Social and Information Networks

University of Toronto CSC303
Winter/Spring 2019

Week 3: January 21,23 (2019)

Announcements

My office hours will be Monday 1-2, and Wednesdays 4-5 or by appointment, or by dropping in and taking your chances.

I have posted a third question for Assignment 1 and I plan to add one or two more question this week.

Week 3: Todays agenda

- Last weeks lectures: We discussed chapter 3 of the EK text. We introduced a number of basic graph-theoretic concepts motivated by social networks (e.g., triadic closure, and the span, embeddedness and dispersion of an edge). A major theme of the chapter was the distinction between strong and weak ties, and the strength of weak ties.
- This week we will discuss chapter 4 of the text on Networks in their surrounding contexts. In particular, we will discuss
  - Homophily
  - the selection vs influence question.
  - Social-affiliation networks; three types of triangle closure
- But first lets finish the discussion of the Sintos and Tsaparas paper. This article is interesting because it gives us another example of how useful (and perhaps private) information can be extracted just from network structure. Moreover, it is an example of where a little knowledge of algorithms is important but one has to be careful about worst case results.

Sintos and Tsaparas: The vertex cover algorithms and the 5 data sets

While there are uncovered edges, the (vertex) greedy algorithm selects a vertex for the vertex cover with maximum current degree. It has worst case $O(\log n)$ approximation ratio. The maximal matching algorithm is a 2-approximation online algorithm that finds an uncovered edge and takes both endpoints of that edge.

Table 1: Datasets Statistics.

Dataset	Nodes	Edges	Weights	Community structure
Actors	1,986	103,121	Yes	No
Authors	3,418	9,908	Yes	No
Les Miserables	77	254	Yes	No
Karate Club	34	78	No	Yes
Amazon Books	105	441	No	Yes

Figure: Weights (respectively, community structure) indicates when explicit edge

End of Wednesday, January 16 Lecture

We ended the lecture on slide 36. In the Monday, January 21 lecture, we will finish up the discussion of the Sintos and Tsaparas paper. I am including the remaining slides for those who want to see some of the results in that paper. I will also post the paper.

Tie strength results in detecting strong and weak ties

Table 2: Number of strong and weak edges for Greedy and MaximalMatching algorithms.

	Greedy		MaximalMatching	
	Strong	Weak	Strong	Weak
Actors	11,184	91,937	8,581	94,540
Authors	3,608	6,300	2,676	7,232
Les Miserables	128	126	106	148
Karate Club	25	53	14	64
Amazon Books	114	327	71	370

Figure: The number of labelled links.

Although the Greedy algorithm has an inferior (worst case) approximation ratio, here the greedy algorithm has better performance than Maximal Matching. (Recall, the goal is to maximize the number of strong ties, or

Results for detecting strong and weak ties

Table 3: Mean count weight for strong and weak edges for Greedy and MaximalMatching algorithms.

	Greedy		MaximalMatching	
	S	W	S	W
Actors	1.4	1.1	1.3	1.1
Authors	1.341	1.150	1.362	1.167
Les Miserables	3.83	2.61	3.87	2.76

Figure: The avergae link weight.

Tie strength results in detecting strong and weak ties normalized by amount of activity

Table 4: Mean Jaccard similarity for strong and weak edges for Greedy and MaximalMatching algorithms.

	Greedy		MaximalMatching	
	S	W	S	W
$Actors$	0.06	0.04	0.06	0.04
$Authors$	0.145	0.084	0.155	0.088

Figure: Normalizing the number of interactions by the amount of activity.

Results for strong and weak ties with respect to known communities

Table 5: Precision and Recall for strong and weak edges for Greedy and MaximalMatching algorithms.

Greedy				
	P_S	R_S	P_W	R_W
Karate Club	1	0.37	0.19	1
Amazon Books	0.81	0.25	0.15	0.69
MaximalMatching				
	P_S	R_S	P_W	R_W
Karate Club	1	0.2	0.16	1
Amazon Books	0.73	0.14	0.14	0.73

Figure: Precision and recall with respect to the known communities.

The meaning of the precision-recall table

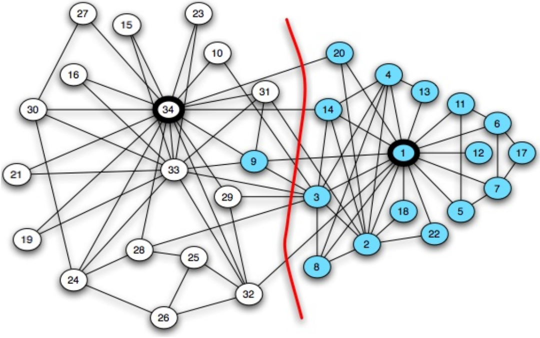
The precision and recall for the weak edges are defined as follows:

$P_W = \frac{|W \cap E_{inter}|}{|W|}$ and $R_W = \frac{|W \cap E_{inter}|}{|E_{inter}|}$

$P_S = \frac{|S \cap E_{intra}|}{|S|}$ and $R_S = \frac{|S \cap E_{intra}|}{|E_{intra}|}$

- Ideally, we want $R_W = 1$ indicating that all edges between communities are weak; and we want $P_S = 1$ idicating that strong edges are all within a community.
- For the Karate Club data set, all the strong links are within one of the two known communities and hence all links between the communities are all weak links.
- For the Amazon Books data set, there are three communities corresponding to liberal, neutral, conservative viewpoints. Of the 22 strong tie edges crossing communities, 20 have one node labeled as neutral and the remaining two inter-community strong ties both deal with the same issue.

A Balanced Min Cut in Graph: Bonding capital of nodes 1 and 34



- Note that node 34 also seems to have bridging capital.
- Wayne Zachary's Ph.D. work (1970-72): observed social ties and rivalries in a university karate club.
- During his observation, conflicts intensified and group split.
- Could the club boundaries be predicted from the network structure?
- Split could almost be explained by minimum cut in social network.

Strong and weak ties in the karate club network

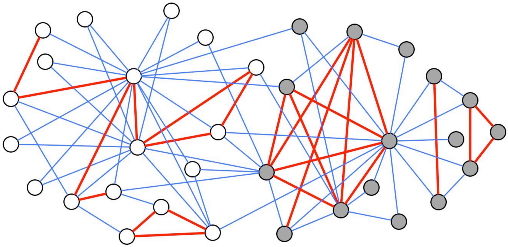


Figure 1: Karate Club graph. Blue light edges represent the weak edges, while red thick edges represent the strong edges.

- Sintos and Tsaparas apply their algorithm to the karate club network.
- Note that all the strong links are within one of the two "computed communities"; that is, links between the communities are all weak links.

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 disappeaance) 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 disappeaance) of an edge.

Yogi Berra(1925-2015):

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

Homophily

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

Homophily

- 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".

Homophily

- 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".
- Why triadic closure? In Chapter 3: some network "intrinsic" reasons (opportunity, trust, incentive) for forming a freindship and now we consider "contextual" reasons for homophily.
- **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 (essentially) immutable factors: race, 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 (essentially) immutable factors: race, 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, political party preference.

Characteristic factors

- Factors which help determine our friendships and relations can be immutable or more transient.
- Some (essentially) immutable factors: race, 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, political party preference.
- Of course, 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

- 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 similarity traits.

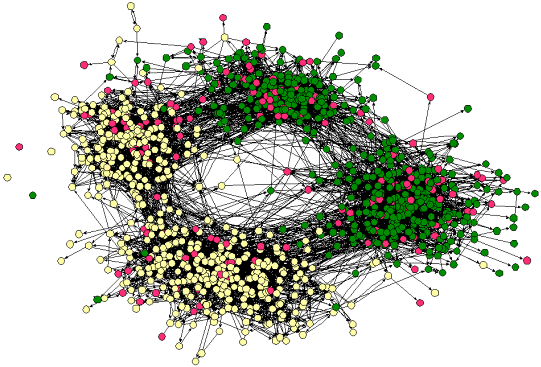
The influence vs selection issue

- 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 similarity 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

- 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 similarity 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.
- And to further complicate 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.

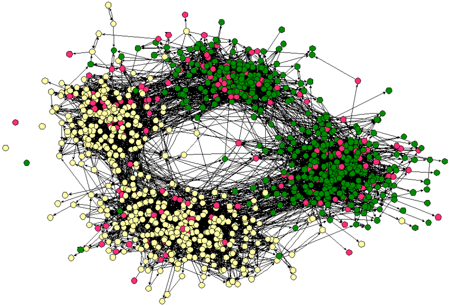
Graphic visualization of homophily



[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 presense or absense 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.

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.

Two interesting 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 dont 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 obsesity 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.

End of Monday, January 21 lecture

We continue with a discussion of social-affiliation networks. We can view joint membership by A and B in an organziation or club as being similar.

Why do we care?

- How do we study the relative interplay (selection vs. social influence) and why do we want to answer this chicken vs. egg type question?
- 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

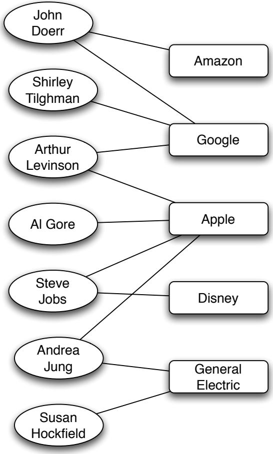


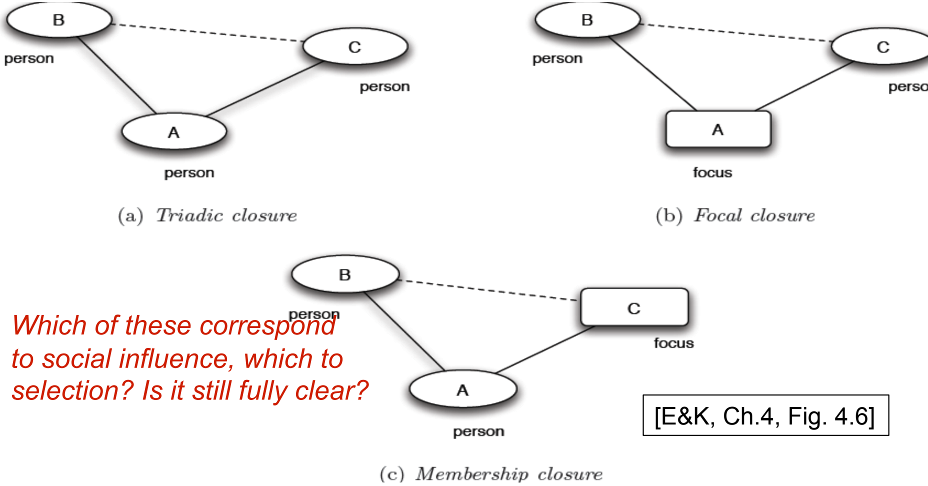
Figure: [E&K, Fig 4.4] One type of affiliation network that has been widely studied is the memberships of people on corporate boards of directors. A very small portion of this network (as of mid-2009) is shown here.

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



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

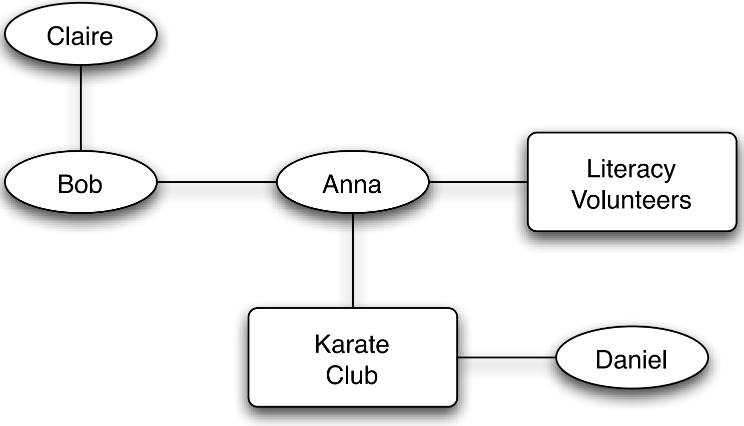


Figure: [E&K, Fig 4.5] In this social-affiliation network, the oval nodes are people and the rectangular nodes are activities. What kinds of triangular closures can occur?

Toy example showing three types of closure

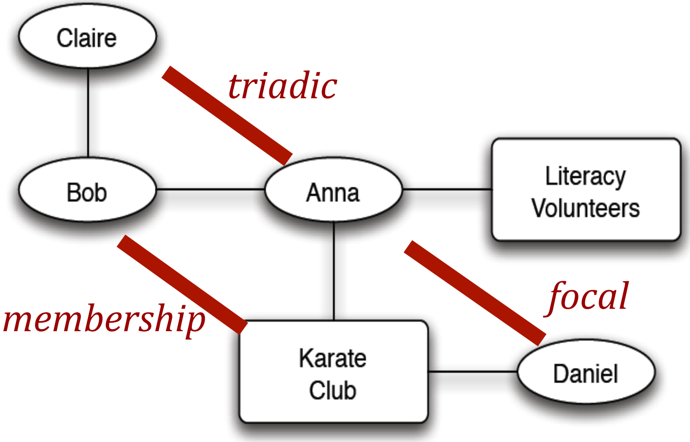


Figure: [E&K, Fig 4.7] We can observe the three types of triangular closures that have occured in some time period.

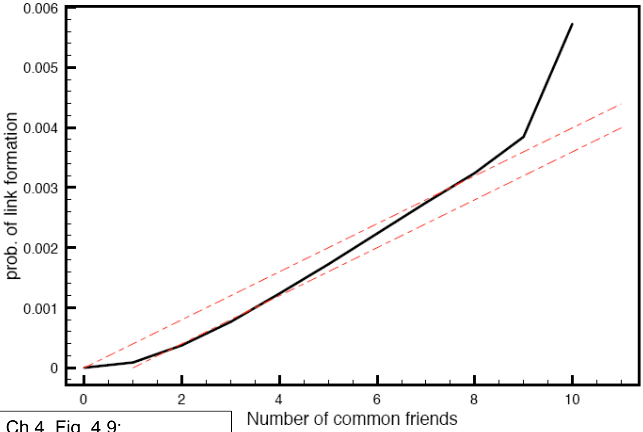
How do we measure extent of these processes?

- 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 60 days) by 22,000 students 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?
- 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.

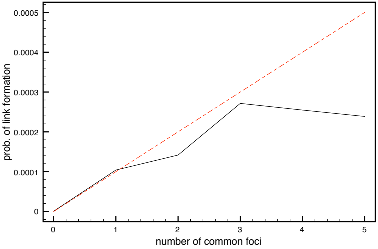


Figure: [E&K, Fig 4.10]

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.

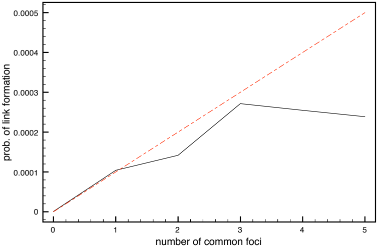


Figure: [E&K, Fig 4.10]

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 p robability of joining the blogging site LiveJournali where "friendship" is self-identified within a user's profile.

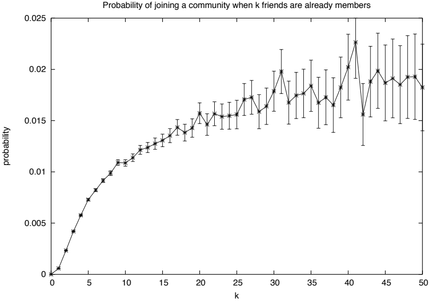


Figure: [E&K, Fig 4.11]

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.

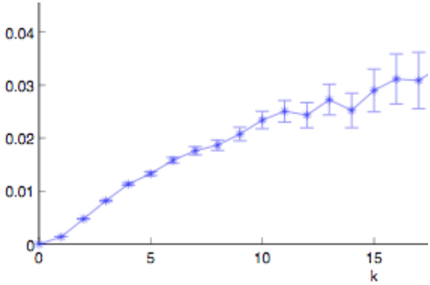


Figure: [E&K, Fig 4.12]

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 interction between a pair (A, B) of editors. This is then averaged over all the (A, B) plots.

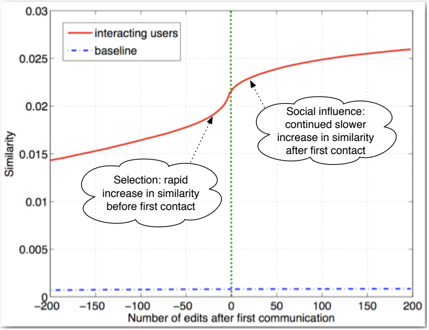


Figure: [E&K, Fig 4.13]

Observations on similarity vs. interactions (Figure 4.13)

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

- The level of similarity is increassing 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 evidence more similarity suggesting some significant similarity factors outside of what is being studied.

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?