Great Ideas in Computing

University of Toronto CSC196
Fall 2021

Week 11 (part 2): November 29-December 3

Announcements and agenda for oend of week 11 and week 12

Announcements

- We hope to have everything graded by the end of the week.
- I know at least one person had to miss quiz 1. If you had some reason for missing any assignment or quiz, or it you had difficulty submitting to markus and submitted a little late, please send me that coprrespondence to be sure I have that properly noted.
- At the end of the Wednesday December 1 class, we discussed the Backstrom and Kleinberg study for discovering which edge represents a romantic relation. We include that discussion in this part of the slides.
- Next week we will then briefly discuss two other studies which again illustrate how graph structure can reveal interesting information.
- Following social networks we will introdue the topic of mechanism design in week 12.

Some additional comments on how graph structure can reveal personal and individual information: Detecting the romantic relation in Facebook

- There is an interesting paper by Backstrom and Kleinberg (http://arxiv.org/abs/1310.6753) on detecting "the" romantic relation in a subgraph of facebook users who specify that they are in such a relationship.
- Backstrom anbd Kleinberg construct two datasets of randomly sampled Facebook users: (i) an extended data set consisting of 1.3 million users declaring a spouse or relationship partner, each with between 50 and 2000 friends and (ii) a smaller data set extracted from neighbourhoods of the above data set (used for the more computationally demanding experimental studies).
- The main experimental results are nearly identical for both data sets.

Detecting the romantic relation (continued)

- They consider various graph strucutral features of edges, including
  1. the *embeddedness* of an edge (A, B) which is the number of mutual friends of A and B.
  2. various forms of a new *dispersion* measure of an edge (A, B) where high dispersion intuitively means that the mutual neighbours of A and B are not "well-connected" to each other (in the graph without A and B).
  3. One definition of dispersion given in the paper is the number of pairs (s, t) of mutual friends of u and v such that $(s, t) \notin E$ and s, t have no common neighbours except for u and v.
- They also consider various "interaction features" including
  1. the number of photos in which both A and B appear.
  2. the number of profile views within the last 90 days.

Embeddedness and disperison example from paper

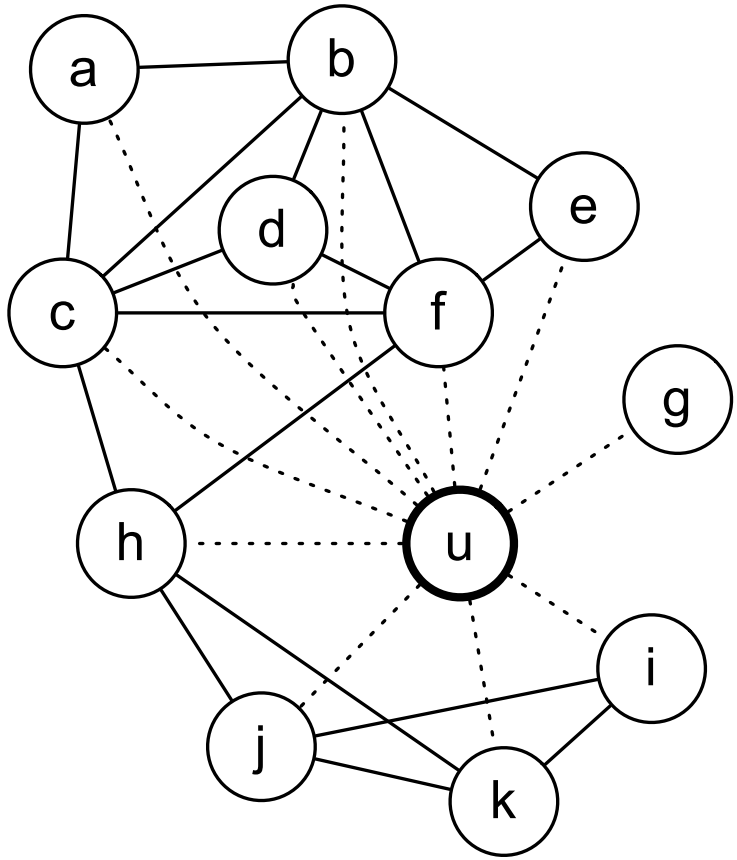


Figure 2. A synthetic example network neighborhood for a user u; the links from u to b, c, and f all have embeddedness 5 (the highest value in this neighborhood), whereas the link from u to h has an embeddedness of 4. On the other hand, nodes u and h are the unique pair of intermediaries from the nodes c and f to the nodes j and k; the u-h link has greater dispersion than the links from u to b, c, and f.

Qualitative results from Backstrom and Kleinberg

- The goal is to predict (for each user in the data set) which of their friendship edges is the romantic relation. Note that each user has between 50 and 2000 friends and assuming say a median of 200 users, a random guess would have prediction accuracy of $1/200 = .5\%$

Qualitative results from Backstrom and Kleinberg

- The goal is to predict (for each user in the data set) which of their friendship edges is the romantic relation. Note that each user has between 50 and 2000 friends and assuming say a median of 200 users, a random guess would have prediction accuracy of $1/200 = .5\%$
- Various disperson measures do better than the embeddedness measure in its ability to predict the correct romantic relationship. Why would high dispersion be a better measure than high embeddedness?

Qualitative results from Backstrom and Kleinberg

- The goal is to predict (for each user in the data set) which of their friendship edges is the romantic relation. Note that each user has between 50 and 2000 friends and assuming say a median of 200 users, a random guess would have prediction accuracy of $1/200 = .5\%$
- Various disperson measures do better than the embeddedness measure in its ability to predict the correct romantic relationship. Why would high dispersion be a better measure than high embeddedness?
- By itself, dispersion outperforms various interaction features.

Qualitative results from Backstrom and Kleinberg

- The goal is to predict (for each user in the data set) which of their friendship edges is the romantic relation. Note that each user has between 50 and 2000 friends and assuming say a median of 200 users, a random guess would have prediction accuracy of $1/200 = .5\%$
- Various disperson measures do better than the embeddedness measure in its ability to predict the correct romantic relationship. Why would high dispersion be a better measure than high embeddedness?
- By itself, dispersion outperforms various interaction features.
- For most measures, performance is better for male users and also better for data when restricted to marriage as the relationship.

Qualitative results from Backstrom and Kleinberg

- The goal is to predict (for each user in the data set) which of their friendship edges is the romantic relation. Note that each user has between 50 and 2000 friends and assuming say a median of 200 users, a random guess would have prediction accuracy of $1/200 = .5\%$
- Various disperson measures do better than the embeddedness measure in its ability to predict the correct romantic relationship. Why would high dispersion be a better measure than high embeddedness?
- By itself, dispersion outperforms various interaction features.
- For most measures, performance is better for male users and also better for data when restricted to marriage as the relationship.
- By combining many features, structural and interaction, the best performance is achieved using machine learning classification algorithms based on these many features.

Qualitative results from Backstrom and Kleinberg

- The goal is to predict (for each user in the data set) which of their friendship edges is the romantic relation. Note that each user has between 50 and 2000 friends and assuming say a median of 200 users, a random guess would have prediction accuracy of $1/200 = .5\%$
- Various disperson measures do better than the embeddedness measure in its ability to predict the correct romantic relationship. Why would high dispersion be a better measure than high embeddedness?
- By itself, dispersion outperforms various interaction features.
- For most measures, performance is better for male users and also better for data when restricted to marriage as the relationship.
- By combining many features, structural and interaction, the best performance is achieved using machine learning classification algorithms based on these many features.
- There are a number of other interesting observations but for me the main result is the predictive power provided by graph structure although there will generally be a limit to what can be learned solely from graph structure.

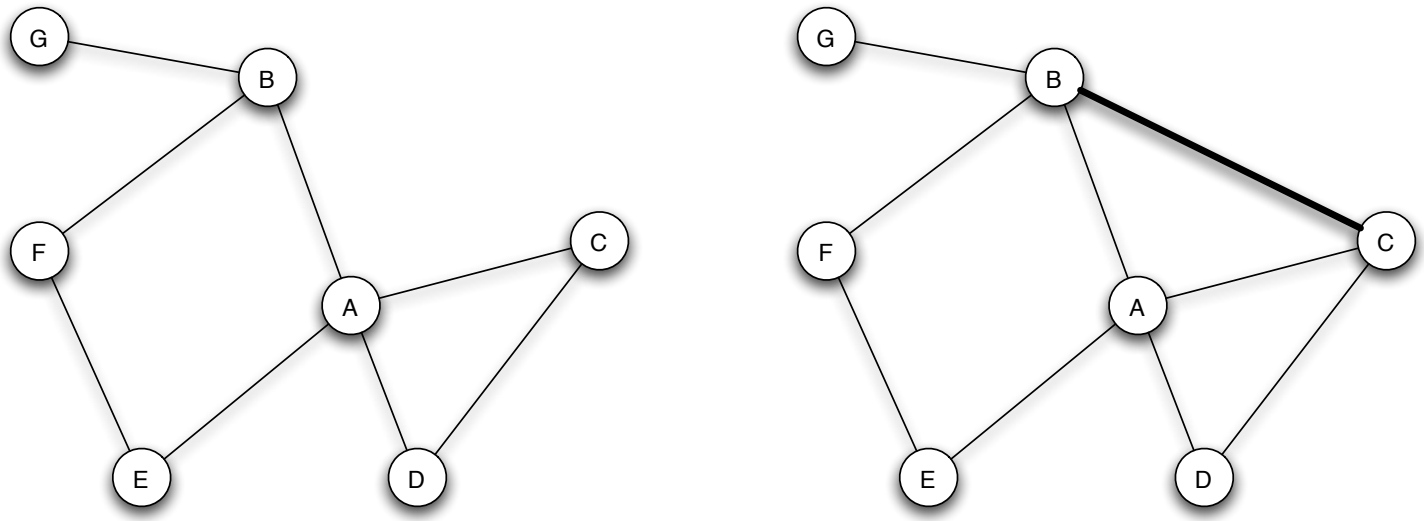
Some experimental results for the fraction of correct predictions

Recall that we argue that the fraction might be .005 when randomly choosing an edge. Do you find anything surprising?

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

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).

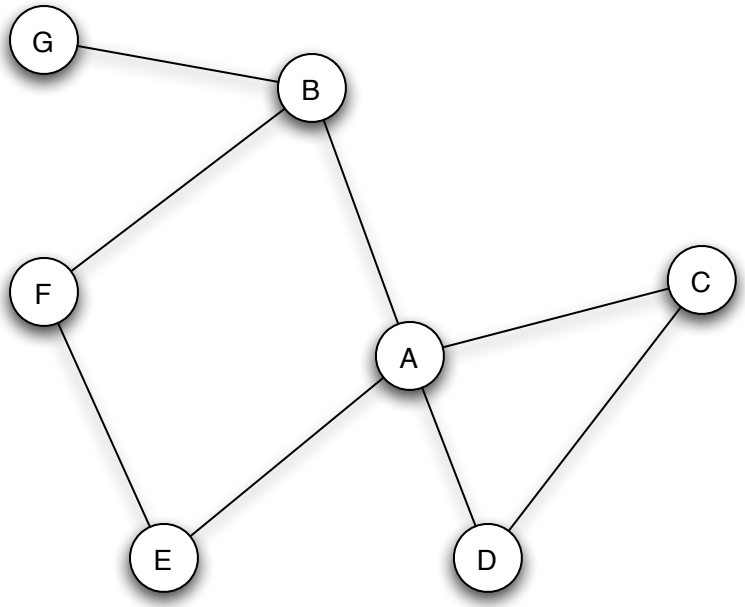
- Let E be the set of an undirected edges of a network graph. (Forgive the abuse of notation where in the previous and next slide E is a node name.) For a node A, the clustering coefficient is the following ratio:

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

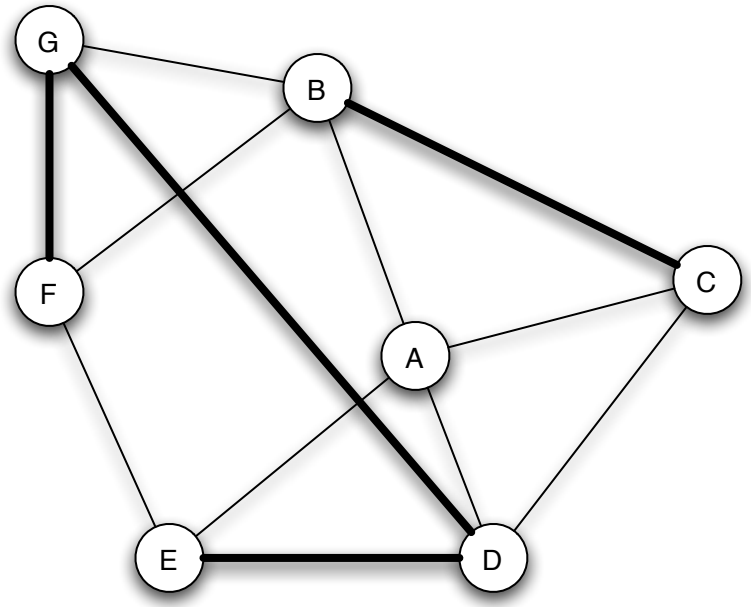
- The numerator is the number of all edges (B, C) in the network such that B and C are adjacent to (i.e. mutual friends of) A.

- 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 $\{D, E\}$). We sometimes refer to a pair of adjacent edges like $(A, B), (A, C)$ as an "open triangle" if (B, C) does not exist.
- 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 contemplating suicide (compared to those with high clustering coeficient). Note:The value of the clustering coefficient is also referred to as the *intransitivity coefficient*.

- They report that " Social network effects for girls overwhelmed other variables in the model and appeared to play an unusually significant role in adolescent female suicidality. These variables did not have a significant impact on the odds of suicidal ideation among boys. "

How can we understand these findings?

Bearman and Moody study continued

- Triadic closure (or lack thereof) can provide some plausible explanation.

Bearman and Moody study continued

- Triadic closure (or lack thereof) can provide some plausible explanation.
  Increased opportunity, trust, incentive ; it can be awkward to have friends (especially good friends with strong ties) who are not themselves friends.

Bearman and Moody study continued

- Triadic closure (or lack thereof) can provide some plausible explanation.
  Increased opportunity, trust, incentive ; it can be awkward to have friends (especially good friends with strong ties) who are not themselves friends.
  As far as I can tell, no conclusions are being made about why there is such a difference in gender results.

The study by Bearman and Moody is quite careful in terms of identifying many possible factors relating to suicidal thoughts. Clearly there are many factors involved but the fact that network structure is identified as such an important factor is striking.

Bearman and Moody factors relating to suicidal thoughts

TABLE 3—Logistic Regression of Suicide Attempts, Among Adolescents With Suicidal Ideation, on Individual, School, Family and Network Characteristics

	Suicide Attempts, OR (95% CI)	
	Males	Females
Demographic		
Age	0.956 (0.808, 1.131)	0.920 (0.810, 1.046)
Race/ethnicity		
Black	0.872 (0.414, 1.839)	1.086 (0.680, 1.736)
Other	1.069 (0.662, 1.728)	1.134 (0.810, 1.586)
Socioeconomic status	0.948 (0.872, 1.031)	1.008 (0.951, 1.069)
School and community		
Junior high school	1.588 (0.793, 3.180)	1.271 (0.811, 1.993)
Relative density	0.049 (0.005, 0.521)	0.415 (0.086, 1.996)
Plays team sport	0.985 (0.633, 1.532)	1.020 (0.763, 1.364)
Attachment to school	1.079 (0.823, 1.414)	1.066 (0.920, 1.235)
Religion		
Church attendance	0.975 (0.635, 1.496)	0.818 (0.618, 1.082)
Family and household		
Parental distance	0.925 (0.681, 1.256)	0.955 (0.801, 1.139)
Social closure	1.004 (0.775, 1.299)	0.933 (0.781, 1.115)
Stepfamily	1.058 (0.617, 1.814)	1.368 (0.967, 1.935)
Single-parent household	1.142 (0.698, 1.866)	1.117 (0.800, 1.560)
Gun in household	1.599 (1.042, 2.455)	1.094 (0.800, 1.494)
Family member attempted suicide	1.712 (0.930, 3.150)	1.067 (0.689, 1.651)
Network		
Isolation	0.767 (0.159, 3.707)	1.187 (0.380, 3.708)
Intransitivity index	0.444 (0.095, 2.085)	1.076 (0.373, 3.103)
Friend attempted suicide	1.710 (1.095, 2.671)	1.663 (1.253, 2.207)
Trouble with people	1.107 (0.902, 1.357)	1.119 (0.976, 1.284)
Personal characteristics		
Depression	1.160 (0.960, 1.402)	1.130 (0.997, 1.281)
Self-esteem	1.056 (0.777, 1.434)	0.798 (0.677, 0.942)
Drunkenness frequency	1.124 (0.962, 1.312)	1.235 (1.115, 1.368)
Grade point average	0.913 (0.715, 1.166)	0.926 (0.781, 1.097)
Sexually experienced	1.323 (0.796, 2.198)	1.393 (0.990, 1.961)
Homosexual attraction	1.709 (0.921, 3.169)	1.248 (0.796, 1.956)
Forced sexual relations		1.081 (0.725, 1.613)
No. of fights	0.966 (0.770, 1.213)	1.135 (0.983, 1.310)
Body mass index	0.981 (0.933, 1.032)	1.014 (0.982, 1.047)
Response profile (n = 1/n = 0)	139/493	353/761
F statistic	1.84 ($P = .0170$)	2.88 ($P < .0001$)

Note. OR = odds ratio; CI = confidence interval. Logistic regressions; standard errors corrected for sample clustering and stratification on the basis of region, ethnic mix, and school type and size.

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 be reduced (preserving approximations) to the vertex cover problem which can be approximated within a factor of 2.
- Their computational results are validated against the 5 networks where the strength of ties is known from the given data. Notably their worst case approximation algorithm (via the reduction) lead to reasonably good results achieved for the 5 real data networks.

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.

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 equivalently minimize the number of weak ties.)

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

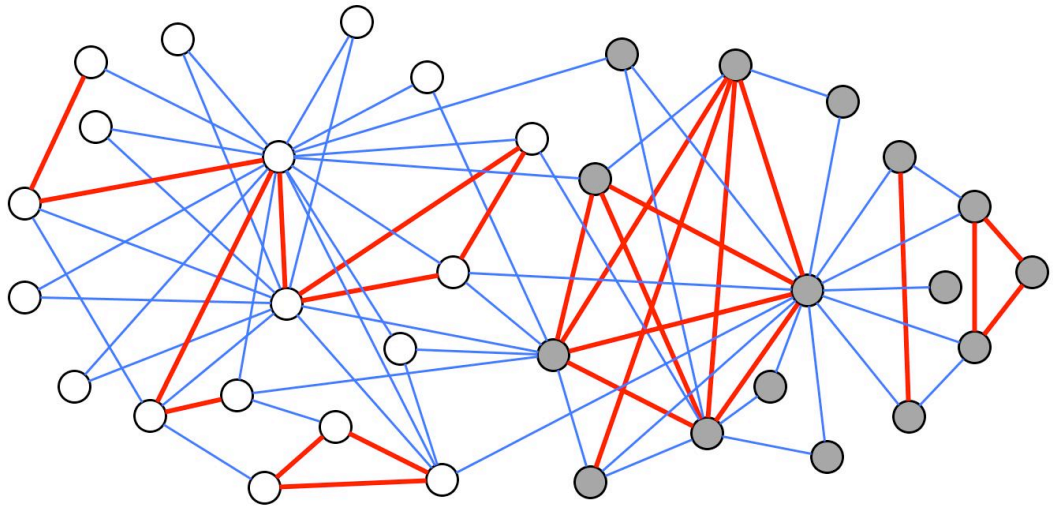


Figure 1: Karate Club graph. Blue light edges represent the weak edges, while red thick edges represent the strong edges.

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