# Social and Information Networks

University of Toronto CSC303
Winter/Spring 2024

Week 2: Jan 15-19

# Survey Results: Hello! It's a pleasure to meet you :)

- Hi! :D
- Hi!!
- Hi!
- hi
- hi :)
- hello! :)
- Hi :)
- hello!

# Office Hours

- Based on the survey results:
  - In-person+Zoom, Wed. 5-6pm, BA2272 (Note: This room is *inside* of BA2270, the Help Centre)
  - Zoom-only, Thursday 10-11pm
  - As always, by appointment

At which of the following times would you be able to attend in-person office hours?

If things go well, these office hours will also be accessible via Zoom.

| | | |
|---|---|---|
| **Monday 17:00-18:00** | 45 respondents | 68 % |
| Wednesday 17:00-18:00 | 50 respondents | 76 % |
| No Answer | 8 respondents | 12 % |

# Goals and Interests

- Far too many to summarize on a slide!
- I'm happy to be teaching a diverse bunch; with interests ranging encompassing pure curiosity, ethics & social justice, graph theory, research, applied applications, to just finishing up their degree :)
- This class is very much introductory, so we'll be touching upon a little bit of everything briefly

# Goals and Interests

- Far too many to summarize on a slide!
- I'm happy to be teaching a diverse bunch; with interests ranging encompassing pure curiosity, ethics & social justice, graph theory, research, applied applications, to just finishing up their degree :)
- This class is very much introductory, so we'll be touching upon a little bit of everything briefly
- Sadly, game theory will not *really* be covered, see CSC304 though if you're interested! :)
- Similarly, the course doesn't have a dedicated ethics component (though it should!), and we won't be covering computer networks (though we will look at webpages and hyperlinks)

# Zoom & Recordings

- Everyone loves 'em, from a fall-back position, to conflicting work obligations or co-op terms, to studying and less stressful notetaking, etc...
- I will do my utmost to keep Zoom access & recordings going

If lectures were made available on Zoom, would you be interested or likely to use this modality?

| | | |
|---|---|---|
| Yes | 37 respondents | 50 % |
| Maybe | 8 respondents | 11 % |
| Only if unable to attend lecture | 25 respondents | 34 % |
| No | 1 respondent | 1 % |
| No Answer | 3 respondents | 4 % |

# Zoom & Recordings

- Everyone loves 'em, from a fall-back position, to conflicting work obligations or co-op terms, to studying and less stressful notetaking, etc...
- I will do my utmost to keep Zoom access & recordings going

If lectures were made available on Zoom, would you be interested or likely to use this modality?

| | | |
|---|---|---|
| Yes | 37 respondents | 50 % |
| Maybe | 8 respondents | 11 % |
| Only if unable to attend lecture | 25 respondents | 34 % |
| No | 1 respondent | 1 % |
| No Answer | 3 respondents | 4 % |

- "The audio quality on zoom and the recordings are better in the moments when you talk at a lower volume [...]"
  - Yes, I think I need to adjust the gain on the room mic, I'll try tweaking it and see what happens

# Suggestions for next year(s)?

- "Upload the course note please"
    - There are no course notes, sorry :(
        - ★ Try the slides, and textbook

# Suggestions outside my control

- "it is very helpful that tutorial can also be available via zoom"
  - I'm sorry, we'll have OCCS recordings, but the tutorials won't be available via Zoom

# Suggestions I will do my utmost

- "I wish there will be more hints and help when it comes to assignments."
  - I try to provide hints where needed, but I makes mistakes too. Head on down to office hours! I'm always happy to chat about what you've tried so far :)
- "[...] more practice for course material [...]", "liked when previous courses include a few simple practice questions for lectures that help with test preparation or overall understanding."
  - Most weeks there is an optional Quercus Quiz, done in tutorial if there's time
  - There'll be practice questions posted and taken up in tutorial – I can post them earlier?
  - Beyond that, I can work towards increasing the number of practice questions, but it takes time :(
  - Check the textbook for more practice questions! It won't cover everything, but it'll cover most topics

# Suggestions I will do my utmost

- “clear lectures were the most helpful [...] When things are explained clearly and at at a digestible pace”
  - Absolutely – if you have any feedback on pacing, if it’s too fast or too slow on a particular day or subject, then please, please, please do let me know
- “assignments be independent of each other and their handouts be enough to guide on what all to code for”
  - Assignments are independent of each other, there’s no coding (sorry)

# Mon. Jan 15th: Announcements & Corrections

- A0 Q6b has been corrected to include symmetry as part of the definition of a positive semi-definite matrix (therefore, any eigenvalues will be real). You don't need to consider complex values for this question.
- Tutorials start this Friday :)

# This week's high-level learning goals

By the end of the week you'll be able to:

- Explain the concept of the strength of weak ties, and its connection with Strong Triadic Closure (STC) and local bridges

# This week's high-level learning goals

By the end of the week you'll be able to:

- Explain the concept of the strength of weak ties, and its connection with Strong Triadic Closure (STC) and local bridges
- Define triadic closure
  - Measure it with the clustering coefficient
  - Explain its driving forces

# This week's high-level learning goals

By the end of the week you'll be able to:

- Explain the concept of the strength of weak ties, and its connection with Strong Triadic Closure (STC) and local bridges
- Define triadic closure
  - Measure it with the clustering coefficient
  - Explain its driving forces
- Recall Granovetter's Thesis
  - Explain strong & weak ties, and apply to scenarios
  - Define and identify (local) bridges
  - Define Strong Triadic Closure, and prove its implications
  - Define neighbourhood overlap and explain its connection to local bridges

# This week's high-level learning goals

By the end of the week you'll be able to:

- Explain the concept of the strength of weak ties, and its connection with Strong Triadic Closure (STC) and local bridges
- Define triadic closure
  - Measure it with the clustering coefficient
  - Explain its driving forces
- Recall Granovetter's Thesis
  - Explain strong & weak ties, and apply to scenarios
  - Define and identify (local) bridges
  - Define Strong Triadic Closure, and prove its implications
  - Define neighbourhood overlap and explain its connection to local bridges
- Define social capital
  - Define, recognize, differentiate, and produce examples of bonding and bridging capital
  - Explain its relationship with graph structure

# This week's high-level learning goals

By the end of the week you'll be able to:

- Explain the concept of the strength of weak ties, and its connection with Strong Triadic Closure (STC) and local bridges
- Define triadic closure
  - Measure it with the clustering coefficient
  - Explain its driving forces
- Recall Granovetter's Thesis
  - Explain strong & weak ties, and apply to scenarios
  - Define and identify (local) bridges
  - Define Strong Triadic Closure, and prove its implications
  - Define neighbourhood overlap and explain its connection to local bridges
- Define social capital
  - Define, recognize, differentiate, and produce examples of bonding and bridging capital
  - Explain its relationship with graph structure
- Predict strong edges in a graph
  - Execute the Sintos & Tsaparas algorithm
  - Execute Rozenshtein algorithm
  - Compare these algorithms

# 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.

# 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).

# Chapter 3: Strong and Weak Ties

There are two themes that run throughout this chapter.

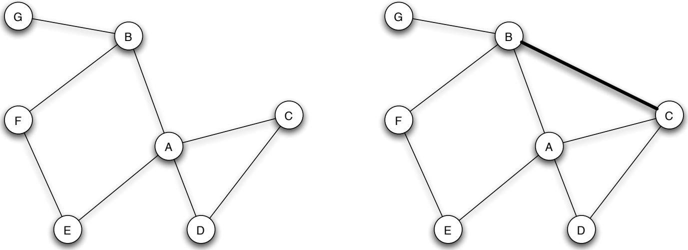
1. Strong vs. weak ties and "the strength of weak ties" is the specific defining theme of the chapter. The chapter also starts a discussion of how networks evolve.

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)

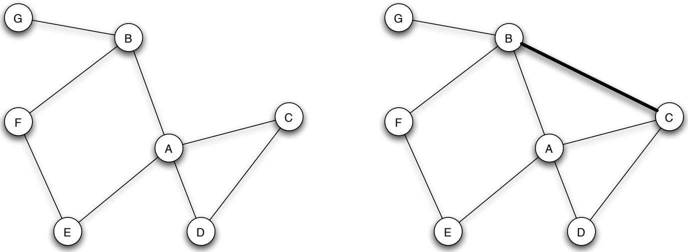



(a) *Before B-C edge forms.* (b) *After B-C edge forms.*

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

- Triadic closure: mutual "friends" of say $A$ are more likely (than "normally") to become friends over time.

# 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 neighbour $A$. [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).

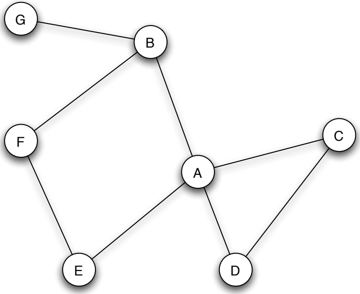
- Let $E$ be the set of an undirected edges of a network graph. (Forgive the abuse of notation where in the previous and next slide $E$ is a node name.) For a node $A$, the clustering coefficient is the following ratio:

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

- 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 $A$.

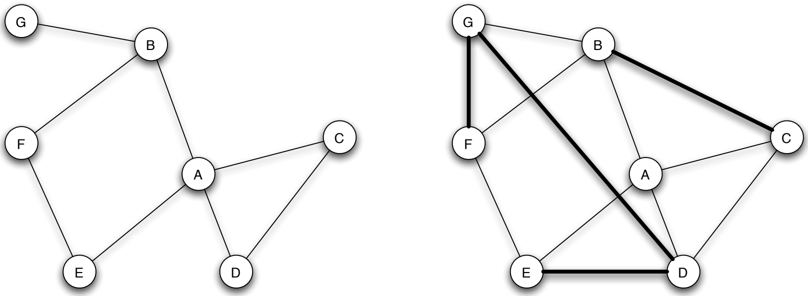
# Example of clustering coefficient




(a) *Before new edges form.*

- 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 $\{D, E\}$)

# 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 $3/6$ (because there are three edges $(B, C)$, $(C, D)$, and $(D, E)$).

# Driving forces behind Triadic Closure

- Social psychology suggests: Increased opportunity, incentive, and trust

# Driving forces behind Triadic Closure

- Social psychology suggests: Increased opportunity, incentive, and trust

- It also predicts that having friends (especially good friends with strong ties) who are not themselves friends causes *latent stress*

# 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 among teenage girls implies a higher probability of contemplating suicide (compared to those with high clustering coefficient). 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 2—Logistic Regression of Suicidal Ideation on Individual, School, Family, and Network Characteristics**

| | Suicide Ideation Among Adolescents, OR (95% CI) | |
|---|---|---|
| | Males | Females |
| Demographic | | |
| Age | 1.031 (0.951, 1.118) | 0.885 (0.830, 0.944) |
| Race/ethnicity | | |
| Black | 0.864 (0.628, 1.187) | 0.873 (0.685, 1.112) |
| Other | 1.079 (0.852, 1.367) | 1.190 (0.986, 1.436) |
| Socioeconomic status | 1.017 (0.979, 1.057) | 1.000 (0.970, 1.031) |
| School and community | | |
| Junior high school | 1.281 (0.938, 1.751) | 0.808 (0.637, 1.023) |
| Relative density | 1.061 (0.375, 2.999) | 0.333 (0.142, 0.783) |
| Plays team sport | 0.831 (0.685, 1.008) | 1.164 (0.999, 1.357) |
| Attachment to school | 0.994 (0.891, 1.109) | 0.952 (0.871, 1.041) |
| Religion | | |
| Church attendance | 0.822 (0.683, 0.989) | 1.008 (0.863, 1.176) |
| Family and household | | |
| Parental distance | 1.573 (1.361, 1.818) | 1.743 (1.567, 1.939) |
| Social closure | 0.904 (0.805, 1.015) | 1.012 (0.921, 1.111) |
| Stepfamily | 1.101 (0.870, 1.394) | 0.998 (0.821, 1.212) |
| Single-parent household | 1.212 (0.959, 1.533) | 1.119 (0.930, 1.345) |
| Gun in household | 1.329 (1.083, 1.630) | 1.542 (1.288, 1.848) |
| Family member attempted suicide | 2.136 (1.476, 3.092) | 1.476 (1.120, 1.943) |
| Network | | |
| Isolation | 0.665 (0.307, 1.445) | 2.010 (1.073, 3.765) |
| Intransitivity index | 0.747 (0.358, 1.558) | 2.198 (1.221, 3.956) |
| Friend attempted suicide | 2.725 (2.187, 3.395) | 2.374 (2.019, 2.791) |
| Trouble with people | 0.999 (0.912, 1.095) | 1.027 (0.953, 1.106) |
| Personal characteristics | | |
| Depression | 1.632 (1.510, 1.765) | 1.445 (1.348, 1.549) |
| Self-esteem | 0.811 (0.711, 0.925) | 0.808 (0.730, 0.894) |
| Drunkenness frequency | 1.112 (1.041, 1.187) | 1.114 (1.039, 1.194) |
| Grade point average | 1.061 (0.948, 1.188) | 0.993 (0.905, 1.089) |
| Sexually experienced | 1.201 (0.972, 1.484) | 0.993 (0.823, 1.198) |
| Homosexual attraction | 1.385 (1.015, 1.891) | 1.544 (1.155, 2.063) |
| Forced sexual relations | | 1.873 (1.435, 2.445) |
| No. of fights | 1.017 (0.924, 1.120) | 1.142 (1.046, 1.246) |
| Body mass index | 1.004 (0.983, 1.026) | 1.027 (1.010, 1.044) |
| Response profile (n = 1/n = 0) | 632/5867 | 1114/5852 |
| F statistic | 17.08 ($P < .0001$) | 16.28 ($P < .0001$) |

*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.

# Recap

- Triadic closure
  - Definition
  - Clustering coefficient
  - Driving forces

# 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?

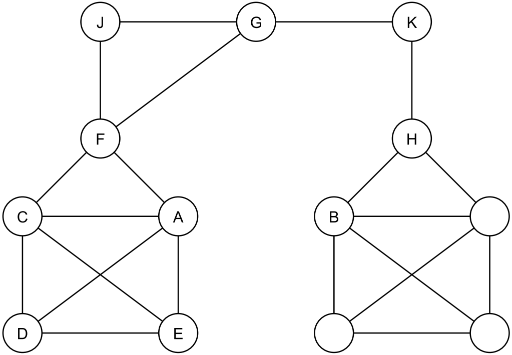
- To gain some understanding of this phenomena, we need some additional concepts relating to structural properties of a graph.

**Recall**

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

# Bridges

- 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.



# Local bridges

- A bridge is an edge whose removal disconnects a connected component in a graph.
  - We expect large social networks to have few bridges.

# Local bridges

- A bridge is an edge whose removal disconnects a connected component in a graph.
  - We expect large social networks to have 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.

# Local bridges

- A bridge is an edge whose removal disconnects a connected component in a graph.
  - We expect large social networks to have 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.
  - 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 bridges

- A bridge is an edge whose removal disconnects a connected component in a graph.
  - We expect large social networks to have 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.
  - 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.
  - By convention, nodes with no path between them have infinite span
    - ★ Question: Are bridges also local bridges?

# Local bridges

- A bridge is an edge whose removal disconnects a connected component in a graph.
    - We expect large social networks to have 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.
    - 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.
    - By convention, nodes with no path between them have infinite span
        - Question: Are bridges also local bridges?
- Aside: Span can be used to define *dispersion measures* (see the Backstrom and Kleinberg article regarding Facebook relations). Specifically, we can use the span between mutual friends of $A$ and $B$ when the nodes $A$ and $B$ are removed from the graph.

# Local bridge $(A, B)$

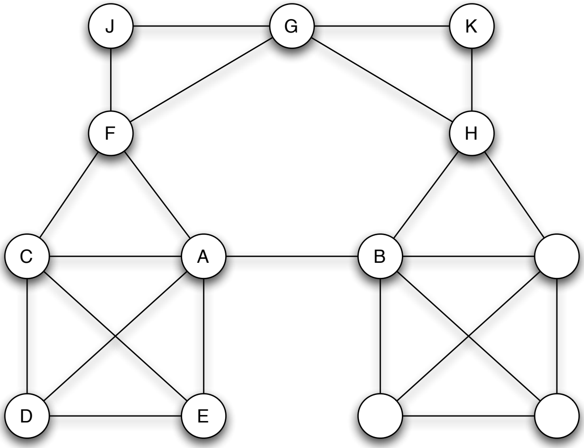



**Figure:** The edge $(A, B)$ is a local bridge of span 4, since the removal of this edge would increase the distance between $A$ and $B$ to 4. [E&K Figure 3.4]

# 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.

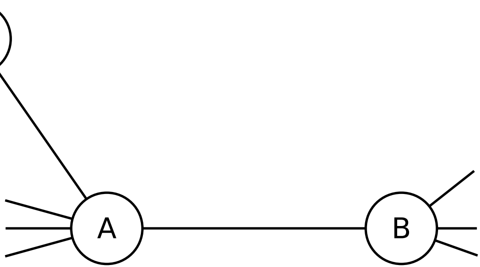
**Theorem**

*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.

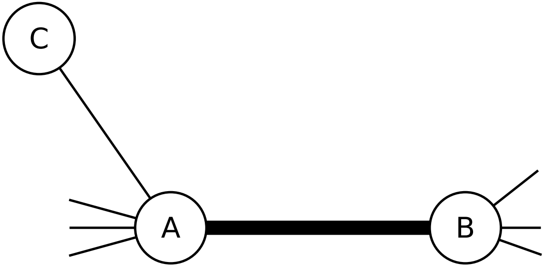
# Triadic closure and local bridges

- Assume Strong Triadic Closure (STC) holds on the graph.
- Let $A$ by any node involved in at least two strong edges and a local bridge. Let $(A, B)$ be a local bridge.



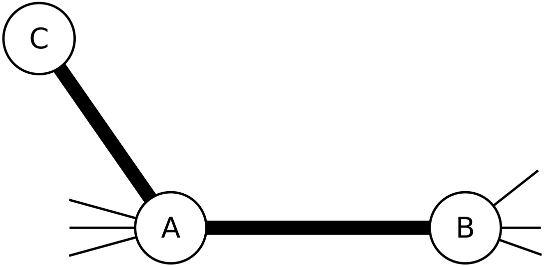
# Triadic closure and local bridges

- Let's assume for contradiction that $(A, B)$ is strong



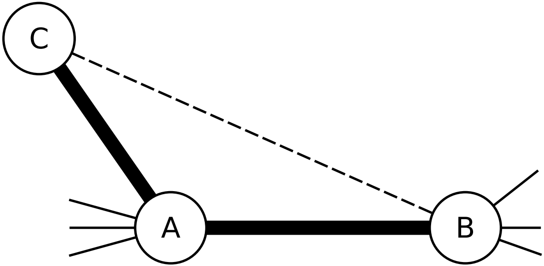
# Triadic closure and local bridges

- Let's assume for contradiction that $(A, B)$ is strong



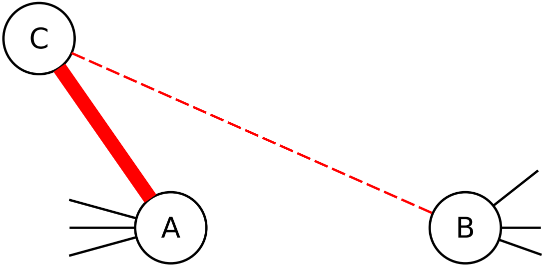
# Triadic closure and local bridges

- Let's assume for contradiction that $(A, B)$ is strong



# Triadic closure and local bridges

- Let's assume for contradiction that $(A, B)$ is strong



# 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, ..."

# 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 (STC) property can help in determining whether a link is a strong or weak tie.

  `www.cs.uoi.gr/~tsap/publications/frp0625-sintos.pdf`

  We will discuss this paper later in the lecture.

# 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 (STC) property can help in determining whether a link is a strong or weak tie.
  `www.cs.uoi.gr/~tsap/publications/frp0625-sintos.pdf`

  We will discuss this paper later in the lecture.
- Later we'll discuss Rozenshtein et al [2019]. They assume the existence of *known communities*, and then their goal is to label all edges so as minimize the number of *open triangles violating the STC property* subject to all communities being connected using only strong edges.
  - This work is inspired by the Sintos and Tsaparas [2014] results for inferring the strength of ties, and an earlier [2013] paper by Angluin et al for minimizing the number of edges needed to maintain "communities"

# Wed. Jan 17th: Announcements & Corrections

- Office hours today after lecture – BA2272
- The first tutorial is this Friday!
- First practice quiz will be released this Friday – you'll get a reminder, and (time permitting) the chance to work on it together, in tutorial

# 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 romantic relationship (especially for marriage) than embeddedness?

# Large scale experiment relating tie strength and bridges

- 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.

# Large scale experiment relating tie strength and bridges

- 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.

# Large scale experiment relating tie strength and bridges

- 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.

# Large scale experiment relating tie strength and bridges

- 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.

# Large scale experiment relating tie strength and bridges

- 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 neighbourhood overlap of an edge $(A, B)$ defined as the ratio

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

- Question: What does a neighbourhood overlap of zero mean?

# Large scale experiment relating tie strength and bridges

- 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 neighbourhood overlap of an edge $(A, B)$ defined as the ratio

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

- Question: What does a neighbourhood overlap of zero mean? Local bridge!

# Large scale experiment relating tie strength and bridges

- 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 neighbourhood overlap of an edge $(A, B)$ defined as the ratio

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

- Question: What does a neighbourhood overlap of zero mean? Local bridge!
- Question: What relationship would we expect between tie-strength & neighbourhood overlap?

# Onnela et al. experiment

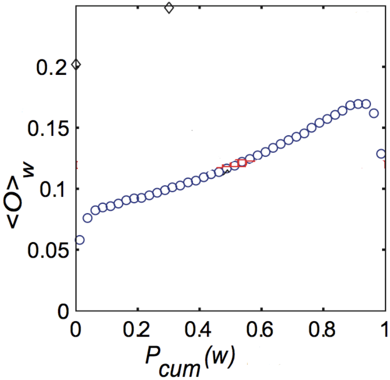



**Figure:** A plot of the neighbourhood overlap of edges as a function of their percentile in the sorted order of all edges by tie strength. [E&K Fig 3.7]

- 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 overlap 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.

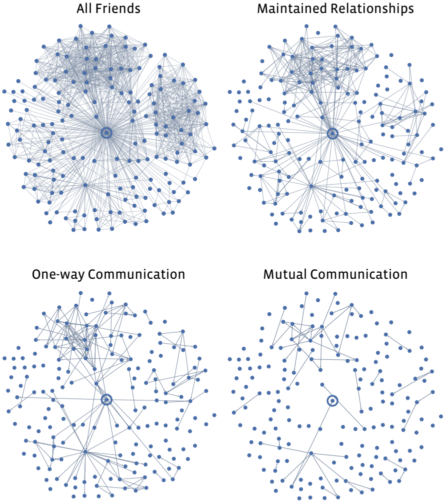
# Strong vs. weak ties in large online social networks (Facebook and Twitter)

- 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.

# 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

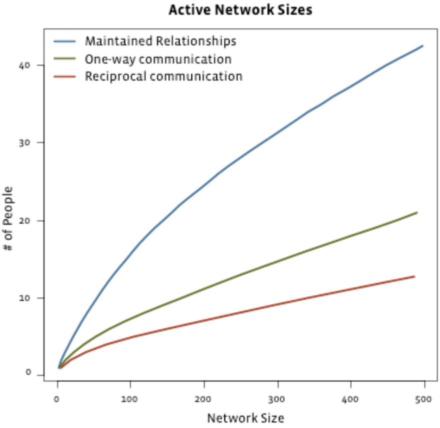



**Figure:** The number of links corresponding to maintained relationships, one-way communication, and reciprocal communication as a function of the total neighbourhood size for users on Facebook. [Figure 3.9, textbook]

# Twitter:Limited Strong Ties vs Followers

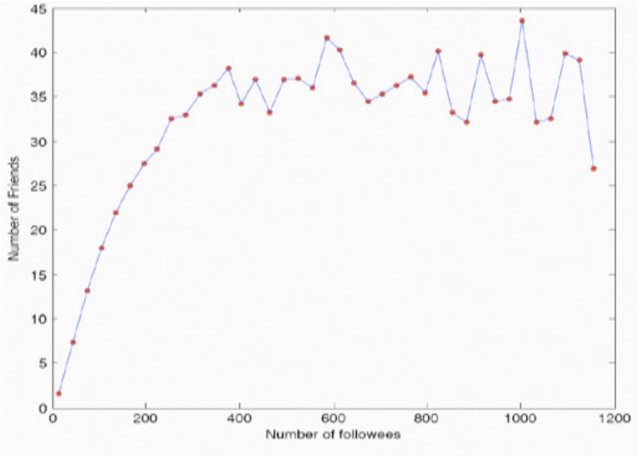



**Figure:** The total number of a user's strong ties (defined by multiple directed messages) as a function of the number of followees he or she has on Twitter. [Figure 3.10, textbook]

# 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.

# 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.

  What's so special about these maintained relationships?

# 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.

  What's so special about these maintained relationships?

- "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 neighbourhood."

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

# Recap

- Granovetter's Thesis
  - Strong & Weak Ties
  - Bridges
  - Strong Triadic Closure and it's implications

# Social capital (as discussed in section 3.5 of EK text)

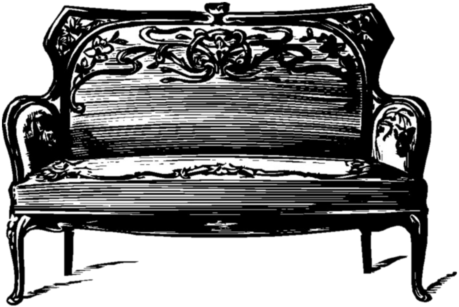
Social capital is a term in increasingly widespread use, but it is a famously difficult one to define.

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

# Social capital (as discussed in section 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, cultural, physical etc...) all of which serve as tangible or intangible resources that can be mobilized to accomplish tasks.



# Social capital (as discussed in section 3.5 of EK text)

A 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 group's members — or whether it is also based on the interactions of the group with the outside world.

A person can have more or less social capital depending on his or her position in the underlying social structure or network.

# “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

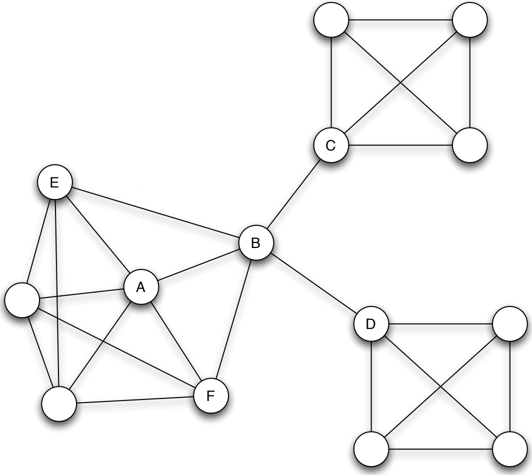
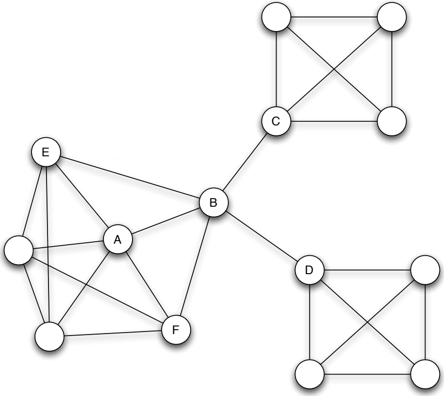


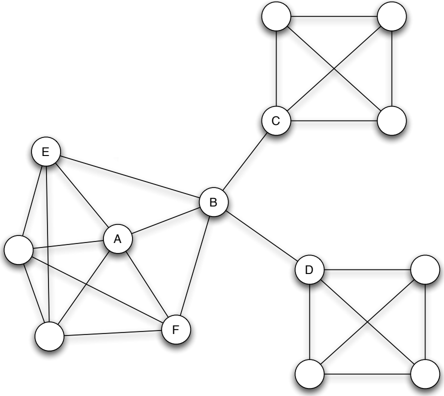

**Figure:** The contrast between densely-knit groups and boundary-spanning links is reflected in the different positions of central node *A* and gatekeeper node *B* in the underlying social network. [Fig 3.11, textbook]

# Social capital of nodes $A$ and $B$



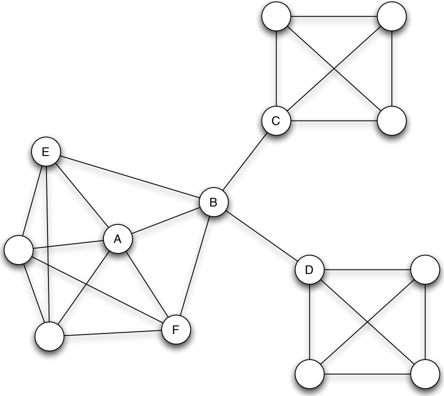

- The edges adjacent to node $A$ all have high embeddedness. Visually, $A$ is a central node in a tightly-knit cluster.
  - The social capital that $A$ has is "bonding capital": the actions of $A$ can (for example) induce norms of behaviour because of the trust in $A$.

# Social capital of nodes $A$ and $B$




- In contrast to $A$, node $B$ is a bridge to other parts of the network.
  - its social capital is in the form of "brokerage" or "bridging capital": $B$ can play the role of a "gatekeeper" (of information and ideas) between different parts of the network, and this position can lead to creativity stemming from the synthesis of ideas.

# Social capital of nodes $A$ and $B$




- In contrast to $A$, node $B$ is a bridge to other parts of the network.
  - its social capital is in the form of "brokerage" or "bridging capital": $B$ can play the role of a "gatekeeper" (of information and ideas) between different parts of the network, and this position 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.

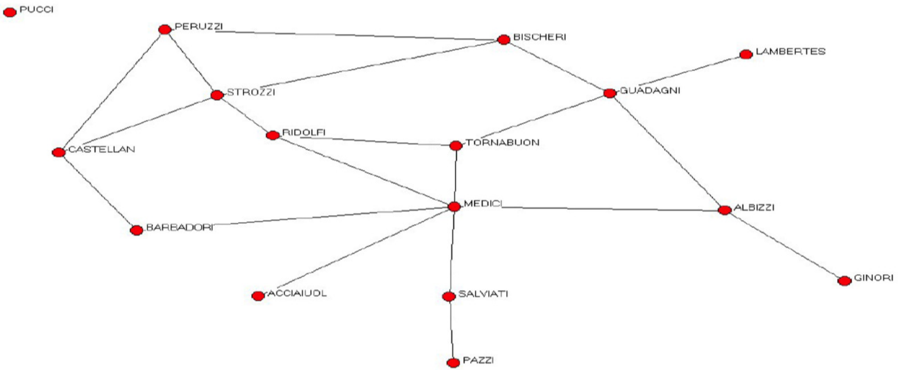
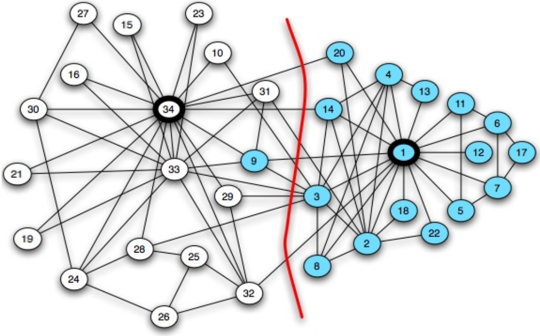


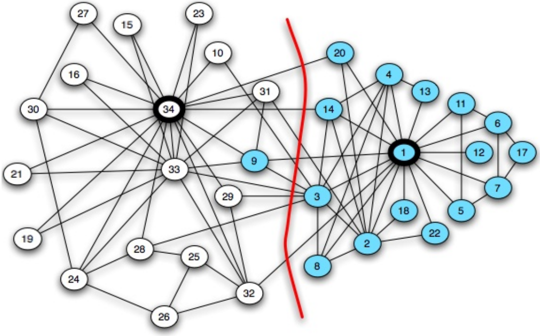

**Figure:** see [Jackson, Ch 1]

# 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?

# 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.

# Recap

- Social Capital
  - Definition
  - Bonding vs. Bridging capital
  - Relationship with graph structure

# Mon. Jan 22nd: Announcements & Corrections

- Some questions: Tutorial slides are on the course website, and the OCCS recordings are up
  - You may need to clear caches & reload the page if you don't see slides on the course site
- Assignment 0 is due this Thursday

# 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.

# 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

- They proved that we cannot efficiently solve this problem, and must settle for approximations.
  - Aside: They showed *NP* hardness by reducing the max clique problem to the above maximization problem.

# The computational problem in identifying strong vs weak ties

- They proved that we cannot efficiently solve this problem, and must settle for approximations.
  - Aside: They showed $NP$ hardness by reducing the max clique problem to the above maximization problem.
- Note that while the Karate Club network has only $m = 78$ edges, a brute force search would require trying $2^{78} \approx 3 \times 10^{23}$ solutions!

# The computational problem in identifying strong vs weak ties

- They proved that we cannot efficiently solve this problem, and must settle for approximations.
  - Aside: They showed *NP* hardness by reducing the max clique problem to the above maximization problem.
- Note that while the Karate Club network has only $m = 78$ edges, a brute force search would require trying $2^{78} \approx 3 \times 10^{23}$ solutions!
- They also showed that we cannot efficiently approximate the maximization problem within a fixed factor of the best solution, but we can efficiently approximate the minimization problem within a fixed factor!
  - Aside: The reduction preserves the approximation ratio, so it is *NP*-hard to approximate the maximization problem with a factor of $n^{1-\epsilon}$.
  - Aside: The minimization problem can be reduced (preserving approximations) to the vertex cover problem, which can be approximated within a factor of 2.

# The vertex cover algorithms and the 5 data sets

- Their computational results are validated against the 5 networks where the strength of ties is known from the given data.
  - Their approach works by converting the problem into a vertex cover problem. (You'll see this in tutorial!)

# The vertex cover algorithms and the 5 data sets

- Their computational results are validated against the 5 networks where the strength of ties is known from the given data.
  - Their approach works by converting the problem into a vertex cover problem. (You'll see this in tutorial!)
  - They tested several vertex cover approximation algorithms
  - Notably their worst case approximation algorithm lead to reasonably good results for the 5 real data networks.

# The vertex cover algorithms and the 5 data sets

- Their computational results are validated against the 5 networks where the strength of ties is known from the given data.
  - Their approach works by converting the problem into a vertex cover problem. (You'll see this in tutorial!)
  - They tested several vertex cover approximation algorithms
  - Notably their worst case approximation algorithm lead to reasonably good results for the 5 real data networks.
    - Just because an approximation algorithm has a bad worst case, it doesn't mean that said worst case is likely to appear in your 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 (Aside: worst case $O(\log n)$ approximation ratio). The maximal matching algorithm repeatedly finds an uncovered edge and takes both endpoints of that edge (Aside: 2-approximation).

**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.

Aside: Despite the Greedy algorithm having 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 average link weight.

Question: Is there a problem with average edge strength?

# 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 average link weight.

Question: Is there a problem with average edge strength? Easy to skew average if weights have high variance

# 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:** Using a normalized edge weight based on activity

$$w((a,b)) = \frac{\mathrm{works}(a) \cap \mathrm{works}(b)}{\mathrm{works}(a) \cup \mathrm{works}(b)} \in [0,1]$$

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

# 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 all within a community.

# Some observations on the predictions

- 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.

# Some observations on the predictions

- 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, edges are co-purchases, and there are three communities corresponding to liberal, neutral, conservative viewpoints. Of the strong edges predicted, only 22 cross communities:
  - 20 cross-community strong edges have one node labelled as neutral.
  - the rest are between books dealing with the same issue.

# Strong and weak ties in the karate club network

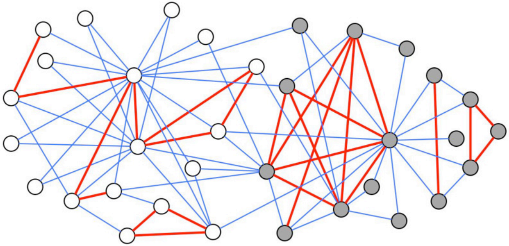


**Figure 1: Karate Club graph. Blue light edges represent the weak edges, while red thick edges represent the strong edges.**

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

# The Rozenshtein et al study

- The Rozenshtein et al. approach assumes a known set of communities (in addition to the unlabelled network).
  - Therefore not directly comparable to Sintos-Tsaparas study!
  - Informally, want a good labelling that preserves the given communities – i.e. communities being strongly connected using strong ties.

# The Rozenshtein et al study

- The Rozenshtein et al. approach assumes a known set of communities (in addition to the unlabelled network).
  - Therefore not directly comparable to Sintos-Tsaparas study!
  - Informally, want a good labelling that preserves the given communities – i.e. communities being strongly connected using strong ties.

- More specifically, each community should be a connected subgraph using only strong edges.
  - i.e., should be a path using only strong edges between any two nodes in a community, and every node in the path is also in the community.

# The Rozenshtein et al study

- The Rozenshtein et al. approach assumes a known set of communities (in addition to the unlabelled network).
  - Therefore not directly comparable to Sintos-Tsaparas study!
  - Informally, want a good labelling that preserves the given communities – i.e. communities being strongly connected using strong ties.
- More specifically, each community should be a connected subgraph using only strong edges.
  - i.e., should be a path using only strong edges between any two nodes in a community, and every node in the path is also in the community.
- 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 (the goal of Sintos and Tsaparas) and only preserving strong connectivity within communities (the goal of Angluin et al.).

# The Karate club figure in Rozenshtein et al

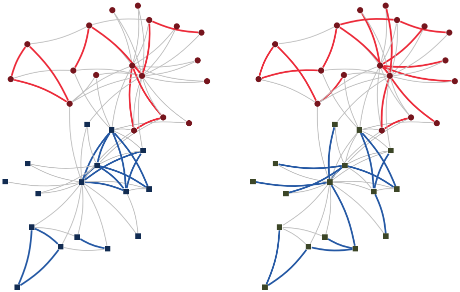


**Figure 1: Strong edges in the Karate-club dataset inferred by the algorithm of Sintos and Tsaparas [27] (left) and our method (right) using two teams. The colors of the edges and the vertices depict the two teams.**

Note: the vertices are coloured 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 ensure the community connectivity constraint).

# Rozenshtein et al. objective and a greedy algorithm

- Rozenshtein et al. objective is to minimize the number of STC violations subject to the constraint that every user-specified community remains connected using only strong ties (Aside: an NP-hard problem).

# Rozenshtein et al. objective and a greedy algorithm

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

# Rozenshtein et al. objective and a greedy algorithm

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

*********************************

Start with all edges labelled as strong.
Find an edge $e \in E$ that we haven't previously considered, and that is causing the most STC violations (i.e., whose removal would minimize the number of STC violations). If there are no violations then we're done.
Else, if making the edge weak would violate the community constraint then the edge stays strong and we never again consider this edge.
Otherwise the edge becomes weak and we never again consider this edge.
*********************************

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

---

# Rozenshtein et al objective and a greedy algorithm

- More rigorous pseudo-code can be found below, where $vio : \mathcal{P}(E) \rightarrow \mathbb{N}$ 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 2** Greedy Rozenshtein Algorithm

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

- Equivalently, choose $e \in A$ responsible for the most STC violations

# 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 earlier (maximize strong edges while satisfying STC).

# Understanding the table of results in Rozenshtein

- By design, Angluin et al. and Rozenshtein et al. ensure 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.

# Understanding the table of results in Rozenshtein

- By design, Angluin et al. and Rozenshtein et al. ensure 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 ensures no STC violations and hence $b_S = 0$ whereas $b$ is not 0 but is perhaps surprisingly small.

# Understanding the table of results in Rozenshtein

- By design, Angluin et al. and Rozenshtein et al. ensure 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 ensures 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 Rozenshtein et al. ensure 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 ensures 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 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 "betweeness" 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 betweeness are used to partition the graph into smaller components and eventually communities. They describe the Givan-Newman algorithm for identifying edges of high betweeness.

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 similarities (where a similarity function is known).

# This week's high-level learning goals

With practice, you'll be able to:

- Explain the concept of the strength of weak ties, and its connection with Strong Triadic Closure (STC) and local bridges

# This week's high-level learning goals

With practice, you'll be able to:

- Explain the concept of the strength of weak ties, and its connection with Strong Triadic Closure (STC) and local bridges
- Define triadic closure
  - Measure it with the clustering coefficient
  - Explain its driving forces

# This week's high-level learning goals

With practice, you'll be able to:

- Explain the concept of the strength of weak ties, and its connection with Strong Triadic Closure (STC) and local bridges
- Define triadic closure
  - Measure it with the clustering coefficient
  - Explain its driving forces
- Recall Granovetter's Thesis
  - Explain strong & weak ties, and apply to scenarios
  - Define and identify (local) bridges
  - Define Strong Triadic Closure, and prove its implications
  - Define neighbourhood overlap and explain its connection to local bridges

# This week's high-level learning goals

With practice, you'll be able to:

- Explain the concept of the strength of weak ties, and its connection with Strong Triadic Closure (STC) and local bridges
- Define triadic closure
  - Measure it with the clustering coefficient
  - Explain its driving forces
- Recall Granovetter's Thesis
  - Explain strong & weak ties, and apply to scenarios
  - Define and identify (local) bridges
  - Define Strong Triadic Closure, and prove its implications
  - Define neighbourhood overlap and explain its connection to local bridges
- Define social capital
  - Define, recognize, differentiate, and produce examples of bonding and bridging capital
  - Explain its relationship with graph structure

# This week's high-level learning goals

With practice, you'll be able to:

- Explain the concept of the strength of weak ties, and its connection with Strong Triadic Closure (STC) and local bridges
- Define triadic closure
  - Measure it with the clustering coefficient
  - Explain its driving forces
- Recall Granovetter's Thesis
  - Explain strong & weak ties, and apply to scenarios
  - Define and identify (local) bridges
  - Define Strong Triadic Closure, and prove its implications
  - Define neighbourhood overlap and explain its connection to local bridges
- Define social capital
  - Define, recognize, differentiate, and produce examples of bonding and bridging capital
  - Explain its relationship with graph structure
- Predict strong edges in a graph
  - Execute the Sintos & Tsaparas algorithm
  - Execute Rozenshtein algorithm
  - Compare these algorithms