Social and Information Networks

University of Toronto CSC303
Winter/Spring 2021

Week 3: January 25-29 (2021)

Mon. Jan 24: Announcements and Corrections

- Assignment 1 has been released and is on the course website
  - We will cover the material needed for Q3d, Q4, Q5, and Q6 this week

Mon. Jan 24: Announcements and Corrections

- Assignment 1 has been released and is on the course website
  - We will cover the material needed for Q3d, Q4, Q5, and Q6 this week
- First post on the discussion board!

Mon. Jan 24: Announcements and Corrections

- Assignment 1 has been released and is on the course website
  - We will cover the material needed for Q3d, Q4, Q5, and Q6 this week
- First post on the discussion board!

- There you can find a clarification for Q1

The Rozenshtein et al study

As stated last week. Rozenshtein et al approach assumes a known set of communities (in addition to the unlabeled network) and hence it is not directly comparable to Sintos-Tsaparas study. Informally, they want to provide a good labeling while preserving communities – i.e. communities being strongly connected using strong ties.

They provide experimental results for 10 different data sets (where they can naturally define communities). Their goal is to provide a compromise between preventing STC violations (as in the goal of Sintos and Tsaparas) and only preserving strong connectivity within communities (which is the goal of Angluin et al.).

Rozenshtein et al objective and a greedy algorithm

The objective in Rozenshtein et al is to minimize the number of STC violations subject to the constraint that every user-specified community remains connected using only strong ties. This is an NP-hard problem. This is equivalent to maximizing the number of open triangles in the graph that satisfy STC under the community constraint. The maximization problem can be approximated to within a multiplicative factor of $k + 1$ by their greedy algorithm below, where k is the number of communities.

Their greedy algorithm works as follows:
****

Start with all edges labeled as strong.
Find an edge $e \in E$ that is causing the most STC violations (that is, whose removal would minimize the number of STC violations). If there are no violations then we're done. Otherwise, if that edge's removal would violate the community constraint then the edge stays strong and we never again consider this edge.
Otherwise the edge becomes weak and $E := E \setminus \{e\}$
****

Rozenshtein et al objective and a greedy algorithm

- More rigorous pseudo-code can be found below, where $vio : \mathcal{P}(E) \to \mathbb{N}$ is the number of open triangles in the original graph that *violate* the STC if the input edges are labelled as strong
- The code returns S, the edges that should be made strong

Algorithm 1 Greedy Rozenshtein Algorithm

```
$S \leftarrow E; A \leftarrow E;$
**while** $A \neq \emptyset$ and $vio(S) \neq 0$ **do**
    $e = \arg\min_{e \in A} vio(S \setminus \{e\});$
    **if** $e$ is part of an open triangle that violates STC and $S \setminus \{e\}$ satisfies strong connectivity constraints **then**
        $S \leftarrow S \setminus \{e\}$
    **end if**
    $A \leftarrow A \setminus \{e\}$
**end while**
**return** $S$
```

The Karate club figure in Rozenshtein et al

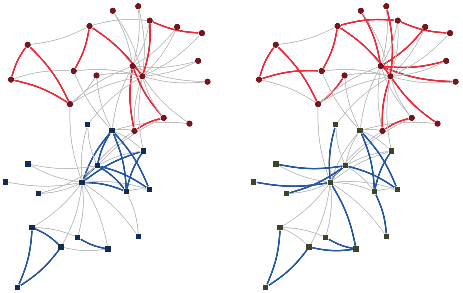


Figure 1: Strong edges in the Karate-club dataset inferred by the algorithm of Sintos and Tsaparas [27] (left) and our method (right) using two teams. The colors of the edges and the vertices depict the two teams.

Note: the vertices are colored according to the two known communities. Sintos and Tsaparas do not know about the communities. We expect that the Rozenshtein et al greedy algorithm would “usually” have more strong edges (to insure the community connectivity constraint).

Comparative statistics in Rozenshtein et al paper

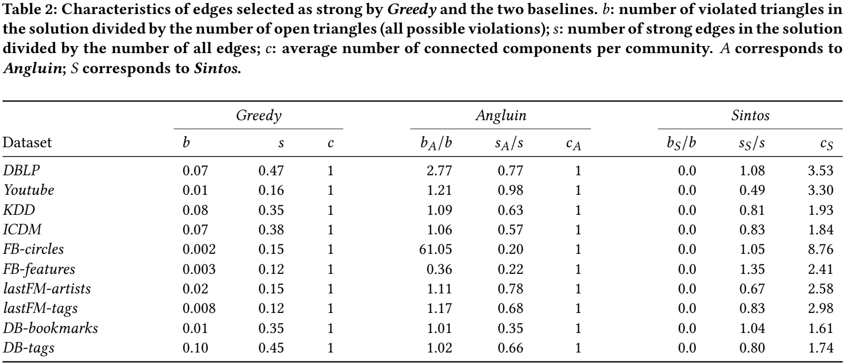


Table 2: Characteristics of edges selected as strong by *Greedy* and the two baselines. b: number of violated triangles in the solution divided by the number of open triangles (all possible violations); s: number of strong edges in the solution divided by the number of all edges; c: average number of connected components per community. A corresponds to *Angluin*; S corresponds to *Sintos*.

	Greedy			*Angluin*			*Sintos*		
Dataset	b	s	c	b_A/b	s_A/s	c_A	b_S/b	s_S/s	c_S
DBLP	0.07	0.47	1	2.77	0.77	1	0.0	1.08	3.53
Youtube	0.01	0.16	1	1.21	0.98	1	0.0	0.49	3.30
KDD	0.08	0.35	1	1.09	0.63	1	0.0	0.81	1.93
ICDM	0.07	0.38	1	1.06	0.57	1	0.0	0.83	1.84
FB-circles	0.002	0.15	1	61.05	0.20	1	0.0	1.05	8.76
FB-features	0.003	0.12	1	0.36	0.22	1	0.0	1.35	2.41
lastFM-artists	0.02	0.15	1	1.11	0.78	1	0.0	0.67	2.58
lastFM-tags	0.008	0.12	1	1.17	0.68	1	0.0	0.83	2.98
DB-bookmarks	0.01	0.35	1	1.01	0.35	1	0.0	1.04	1.61
DB-tags	0.10	0.45	1	1.02	0.66	1	0.0	0.80	1.74

- Greedy is the algorithm from the previous slide (minimize STC violations while strongly connecting communities).
- Angluin seeks to make all communities internally strongly connected using the minimal number of strong edges (STC is ignored).
- Sintos is the algorithm discussed last week (maximize strong edges while satisfying STC).

Understanding the table of results in Rozenshtein

- By design, Angluin et al. and Rozenstein et al. insure that the given communities remain connected by strong edges and hence $c = c_A = 1$ whereas c_S can be large (namely 8.76 for the FB-circles date set), indicating how disconnected the communities become wrt. strong edges.
- By design, Sintos and Tsaparas insures no STC violations and hence $b_S = 0$ whereas b is not 0 but is perhaps surprisingly small.
- The column that does seem surprising is the reporting of $\frac{s_S}{s}$ which is the ratio $\frac{\text{strong edges in Sinitos}}{\text{strong edges in Rozenstein}}$. As we said when looking at the Karate figure, we would expect that "usually" the Rozenshtein et al algorithm would produce more strong edges. But note that for some data sets, the ratio is great than 1.

Understanding the table of results in Rozenshtein

- By design, Angluin et al. and Rozenstein et al. insure that the given communities remain connected by strong edges and hence $c = c_A = 1$ whereas c_S can be large (namely 8.76 for the FB-circles date set), indicating how disconnected the communities become wrt. strong edges.
- By design, Sintos and Tsaparas insures no STC violations and hence $b_S = 0$ whereas b is not 0 but is perhaps surprisingly small.
- The column that does seem surprising is the reporting of $\frac{s_S}{s}$ which is the ratio $\frac{\text{strong edges in Sinitos}}{\text{strong edges in Rozenstein}}$. As we said when looking at the Karate figure, we would expect that "usually" the Rozenshtein et al algorithm would produce more strong edges. But note that for some data sets, the ratio is great than 1. How can this happen?

A comment about computational complexity and efficient algorithms

The studies by Sintos and Tsaparas, and that of Rozenshtein et al demonstrate some not so uncommon phenomena:

1. While two optimization problem may be equivalent from the viewpoint of optimality, they can be dramatically different from the viewpoint of approximation.
2. Often a simple greedy algorithm will provide a good approximation, sometime theoretically but more often “in practice”.

Comments on tightly knit communities

As we mentioned and as the EK text emphasizes (see section 3.6) , it is an interesting question as to how to define and efficiently find tightly knit communities.

Section 3.6 argues why cannot rely on the existence of a local bridge to help identify a community. Rather, a notion "betweenness" of an edge is defined which is based on the amount of traffic or flow through that edge. (Recall the Florentine marriages and centrality.) Edges of high betweenness are used to partition the graph into smaller components and eventually communities. They describe the Givan-Newman algorithm for identifying edges of high betweenness.

Other approaches to finding communities include finding dense subgraphs, subgraphs connected via strong edges (when the strength of edges is known to some extent), and subgraphs where vertices have high correlation coefficients.

Chapter 4: The context of network formation

- In this chapter, we study social networks within their context, considering factors outside of the nodes and edges of the network that impact how the network structure evolves.

- The chapter introduces a very important (and often controversial) issue, namely the relative roles of selection (similarity) vs influence in social relations.

- As we have already noted, Easley and Kleinberg have already indicated that there is a limit to what one can understand just in terms of the network structure.

Word of caution from Chapter 3 repeated

Easley and Kleinberg (end of Section 3.3):

> *Given the size and complexity of the (who call whom) network, we cannot simply look at the structure. . . Indirect measures must generally be used and, because one knows relatively little about the meaning or significance of any particular node or edge, it remains an ongoing research challenge to draw richer and more detailed conclusions. . .*

We should also add that we may know very little about the reasons for the formation (or disappearance) of an edge.

Word of caution from Chapter 3 repeated

Easley and Kleinberg (end of Section 3.3):

> *Given the size and complexity of the (who call whom) network, we cannot simply look at the structure. . . Indirect measures must generally be used and, because one knows relatively little about the meaning or significance of any particular node or edge, it remains an ongoing research challenge to draw richer and more detailed conclusions. . .*

We should also add that we may know very little about the reasons for the formation (or disappearance) of an edge.

Unknown:

> *In theory there is no difference between theory and practice. In practice there is.*

Homophily

- Why triadic closure? In Chapter 3: some network "intrinsic" reasons (opportunity, trust, incentive) for forming a friendship and now we consider "contextual" reasons

Homophily

- Why triadic closure? In Chapter 3: some network "intrinsic" reasons (opportunity, trust, incentive) for forming a friendship and now we consider "contextual" reasons
- Homophily: we tend to be similar to our friends.

Homophily

- Why triadic closure? In Chapter 3: some network "intrinsic" reasons (opportunity, trust, incentive) for forming a friendship and now we consider "contextual" reasons

- Homophily: we tend to be similar to our friends.

- This observation is captured in various writings and proverbs perhaps most notably by "Birds of a feather flock together" suggesting that friendships (and membership in groups) are selectively formed due to similar interests.

- In contrast we also have "opposites attract" but the quote might better be "opposites attract but the like-minded last".

- **Note:** But to what extent do we adopt similar interests based on friendship rather than conversely?

Characteristic factors

- Factors which help determine our friendships and relations can be immutable or more transient.

- Some (typically) immutable factors: ethnicity, birth date, gender; religion, height. What other such (mainly permanent) factors exist?

Characteristic factors

- Factors which help determine our friendships and relations can be immutable or more transient.
- Some (typically) immutable factors: ethnicity, birth date, gender; religion, height. What other such (mainly permanent) factors exist?
- Some more mutable (often related) factors: membership in clubs or courses, educational level, recreational interests, professional interests, income level, residential neighbourhood

Characteristic factors

- Factors which help determine our friendships and relations can be immutable or more transient.

- Some (typically) immutable factors: ethnicity, birth date, gender; religion, height. What other such (mainly permanent) factors exist?

- Some more mutable (often related) factors: membership in clubs or courses, educational level, recreational interests, professional interests, income level, residential neighbourhood

- Of course, immutable factors can and do influence mutable factors.

The Schelling Model

A dramatic example of homophily (i.e., we tend to live "adjacent" to people similar to ourselves) can be found at the end of Chapter 4. The chapter ends with a discussion of Schelling's segregation model that provides an explanation as to how racial neighbourhood segregation can evolve when driven by individuals wanting to be near "people similar to themselves". Schelling's model and his simulations led him to a fundamental observation:

Segregation can and will happen even if there is no explicit individual desire to avoid (say) people of a different race. All that is needed is some desire to be near enough similar people.

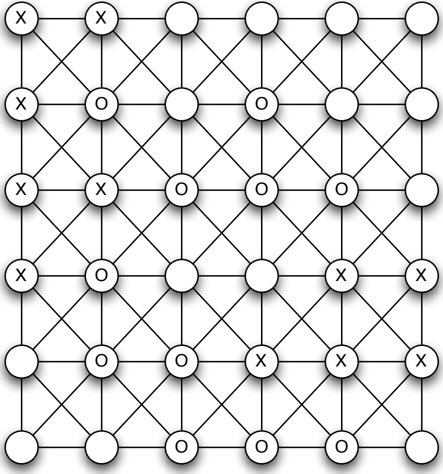
This observation isn't restricted to racial segregation.

In addition to the importance of this fundamental observation, the model provides an interesting study of network dynamics, homophily driven by selection, and how local decisions lead to global structure in a network.

Importantly, Schelling's model does not preclude the presence of other economic and political factors, nor does it preclude explicit racism!

The Schelling model

The model itself is quite simple but still hard to analyze analytically. In this model, we view two classes of individuals (X and O) living in a grid. More specifically, individuals occupy some subset of the nodes as depicted in figure 4.15 of the text.



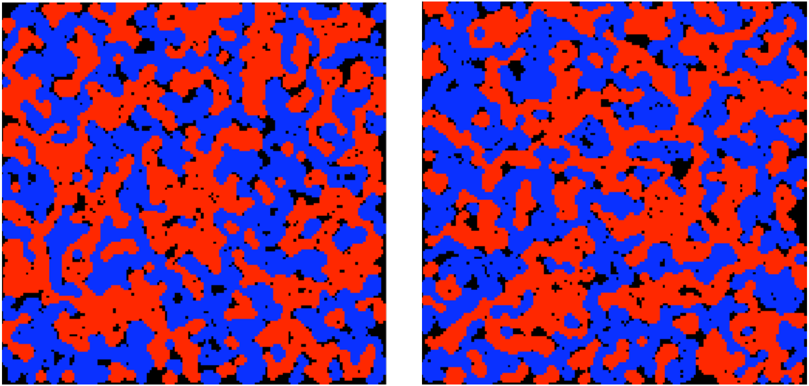
The dynamics of the Schelling model

Schelling then hypothesizes that every individual wants to have at least some threshold t of their immediate neighbours be of the same class. When a individual's threshold is not met, they move. There are different versions of the model depending on the order in which individuals move and where they randomly move to in order to satisfy the desire for similarity. The claim is that the results do not change qualitatively.

To observe the dynamics, simulations of the network are conducted for different threshold values. What is very apparent is the segregated structure of the network as it evolves.

The specific gird is a 150 by 150 grid (i.e., 12,500 cells, with 10,000 cells occupied) with both groups equally represented. The following figures show the results for thresholds $t = 3$ (i.e. an individual desires less than a majority of his/her neighbours to be similar) and $t = 4$.

Simulations for $t = 3$

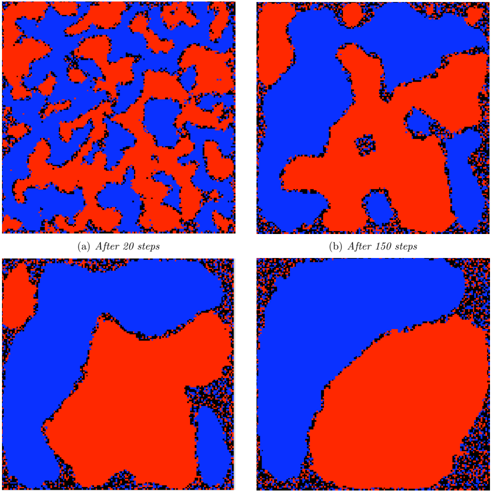


(a) *A simulation with threshold 3.*

(b) *Another simulation with threshold 3.*

Figure 4.17: Two runs of a simulation of the Schelling model with a threshold t of 3, on a 150-by-150 grid with $10,000$ agents of each type. Each cell of the grid is colored red if it is occupied by an agent of the first type, blue if it is occupied by an agent of the second type, and black if it is empty (not occupied by any agent).

Simulation for $t = 4$



(a) *After 20 steps*

(b) *After 150 steps*

(c) *After 350 steps*

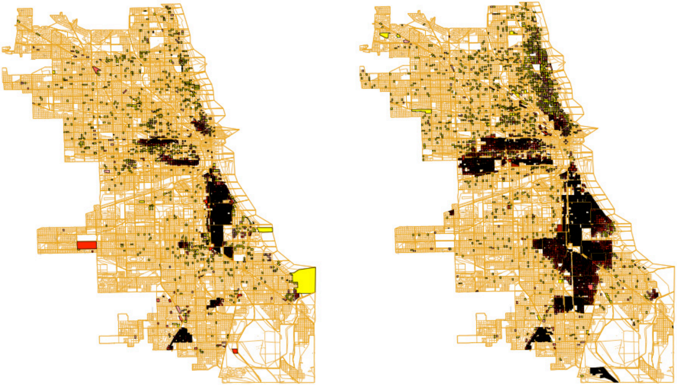
(d) *After 800 steps*

Figure 4.19: Four intermediate points in a simulation of the Schelling model with a threshold t of 4, on a 150-by-150 grid with $10,000$ agents of each type. As the rounds of movement progress, large homogeneous regions on the grid grow at the expense of smaller, narrower regions.

Some concluding comments on the Schelling study

- The model is not constructed so as to build in segregation. More specifically, the model allows for stable configurations that are well integrated.
- However, given a random starting configuration, the simulations show that people will gravitate to a segregated structure.
- There is a compounding effect of the model dynamics. Namely, when one person leaves, it can result in other neighbours falling below their threshold and hence a new desire to leave the current location has been created.
- The word segregation is a term with a very negative connotation due to the use of the term with respect to racial (e.g., Jim Crow legislation in the US) and religious segregation (e.g., ghettos in Europe) which was forced by governments.
- What if we used the word ”clustered” instead of segregated? Do we think that neighbourhoods that are concentrated along say ethnic lines is a bad thing? Be careful to distinguish between the model's assumptions and reality.

The reality of neighbourhood segregation in Chicago (1970s)



(a) *Chicago, 1940*

(b) *Chicago, 1960*

And how integrated or segregated is Toronto?

It seems much easier to talk about Chicago (as we all know about racial segregation in the US) but perhaps more difficult to talk about Canada and Toronto.

At some level (i.e., Metro Toronto), Toronto may be the most ethnically diverse city as is claimed. But at a more detailed level, many neighbourhoods are far from being "integrated". I am posting a September 2018 newspaper article and a February 2019 talk by David Hulchanski (UT Faculty of Social Work) that describes the changes in income levels and neighbourhoods in Toronto. The title of the article more or less summarizes his major conclusion: "Toronto is segregated by race and income. And the numbers are ugly".

I call attention to this (and other similar studies) to indicate that while homophily is a factor (especially with regard to ethnicity), there are clearly many other factors that are prevalent and arguably dominant.

The influence vs selection issue

- Thusfar, we've seen that immutable factors can and do influence mutable factors.

The influence vs selection issue

- Thusfar, we've seen that immutable factors can and do influence mutable factors.
- Furthermore, one's friendships can and do influence mutable factors such as say recreational interests.

The influence vs selection issue

- Thusfar, we've seen that immutable factors can and do influence mutable factors.
- Furthermore, one's friendships can and do influence mutable factors such as say recreational interests.
- So the selection vs influence issue can be seen as the relative extent to which our friendships are formed selectively due to similarity vs friendships influencing our interests and other traits.

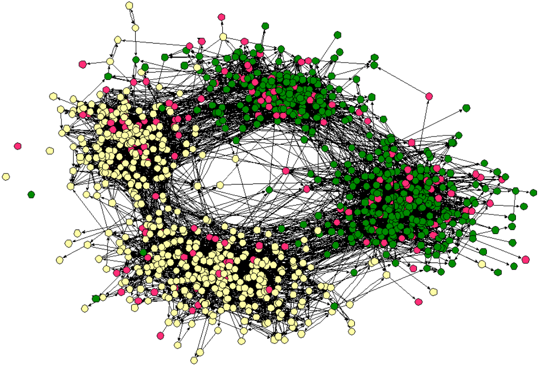
The influence vs selection issue

- Thusfar, we've seen that immutable factors can and do influence mutable factors.
- Furthermore, one's friendships can and do influence mutable factors such as say recreational interests.
- So the selection vs influence issue can be seen as the relative extent to which our friendships are formed selectively due to similarity vs friendships influencing our interests and other traits.
- Homophily (which we will use just to note the correlation between friendships and similarity) can be more easily attributed (directly or indirectly) to similarity leading to friendships when similarity factors are immutable or not easily changeable. The issue becomes much less clear and sometimes quite controversial when the similarity factors are mutable.

The influence vs selection issue

- Further complicating matters, the “environment” of various (perhaps unobserved) external events or hidden influences can also impact one’s friendships and/or interests and affiliations.
- For example, Alice and Bob are not friends nor have any interest in political issues. Then a popular entertainer is performing in a rally for a political candidate. Alice and Bob meet at the event and become friends as well as becoming more politically involved.

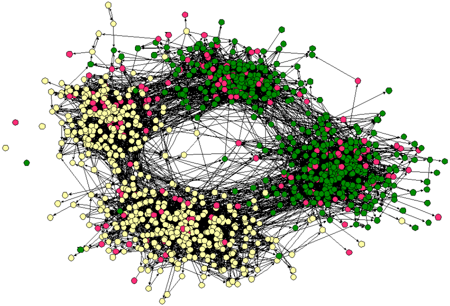
Graphic visualization of homophily



[Fig. 4.1, textbook]

- Homophily can divide a social network into densely-connected, homogeneous parts that are weakly connected to each other.
- In this social network from a town's middle school and high school, two divisions are apparent: one based on race (students of different races drawn as differently-colored circles), and the other based on friendships in the middle and high schools.

Comments on figure 4.1



[Fig. 4.1, textbook]

- Such a visualization is not at a scale that one can see most of the individual relations. The visualization clearly shows homophily based on race and the junior/senior high split (both immutable factors).
- We can measure the extent of homophily (as we will next see) but observing any such phenomena (even for immutable factors) is just the starting point in truly understanding the phenomena.
- The figure does show some detailed information; i.e. individuals without any friends (isolated nodes) or with few friends (low degree).

Measuring homophily

- As mentioned before, when networks are large (and/or when homophily is less dramatic) it is difficult if not impossible to visualize various aspects of a network and so one needs a measure of homophily (whatever the cause or the consequence of the network).

- Suppose we wish to study the likelihood of friendships according to some factor (with say two values) such as gender. (Recall Moreno's sociograms regarding seating preferences in elementary school.)

- Think Big!: Lets think in terms of large social networks where the presence or absence of a given individual will not have any noticeable impact on the probability of any phenomena.

Thought experiment

- What would it mean to say that a social network does or does not exhibit homophily according to some factor such as gender?

- Consider a given network where the fraction (i.e. probability) of males is p and the fraction of females is q.
  - Consider a given edge (u, v) in the network.
  - If gender has no correlation with relations, then the probability that the genders of u and v are different is $2pq$. Why?

Thought experiment

- What would it mean to say that a social network does or does not exhibit homophily according to some factor such as gender?

- Consider a given network where the fraction (i.e. probability) of males is p and the fraction of females is q.
  - Consider a given edge (u, v) in the network.
  - If gender has no correlation with relations, then the probability that the genders of u and v are different is $2pq$. Why?

- This leads to a homophily test: If the actual fraction of cross-gender edges is "significantly less than" $2pq$ then there is evidence for homophily.

Thought experiment

- What would it mean to say that a social network does or does not exhibit homophily according to some factor such as gender?

- Consider a given network where the fraction (i.e. probability) of males is p and the fraction of females is q.
  - Consider a given edge (u, v) in the network.
  - If gender has no correlation with relations, then the probability that the genders of u and v are different is $2pq$. Why?

- This leads to a homophily test: If the actual fraction of cross-gender edges is "significantly less than" $2pq$ then there is evidence for homophily.

  What would this say about same gender (male-male) or (female-female) edges?

Thought experiment

- What would it mean to say that a social network does or does not exhibit homophily according to some factor such as gender?

- Consider a given network where the fraction (i.e. probability) of males is p and the fraction of females is q.
  - Consider a given edge (u, v) in the network.
  - If gender has no correlation with relations, then the probability that the genders of u and v are different is $2pq$. Why?

- This leads to a homophily test: If the actual fraction of cross-gender edges is "significantly less than" $2pq$ then there is evidence for homophily.
  What would this say about same gender (male-male) or (female-female) edges?

- Clearly the meaning of an edge is an essential aspect of any study; e.g. consider the difference between an edge representing collaboration in a course project vs an edge meaning a romantic relationship.

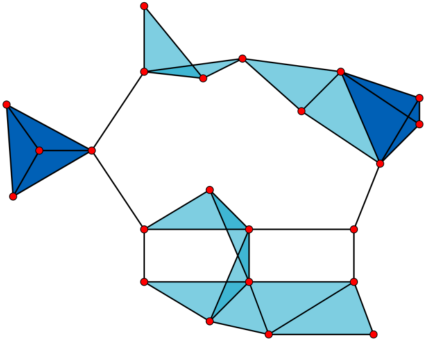
Fri. Jan 29: Announcements and Corrections

- You still have the opportunity to be a volunteer note taker! Please see the announcement on the course website for details

Fri. Jan 29: Announcements and Corrections

- You still have the opportunity to be a volunteer note taker! Please see the announcement on the course website for details

- Participation Quiz #2
  - Typo in Q1, should be referring to A and not B. It's now been fixed
  - Definition of a clique: For an undirected graph $G = (V, E)$, $V' \subseteq V$ is a clique if $\forall a, b \in V' : a \neq b \implies (a, b) \in E$



Fri. Jan 29: Announcements and Corrections

- Please confirm that you can log-in to MarkUs and can see A1
- A1 can be submitted as any *legible* PDF. I strongly recommend LaTeX, and the .tex for A1 is available on Quercus under the Files tab.

Fri. Jan 29: Announcements and Corrections

- Please confirm that you can log-in to MarkUs and can see A1
- A1 can be submitted as any *legible* PDF. I strongly recommend LaTeX, and the .tex for A1 is available on Quercus under the Files tab.
- Today we should be covering the last of the material needed to complete A1 – don't forget it's due Feb 12!

Fri. Jan 29: Announcements and Corrections

- Please confirm that you can log-in to MarkUs and can see A1
- A1 can be submitted as any *legible* PDF. I strongly recommend LaTeX, and the .tex for A1 is available on Quercus under the Files tab.
- Today we should be covering the last of the material needed to complete A1 – don't forget it's due Feb 12!

Fri. Jan 29: Announcements and Corrections

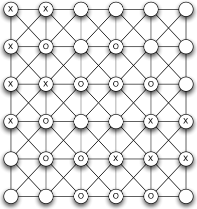
- There was a question last class about more general forms of the Schelling Model
- Rogers & McKane published a work accessible here `https://arxiv.org/abs/1104.1971` defining a more general model, analyzing a simplified model, and demonstrating similar behaviour to multiple other variants
- They assumed individuals lived on an arbitrary undirected graph with nodes $V = \{1 \dots n\}$
- $\sigma \in \{-1, 0, 1\}^n$ defined the state of their world
  - -1 signifies occupied by individual of type A
  - 0 signifies empty
  - 1 signifies occupied by individual of type B
- $\sigma^{(ij)}$ is σ with the entries i and j swapped
- They assumed that each timestep a pair of nodes is chosen at random and their occupants swapped – this is done with probability T_{ij}, subject to some constraints
- $P(\sigma, t+1) = \sum_{ij} P(\sigma^{(ij)}, t) T_{ij}(\sigma^{(ij)})$

Fri. Jan 29: Announcements and Corrections

- They compared their analysis to various pre-existing models that could be expressed in this general form
- The graphs considered were:

Fri. Jan 29: Announcements and Corrections

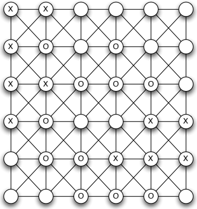
- They compared their analysis to various pre-existing models that could be expressed in this general form
- The graphs considered were:



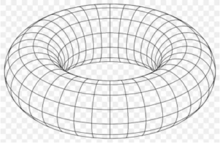
Grid; E&K Fig 4.15

Fri. Jan 29: Announcements and Corrections

- They compared their analysis to various pre-existing models that could be expressed in this general form
- The graphs considered were:



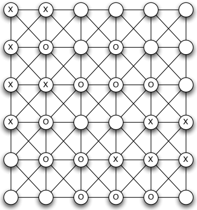
Grid; E&K Fig 4.15



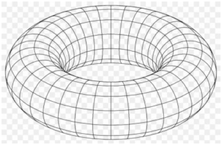
Toroidal Grid. Image by Adith George 1840427, CC BY-SA 4.0, via Wikimedia Commons

Fri. Jan 29: Announcements and Corrections

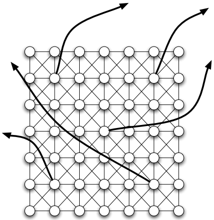
- They compared their analysis to various pre-existing models that could be expressed in this general form
- The graphs considered were:



Grid; E&K Fig 4.15



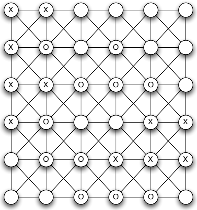
Toroidal Grid. Image by Adith George 1840427, CC BY-SA 4.0, via Wikimedia Commons



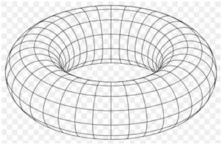
Small world network. E&K Fig 20.3

Fri. Jan 29: Announcements and Corrections

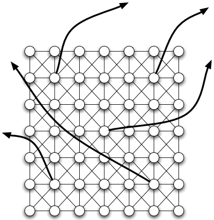
- They compared their analysis to various pre-existing models that could be expressed in this general form
- The graphs considered were:



Grid; E&K Fig 4.15



Toroidal Grid. Image by Adith George 1840427, CC BY-SA 4.0, via Wikimedia Commons



Small world network. E&K Fig 20.3

- Small World Networks! Coming soon to a lecture near you ;)

Reviewing selection vs social influence

- With immutable factors (such as race and for the most part gender), when we observe evidence of homophily, we often attribute increased friendships to selection, which is the tendency to form friendships with others who are like you in some way(s). (But note that race often correlates with neighbourhoods or academic programs.)

Reviewing selection vs social influence

- With immutable factors (such as race and for the most part gender), when we observe evidence of homophily, we often attribute increased friendships to selection, which is the tendency to form friendships with others who are like you in some way(s). (But note that race often correlates with neighbourhoods or academic programs.)
- But when considering more mutable factors, there is a feedback between similar characteristics and social links.
  - To what extent does behaviour get modified by our social network?
  - That is, to what extent is social influence determining interests and behaviour?
- Of course, both selection and social influence can be interacting in the same social network. How does one understand the relative interplay?

Reviewing selection vs social influence

- With immutable factors (such as race and for the most part gender), when we observe evidence of homophily, we often attribute increased friendships to selection, which is the tendency to form friendships with others who are like you in some way(s). (But note that race often correlates with neighbourhoods or academic programs.)
- But when considering more mutable factors, there is a feedback between similar characteristics and social links.
  - To what extent does behaviour get modified by our social network?
  - That is, to what extent is social influence determining interests and behaviour?
- Of course, both selection and social influence can be interacting in the same social network. How does one understand the relative interplay?

Longitudinal studies may make it possible to see the behavioral changes that occur after changes in an individuals network connections, as opposed to the changes to the network that occur after an individual changes his or her behavior.

A study of similarity and interaction

We point ahead to a study by Crandall et al [2008] that suggests that in certain settings, it may be possible to gain some insight into the selection vs influence issue. We return to this study later in lecture.

Using Wikipedia data, the text presents one study that speaks to the manner in which selection and influence combine to result in observed homophily. The nodes are Wikipedia editors, and edges correspond to communication via a user-talk page for a Wikipedia page. So we know what the graph means and can observe the emergence of edges over time.

The study defines a numerical similarity measure between two users A and B as a small variation on the following ratio which is analogous to the way neighbourhood overlap was defined:

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

Fortunately, every action on Wikipedia is recorded and time-stamped so it is possible to conduct a meaningful longitudinal study by looking at each "time step" defined by an "action" by either of the editors where an action is either an article edit, or a communication.

Average level of similarity before and after the first Wikipedia communication

The figure below plots the level of similarity as a function of the number of edits before and after the first communication. Time 0 is defined to be the time of the first interaction between a pair (A, B) of editors. This is then averaged over all the (A, B) plots.

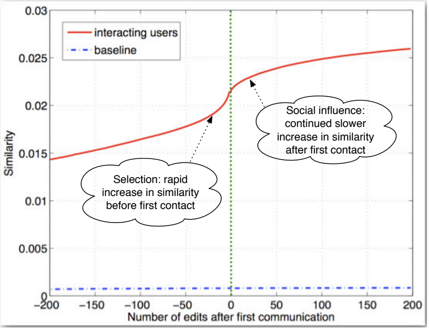


Figure: [E&K, Fig 4.13]

Two interesting and opposing longitudinal studies

- In academic success (or drug usage) in teenage friendship networks, Cohen (1977) and Kandel (1978) claim that peer pressure (i.e. social influence) is less a factor here than previously believed. We can speculate that (for example) similar family environments is a significant determining factor for such behaviour amongst friends.

- In contrast to the above example, in a controversial report on obesity patterns of 32,000 people observed over a 32 year period, Christakis and Fowler (2007) claim: obesity or keeping fit is (perhaps surprisingly) to some extent a contagious disease spread within a social network. "You don't necessarily catch it from your friends the way you catch the flu, but it nonetheless can spread through the underlying social network via the mechanism of social influence." (Later in the course we will discuss models for the spread of influence in a network.)

Why the obesity homophily?

- Three possibilities identified by Christakis and Fowler:
  1. [1] selection
  2. [2] homophily being driven by other factors that correlate with obesity (e.g. poverty)
  3. [3] the social influence of peer pressure say as in the case of drug use or academic performance or fitness.

- Christakis and Fowler conclude that even accounting for [1] and [2], social influence is a significant factor.

  Aside: I am not sure as to the extent that they consider the relative role of genetics vs diet.

- Once again, we caution that observing homophily is clearly only a starting point.

Why do we care?

- Why do we care about the relative interplay (selection vs. social influence) and how could we model this?

Why do we care?

- Why do we care about the relative interplay (selection vs. social influence) and how could we model this?

- If indeed social influence is a significant factor, then targeting key individuals and trying to modify undesirable behaviour (or promote positive behaviour) can be effective since we are then viewing such behaviour as a process of influence spread.

- If not, focusing on a few individuals will at best change the behaviour of a few individuals.

Social-affiliation networks: incorporating context into the network

- Up to now we have viewed contextual (mutable and immutable) factors that affect the formation of links to be outside of the social network being considered.
- Section 4.3 discusses how to include context in the network so as to have a common framework for studying the interplay between the extent of (social) triadic closure (common friendships induce new friendships), homophily determined by selection, and mutual activity determined by social influence.

Social-affiliation networks: incorporating context into the network

- Up to now we have viewed contextual (mutable and immutable) factors that affect the formation of links to be outside of the social network being considered.
- Section 4.3 discusses how to include context in the network so as to have a common framework for studying the interplay between the extent of (social) triadic closure (common friendships induce new friendships), homophily determined by selection, and mutual activity determined by social influence.
- Let's consider the (mutable) context of affiliation in a group/participation in an activity. Such an activity is referred to as a foci, a focal point for social interaction.

Social-affiliation networks: incorporating context into the network

- Up to now we have viewed contextual (mutable and immutable) factors that affect the formation of links to be outside of the social network being considered.
- Section 4.3 discusses how to include context in the network so as to have a common framework for studying the interplay between the extent of (social) triadic closure (common friendships induce new friendships), homophily determined by selection, and mutual activity determined by social influence.
- Let's consider the (mutable) context of affiliation in a group/participation in an activity. Such an activity is referred to as a foci, a focal point for social interaction.
- We incorporate such foci into social networks by considering a focus to be a different type of node, distinct from a node representing an individual. We first consider a pure affiliation network, an example being of which we have already seen in a bipartite graph with individuals and corporate boards.

Example of a pure affiliation network

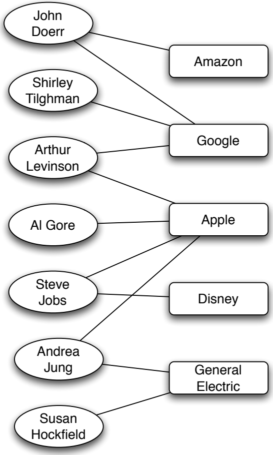


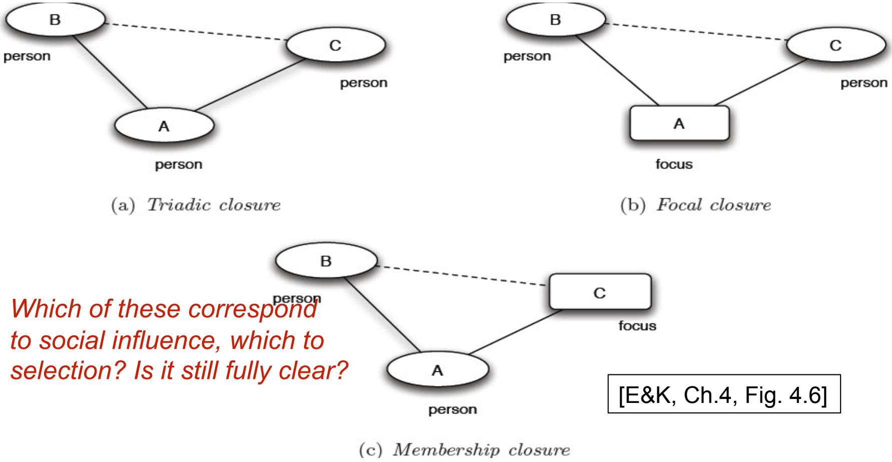
Figure: [E&K, Fig 4.4] One type of affiliation network that has been widely studied is the memberships of people on corporate boards of directors. A very small portion of this network (as of mid-2009) is shown here.

Social-affiliation networks continued

We can then combine the people-people edges of a social network with the people-focus edges of an affiliation network to form a social-affiliation network. Within such a combined network, we can discuss three types of graph triangle closures:

- triadic closure as introduced in chapter 3 where common friends of one or more individuals become friends
- focal closure where individuals become friends based on their common interest(s)
- membership closure where an individual joins an activity because a friend (or a group of friends) is (are) already in that activity

Three types of closure



(a) *Triadic closure*

(b) *Focal closure*

(c) *Membership closure*

Which of these correspond to social influence, which to selection? Is it still fully clear?

[E&K, Ch.4, Fig. 4.6]

Figure: [E&K, Fig 4.6] Three types of closure

Toy example of a social-affiliation network

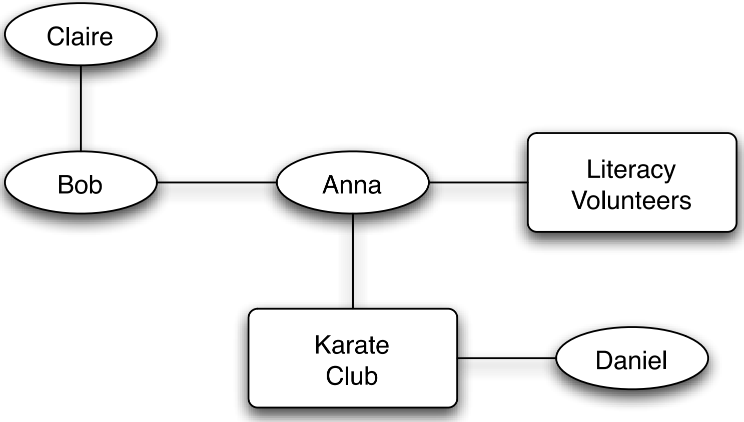


Figure: [E&K, Fig 4.5] In this social-affiliation network, the oval nodes are people and the rectangular nodes are activities. What kinds of triangular closures can occur?

Toy example showing three types of closure

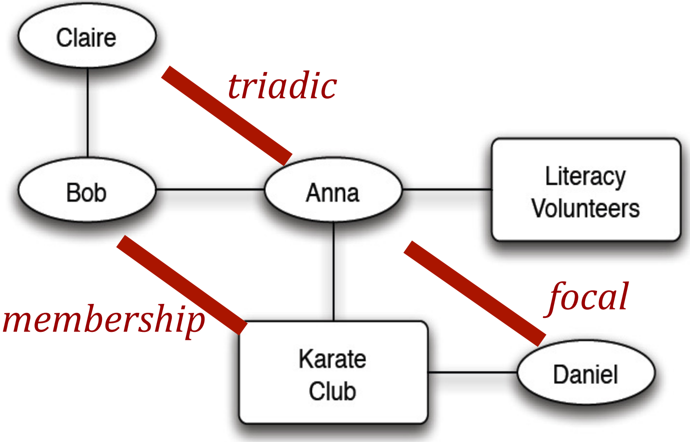


Figure: [E&K, Fig 4.7] We can observe the three types of triangular closures that have occurred in some time period.

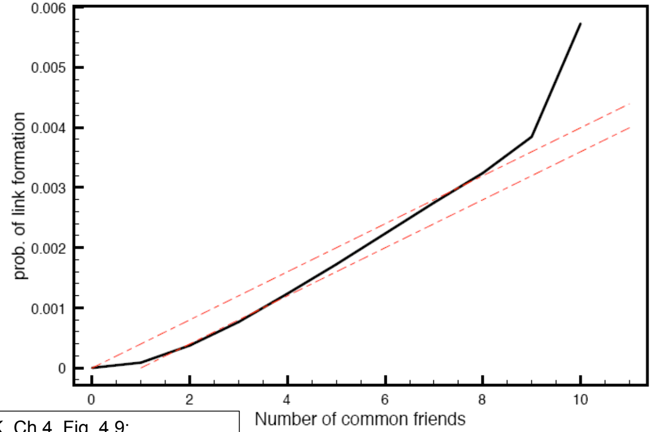
Empirically measuring these processes

- Closure is inherently dynamic
  - So we need to take snapshots of the network at different times to see how the relationships evolve and to what extent each form of closure occurs
  - If common friends or common interests are causing new links (i.e., closures) then the more friends or interests in common, the more we should see this effect.
- We briefly look at a couple studies stemming from online interactions, but realize the usual warning about limitations of such studies
  - As in all modeling we may be missing many factors
  - The timing of the snapshots may influence results
  - These particular studies look at link formation, but not link dissolution. What would the network look like if links formed but never dissolved?

Triadic closure: dependence on number mutual friends

- Email exchanges (over a year) by 45,000 students & staff in large US university [Kossinets, Watts 2006]
- "Friends" defined as two-way email communication (prev. 60 days)
- Measure probability $T(k)$ of a new friendship emerging between a pair of students as a function of the number k of mutual friends
- That is, the probability of it happening in any given day (averaging over many such pairs)
- Compare data (black) with baseline theoretical model (red) baseline: assume any single mutual friend will generate a new friendship with probability p and that this will happen *independently* for each common friend. Thus $T(k) = 1 - (1-p)^k$ Why?
- For small p, $(1-p)^k \approx 1 - pk$ so that $T(k) \approx pk$.

Probability (per-day) of triadic closure as a function of the number of common friends



[E&K, Ch.4, Fig. 4.9;
from Kossinets and Watts, 2006]

Figure: [E&K, Fig 4.9]

Observations

- Data does not show much more propensity for friendship when going from zero to one mutual friend.
  - The second dashed red line shifts the curve over by one friend so as to better compare the actual data and baseline model.
  - Why no major impact with one common friend?

Observations

- Data does not show much more propensity for friendship when going from zero to one mutual friend.
  - The second dashed red line shifts the curve over by one friend so as to better compare the actual data and baseline model.
  - Why no major impact with one common friend?
- Increasing from 1 to 9 friends shows linear curve (greater slope than baseline)
- A sharp difference going beyond 9 friends
  - The theoretical model (and its assumption of independence) no longer supported.
  - Is there some threshold of mutual friends which escalates the pressure for triadic closure?

Exercise: translate per-day probability into per-month or per-year probability

Probability of focal closure as a function of the number of common classes

Kossinetts and Watts also studied focal closure where a focus means a class in which a student is enrolled.

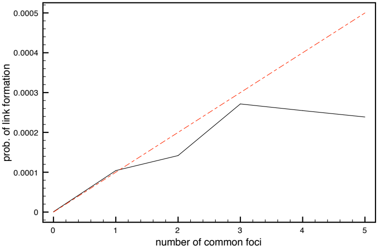


Figure: [E&K, Fig 4.10]

Clearly the theory and the actual data do not correspond especially when considering students going from 3 to 4 common classes. Can you speculate on a reason?

Probability of focal closure as a function of the number of common classes

Kossinetts and Watts also studied focal closure where a focus means a class in which a student is enrolled.

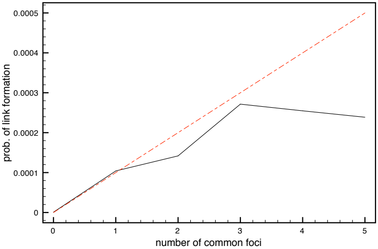


Figure: [E&K, Fig 4.10]

Clearly the theory and the actual data do not correspond especially when considering students going from 3 to 4 common classes. Can you speculate on a reason? If you haven't formed a friendship having attend 3 classes together, then perhaps there is a reason?

Probability of membership closure as a function of the number of common friends

The text presents two studies of membership closure where there is data concerning both person-to-person interactions and person-foci affiliations. The first study shows the probability of joining a community in the blogging site LiveJournal where “friendship” is self-identified within a user’s profile.

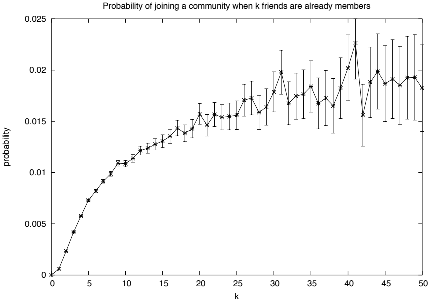


Figure: [E&K, Fig 4.11]

Second study of membership closure as a function of the number of common friends

The second study concerns Wikipedia editors and foci are specific Wikipedia pages. Here “friendship” is defined as having communicated together on a user-talk page and membership in a foci corresponds to having edited a Wikipedia page.

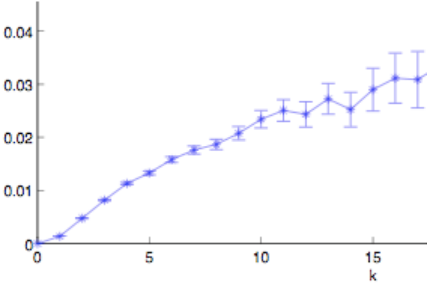


Figure: [E&K, Fig 4.12]

The interplay between selection and influence

Using the same Wikipedia data as in the previous focal closure example, The text presents one study that speaks to the manner in which selection and influence combine to result in observed homophily. Once again, the nodes are Wikipedia editors, the foci are articles, and edges correspond to communication via a user-talk page.

In addition, the study defines a numerical similarity measure between two users A and B as a small variation on the following ratio which is analogous to the way neighbourhood overlap was defined:

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

Fortunately, every action on Wikipedia is recorded and time-stamped so it is possible to conduct a meaningful longitudinal study by looking at each "time step" defined by an "action" of an editor where an action is either an article edit, or a communication.

Average level of similarity before and after the first Wikipedia communication

The figure below plots the level of similarity as a function of the number of edits before and after the first communication. Time 0 is defined to be the time of the first interaction between a pair (A, B) of editors. This is then averaged over all the (A, B) plots.

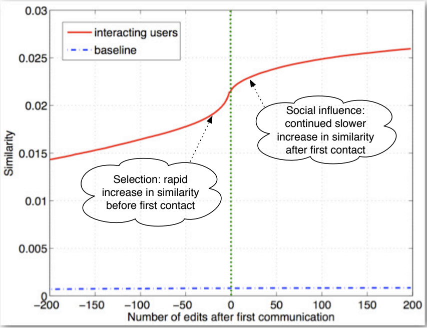


Figure: [E&K, Fig 4.13]

Observations on similarity vs. interactions (Figure 4.13)

There are a number of interesting observations and caveats regarding Figure 4.13. First some notable observations.

- The level of similarity is increasing over “time” before and after the first interaction.
- The steepest increase in similarity occurs just before the first interaction suggesting that selection is playing a pronounced role in forming this “friendship link” in the networks that are being dynamically created.
- The bottom dashed line indicates the level of similarity for those who never communicate. Clearly those who eventually interact are more similar.

Some caveats

- Like any averaging of individual data, we cannot say why any particular pair of editors have decided to communicate.
- Because the defined time 0 corresponds to different moments in “real time” for each pair, we cannot understand to what extent real time events may also be a factor leading communication.
- In this study, links are never eliminated. Other “fully dynamic” network settings would have node and/or links that are not permanent.
- The biggest question about such a study is the extent to which any observations may or may not extend to different settings. In what settings do we have the same kind of detailed time stamping of events?