Social and Information Networks

University of Toronto CSC303
Winter/Spring 2019

Week 2: January 14,16 (2019)

Todays agenda

- Last weeks lectures and tutorial: a number of basic graph-theoretic concepts:
  - undirected vs directed graphs
  - unweighted and vertex/edge weighted graphs
  - paths and cycles
  - breadth first search
  - connected components and strongly connected components
  - giant component in a graph
  - bipartite graphs

- This lecture: Chapter 3 of the textbook on "Strong and Weak Ties".

- But let's first briefly return to the "romantic relations" table to see how graph structure may or may not align with our understanding of sociological phenomena.

The dispersion of edge

- In their paper, Backstrom and Kleinberg introduce various *dispersion measures* for an edge. The general idea of the dispersion of an edge (u, v) is that mutual neighbours of u and v should not be well "well-connected" to one another.
- They consider dense subgraphs of the Facebook network (where a known relationship has been stated) and study how well certain dispersion measures work in predicting a relationship edge (marriage, engagement, relationship) when one is known to exist.
- Another problem is to discover which individuals have a romantic relationship. Note that discovering the romantic edge or if one exists raises questions of the privacy of data.
- They compare their dispersion measures against the *embeddedmess* of an edge (to be defined) and also against certain semantic interaction features (e.g., appearance together in a photo, viewing the profile of a neighbour).

Some experimental results for the fraction of correct predictions

Recall that we argued (since the median number of friends was 200), that the fraction might be around .005 when randomly choosing an edge to be the edge to be link to the significant partner. Do you find anything surprising in this table?

type	embed	rec.disp.	photo	prof.view.
all	0.247	0.506	0.415	0.301
married	0.321	0.607	0.449	0.210
married (fem)	0.296	0.551	0.391	0.202
married (male)	0.347	0.667	0.511	0.220
engaged	0.179	0.446	0.442	0.391
engaged (fem)	0.171	0.399	0.386	0.401
engaged (male)	0.185	0.490	0.495	0.381
relationship	0.132	0.344	0.347	0.441
relationship (fem)	0.139	0.316	0.290	0.467
relationship (male)	0.125	0.369	0.399	0.418

type	max. struct.	max. inter.	all. struct.	all. inter.	comb.
all	0.506	0.415	0.531	0.560	0.705
married	0.607	0.449	0.624	0.526	0.716
engaged	0.446	0.442	0.472	0.615	0.708
relationship	0.344	0.441	0.377	0.605	0.682

Chapter 3: Strong and Weak Ties

There are two themes that run throughout this chapter.

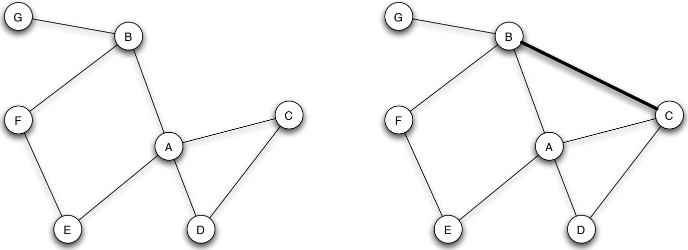
1. Strong vs. weak ties and "the strength of weak ties" is the specific defining theme of the chapter. The chapter also starts a discussion of how networks evolve.

2. The larger theme is in some sense "the scientific method".
   - Formalize concepts, construct models of behaviour and relationships, and test hypotheses.
   - Models are not meant to be the same as reality but to abstract the important aspects of a system so that it can be studied and analyzed.
   - See the discussion of the strong triadic closure property in section 3.2 of text (pages 53 and 56 in my online copy).

Informally

- strong ties: stronger links, corresponding to friends
- weak ties: weaker links, corresponding to acquaintances

Triadic closure (undirected graphs)



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

Figure: The formation of the edge between B and C illustrates the effects of triadic closure, since they have a common neighbor A. [E&K Figure 3.1]

- Triadic closure: mutual "friends" of say A are more likely (than "normally") to become friends over time.
- How do we measure the extent to which triadic closure is occurring?
- How can we know why a new friendship tie is formed? (Friendship ties can range from just knowing someone to a true friendship .)

Measuring the extent of triadic closure

- The clustering coefficient of a node A is a way to measure (over time) the extent of triadic closure (perhaps without understanding why it is occurring).

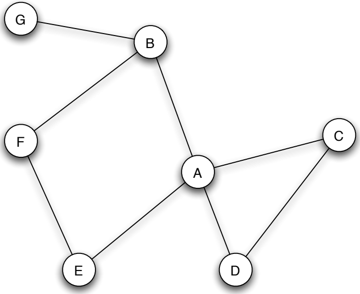
- Let E be the set of an undirected edges of a network graph. (Forgive the abuse of notation where in the previous and next slide E is a node name.) For a node A, the clustering coefficient is the following ratio:

$$\frac{\left|\left\{(B,C) \in E : (B,A) \in E \text{ and } (C,A) \in E\right\}\right|}{\left|\left\{\{B,C\} : (B,A) \in E \text{ and } (C,A) \in E\right\}\right|}$$

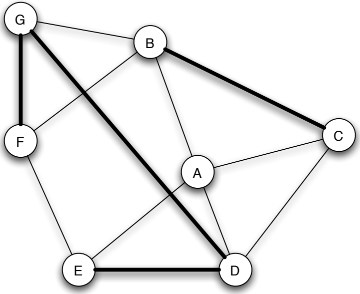
- The numerator is the number of all edges (B,C) in the network such that B and C are adjacent to (i.e. mutual friends of) A.

- The denominator is the total number of all unordered pairs $\{B,C\}$ such that B and C are adjacent to A.

Example of clustering coefficient



(a) *Before new edges form.*



(b) *After new edges form.*

- The clustering coefficient of node A in Fig. (a) is $1/6$ (since there is only the single edge (C, D) among the six pairs of friends: $\{B, C\}$, $\{B, D\}$, $\{B, E\}$, $\{C, D\}$, $\{C, E\}$, and $\{D, E\}$)
- The clustering coefficient of node A in Fig. (b) increased to $1/2$ (because there are three edges (B, C), (C, D), and (D, E)).

Interpreting triadic closure

- Does a low clustering coefficient suggest anything?

Interpreting triadic closure

- Does a low clustering coefficient suggest anything?

- Bearman and Moody [2004] reported finding that a low clustering coefficient amongst teenage girls implies a higher probability of suicide (compared to those with high clustering coeficient). How can we understand this finding?

Interpreting triadic closure

- Does a low clustering coefficient suggest anything?

- Bearman and Moody [2004] reported finding that a low clustering coefficient amongst teenage girls implies a higher probability of suicide (compared to those with high clustering coeficient). How can we understand this finding?

- Why triadic closure?

Interpreting triadic closure

- Does a low clustering coefficient suggest anything?

- Bearman and Moody [2004] reported finding that a low clustering coefficient amongst teenage girls implies a higher probability of suicide (compared to those with high clustering coeficient). How can we understand this finding?

- Why triadic closure?
Increased opportunity, trust, incentive ; it can be awkward to have good friends (i.e. with strong ties) who are not themselves friends.

Granovetter's thesis: the strength of weak ties

- In 1960s interviews: Many people learn about new jobs from personal contacts (which is not surprising) and often these contacts were acquaintances rather than friends. Is this surprising?

Granovetter's thesis: the strength of weak ties

- In 1960s interviews: Many people learn about new jobs from personal contacts (which is not surprising) and often these contacts were acquaintances rather than friends. Is this surprising?
  Upon a little reflection, this intuitively makes sense.

Granovetter's thesis: the strength of weak ties

- In 1960s interviews: Many people learn about new jobs from personal contacts (which is not surprising) and often these contacts were acquaintances rather than friends. Is this surprising?
  Upon a little reflection, this intuitively makes sense.

- The idea is that weak ties link together "tightly knit communities", each containing a large number of strong ties.

Granovetter's thesis: the strength of weak ties

- In 1960s interviews: Many people learn about new jobs from personal contacts (which is not surprising) and often these contacts were acquaintances rather than friends. Is this surprising?
  Upon a little reflection, this intuitively makes sense.

- The idea is that weak ties link together "tightly knit communities", each containing a large number of strong ties.

- Can we say anything more quantitative about such phenomena?

- To gain some understanding of this phenomena, we need some additional concepts relating to structural properties of a graph.

Recall

- strong ties: stronger links, corresponding to friends
- weak ties: weaker links, corresponding to acquaintances

Bridges and local bridges

- One measure of connectivity is the number of edges (or nodes) that have to be removed to disconnect a graph.
- A bridge (if one exists) is an edge whose removal will disconnect a connected component in a graph.
- We expect that large social networks will have a "giant component" and few bridges.

Bridges and local bridges

- One measure of connectivity is the number of edges (or nodes) that have to be removed to disconnect a graph.
- A bridge (if one exists) is an edge whose removal will disconnect a connected component in a graph.
- We expect that large social networks will have a "giant component" and few bridges.
- A local bridge is an edge (A, B) whose removal would cause A and B to have graph distance (called the span of this edge) greater than two. Note: span is a *dispersion measure*, as introduced in the Backstrom and Kleinberg article regarding Facebook relations.
- A local bridges (A,B) plays a role similar to bridges providing access for A and B to parts of the network that would otherwise be (in a useful sense) inaccessible.

Local bridge (A, B)

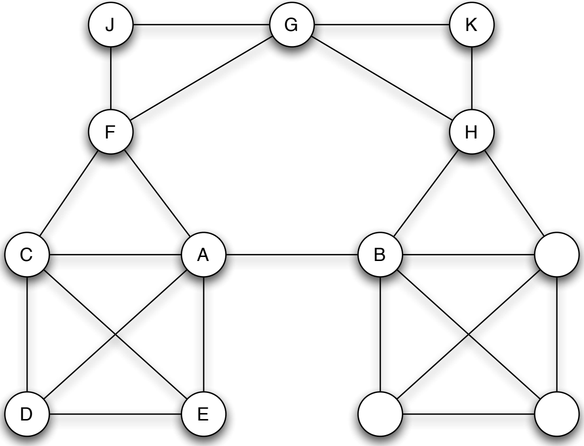


Figure: The edge (A, B) is a local bridge of span 4, since the removal of this edge would increase the distance between A and B to 4. [E&K Figure 3.4]

Strong triadic closure property: connecting tie strength and local bridges

Strong triadic closure property

Whenever (A, B) and (A, C) are strong ties, then there will be a tie (possibly only a weak tie) between B and C.

- Such a strong property is not likely true in a large social network (that is, holding for every node A)
- However, it is an abstraction that may lend insight.

Strong triadic closure property: connecting tie strength and local bridges

Strong triadic closure property

Whenever (A, B) and (A, C) are strong ties, then there will be a tie (possibly only a weak tie) between B and C.

- Such a strong property is not likely true in a large social network (that is, holding for every node A)
- However, it is an abstraction that may lend insight.

Theorem

Assuming the strong triadic closure property, for a node involved in at least two strong ties, any local bridge it is part of must be a weak tie.

Informally, local bridges must be weak ties since otherwise strong triadic closure would produce shorter paths between the end points.

Triadic closure and local bridges

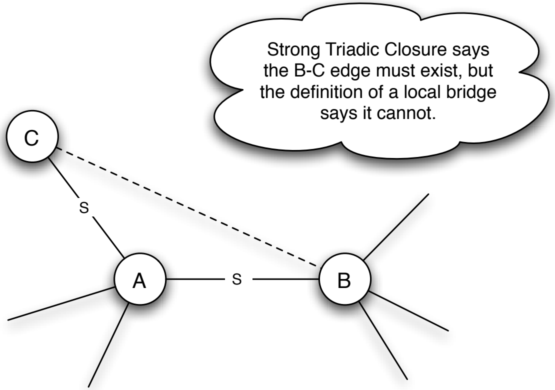


Figure 3.6: If a node satifies Strong Triadic Closure and is involved in at least two stron
ties, then any local bridge it is involved in must be a weak tie. The figure illustrates th
reason why: if the A-B edge is a strong tie, then there must also be an edge between B an
C, meaning that the A-B edge cannot be a local bridge.

Strong triadic closure property continued

- Again we emphasize (as the text states) that "Clearly the strong triadic closure property is too extreme to expect to hold across all nodes ... But it is a useful step as an abstraction to reality, ..."
- Sintos and Tsaparas give evidence that assuming the strong triadic closure property can help in determining whether a link is a strong or weak tie.
  (http:/(www.cs.uoi.gr/∼tsap/publications/frp0625-sintos.pdf)

  We will discuss this paper later in the lecture.
- A followup and refinement of the Sintos and Tsaparas 2014 experiments can be found in a relatively a 2018 paper by Adriaens et al
  (http://http://www.mlgworkshop.org/2018/papers/MLG2018_paper_28.

Embeddedness of an edge

Just as there are many specific ways to define the dispersion of an edge, there are different ways to define the embeddedness of an edge.

The general idea is that embeddedness of an edge (u, v) should capture how much the social circles of u and v "overlap". The next slide will use a particular definition for embeddedness.

Why might dispersion be a better discriminator of a romantiic relationship (especially for marriage) than embeddedness?

Large scale experiment supporting strength of weak ties and triadic closure

- Onnela et al. [2007] study of who-talks-to-whom network maintained by a cell phone provider. Large network of cell users where an edge exists if there existed calls in both directions in 18 weeks.
- First observation: a giant component with 84% of nodes.
- Need to quantify the tie strength and the closeness to being a local bridge.
- Tie strength is measured in terms of the total number of minutes spent on phone calls between the two end of an edge.
- Closeness to being a local bridge is measured by the neighborhood overlap of an edge (A, B) defined as the ratio

$$\frac{\text{number of nodes adjacent to both } A \text{ and } B}{\text{number of nodes adjacent to at least one of } A \text{ or } B}$$

- Local bridges are precisely edges having overlap 0.
- The numerator is the embeddedness of the edge.

Onnela et al. experiment

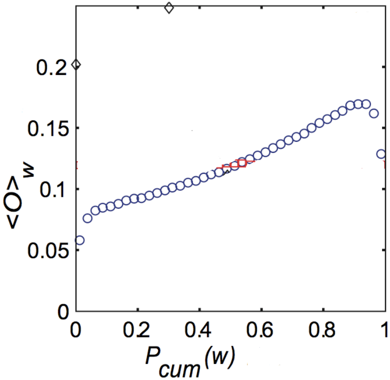


Figure: A plot of the neighborhood overlap of edges as a function of their percentile in the sorted order of all edges by tie strength. [E&K Fig 3.7]

- The figure shows the relation between tie strength and overlap.
- Quantitative evidence supporting the theorem: as tie strength decreases, the overlap decreases; that is, weak ties are becoming "almost local bridges" having ovelap almost equal to 0.

End of Monday, January 14 Lecture

We ended the lecture on slide 18.

Todays agenda

1. Continue the discussion of Chapter 3: strong vs weak ties and the strength of weak ties
   1. Continue with the Onnela et al study
   2. The Marlow et al study of different types of ties in the Facebook and Twitter networks
   3. Social captial: bonding capital and bridging capital
2. The Sintos and Tsaparas experiments

Onnela et al. experiment (shown again)

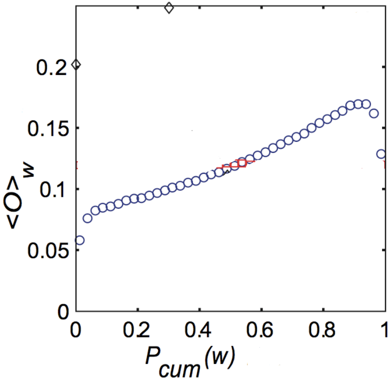


Figure: A plot of the neighborhood overlap of edges as a function of their percentile in the sorted order of all edges by tie strength. [E&K Fig 3.7]

- The figure shows the relation between tie strength and overlap.
- Quantitative evidence supporting the theorem: as tie strength decreases, the overlap decreases; that is, weak ties are becoming "almost local bridges" having ovelap almost equal to 0.

Onnela et al. study continued

To support the hypothesis that weak ties tend to link together more tightly knit communities, Onnela et al. perform two simulations:

1. Removing edges in decreasing order of tie strength, the giant component shrank gradually.

2. Removing edges in increasing order of tie strength, the giant component shrank more rapidly and at some point then started fragmenting into several components.

Word of caution in text regarding such studies

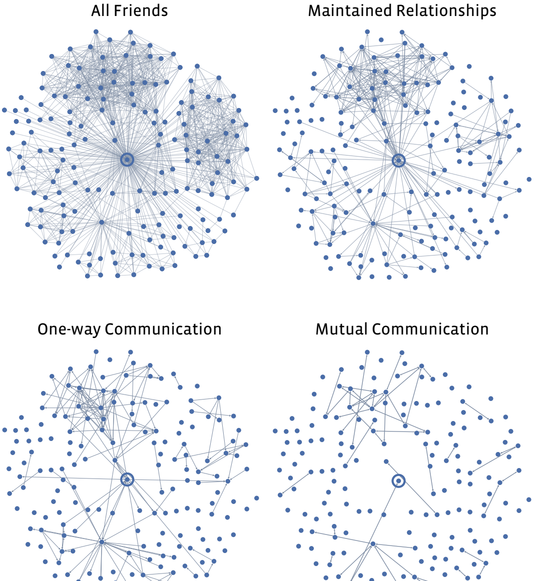
Easley and Kleinberg (end of Section 3.3):

> *Given the size and complexity of the (who calls 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. . .*

Strong vs. weak ties in large online social networks (Facebook and Twitter)

- The meaning of "friend" as in Facebook is not the same as one might have traditionally interpreted the word "friend".
- Online social networks give us the ability to qualify the strength of ties in a useful way.
- For an observation period of one month, Marlow et al. (2009) consider Facebook networks defined by 4 criteria (increasing order of strength): all friends, maintained (passive) relations of following a user, one-way communication, and reciprocal communication.
  1. These networks thin out when links represent stronger ties.
  2. As the number of total friends increases, the number of reciprocal communication links levels out at slightly more than 10.
  3. How many Facebook friends did you have for which you had a reciprocal communication in the last month?

Different Types of Friendships: The neighbourhood network of a sample Facebook individual



A limit to the number of strong ties

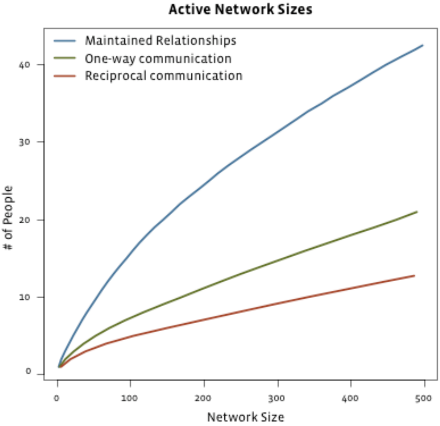


Figure: The number of links corresponding to maintained relationships, one-way communication, and reciprocal communication as a function of the total neighborhood size for users on Facebook. [Figure 3.9, textbook]

Twitter:Limited Strong Ties vs Followers

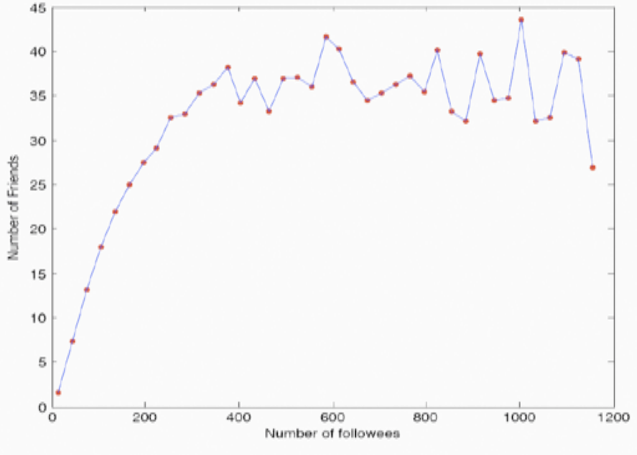


Figure: The total number of a user's strong ties (defined by multiple directed messages) as a function of the number of followees he or she has on Twitter. [Figure 3.10, textbook]

Information spread in a passive network

- The maintained or passive relation network (as in the Facebook network on slide 24) is said to occupy a middle ground between
    1. strong tie network (in which individuals actively communicate), and
    2. very weak tie networks (all "friends") with many old (and inactive) relations.

- "Moving to an environment where everyone is passively engaged with each other, some event, such as a new baby or engagement can propagate very quickly through this highly connect neighborhood."

- We can add that an event might be a political demonstration.

Social capital (as discussed in secrion 3.5 of EK text)

Social capital is a term in increasingly widespread use, but it is a famously difficult one to define.

The term social capital is designed to suggest its role as part of an array of different forms of capital (e.g., economic capital) all of which serve as tangible or intangible resources that can be mobilized to accomplish tasks.

A person can have more or less social capital depending on his or her position in the underlying social structure or network. A second, related, source of terminological variation is based on whether social capital is a property that is purely intrinsic to a group based only on the social interactions among the groups members or whether it is also based on the interactions of the group with the outside world.

“Tightly knit communities” connected by weak ties

- The intuitive concept of tightly knit communities occurs several times in Chapter 3 but is deliberately left undefined.

- In a small network we can sometimes visualize the tightly knit communities but one cannot expect to do this is a large network. That is, we need algorithms and this is the topic of the advanced material in Section 3.6.

"Tightly knit communities" connected by weak ties

- The intuitive concept of tightly knit communities occurs several times in Chapter 3 but is deliberately left undefined.
- In a small network we can sometimes visualize the tightly knit communities but one cannot expect to do this is a large network. That is, we need algorithms and this is the topic of the advanced material in Section 3.6.
- Recalling the relation to weak ties, the text calls attention to how nodes at the end of one (or especially more) local bridges can play a pivotal role in a social network.
- These "gatekeeper nodes" between communities stand in contrast to nodes which sit at the center of a tightly knit community.

Central nodes vs. gatekeepers

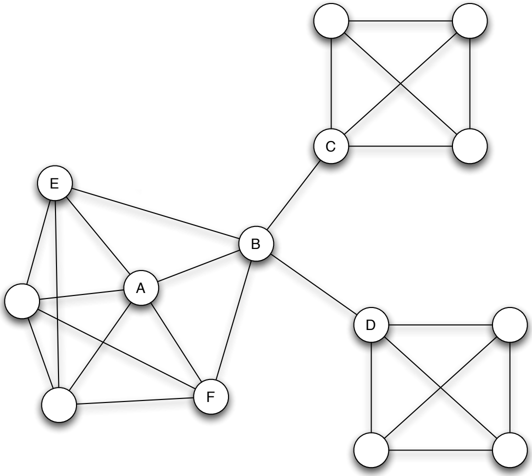


Figure: The contrast between densely-knit groups and boundary-spanning links is reflected in the different positions of central node *A* and gatekeeper node *B* in the underlying social network. [Fig 3.11, textbook]

Social capital of nodes A and B

- The edges adjacent to node A all have high embeddedness. Visually one sees node A as a central node in a tightly-knit cluster. As such, the social capital that A enjoys is its "bonding capital" in that the actions of A can (for example) induce norms of behaviour because of the trust in A.
- In contrast, node B is a bridge to other parts of the network. As such, its social capital is in the form of "brokerage" or "bridging capital" as B can play the role of a "gatekeeper" (of information and ideas) between different parts of the network. Furthermore, being such a gatekeeper can lead to creativity stemming from the synthesis of ideas.
- Some nodes can have both bonding capital and bridging capital.

Florentine marriages: Bridging capital of the Medici

- The Medici are connected to more families, but not by much.
- More importantly: Four of the six edges adjacent to the Medici are bridges or local bridges and the Medici lie on the shortest paths between most pairs of families.

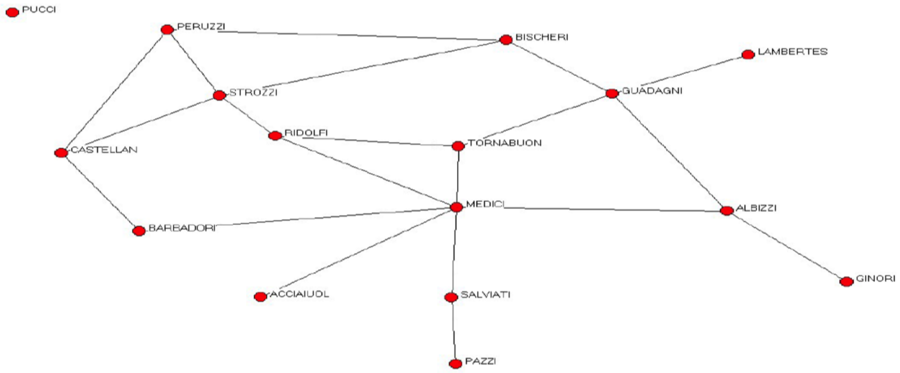
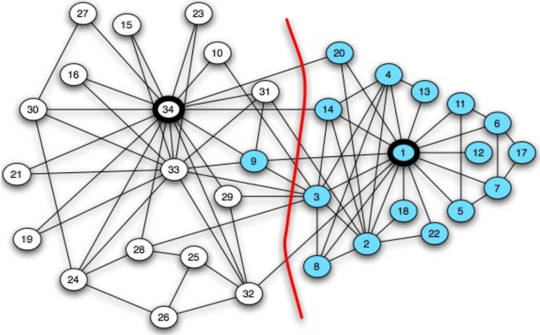


Figure: see [Jackson, Ch 1]

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.

The Sintos and Tsaparas Study

In their study of the strong triadic closure (STC) property, Sintos and Tsaparas study 5 small networks. They give evidence as to how the STC assumption can help determine weak vs strong ties, and how weak ties act as bridges to different communities.

More specifically, for a social network where the edges are not labelled they define the following two computational problems: Label the graph edges (by strong and weak) so as to satisfy the strong triadic closure property and

1. Either maximize the number of strong edges, or equivalently
2. minimize the number of weak edges

The computational problem in identifying strong vs weak ties

- For computational reasons (i.e., assuming $P \neq NP$ and showing NP hardness by reducing the max clique problem to the above maximization problem), it is not possible to efficiently optimize and hence they settle for approximations.
- Note that even for the small Karate Club network having only $m = 78$ edges, a brute force search would require trying 2^{78} solutions. Of course, there may be better methods for any specific network.
- The reduction preserves the approximation ratio, so it is also NP-hard to approximate the maximization problem with a factor of $n^{1-\epsilon}$. However, the minimization problem can can reduced (preserving approximations) to the vertex cover problem which can be approximated within s factor of 2.
- Their computational results are validated against the 5 networks where the strength of ties is known from the given data.

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 weights (resp. a community structure) are known.

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 $P_S = 1$ indicating that strong edges are wll 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.

Strong and weak ties in the karate club network

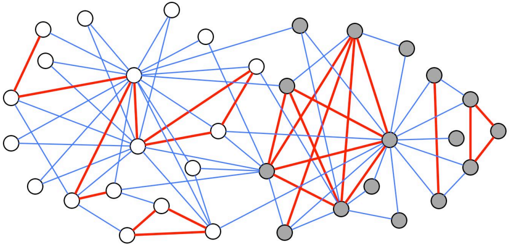


Figure 1: Karate Club graph. Blue light edges represent the weak edges, while red thick edges represent the strong edges.

- Note that all the strong links are within one of the two known communities and hence all links between the communities are weak links.