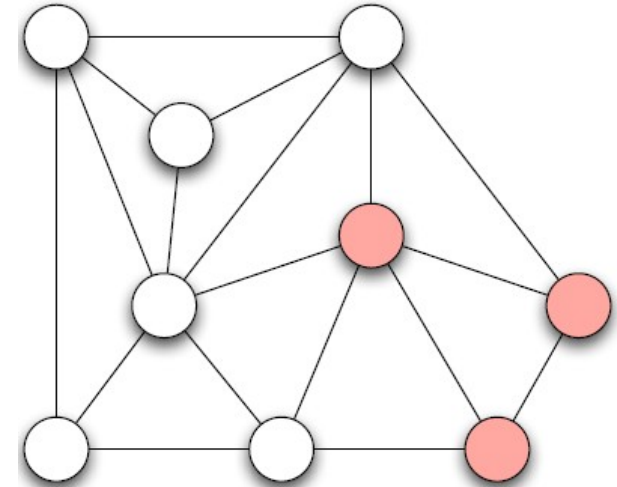
- What does it mean to: “test if this network exhibits homophily according to gender?”



Network of 3 girls and 6 boys!

Measuring Homophily- Cnt.

- Test if this network exhibits homophily according to gender?
- Extreme sense:
 - Edges btw boys
 - Edges btw girls
 - But no cross-gender edges

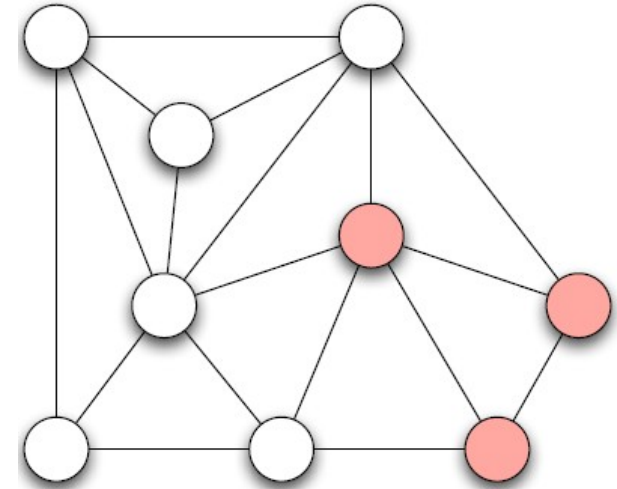


Network of 3 girls and 6 boys!

Boys tend to be friends with boys,
Girls tend to be friends with girls

Measuring Homophily- Cnt.

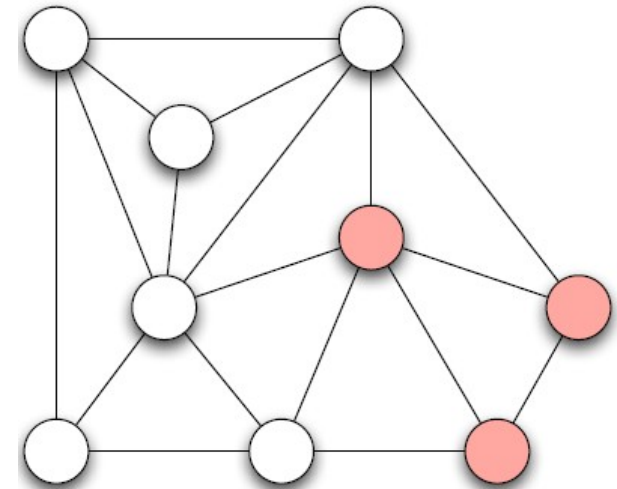
- What does it mean for a network *not* to exhibit homophily by gender?



Network of 3 girls and 6 boys!

Measuring Homophily- Cnt.

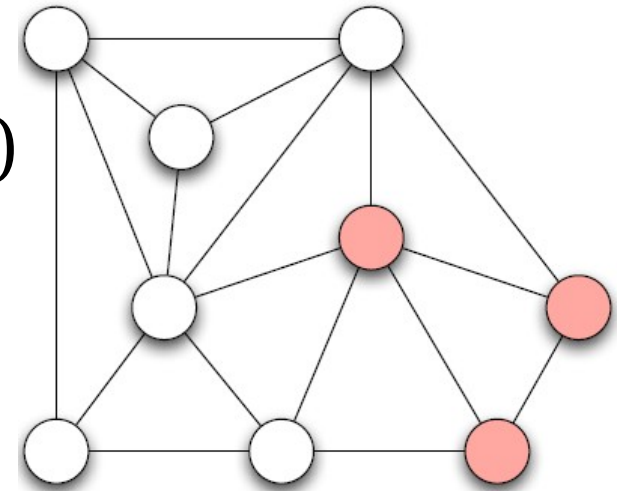
- What does it mean for a network *not* to exhibit homophily by gender?
 - The number of cross-gender edges is not very different from when we randomly assign each node a gender
 - according to the gender balance in the network



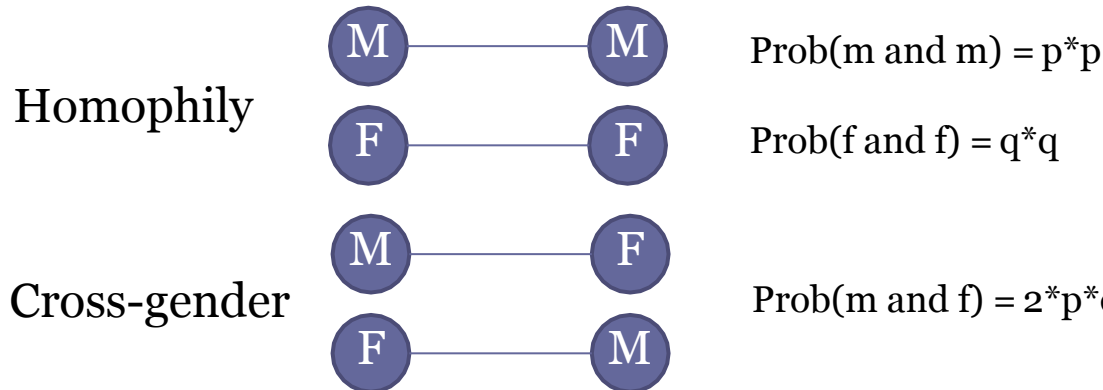
Network of 3 girls and 6 boys!

Measuring Homophily- Cnt.

- p : probability of males ($2/3$)
- $q=1-p$: probability of females ($1/3$)
- For a given edge:
 - if we independently assign each node M with prob p and F with prob q , then



Network of 3 girls and 6 boys!



$5/18 < 2pq = 4/9$ ✓

If the fraction of cross-gender edges is **significantly less than** $2pq$, then there is evidence for homophily!

The probability of cross-gender edge when each node is randomly assigned a gender (according to the gender balance in the original network)

Measuring Homophily- Cnt.

- Does this network exhibit homophily wrt to gender?

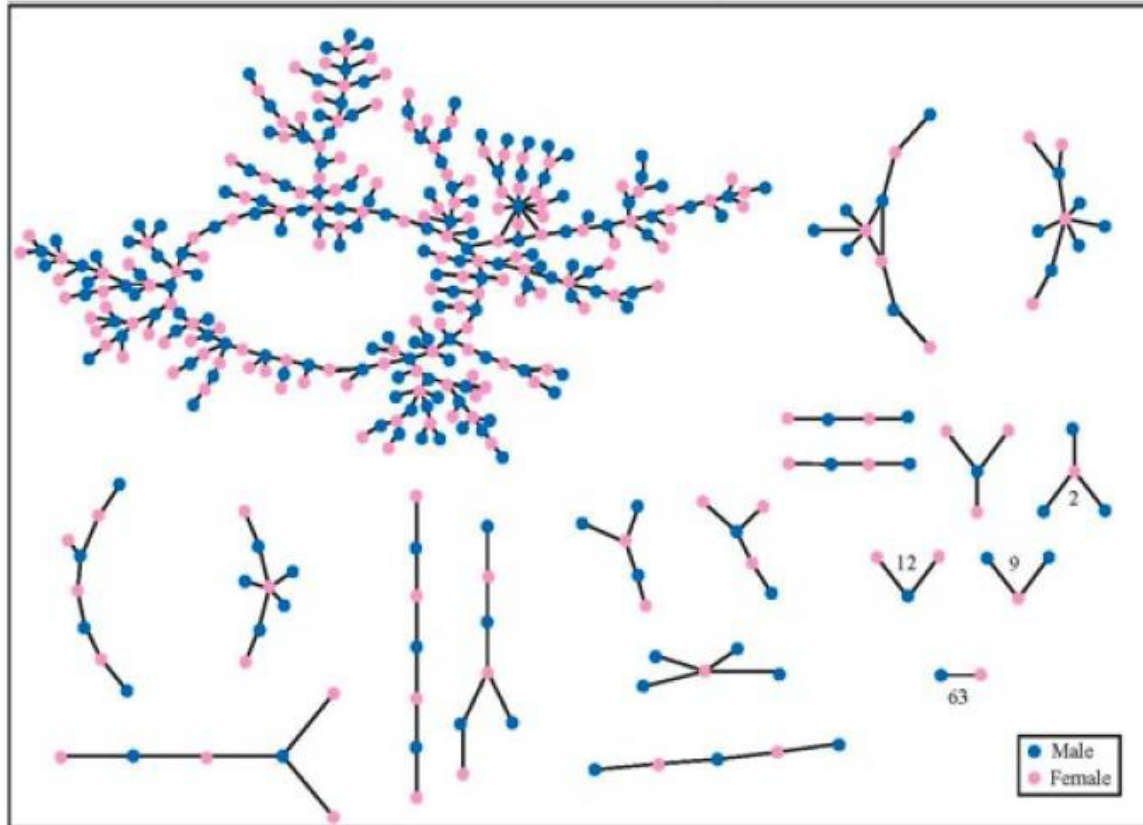


Figure 2.7: A network in which the nodes are students in a large American high school, and an edge joins two who had a romantic relationship at some point during the 18-month period in which the study was conducted [49].

Mechanisms Underlying Homophily

- Homophily has two mechanisms for link formation:
 - Selection
 - Selecting friends with similar characteristics
 - Individual characteristics drive the formation of links
 - Social Influence (socialization)
 - Modify behaviors to make them close to behaviors of friends
 - Existing links influence the individual characteristics of the nodes

Mechanisms Underlying Homophily- Cnt.

- Often times, both Selection and Social Influence apply and interact with each other
 - **Teenager behavior:**
 - Teenagers seek out social circles composed of people like them, and peer pressure causes them to conform to behavioral patterns within their social circles.
 - **Drug use:**
 - If drug use exhibits homophily in a friendship network,
 - people showing a greater likelihood to use drugs when their friends do,
 - We can target certain people and influences them to stop using drugs.

Mechanisms Underlying Homophily- Cnt.

- When Homophily is observed, is it more because of Selection or Social Influence?
 - **Selection: Have they selected people who were already like them?**
 - **Social Influence: Have people adapted their behaviors to become more like their friends?**
- More on this later!

Summary

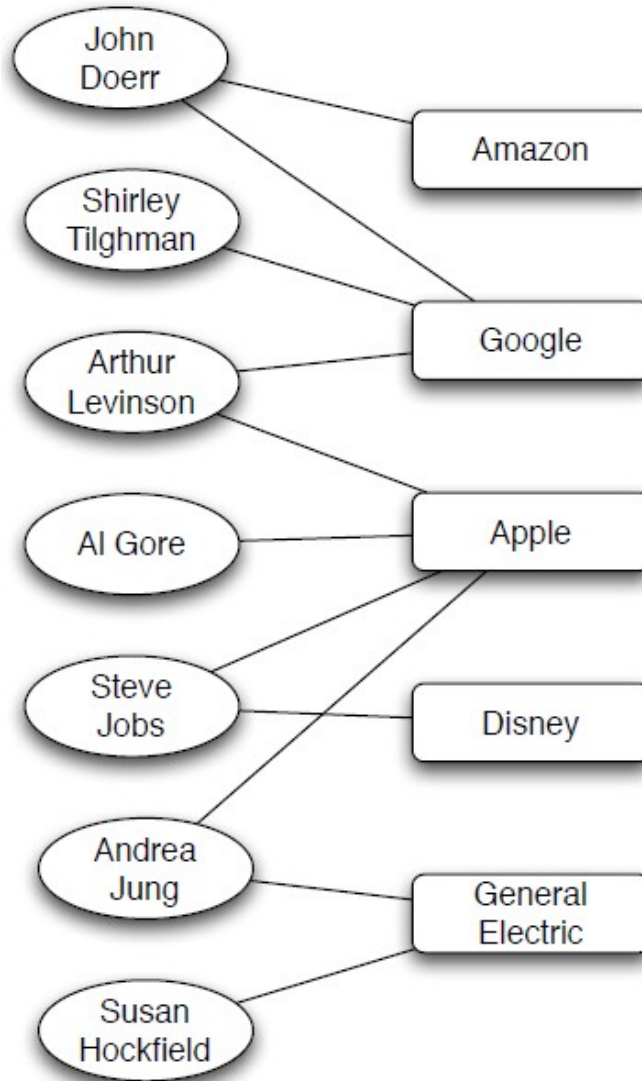
- Homophily links nodes with similar characteristics
- Measuring Homophily
 - compare with random network (generated according to the node characteristics in the original network)
- Selection and social influence determine the formation of links
- Surrounding contexts, forces that exist outside of networks

Affiliation Networks

- Network that contains both original nodes & surrounding contexts such as activities a person takes part in:
 - E.g. being part of a particular company / neighborhood, frequenting a particular place, hobby or interest, etc.
- Refer to activities as **foci: focal points** of social interaction

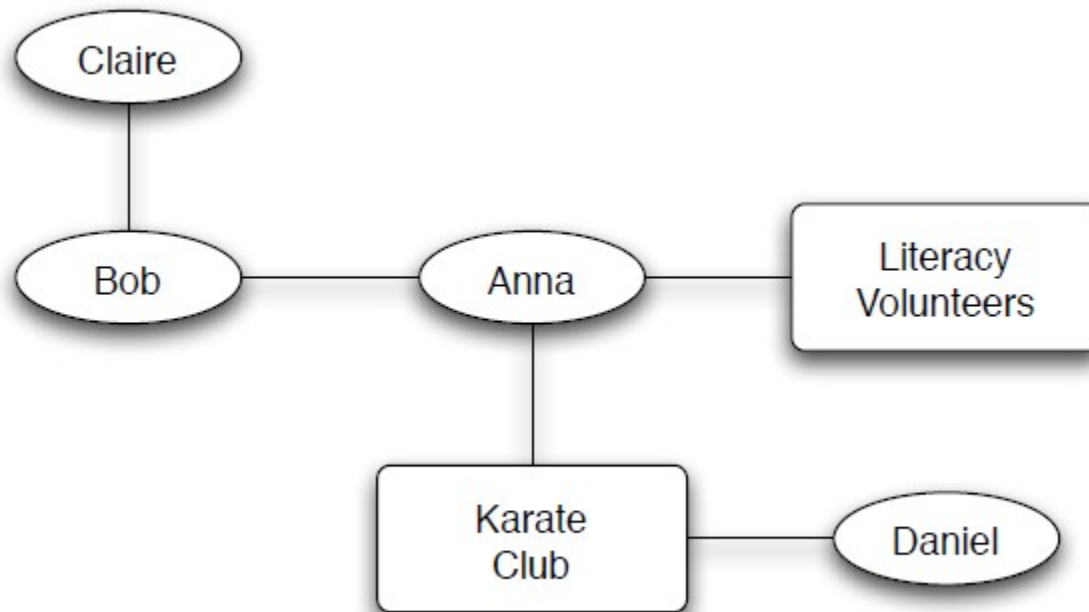
Affiliation Networks- Cnt.

- Bipartite Graph



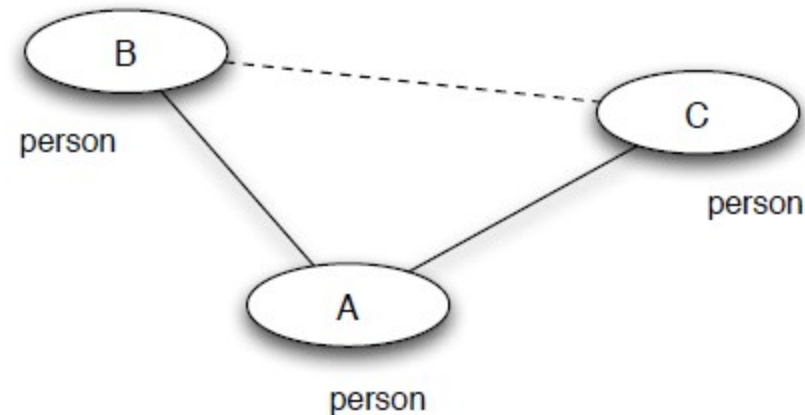
Social-Affiliation Network- Cnt.

- Social-affiliation network contains:
 - a social network of people, and
 - an affiliation network btw people and foci



Social-Affiliation Network- Cnt.

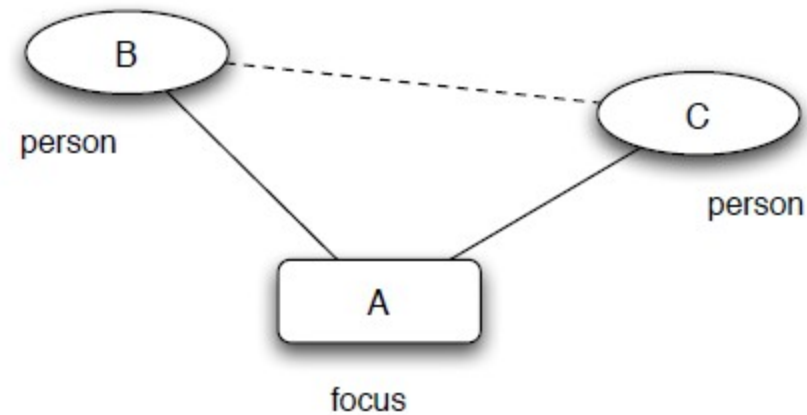
- Different mechanisms for link formation as types of closure processes!
- **Triadic Closure:**
A, B, and C represent people



(a) *Triadic closure*

Social-Affiliation Network- Cnt.

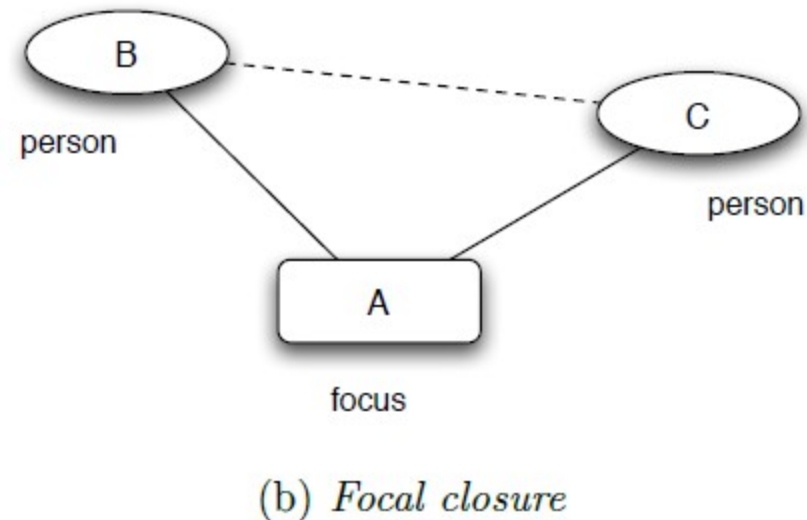
- Different mechanisms for link formation as types of closure processes!
- **Focal Closure:**
B and C people, A focus



(b) *Focal closure*

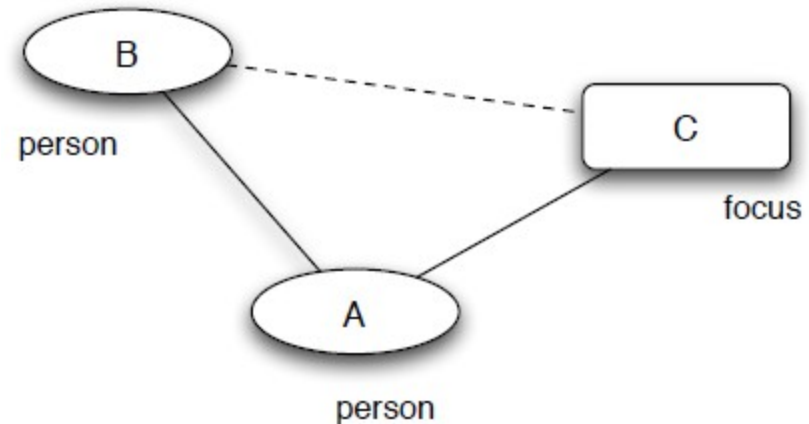
Social-Affiliation Network- Cnt.

- Different mechanisms for link formation as types of closure processes!
- **Focal Closure:**
B and C people, A focus
- **Selection:** B links to similar C (common focus)

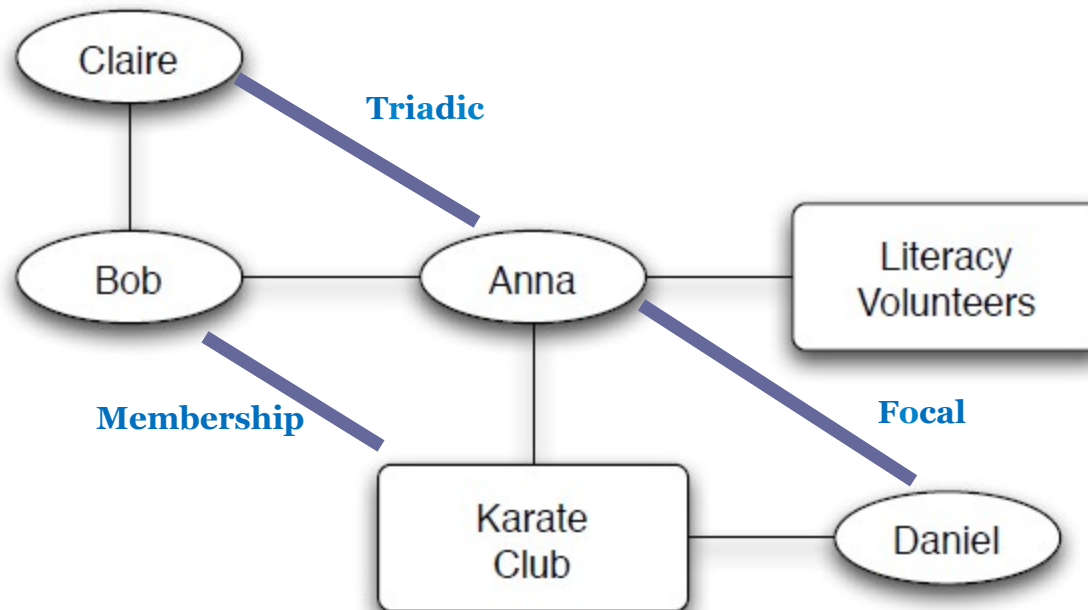


Social-Affiliation Network- Cnt.

- Different mechanisms for link formation as types of closure processes!
- **Membership Closure:**
A and B people, C focus
- **Social influence:** B links to C influenced by A



Social-Affiliation Network- Cnt.



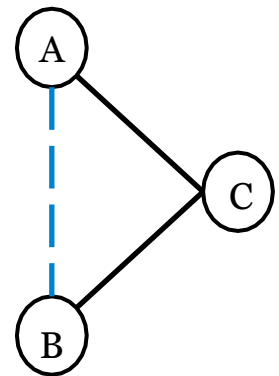
Tracking Link Formation

- Three mechanisms that lead to link formation
 - triadic closure
 - focal closure
 - membership closure

- How can we track link formation in large scale datasets based on e.g. triadic closure?

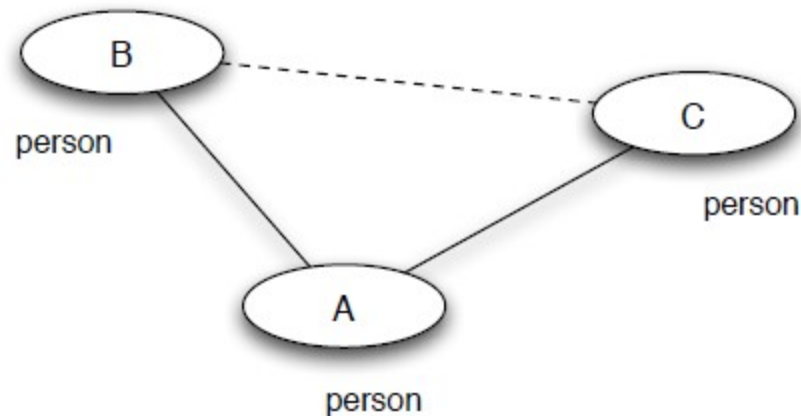
Tracking Link Formation- Cnt.

- Potential solution:
 - Compute link formation probability btw two nodes, if they already have a neighbor in common!
 - What if the nodes have k neighbors in common?



Tracking Triadic Closure

- The probability that 2 people form a link as a function of the number of **neighbors** they have in common.




(a) *Triadic closure*

- Q: How to properly design and conduct such experiments?

Tracking Triadic Closure- Cnt.

Algorithm

- 1) Take 2 snapshots of network at different times:
S(1), **S(2)**.
- 2) For each k , find all pairs of nodes in **S(1)** that are not directly connected but have k common friends.
- 3) Compute $T(k)$ as the fraction of these pairs connected in **S(2)**.


estimate for the probability that a link will form
btw 2 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- Cnt.

- E-mail communication among students
 - who-talks-to-whom network
- 22,000 students
- One-year period
- observations in each snapshot were one day apart (averaged over multiple snapshots)
 - Shows the average probability that 2 people form a link per day, as a function of the number of common friends they have

Tracking Triadic Closure- Cnt.

- **Baseline**

- Assume that each common friend that 2 people have, gives them an independent probability p of forming a link

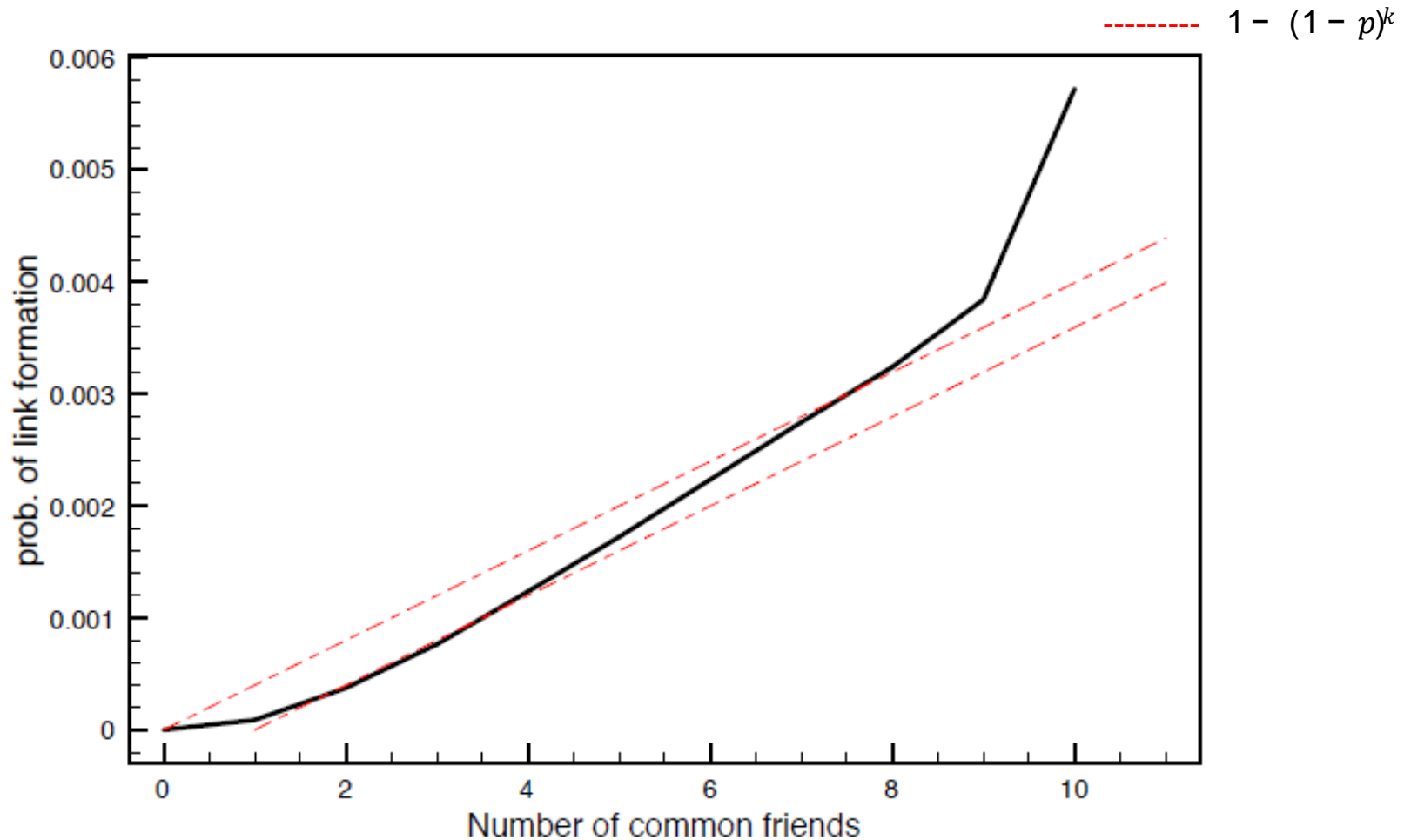
- 2 people have k friends in common \Rightarrow the probability they fail to form a link is:

- $(1 - p)^k$

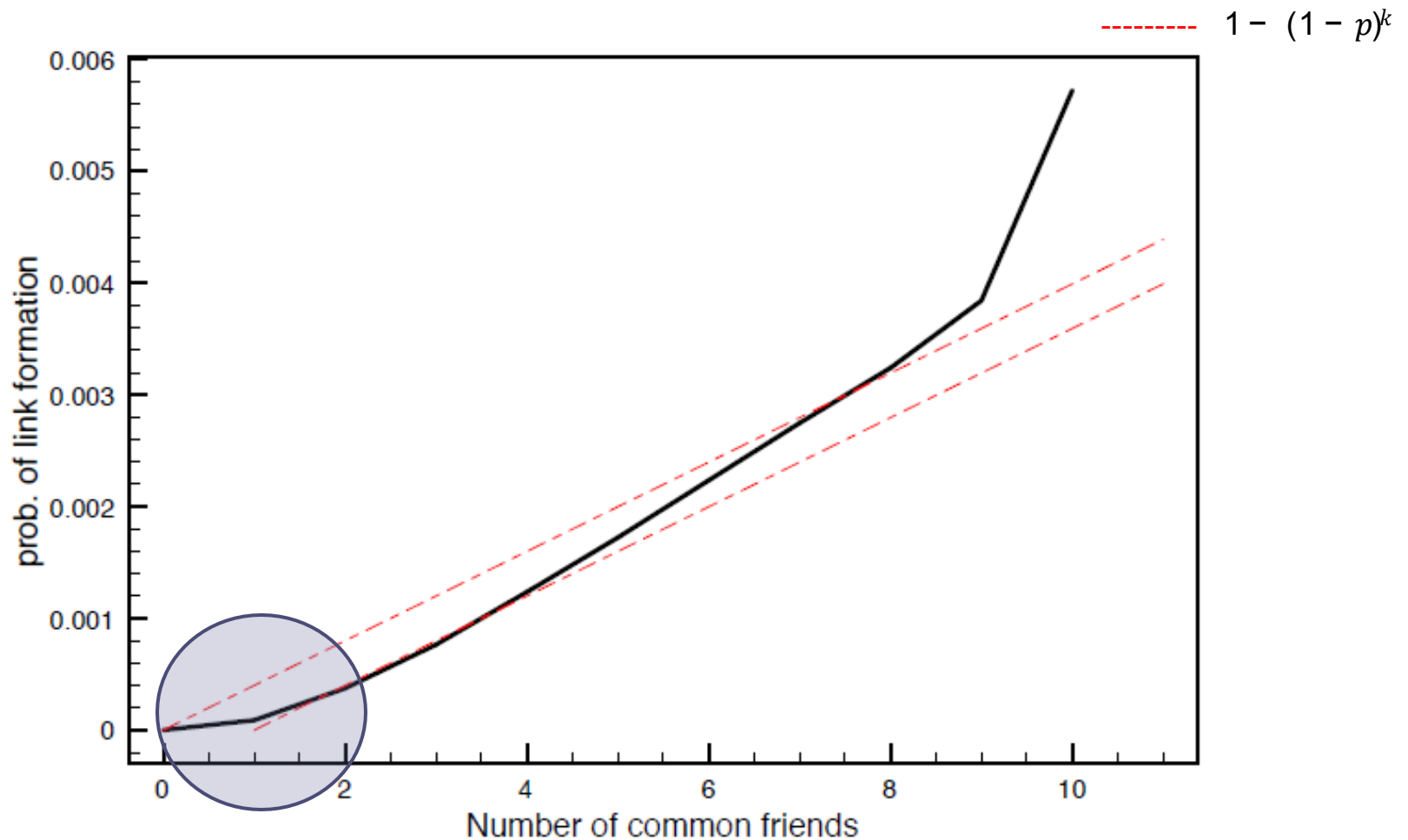
- probability that they form a link is

$$T_{baseline}(k) = 1 - (1 - p)^k$$

Tracking Triadic Closure- Cnt.

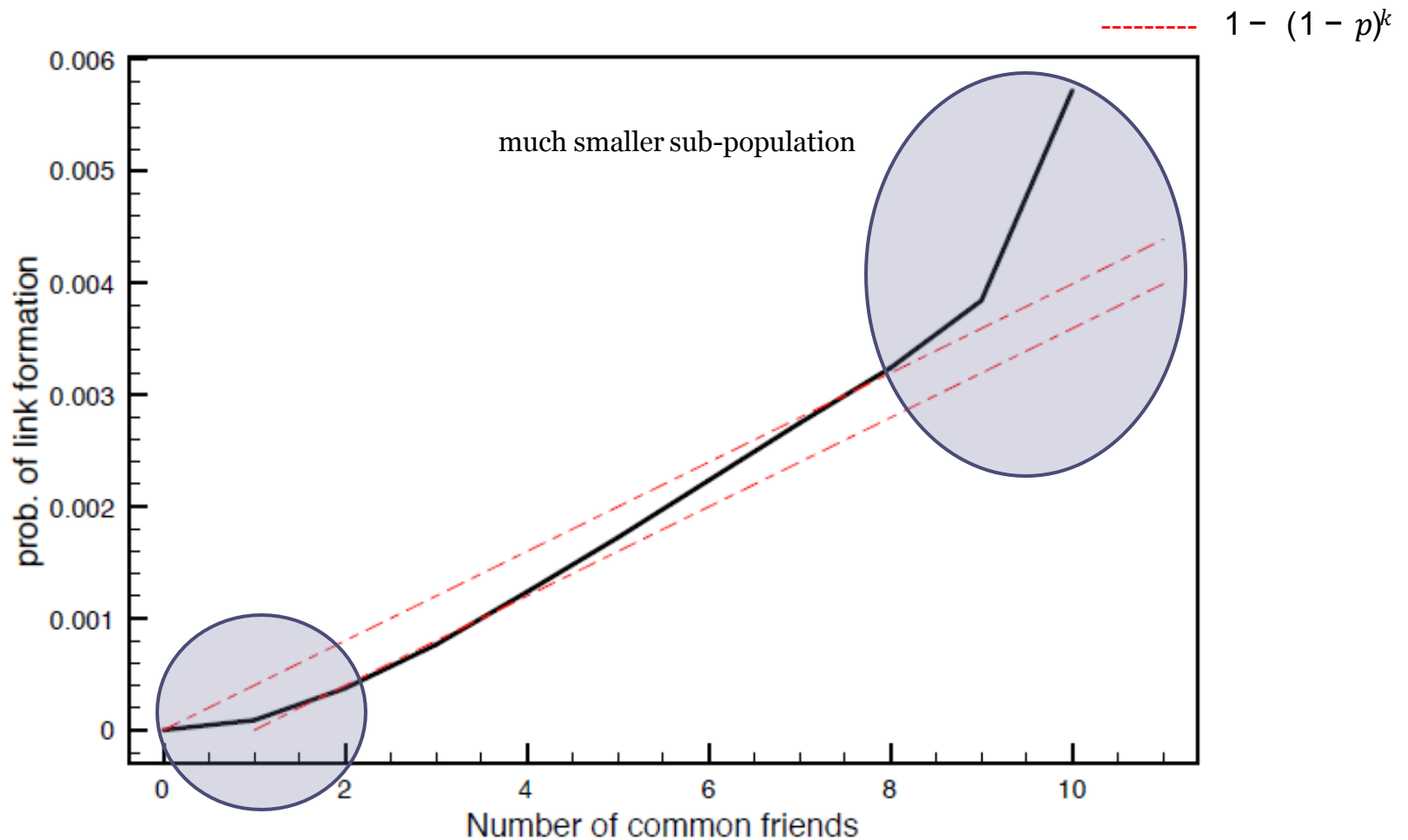


Tracking Triadic Closure- Cnt.



Having 2 common friends produces significantly more than twice the effect on link formation compared to having a single common friend!

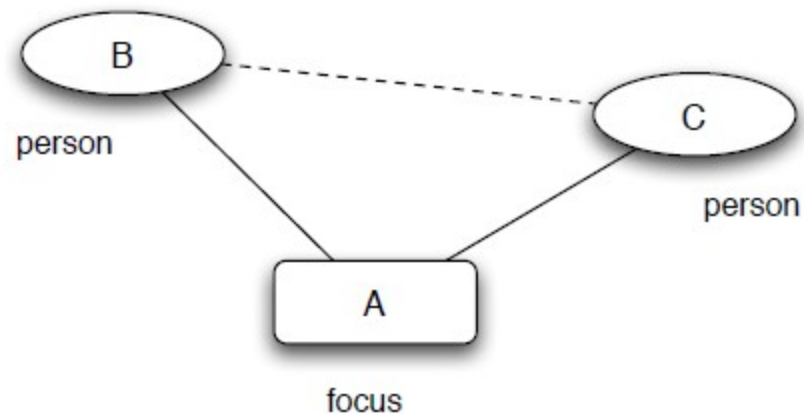
Tracking Triadic Closure- Cnt.



Having 2 common friends produces significantly more than twice the effect on link formation compared to having a single common friend!

Tracking Focal Closure

- The probability that 2 people form a link as a function of the number of **foci** they have in common.

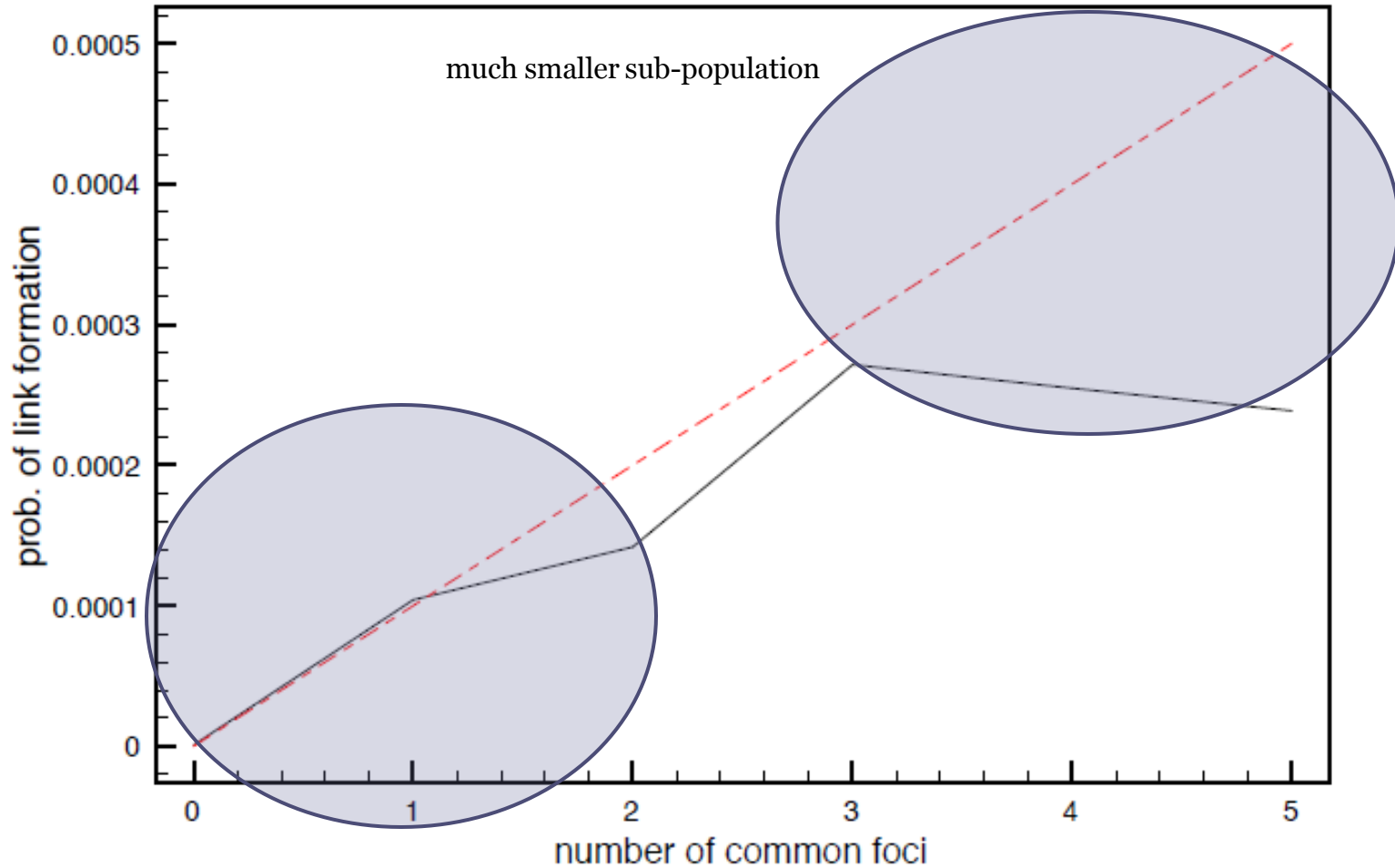


(b) *Focal closure*

Tracking Focal Closure- Cnt.

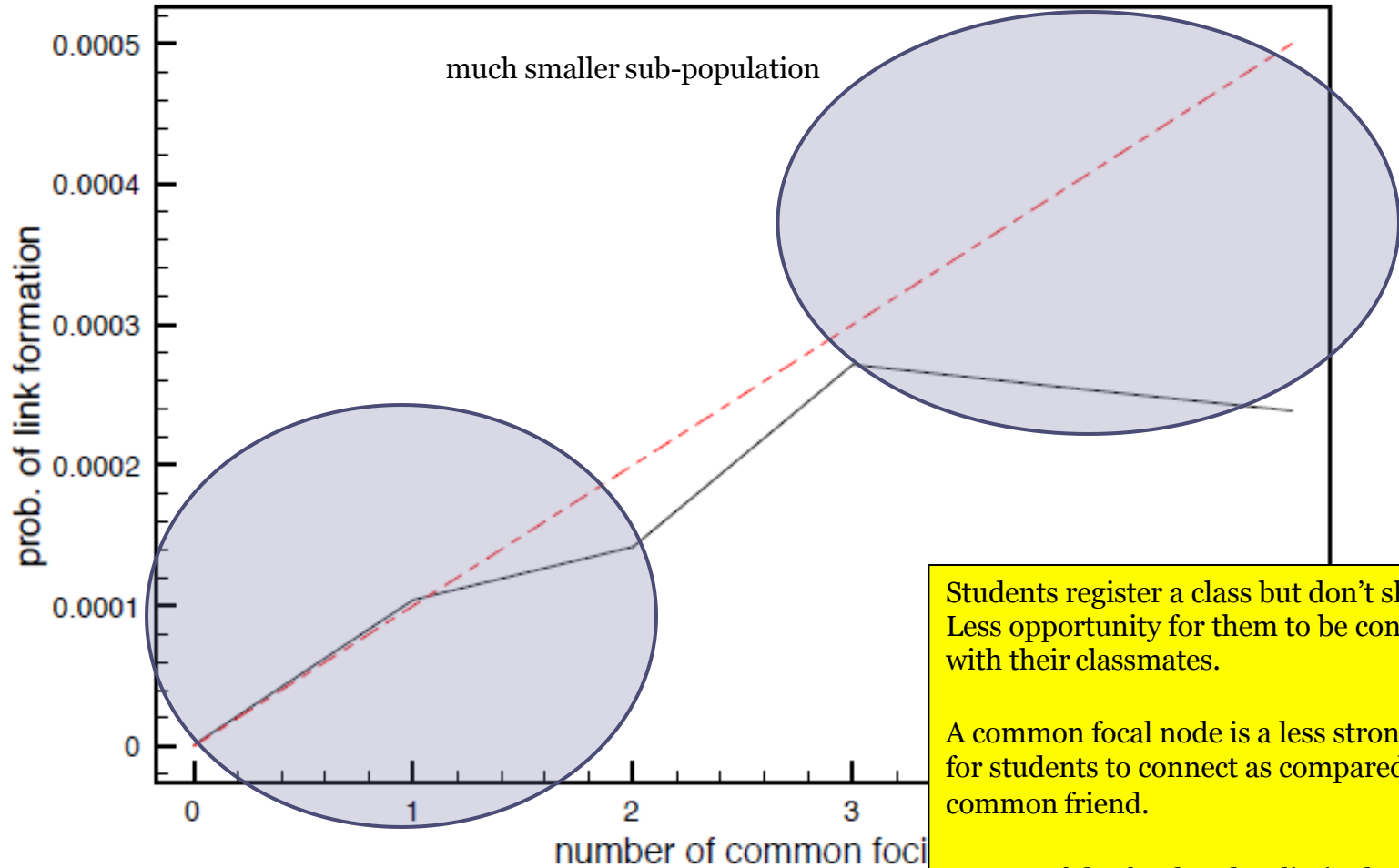
- Supplement university e-mail dataset with information about the class schedules!
 - each class is a focus, and
 - students shared a focus if they had taken a class together.

Tracking Focal Closure- Cnt.



$$P(k=2) < 2 * P(k=1)$$

Tracking Focal Closure- Cnt.



$$P(k=2) < 2 * P(k=1)$$

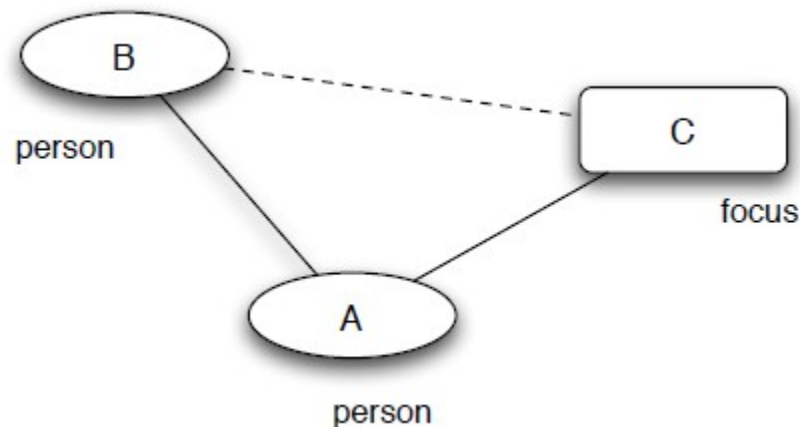
Students register a class but don't show up :
Less opportunity for them to be connected
with their classmates.

A common focal node is a less strong reason
for students to connect as compared to a
common friend.

Nature of the focal nodes: limited number
classes as compared to number of students.

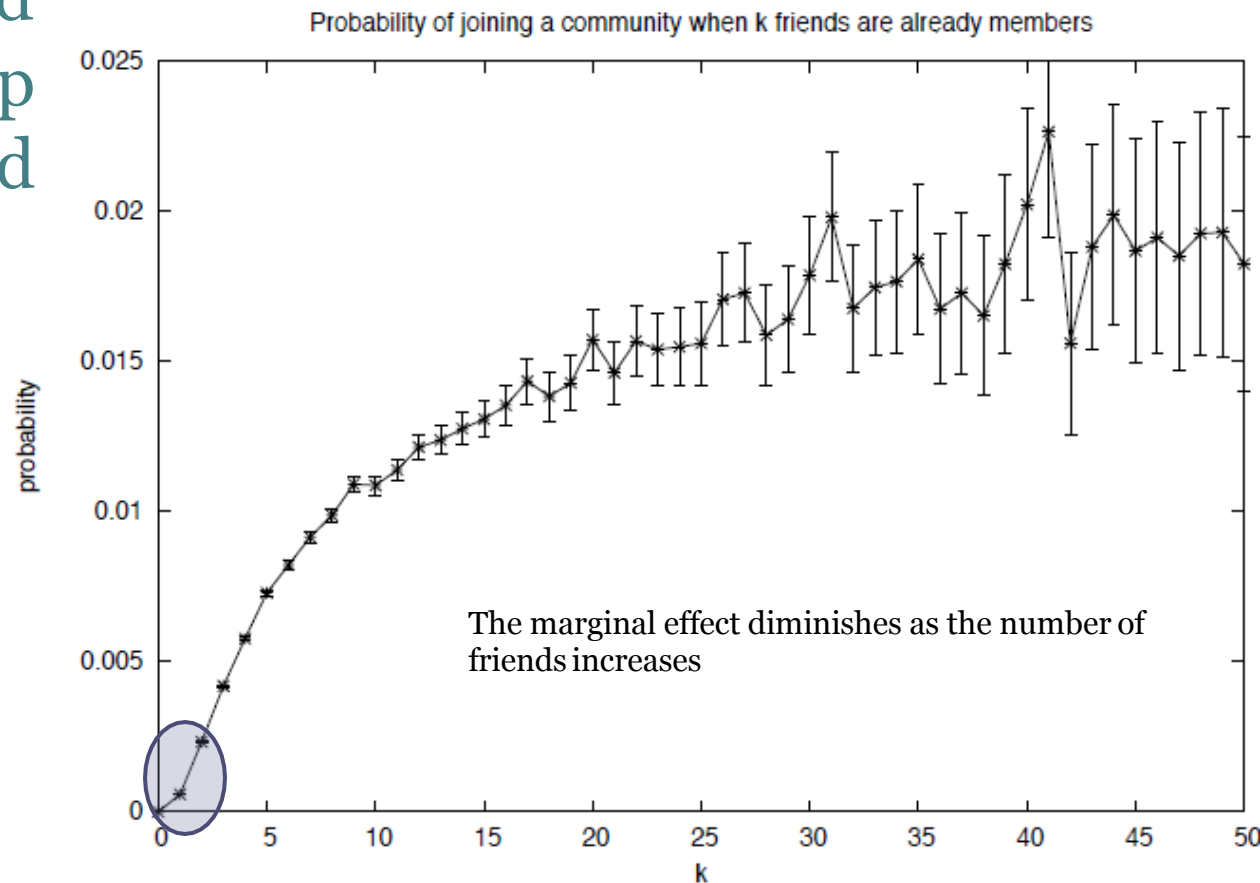
Tracking Membership Closure

- The probability that a person becomes involved with a particular **focus** as a function of the number of friends who are already involved in it?



Tracking Membership Closure- Cnt.

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

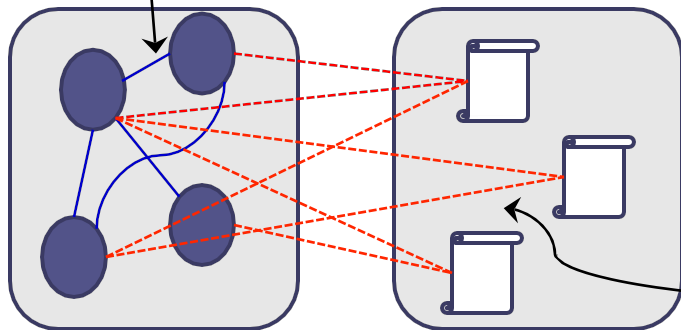


Tracking Membership Closure- Cnt.

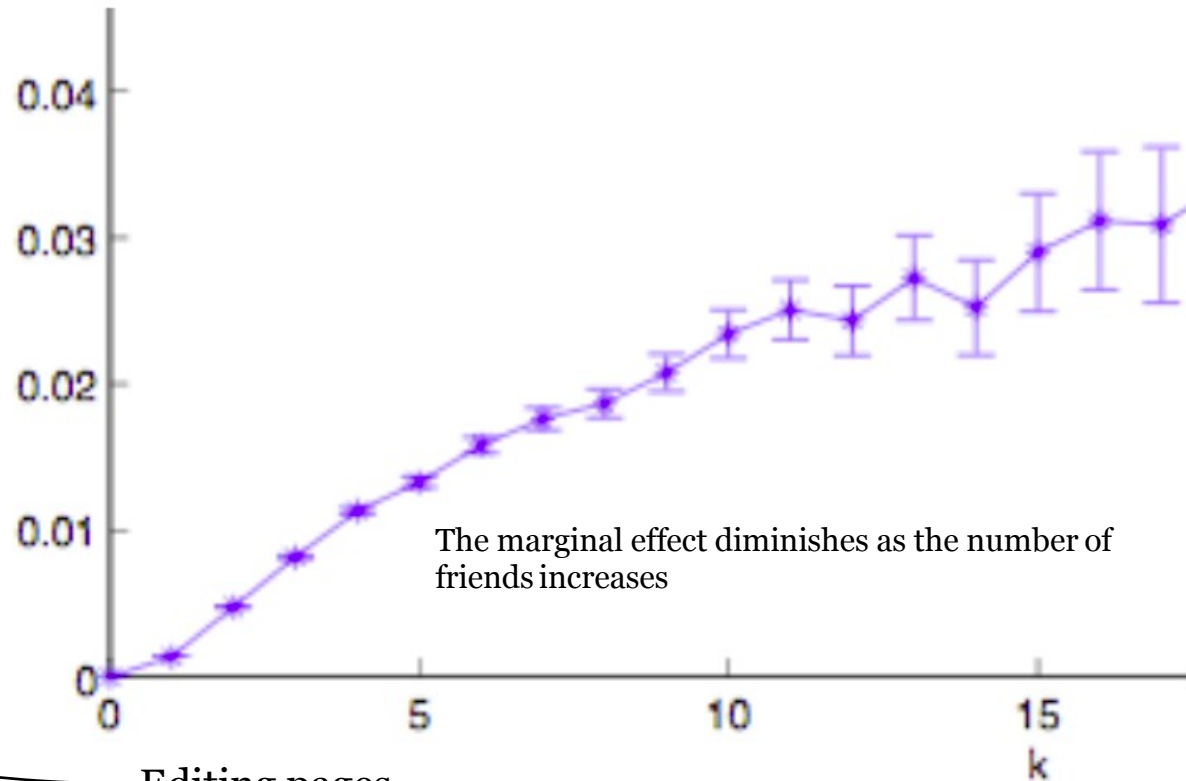
- **Wikipedia Editors**

- social network (link \rightarrow writing on user talk page)
- foci correspond to Wikipedia pages
 - Link \rightarrow editing a page!

Talk pages



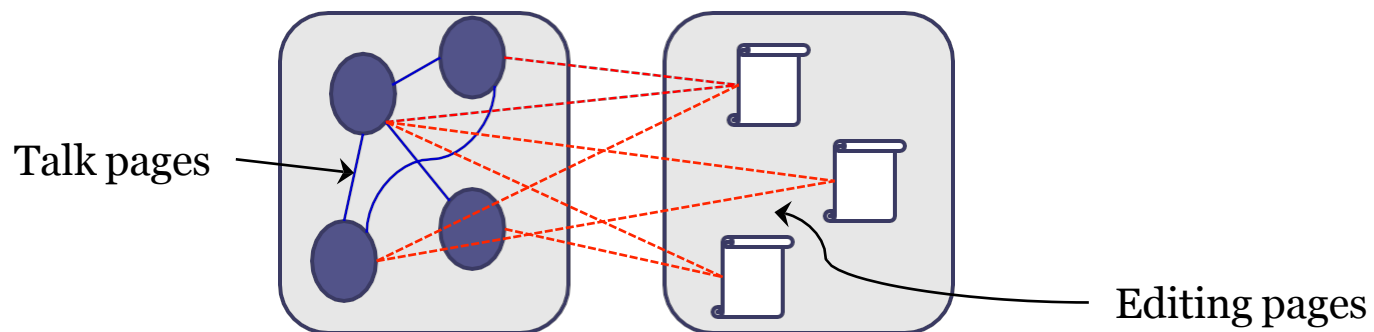
Editing pages



Selection and Social Influence

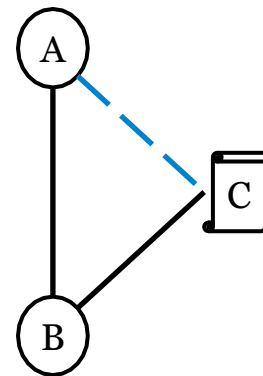
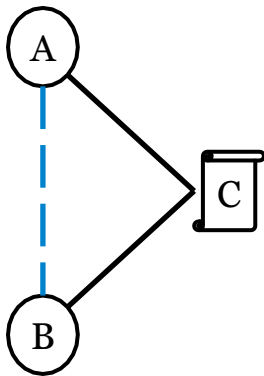
- Interplay btw Selection and Social Influence in producing homophily
 - Similarity btw two Wikipedia editors:

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



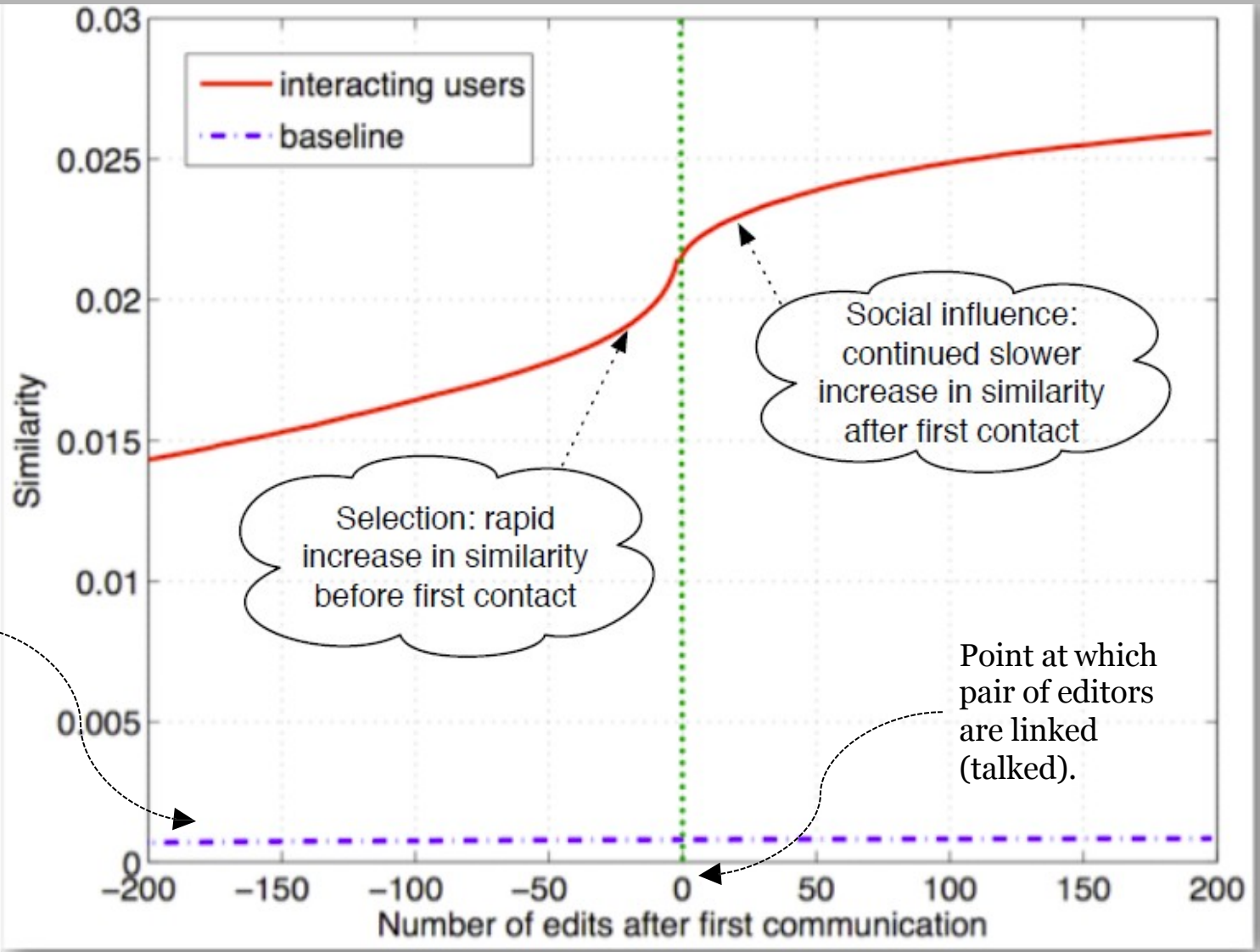
Selection and Social Influence- Cnt.

- Does **homophily (similarity)** arise because
 - editors are forming connections with those who have edited the same articles (**selection**), or
 - is it because editors are led to edit articles by those they talk to (**social influence**)?



Selection and Social Influence- Cnt.

average similarity relative to the time of first interaction, over all pairs of editors who have ever talked



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

Plot the average similarity over all pairs.

similarity of non-interacting pair of editors

homophily is clearly present: pairs of editors who have talked are significantly more similar than those who never talked.

Reading

- Ch.04 Networks in Their Surrounding Context [NCM]
- Empirical analysis of an evolving social network. Kossinets, G. and Watts, D.J. Science 2006.
- Group formation in large social networks: membership, growth, and evolution. Backstrom, L., et al. SIGKDD'06.
- Feedback effects between similarity and social influence in online communities. Crandall, D., et al. SIGKDD'08.