Great Ideas in Computing

University of Toronto CSC196
Fall 2022

Week 12: December 5 - December 7

Announcements

Announcements

- As you know, this is the last week of classes for the fall term. We will not avail ourselves of the makeup Monday class as I am sure everyone would rather have the time to study for exams.
- Please promptly submit any regrade requests for the quiz. In particular if you received any deduction for first normalizing the values in the fair division question, please resubmit.
- I have some quizzes that were not picked up last Wednesday. Any unclaimed quizzes after the last class this Wednesday (December 7) will be given to the undergraduate office.
- The final assignment is due December 5 at 8AM. I am sure that Vignesh will grades them as fast as possible. After Assignment 4 is graded, I will calculate the participation grade and then calculate final grades. I will announce (on Quercus) these grades before submitting. Pleaase notify (by email) within two days if you notice any clerical errors. I am supposed to submit final grades within a week of the end of classes for courses not having a final exam.

This weeks impossible agenda

- We left off having introduced the preferential attachment model for network formation and proven or observed properties for such models. In particular, we mentioned the Barabasi and Albert model.
- We will briefly repeat those slides.
- The main theme today will be how graph structure can reveal personal and individual information as well as communities. In particular, we will discuss
- Floretine marriages and "centrality". Why were the Medici's so influential?
- The Bearman et al study of romatic relations in a US high school which we mentioned briefly before.
- The Backstrom and Kleinberg method for discovering *the* romantic relation in a subgraph of facebook.
- Bearman and Moody discussion of low triadic closure
- Modelling and understanding the small worlds phenomena. The Watts-Strogaatz, and Kleinberg models and analysis.
- More realistic georgraphic models.
- Extending geographic distance to social distance. :q

Looking ahead: The punch line in the Kleinberg-Easley text, and a major theme in the study of social networks (and this course)

The plots in Figure 20.10, and their follow-ups, are thus the conclusion of a sequence of steps in which we start from an experiment (Milgram's), build mathematical models based on this experiment (combining local and long-range links), make a prediction based on the models (the value of the exponent controlling the long-rang links), and then validate this prediction on real data (from LiveJournal and Facebook, after generalizing the model to use rank-based friendship). This is very much how one would hope for such an interplay of experiments, theories, and measurements to play out. But it is also a bit striking to see the close alignment of theory and measurement in this particular case, since the predictions come from a highly simplified model of the underlying social network, yet these predictions are approximately borne out on data arising from real social networks.

[From E&K Ch.20, p.549]

The Barabasi and Albert preferential model

Barabasi and Albert [1999] specified a particular preferential attachment model and conjectured that the vertex degrees satisfy a power law in which the fraction of nodes having degree d is proportional to d^{-3}.

They obtained $\gamma \approx 2.9$ by experiments and gave a simple heuristic argument suggesting that $\gamma = 3$. That is, $P(d)$ is proportional to d^{-3}

Bollobas et al [2001] prove a result corresponding to this conjectured power law. Namely, they show for all $d \leq n^{1/15}$ that the *expected* degree distribution is a power law distribution with $\gamma = 3$ asymptotically (with n) where n is the number of vertices.

Note: It is known that an actual realized distribution may be far from its expectation, However, for small degree values, the degree distribution is close to expectation.

When we say that a distribution $P(d)$ is a power law distribtion this is often meant to be a "with high probaility" whereas many results for networks generated by a preferential attachment process the power law is usually only in expectation.

Proven or observed properties of nodes in a social network generated by preferential attachment models

In addition to the power law phenomena suggesting many nodes with high degree, other properies of social networks have been obseerved such as a relatively large number of nodes u having some or all of properties such as the following: .

- high clustering coefficient defined as : $\frac{(u,v),(u,w),(v,w)\in E}{(u,v),(u,w)\in E}$. That is, mutual friends of u are likely to be friends.
- high centrality ; e,g, nodes on many pairs of shortest paths.

Brautbar and Kearns refer to such nodes (as above) as "interesting indiviudals" and these individuals might be candidates for being "highly influential individuals". Bonato et al [2015] refers to such nodes as the *elites* of a social network

Florentine marriages and "centrality"

- Medici connected to more families, but not by much
- More importantly: lie between most pairs of families
  - shortest paths between two families: coordination, communication
  - Medici lie on 52% of all shortest paths; Guadagni 25%; Strozzi 10%

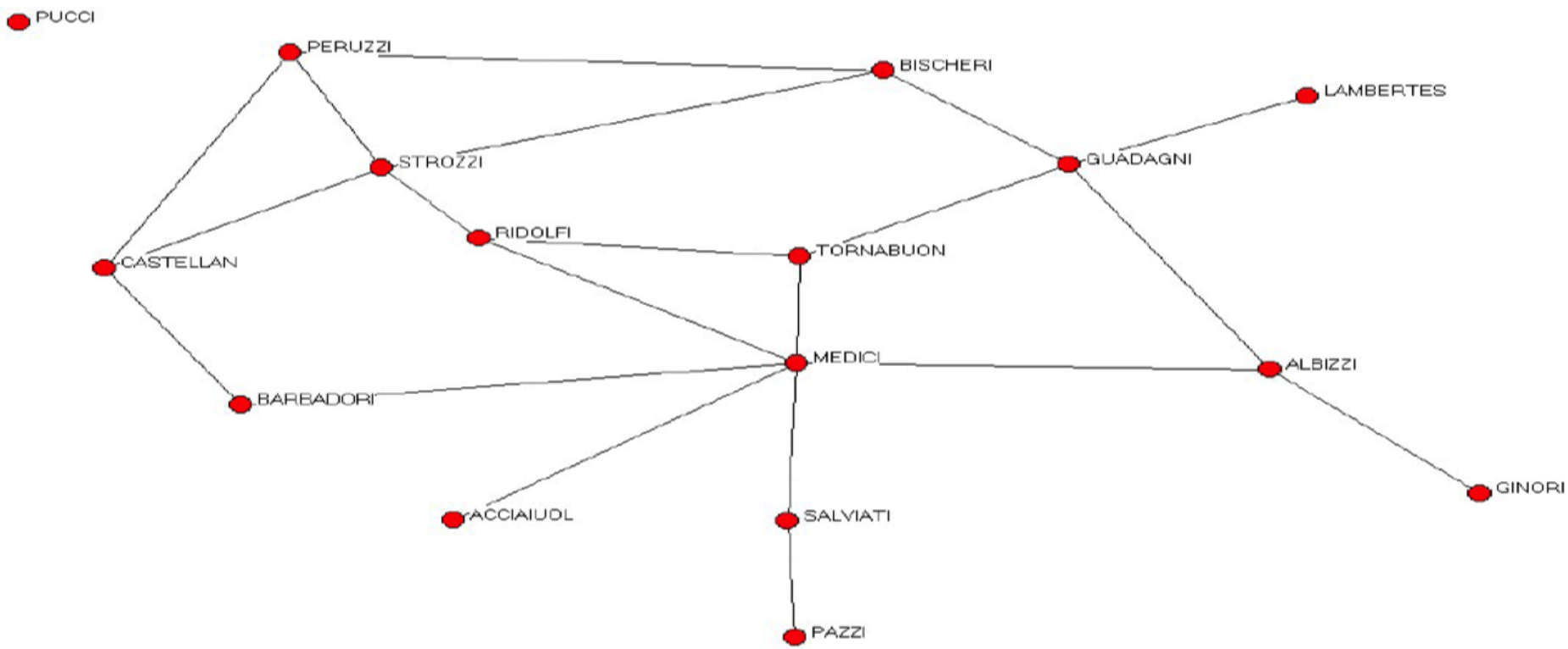


Figure: see [Jackson, Ch 1]

Other proven or observed properties of networks generated by preferentical attachment models

- correlation between the degree of a node u and the degrees of the neighboring nodes.
- the graph has small diameter; suggesting "6 degrees of separation phenomena"
- relatively large dense subgraph communities.
- rapid mixing (for random walks to approach stationary distribution)
- relatively small (almost) *dominating sets*. What do we mean by "almost"?

On my spring 2020 CSC303 web page, I posted a paper by Avin et al (2018) that shows that preferential attachment is the *only* "rational choice" for players (people) playing a simple natural network formation game. It is the rational choice in the sense that the strategy of the players will lead to a unique equilibrium (i.e. no player will want to deviate assumming other players do not deviate). For those intersted, I have posted (in my CSC303 webpage) a number of other papers on elites in a social network and preferential attachment.

The Small World Phenomena

I already mentioned the small worlds phenomena. A mathematical explanation of this phenomena (expecially how one hones in on a target recipient) was given by J. Kleinberg in a network formation model that explicitly forces a power law property.

The small world phenomena suggests that in a connected social network any two individuals are likely to be connected (i.e. know each other indirectly) by a short path. Moreover, such a path can be found in a decentralized manner

Later in these slides we will discuss other power laws with respect to the Kleinberg model and extensions.

Romantic Relationships [Bearman et al, 2004]

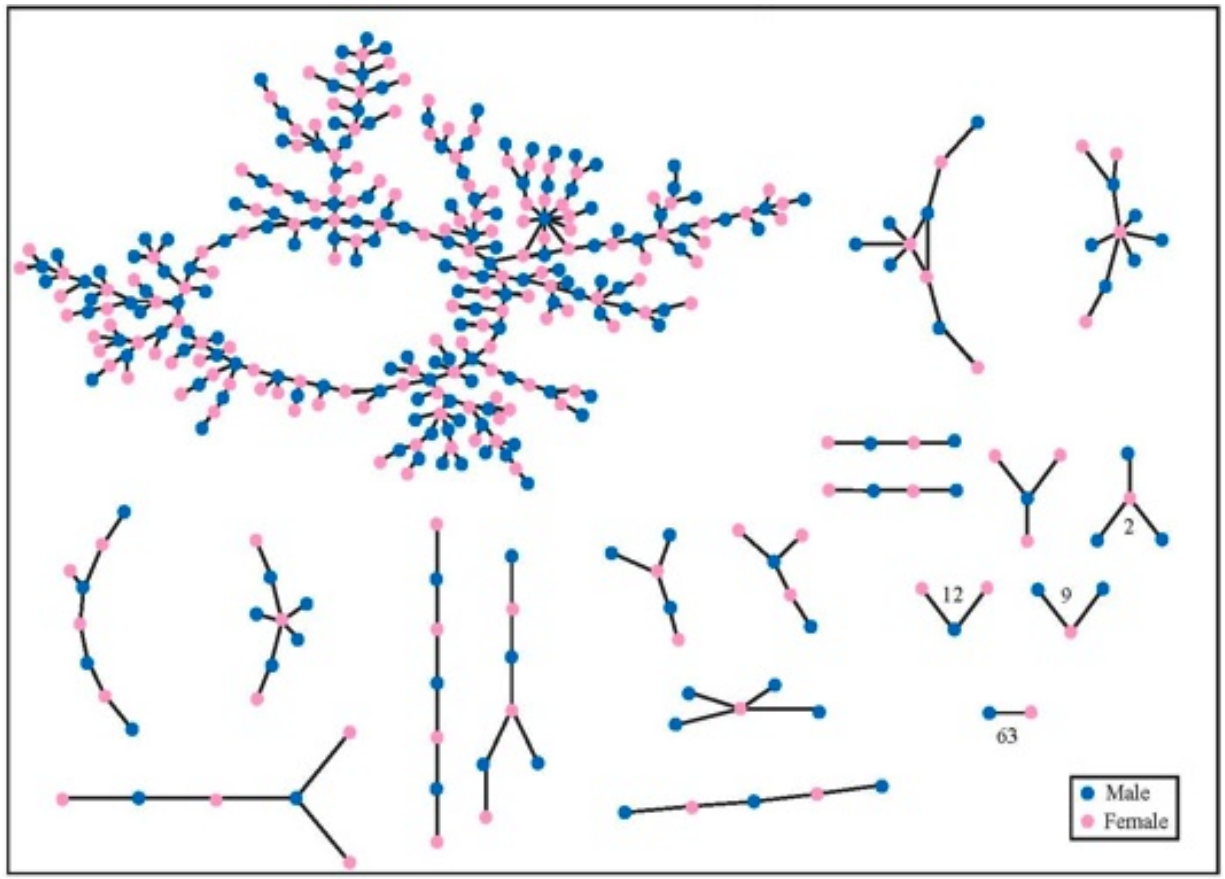
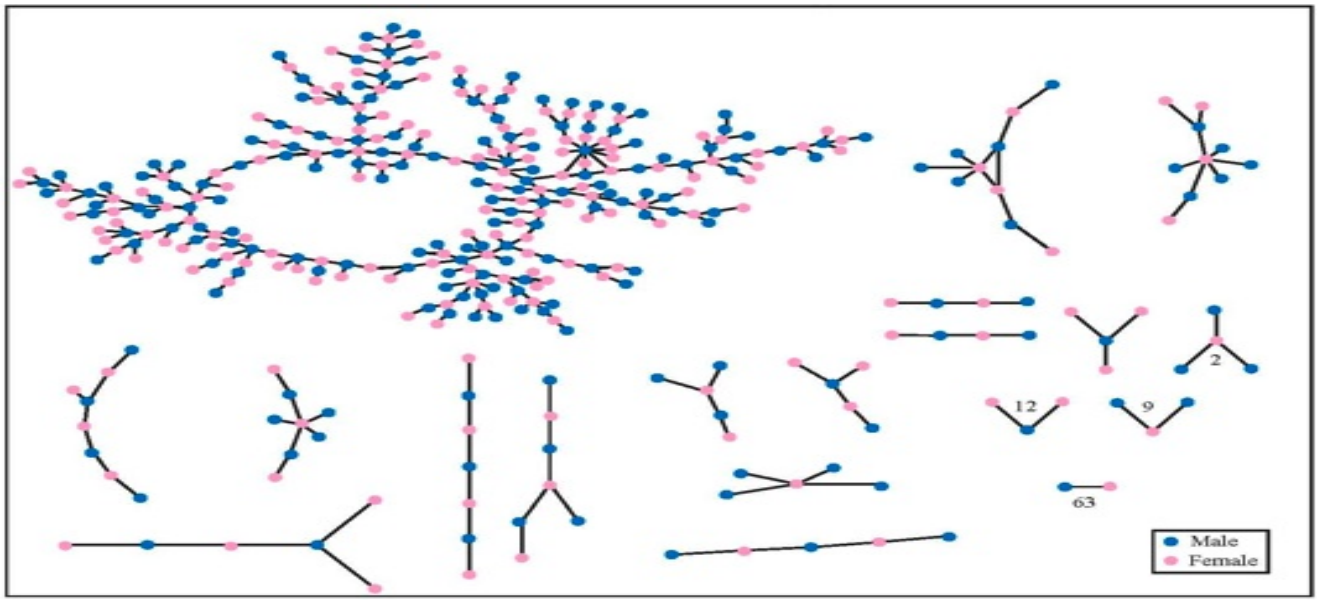


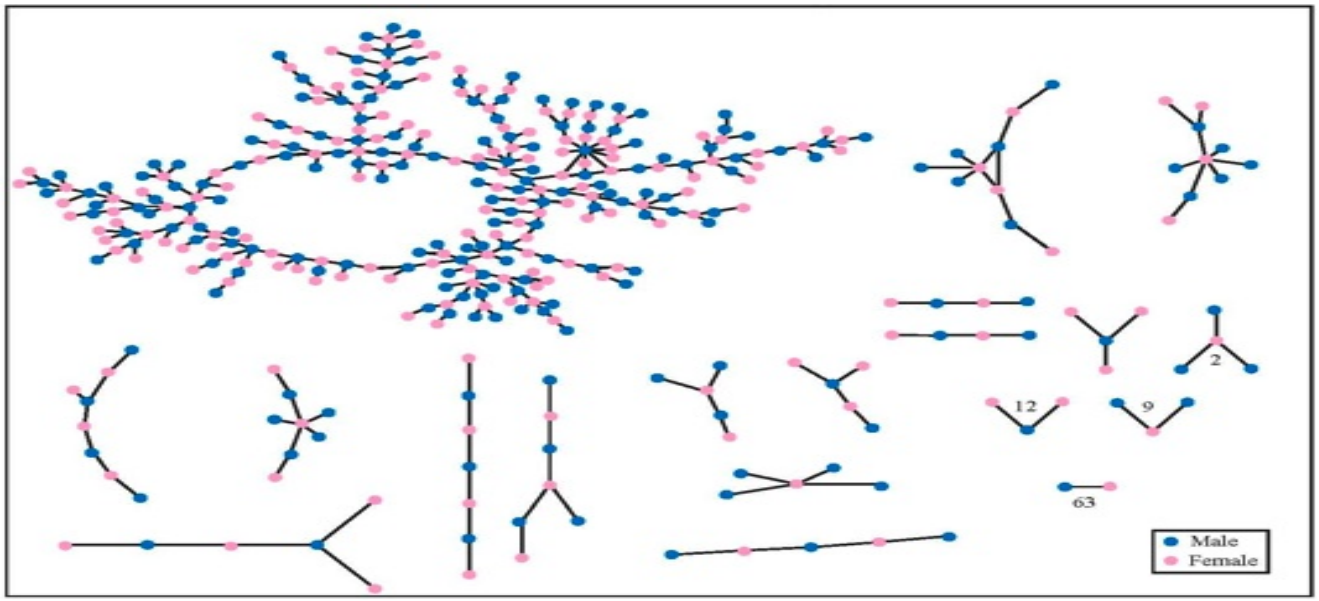
Figure: Dating network in US high school over 18 months.

- Illustrates common "structural" properties of many networks
- What predictions could you use this for?

More basic definitions



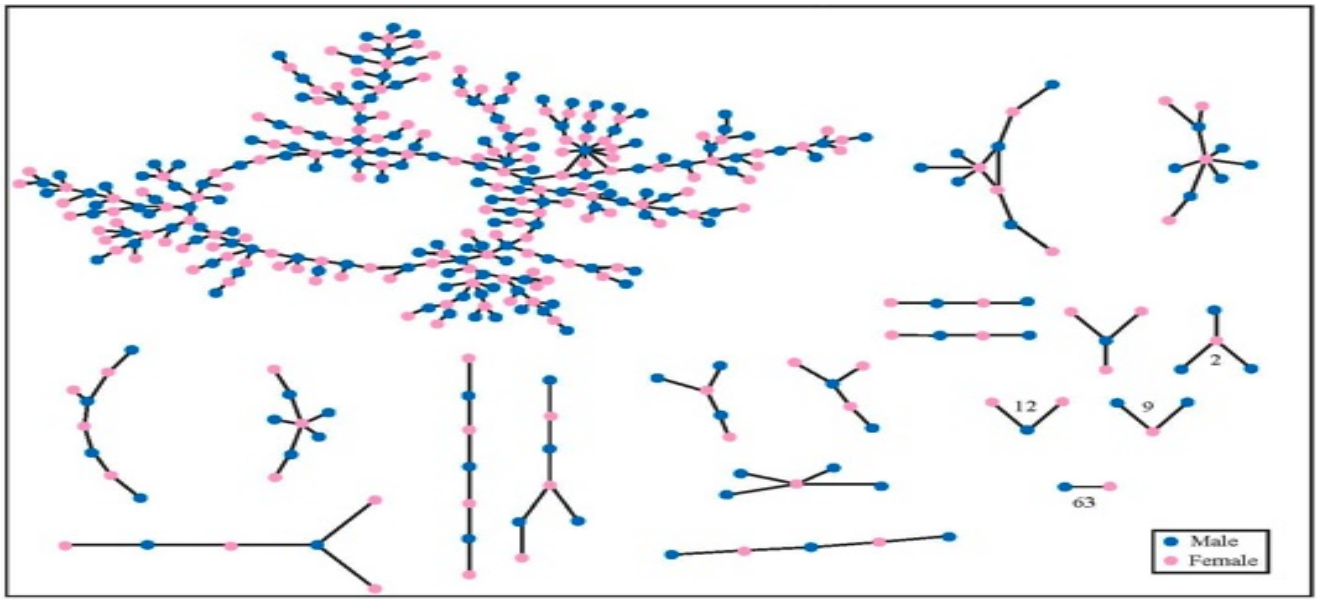
More basic definitions



Observation

Many connected components including one "giant component"

More basic definitions



Observation

Many connected components including one "giant component"

- We will use this same graph to illustrate some other basic concepts.
- A cycle is path $u_1, u_2, \ldots, u_k$ such that $u_1 = u_k$; that is, the path starts and ends at the same node.

More 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

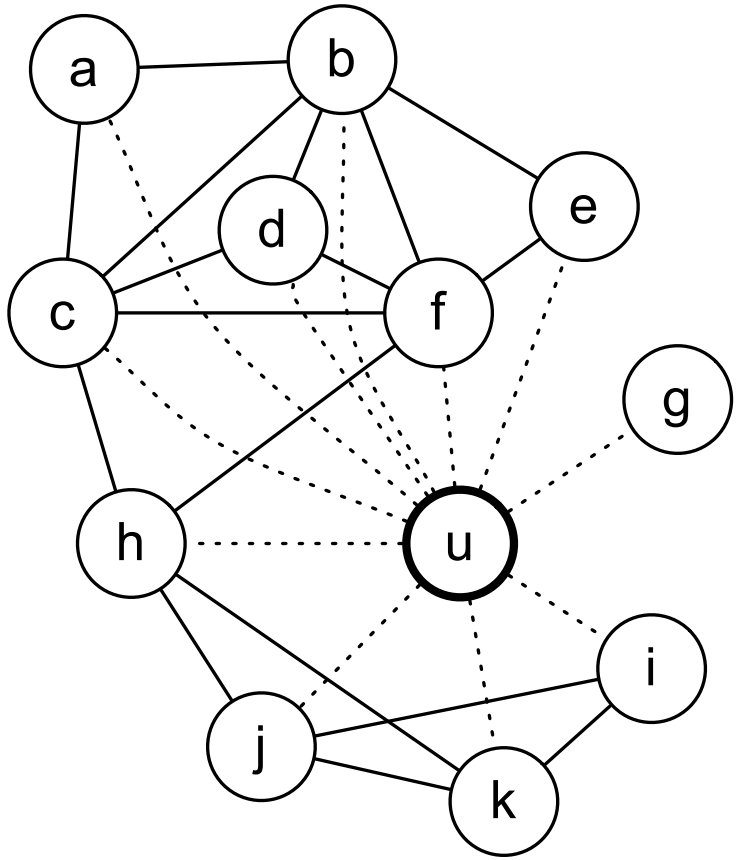


Figure 2. A synthetic example network neighborhood for a user u; the links from u to b, c, and f all have embeddedness 5 (the highest value in this neighborhood), whereas the link from u to h has an embeddedness of 4. On the other hand, nodes u and h are the unique pair of intermediaries from the nodes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

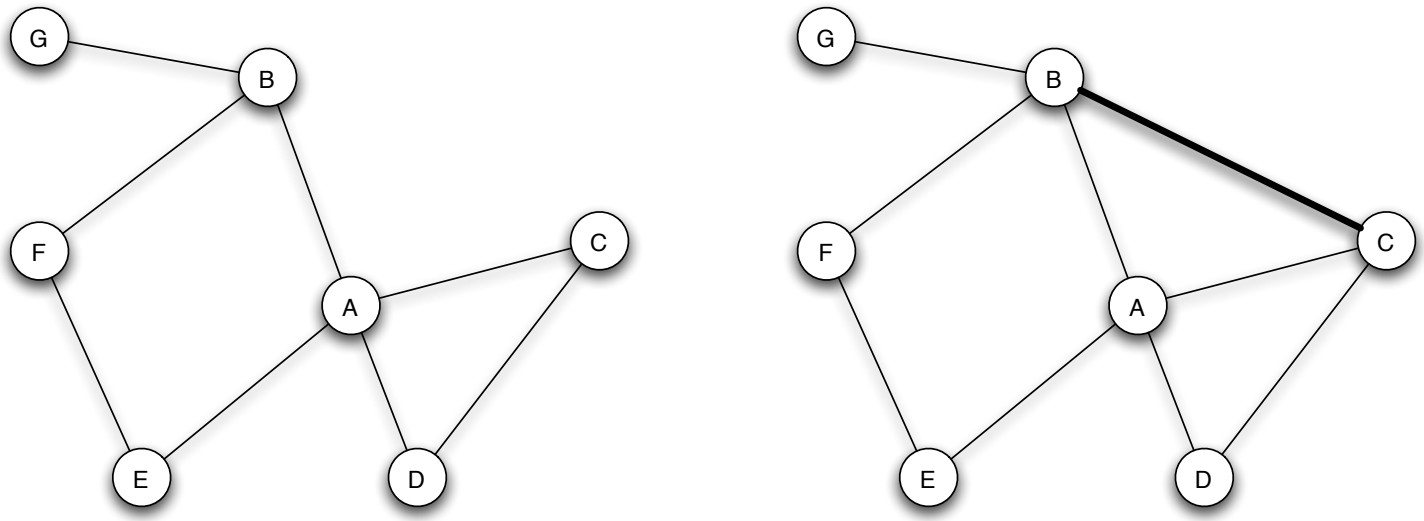
Some experimental results for the fraction of correct predictions

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).

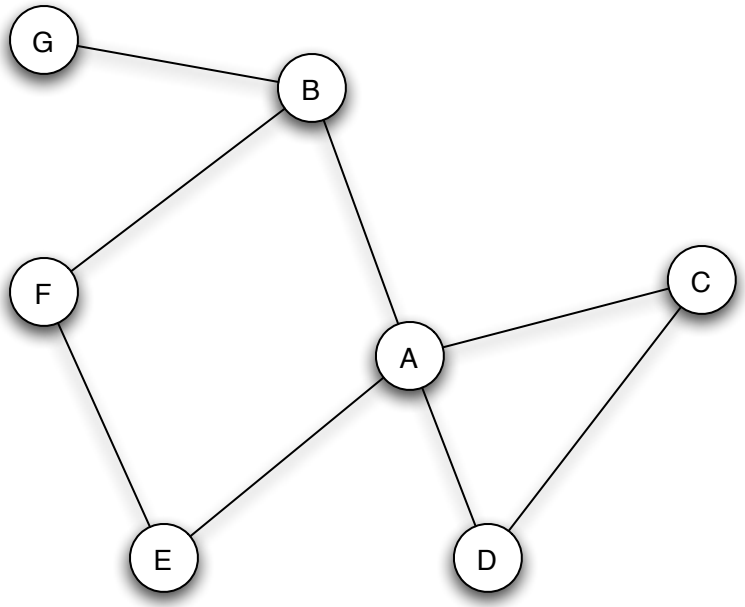
- Let E be the set of an undirected edges of a network graph. (Forgive the abuse of notation where in the previous and next slide E is a node name.) For a node A, the clustering coefficient is the following ratio:

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

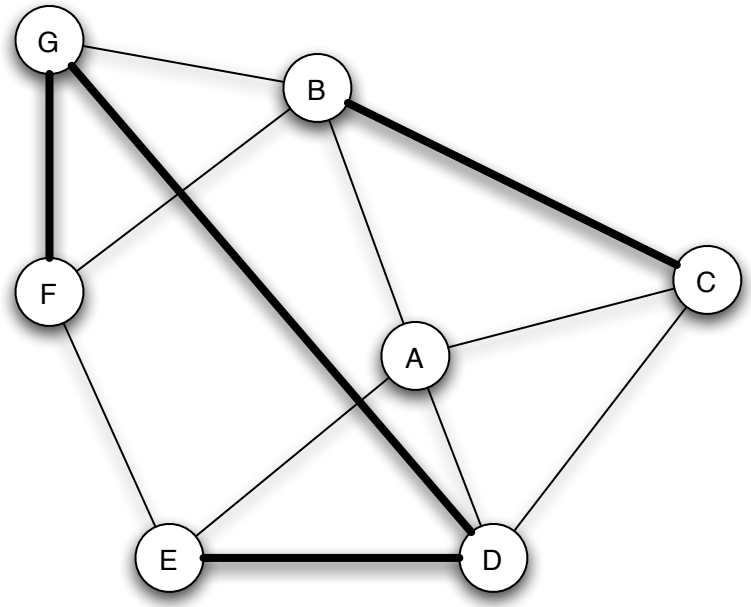
- The numerator is the number of all edges (B, C) in the network such that B and C are adjacent to (i.e. mutual friends of) A.

- 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.

End of Monday, December 5 class

We ended having discussed the Bearman and Moody study which indicates the correlation betwewen low correlation coefficient and suicidal thoughts.

We will continue Wednesday discussing social networks beginning with the strength of weak ties, and the the Sintos and Tsaparas results for infering the strength of ties. If time permits, we conclude with the six degrees of separation phenomena.

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 within that connected component.

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 within that connected component.
- 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)

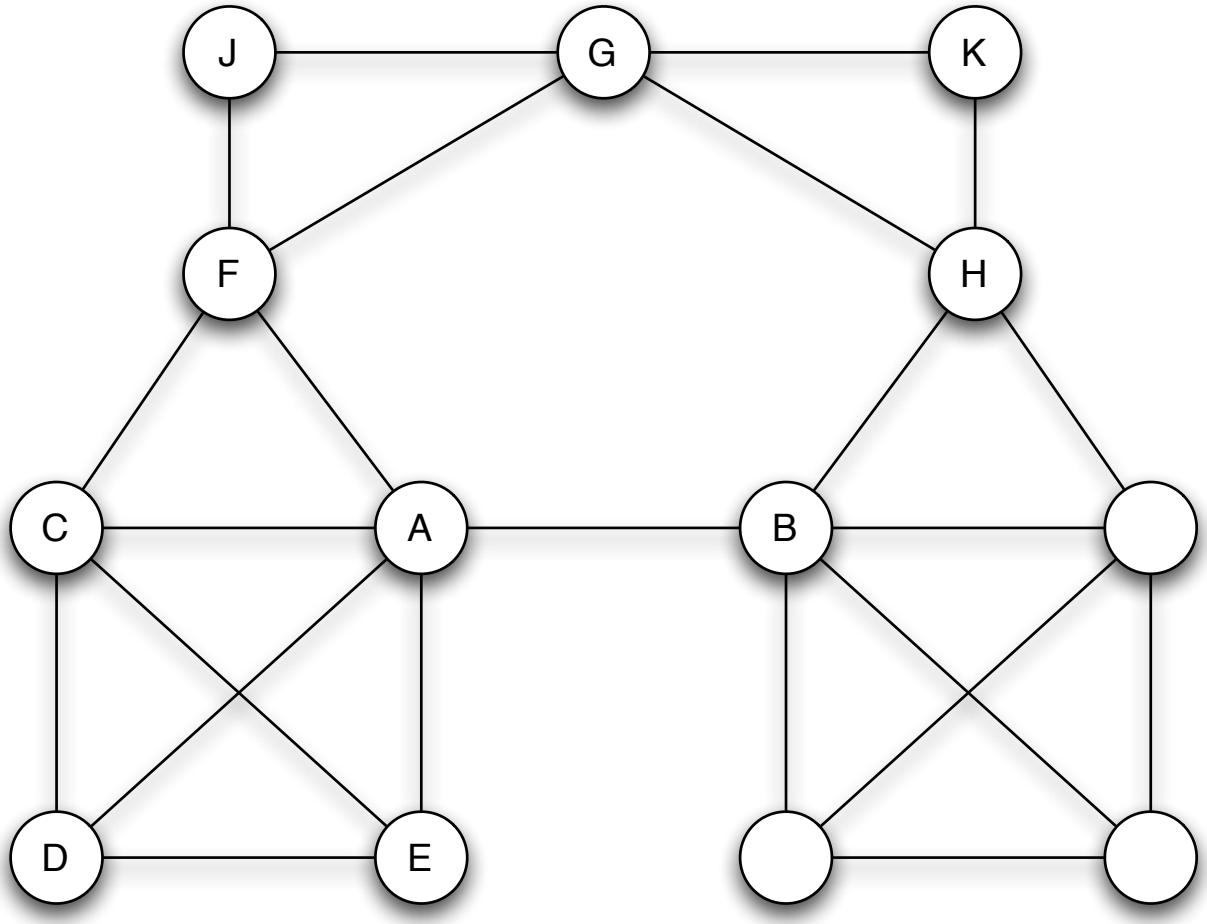


Figure: The edge (A, B) is a local bridge of span 4, since the removal of this edge would increase the distance between A and B to 4. [E&K Figure 3.4]

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

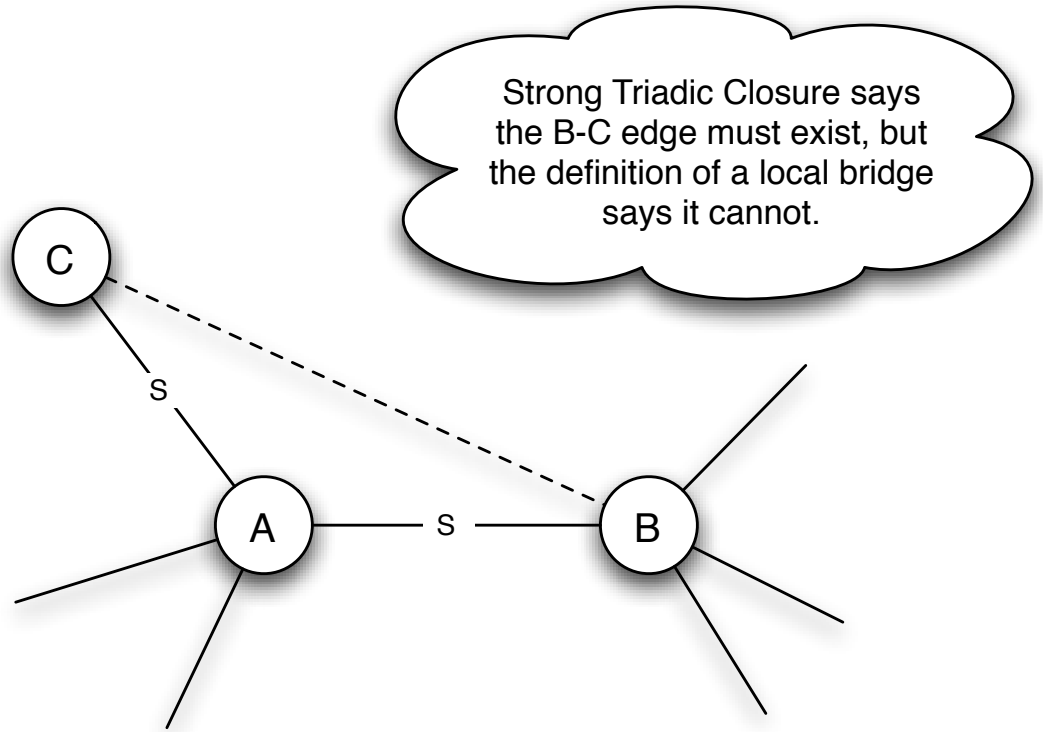


Figure 3.6: If a node satifies Strong Triadic Closure and is involved in at least two strong ties, then any local bridge it is involved in must be a weak tie. The figure illustrates the reason why: if the A-B edge is a strong tie, then there must also be an edge between B and C, meaning that the A-B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

Strong triadic closure property continued

- Again we emphasize (as the EK 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.
  (http:/(www.cs.uoi.gr/∼tsap/publications/frp0625-sintos.pdf)

  We will discuss this paper later today.

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 relating tie strength and and overlap

- Onnela et al. [2007] studied the who-talks-to-whom network maintained by a cell phone provider. iMore specifically, a 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

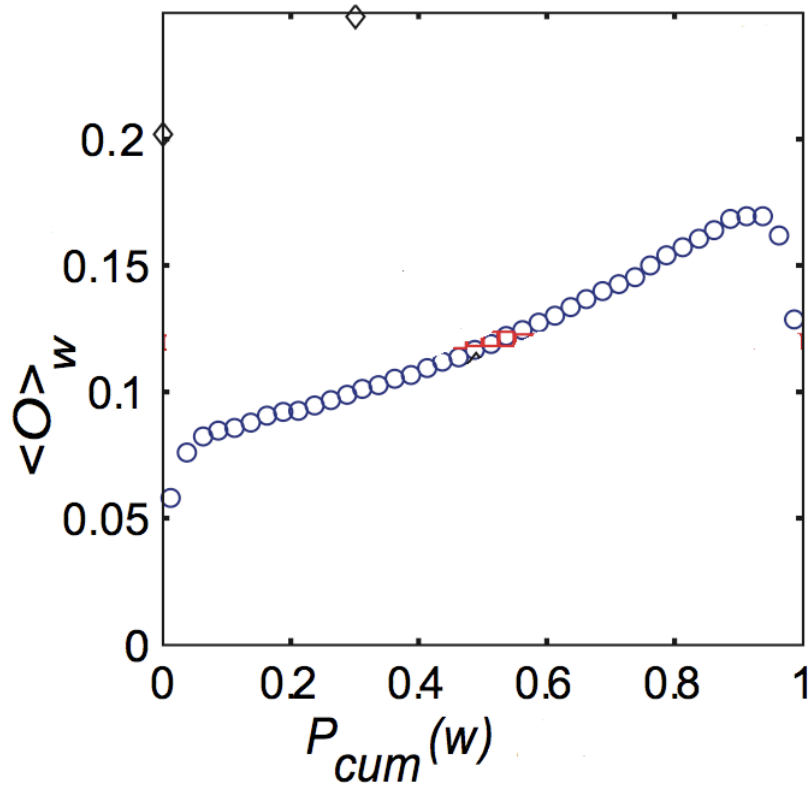


Figure: A plot of the neighborhood overlap of edges as a function of their percentile in the sorted order of all edges by tie strength. [E&K Fig 3.7]

- 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.

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):

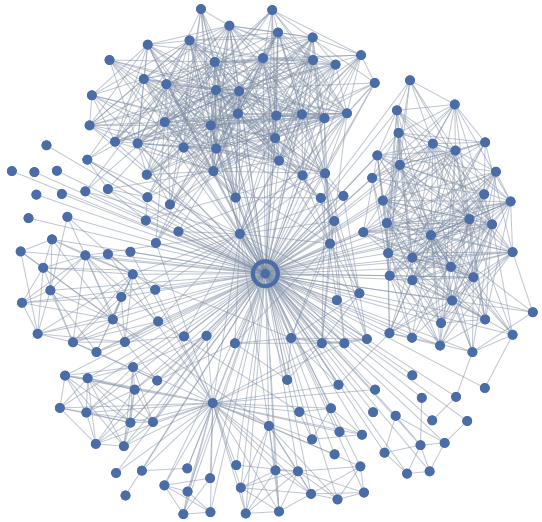
> *Given the size and complexity of the (who calls whom) network, we cannot simply look at the structure. . . Indirect measures must generally be used and, because one knows relatively little about the meaning or significance of any particular node or edge, it remains an ongoing research challenge to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

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

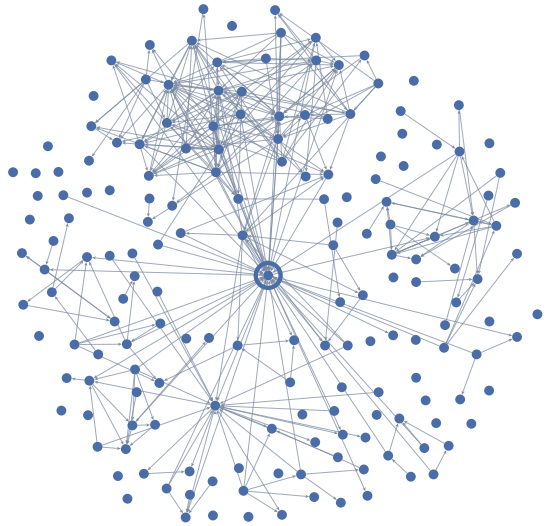
- The meaning of "friend" as in Facebook is not the same as one might have traditionally interpreted the word "friend".
- Online social networks give us the ability to qualify the strength of ties in a useful way.
- For an observation period of one month, Marlow et al. (2009) consider Facebook networks 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

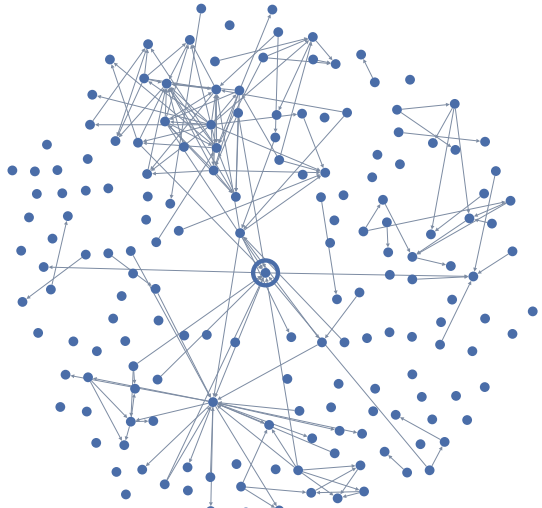
All Friends



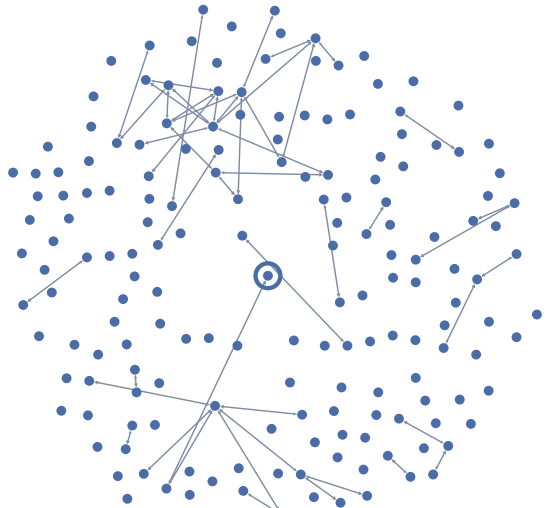
Maintained Relationships



One-way Communication



Mutual Communication



A limit to the number of strong ties

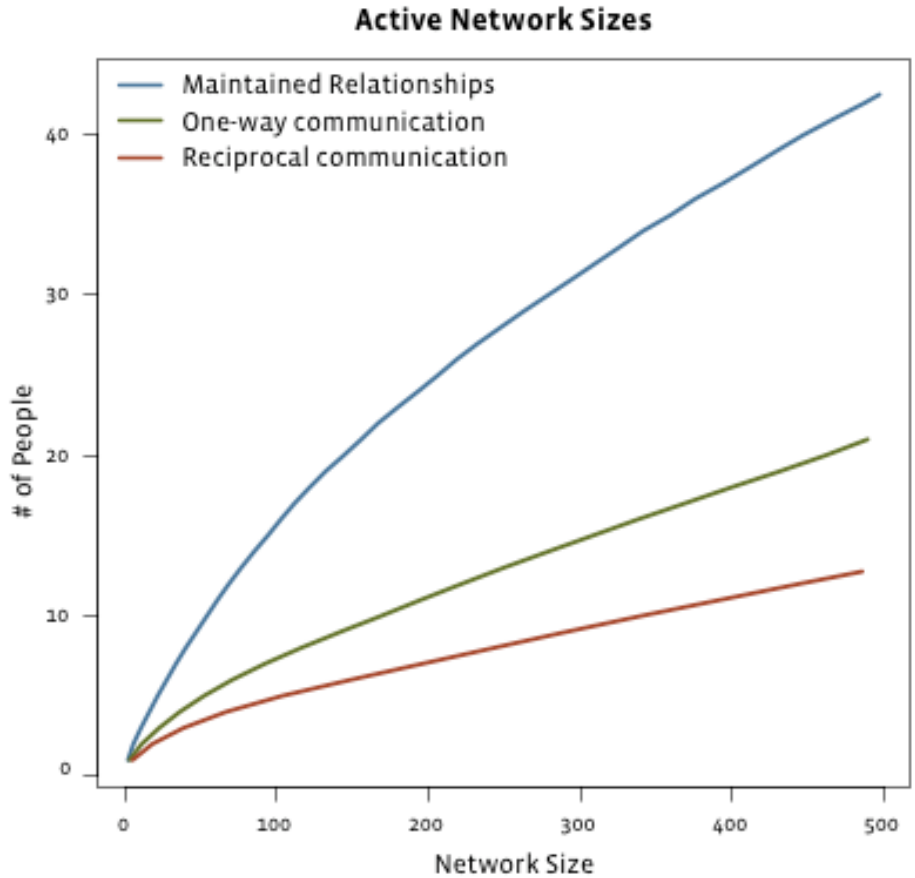


Figure: The number of links corresponding to maintained relationships, one-way communication, and reciprocal communication as a function of the total neighborhood size for users on Facebook. [Figure 3.9, textbook]

Twitter:Limited Strong Ties vs Followers

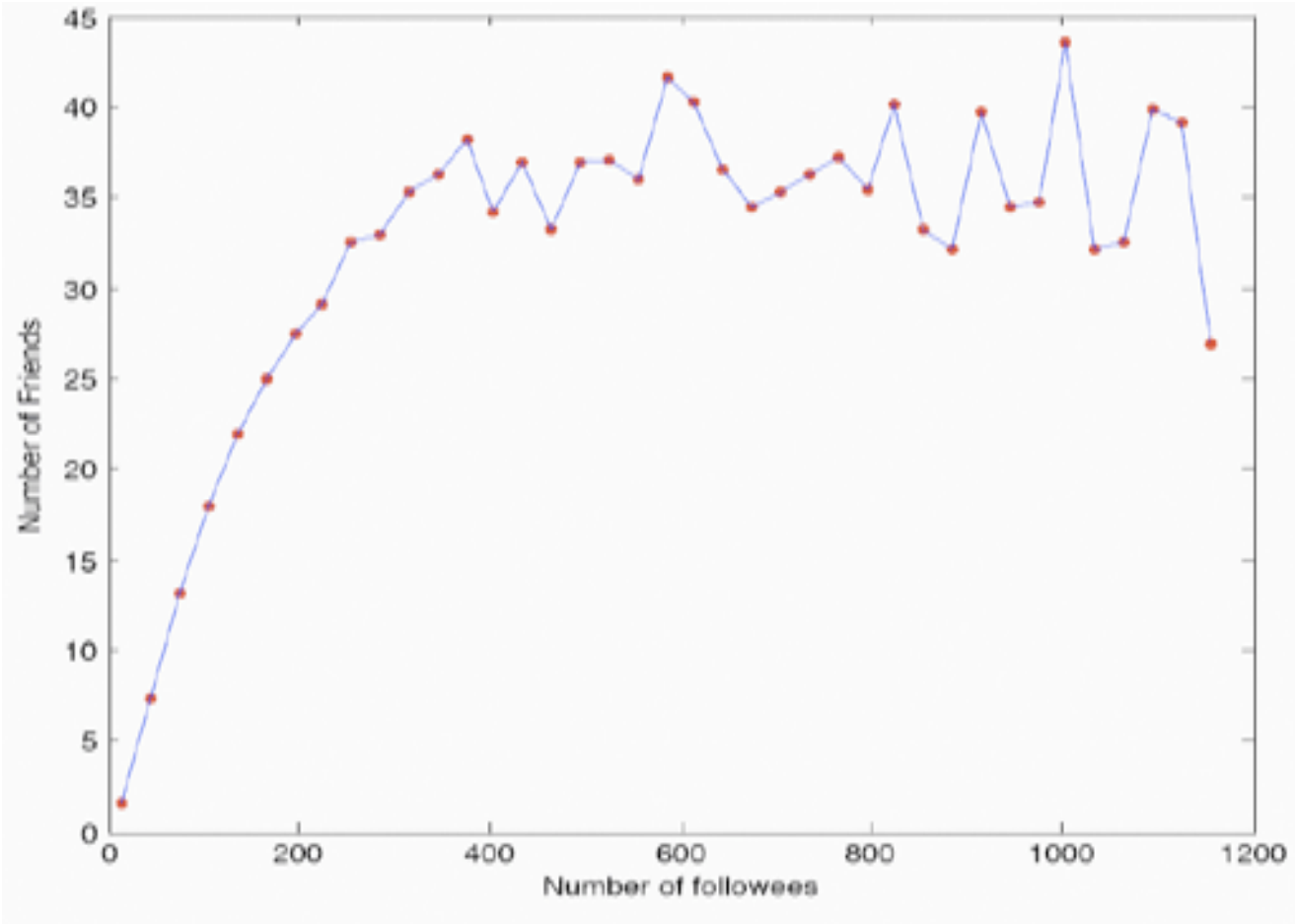


Figure: The total number of a user's strong ties (defined by multiple directed messages) as a function of the number of followees he or she has on Twitter. [Figure 3.10, textbook]

Information spread in a passive network

- The maintained or passive relation network (as in the Facebook network on slide 36) 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

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

"Tightly knit communities" connected by weak ties

- 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.

"Tightly knit communities" connected by weak ties

- 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.
- Recalling the relation to weak ties, the Easley and Kleinberg 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

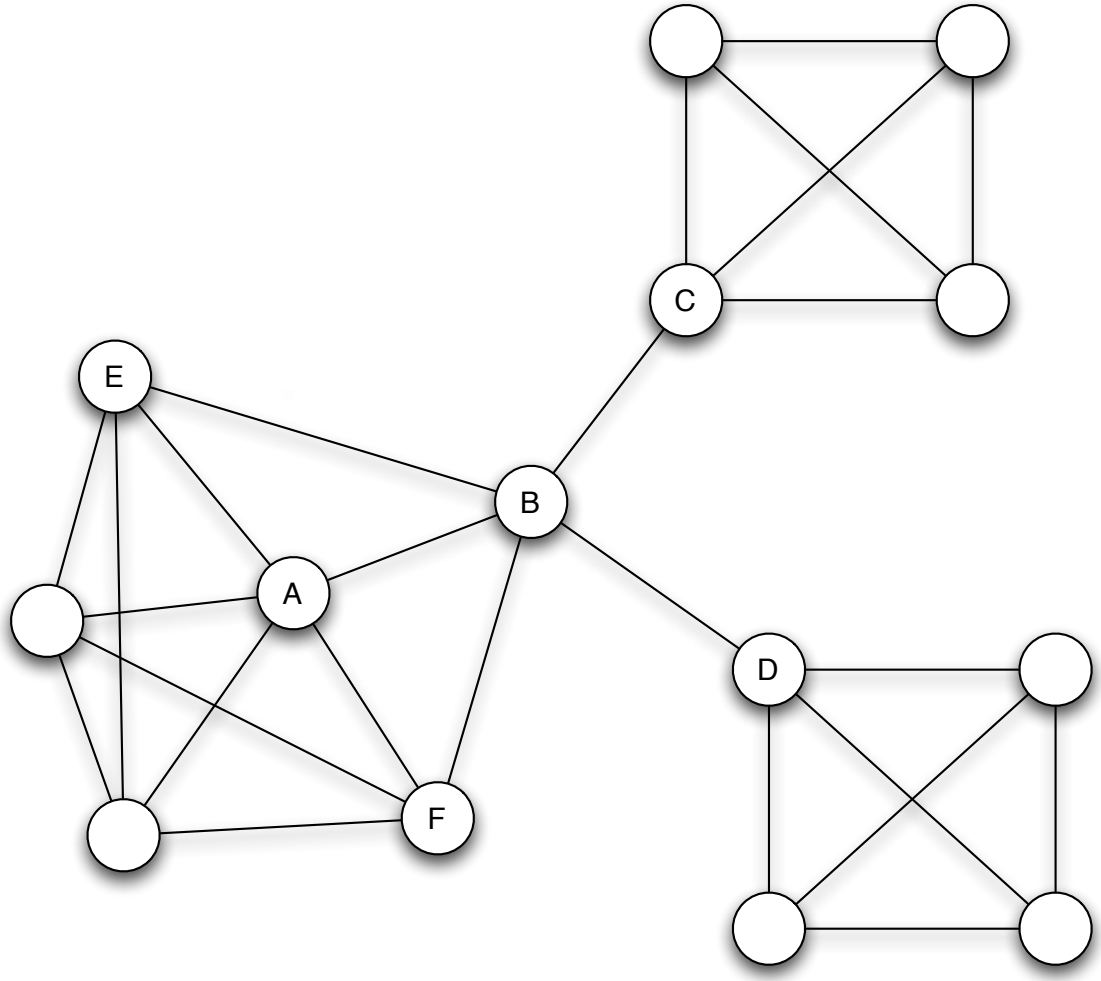


Figure: The contrast between densely-knit groups and boundary-spanning links is reflected in the different positions of central node A and gatekeeper node B in the underlying social network. [Fig 3.11, textbook]

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 again: 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 (as noted before) the Medici lie on the shortest paths between most pairs of families.

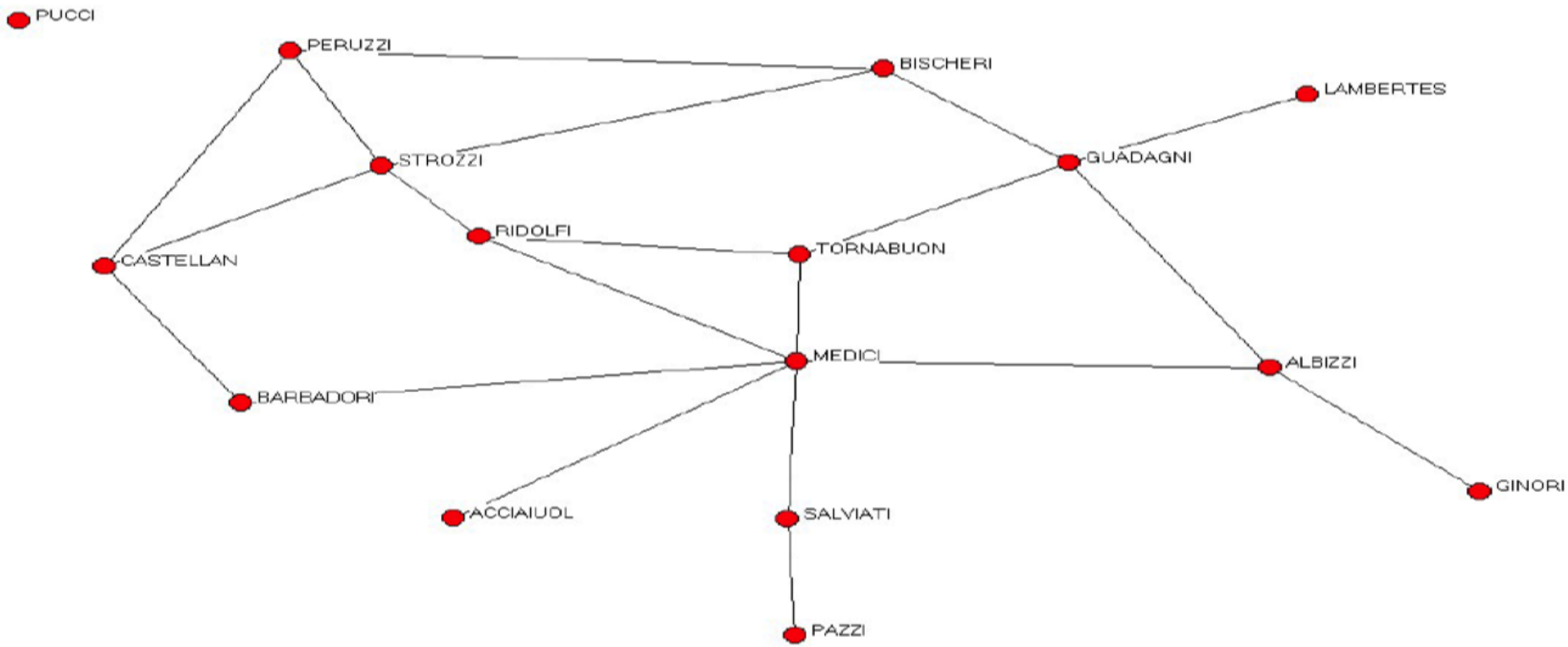
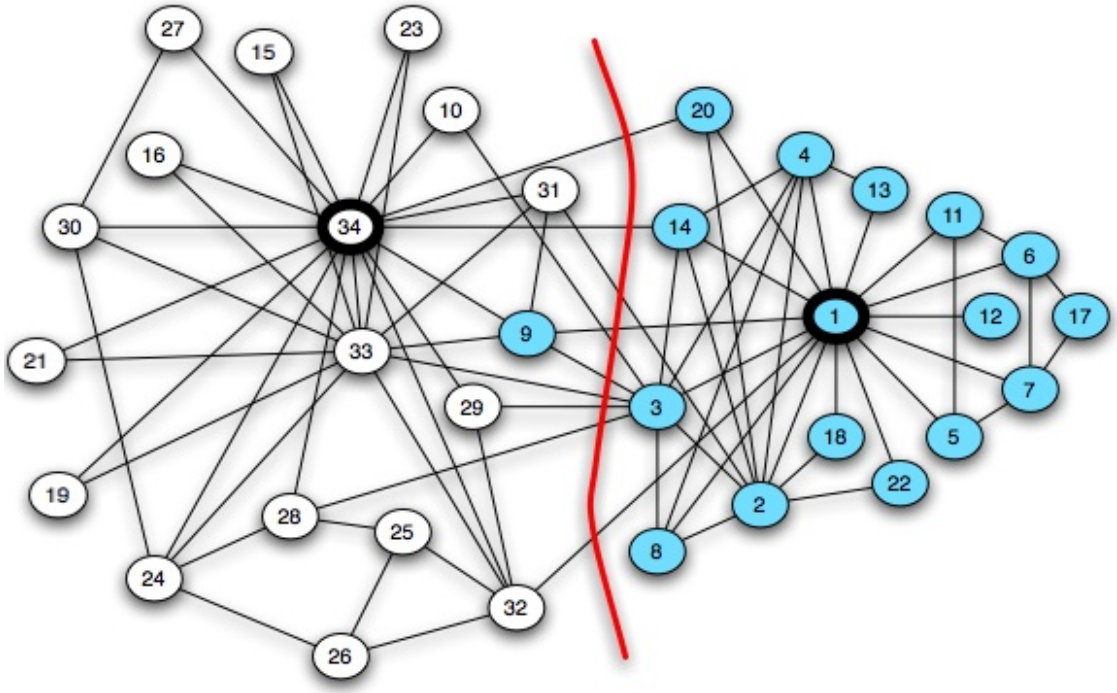


Figure: see [Jackson, Ch 1 in EK text]

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 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. In 3 of these networks, the strength of ties is known from the given data. Their worst case approximation algorithm (via the reduction) leads 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

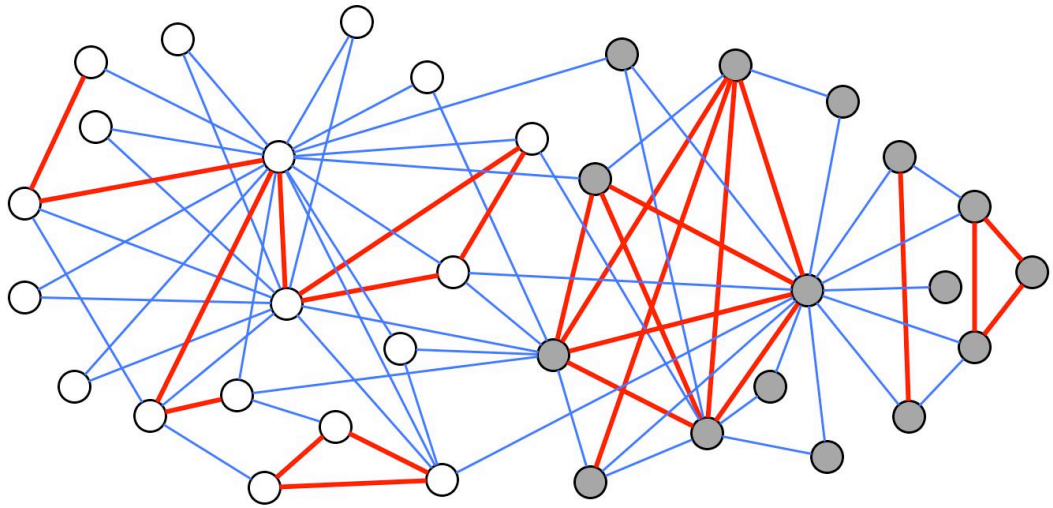


Figure 1: Karate Club graph. Blue light edges represent the weak edges, while red thick edges represent the strong edges.

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

The six degrees of freedom phenomena

There are two basic ways for finding someone in a social network.

- We could ask all of our friends to tell all of their friends to tell all of their friends. . . (i.e. a traditional chain letter) that I am looking for person X.

- Now say assuming your online social network has a "broadcast to all" feature, this can be done easily but it has its drawbacks. Drawbacks?

The six degrees of freedom phenomena

There are two basic ways for finding someone in a social network.

- We could ask all of our friends to tell all of their friends to tell all of their friends. . . (i.e. a traditional chain letter) that I am looking for person X.

- Now say assuming your online social network has a "broadcast to all" feature, this can be done easily but it has its drawbacks. Drawbacks?

- Suppose on the other hand that we want to reach someone and it either costs real money/effort to pass a message (e.g. postal mail) or perhaps I would prefer to not let everyone know that I am looking for person X. And as was pointed out in class, there is also possibly a "social cost" in terms of annoyance to people in the network receiving multiple requestss to pass on a message.

The six degrees of freedom phenomena

There are two basic ways for finding someone in a social network.

- We could ask all of our friends to tell all of their friends to tell all of their friends. . . (i.e. a traditional chain letter) that I am looking for person X.

- Now say assuming your online social network has a "broadcast to all" feature, this can be done easily but it has its drawbacks. Drawbacks?

- Suppose on the other hand that we want to reach someone and it either costs real money/effort to pass a message (e.g. postal mail) or perhaps I would prefer to not let everyone know that I am looking for person X. And as was pointed out in class, there is also possibly a "social cost" in terms of annoyance to people in the network receiving multiple requestss to pass on a message.

- Clearly if everyone cooperates, the broadcast method ensures the shortest path to the intended target X in the leveled tree/graph of reachable nodes.

Reachable nodes without triadic closure

- If there is no triadic closure (i.e. your friends are not mutual friends, etc.), it is easy to see why every path is a shortest path to everyone in the network.
- Consider the number of people that you could reach by a path of length at most t if every person has say at least 5 friends.

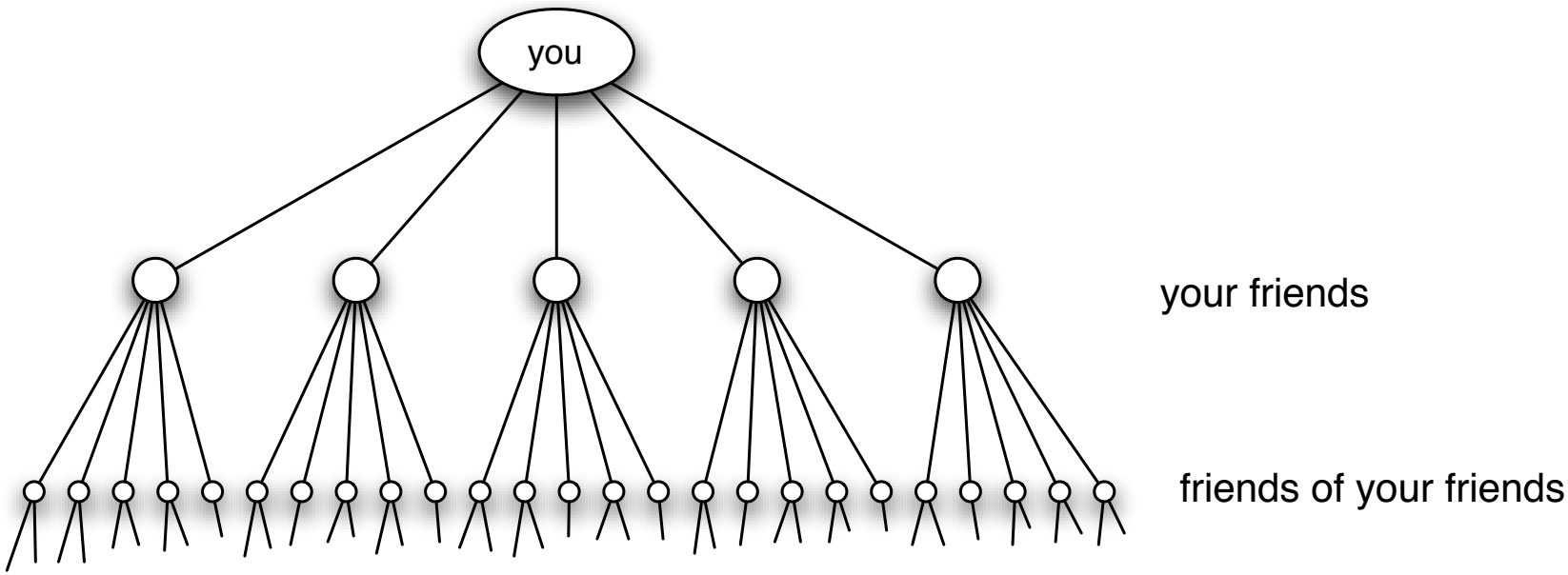


Figure: Pure exponential growth produces a small world [Fig 20.1 (a), E&K]

Reachable nodes with triadic closure

- Given that our friends tend to be mostly contained within a few small communities, the number of people reachable will be much smaller.

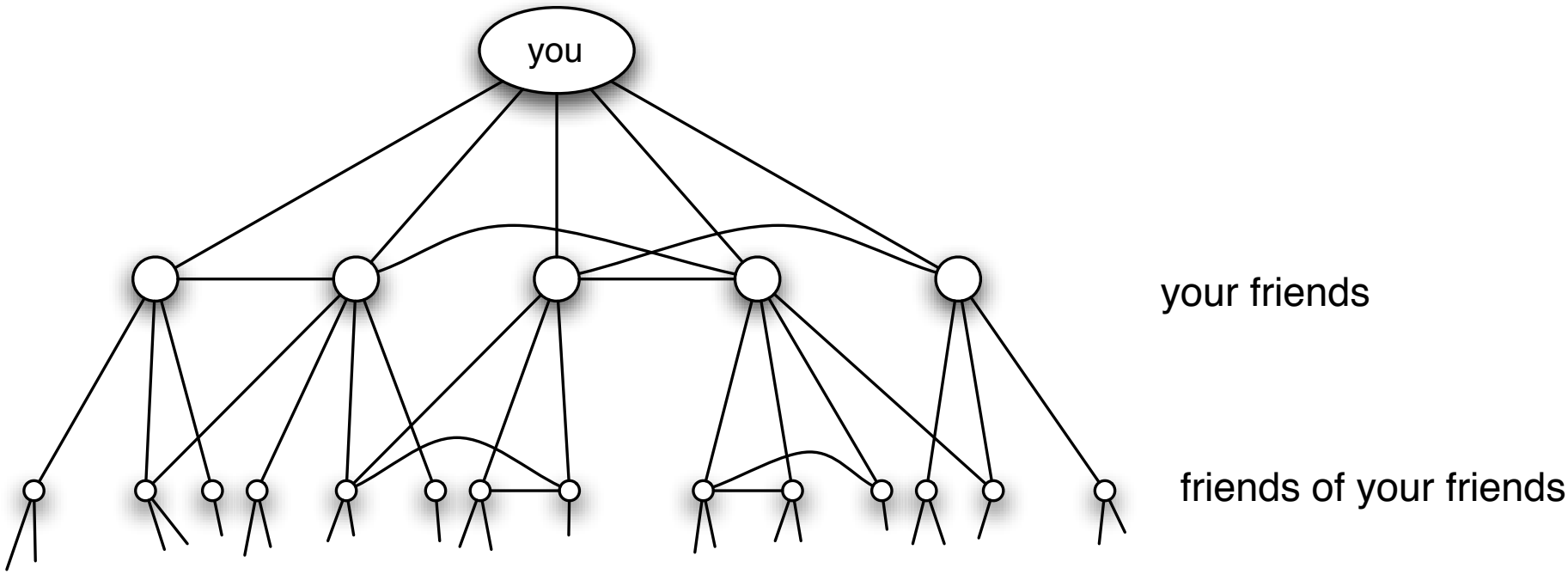


Figure: Triadic closure reduces the growth rate [Fig 20.1 (b), E&K]

The Watts-Strogatz model

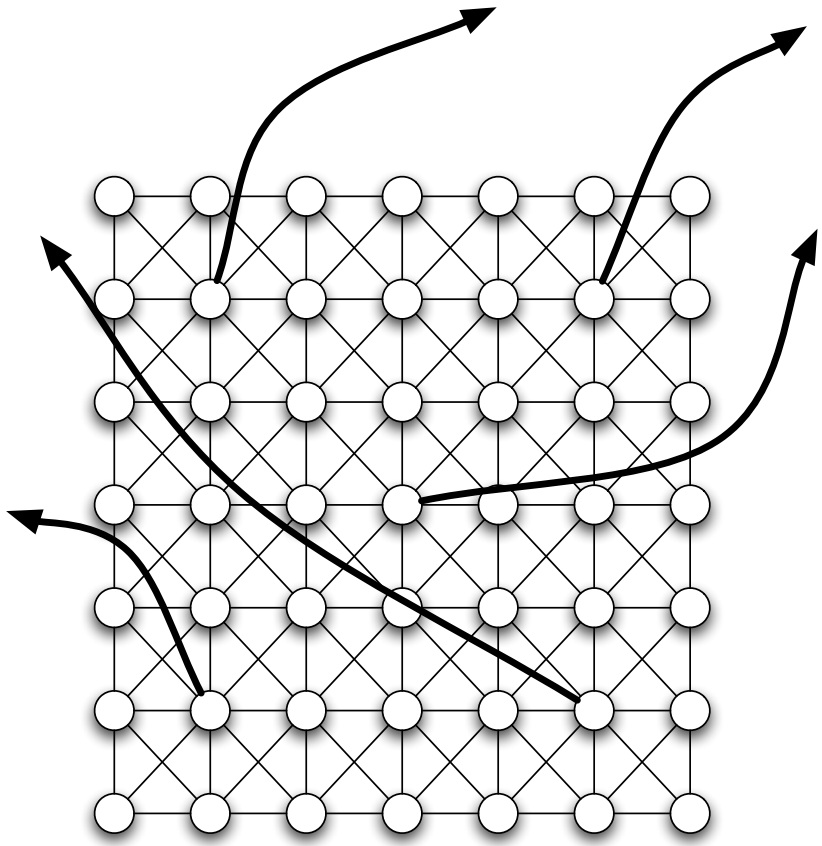
- Is it possible to have extensive triadic closure and still have short paths?

- Homophily is consistent with triadic closure especially for strong ties whereas weak ties can connect different communities and thereby provide the kind of branching that yields short paths to many nodes.

- One stylized model to demonstrate the effect of these different kinds of ties is the Watts-Strogatz model, which considers nodes lying in a two dimensional grid and then having two types of edges:
  - Short-range edges to all nodes within some small distance r. This captures an idealized sense of homophily
  - A small number of random longer-distance edges to other nodes in the network; in fact, one needs very few such random edges to achieve the effect of short paths.

Very few random edges are needed

- A k by k "town" with probability $1/k$ that a person has a random weak tie.
- This would be sufficient to establish short paths.



[Fig 20.3, E&K]

But how does this explain the ability to find people in a decentralized manner

- In the Watts-Strogatz type of model, we can use the random edges (in addition to the short grid edges) and the geometric location of nodes to keep trying to reduce the grid distance to a target node.
  - This is analogous to the Milgram experiment where individuals seem to use geographic information to guide the search.
  - However, completely random edges does no reflect real social networks

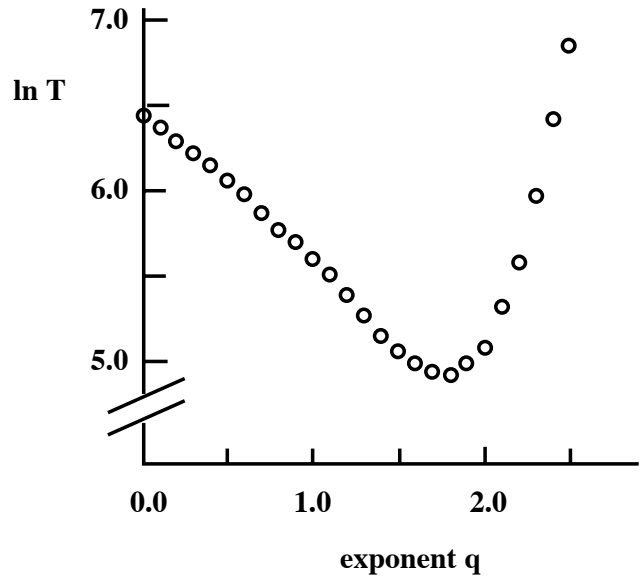
- Furthermore, having uniformly random edges will not work in general as:
  - Completely random edges (i.e. going to a random node anywhere in the network) are too random.
  - A random edge in an $n \times n$ grid is likely to have grid distance $\Theta(n)$.
  - Without some central guidance, such random edges will essentially just have us bounce around the network causing a substantially longer path to the target than the shortest path.

A modification of the model

- Random edges outside of ones "close community" are still more likely to reflect some relation to closeness.
- So assume as in the Watts-Strogatz model, from every node v we have edges to all nodes x within some grid distance r from v.
- And now in addition random edges are generated as follows: we (independently) create an edge from v to w with probability proportional to $d(v,w)^{-q}$ where $d(v,w)$ is the grid distance from v to w and $q \geq 0$ is called the clustering exponent.
- The smaller $q \geq 0$ is, the more completely random is the edge whereas large $q \geq 0$ leads to edges which are not sufficiently random and basically keeps edges within or very close to ones community.
- What is the best choice of $q \geq 0$?

So what is a good or the best choice of the clustering exponent q?

- It turns out that in this 2-dimensional grid model decentralized search works best when $q = 2$. (This is a result that holds and can be proven for the limiting behaviour, in the limit as the network size increases.)



[Fig 20.6, E&K]

- Simulation of decentralized search in the grid-based model with clustering exponent q.
- Each point is the average of 1000 runs on (a slight variant of) a grid with 400 million nodes.
- The delivery time is best in the vicinity of exponent $q = 2$, as expected.
- But even with this number of nodes, the delivery time is comparable over the range between 1.5 and 2.

More precise statements of Kleinberg's results on navigation in small worlds

The Milgram-like experiment

- Consider a grid network and construct (local contact) directed edges from each node u to all nodes v within grid distance $d(u, v) = k > 1$.
- Also probabilistically construct m (long distance) directed edges where each such edge is chosen with probability proportional to $d(v, w)^{-q}$ for $q \geq 0$.
- We think of k and m as constants and consider the impact of the clustering exponent q as the network size n increases.
- We assume that each node knows its location and the location of its adjacent edges and its distance to the location of a target node t.
- The Milgram-like experiment is that each node it *tries* (without knowing the entire network) to move from a node u to a node v that is closest to t (in grid distance).

Reflection on the Kleinberg-Milgram model

As we said at the start of this topic, the real surprise is that a "short" (but not shortest) path is (probably wrt to the randomly generated network) being found by a decentralized search.

It is true that each node will pursue a "greedy strategy" but this is different than say Dijkstra's least cost/distance algorithm which entails a centralized search.

Navigation in small worlds results

Theorem

(J. Kleinberg 2000)

(a) *For $0 \leq q < 2$, the (expected) delivery time T of any "decentralized algorithm" in the $n \times n$ grid-based model is $\Omega\left(n^{\frac{2-q}{3}}\right)$.*

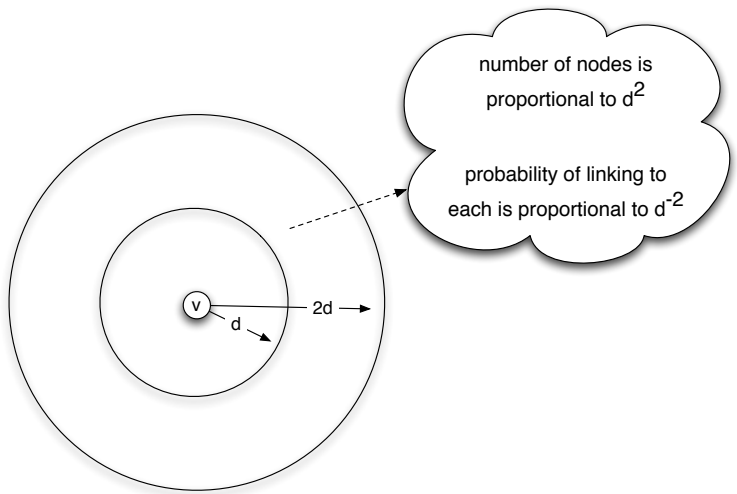
(b) *For $q = 2$, there is a decentralized algorithm with delivery time $O(\log n)$.*

(c) *For $q > 2$, the delivery time of any decentralized algorithm in the grid-based model is $\Omega\left(n^{\frac{q-2}{q-1}}\right)$.*

(The lower bounds in (a) and (c) hold even if each node has an arbitrary constant number of long-range contacts, rather than just one.)

Intuition as to why $q = 2$ is best for the grid

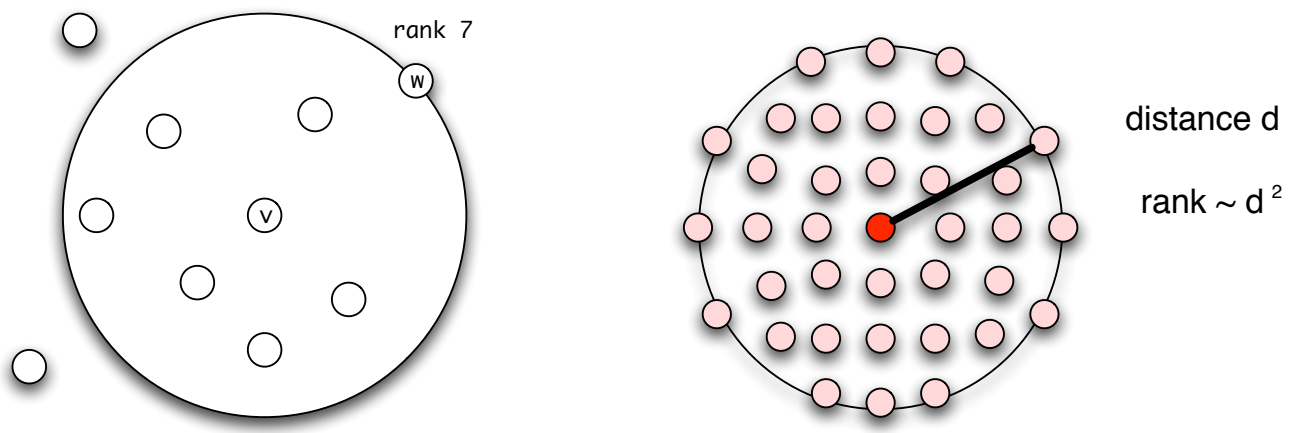
- It is instructive to see why this choice of q provides links at the different "scales of resolution" seen in the Milgram experiment.
- That is, if D is the maximum distance to be travelled, then we would like links with distances between d and $2d$ for all $d < \log D$
- Given that we have a 2-dimensional grid, the number of points with distance say d from a given node v will be $\approx d^2$.
- We are choosing such a node with probability proportional to $1/d^2$ and hence we expect to have a link to some node whose distance from v is between d and $2d$ for all d.



[Fig 20.7, E&K]

More realistic (nonuniformly spread) population data

- In the grid model, the population density is completely uniform which is not what one would expect in real data.
- How can this $1/d^2$ (inverse-square) distribution be modified to account for population densities that are very non-uniform?
- The idea is to replace distance $d(v, w)$ from v to w by the rank of w relative to v.
  - For a fixed v, define the $rank(w)$ to be the number of nodes closer to v than w.
  - In the 2D grid case, when $d(v, w) \sim d$, then $rank(w) \sim d^2$.

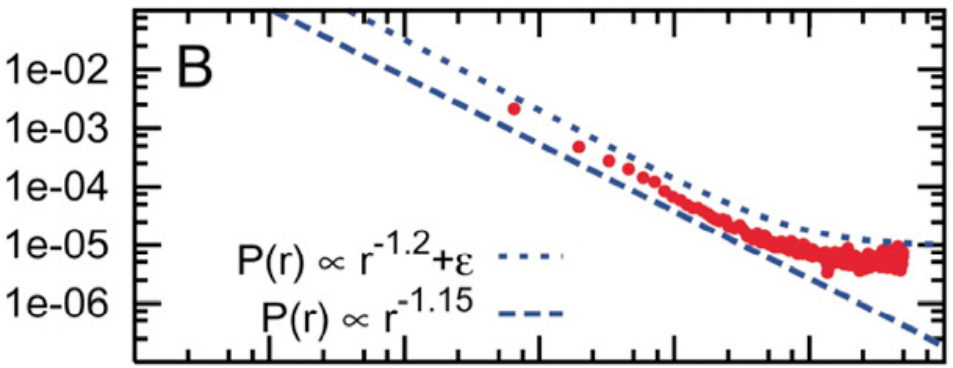


[Fig 20.9, E&K]

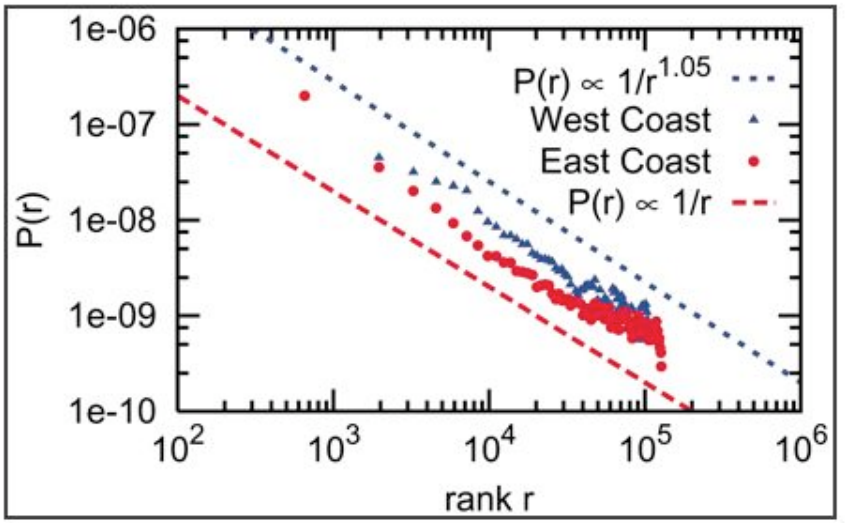
More realistic geographic data continued

- We can then restate the inverse-square distribution by saying that the probability that v links to w is proportional to $1/rank(w)$.
- Using zip code information, for every pair of nodes (500,000 users on the blogging site LiveJournal) one can assign ranks.
- Liben-Nowell et al did such a study in 2005, and then for different rank values examined the fraction f of edges that are actually friends.
- The theory tells us that this fraction f should be a decreasing function proportional to $1/rank$.
- That is, $f \sim rank^{-1}$. Taking logarithms, $\log f \sim (-1) \log rank$.

More realistic (LiveJournal) friendship data



(a) *Rank-based friendship on LiveJournal*



(b) *Rank-based friendship: East and West coasts*

[Fig 20.10, E&K]

- In Figure 20.10 (a), the Lower (upper) line is exponent $= -1.15$ (resp. -1.12).
- In Figure 20.10 (b), the Lower (upper) line is exponent $= -1.05$ (resp. -1). The red data is East Coast data and the blue data is West Coast data.

Liben-Nowell: practice closely matches theory

Liben-Nowell prove that for "essentially" any population density (i.e. no matter where people are located) if links are randomly constructed so that the probability of a friendship is proportional to $rank^{-1}$, then the resulting network is one that can be efficiently searched in a decentralized manner.

That is, Kleinberg's result for the grid generalizes. This is a rather exceptional result in that the abstraction from d^{-2} to $rank^{-1}$ is not at all an obvious generalization.

How surprised should we be that natural populations locate themselves in this probabilistic manner since there is no centralized organizing mechanism that is causing this phenomena?

Liben-Nowell: practice closely matches theory

Liben-Nowell prove that for "essentially" any population density (i.e. no matter where people are located) if links are randomly constructed so that the probability of a friendship is proportional to $rank^{-1}$, then the resulting network is one that can be efficiently searched in a decentralized manner.

That is, Kleinberg's result for the grid generalizes. This is a rather exceptional result in that the abstraction from d^{-2} to $rank^{-1}$ is not at all an obvious generalization.

How surprised should we be that natural populations locate themselves in this probabilistic manner since there is no centralized organizing mechanism that is causing this phenomena?

The EK text refers to a 2008 article by Oscar Sandberg who analyzes a network model where decentralized search takes place which in turn causes links to "re-wire" so as to fascilitate more efficient decentralized search.

It remains an intringing question as to the extent this does happen in social networks and the implicit mechanisms that would cause networks to evolve this way.

IP addresses and the TCP/IP routing protocol

For those taking (or having taken) a computer networks course, you can observe how IP addresses allow the IP transmission protocol to send messages along a decnetralized route.

TCP/IP originated in the earlyh 1980's which is much after Milgram but well before Strogatz and Kleinberg. To what extent was the TCP/IP protocol and IP addresses motivated by Milgram's work?

But perhaps postal codes are the original motivation?

Aside Interesting ideas usually have a history and the best we can do is document some of the major events in the adoption of any important idea.

The Backstrom et al rank-based study

- Backstrom et al study US Facebook 2010 geographic user data.
  1. Roughly 100 million users
  2. About 6% of which enter home address info and of that population about 60% can be parsed into longitude and lattiude information.
  3. This gave a set of 3.5 million users (of which 2.9 million had at least one friend with a well specified address and each of these 2.9 million users had an average of 10 friends with specified addresses resulting in 30.6 million edges.
  4. Although a small part of Facebook, this 2.9 million person "geolocated data set" is sufficently large and representative for experimental study.

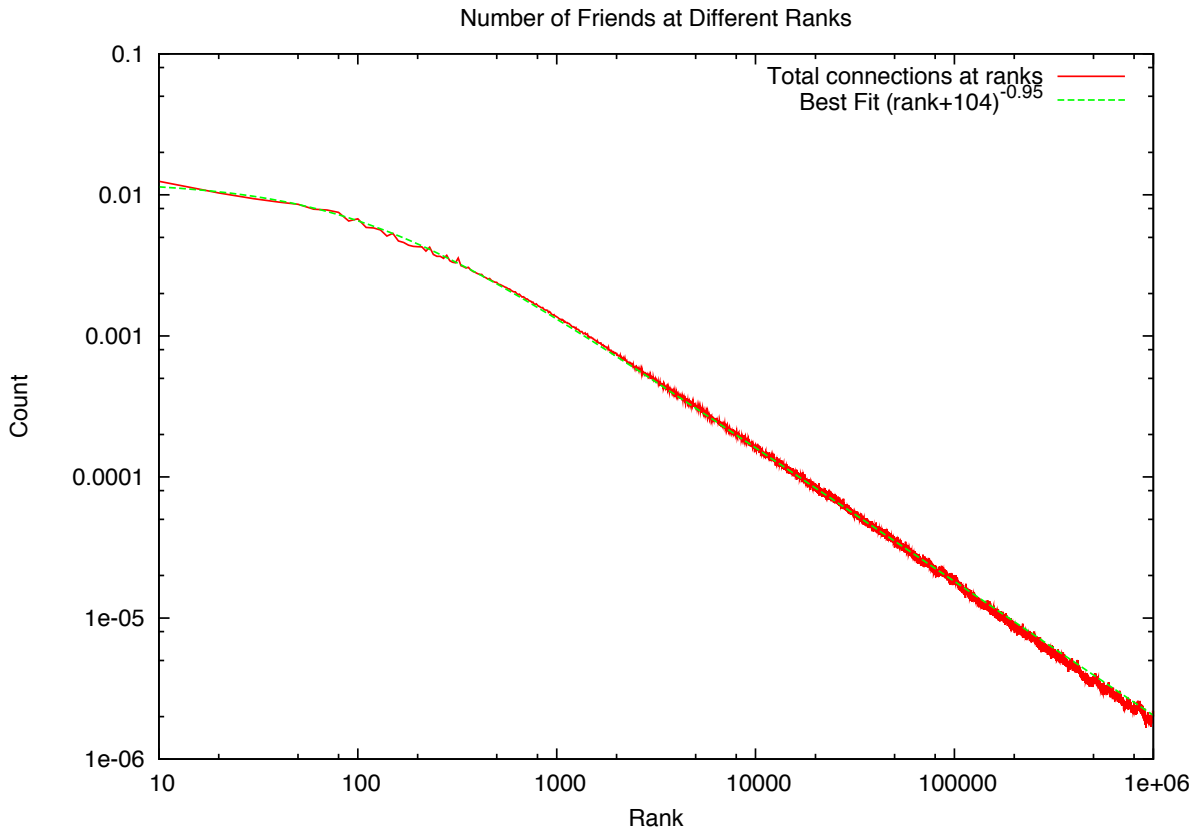
The Backstrom et al rank-based study

- Backstrom et al study US Facebook 2010 geographic user data.
  1. Roughly 100 million users
  2. About 6% of which enter home address info and of that population about 60% can be parsed into longitude and lattiude information.
  3. This gave a set of 3.5 million users (of which 2.9 million had at least one friend with a well specified address and each of these 2.9 million users had an average of 10 friends with specified addresses resulting in 30.6 million edges.
  4. Although a small part of Facebook, this 2.9 million person "geolocated data set" is sufficently large and representative for experimental study.
- They study probability of friendships vs distance and rank and how those probabilities depend on population densities for where people live. This study provides more evidence as to the power law relation between distance/rank and probability ($\approx rank^{-.95}$) of friendship.

The Backstrom et al rank-based study

- Backstrom et al study US Facebook 2010 geographic user data.
  1. Roughly 100 million users
  2. About 6% of which enter home address info and of that population about 60% can be parsed into longitude and lattiude information.
  3. This gave a set of 3.5 million users (of which 2.9 million had at least one friend with a well specified address and each of these 2.9 million users had an average of 10 friends with specified addresses resulting in 30.6 million edges.
  4. Although a small part of Facebook, this 2.9 million person "geolocated data set" is sufficently large and representative for experimental study.
- They study probability of friendships vs distance and rank and how those probabilities depend on population densities for where people live. This study provides more evidence as to the power law relation between distance/rank and probability ($\approx rank^{-.95}$) of friendship.
- Furthermore, they utilize this relationship between friends and distance to create an algorithm that will predict the location of an individual from a small set of users with known locations. They claim their algorithm can predict geographic locations better than using IP information!

Number of friends wrt. rank



[Figure 9 from Backstrom et al]

Predicting locations

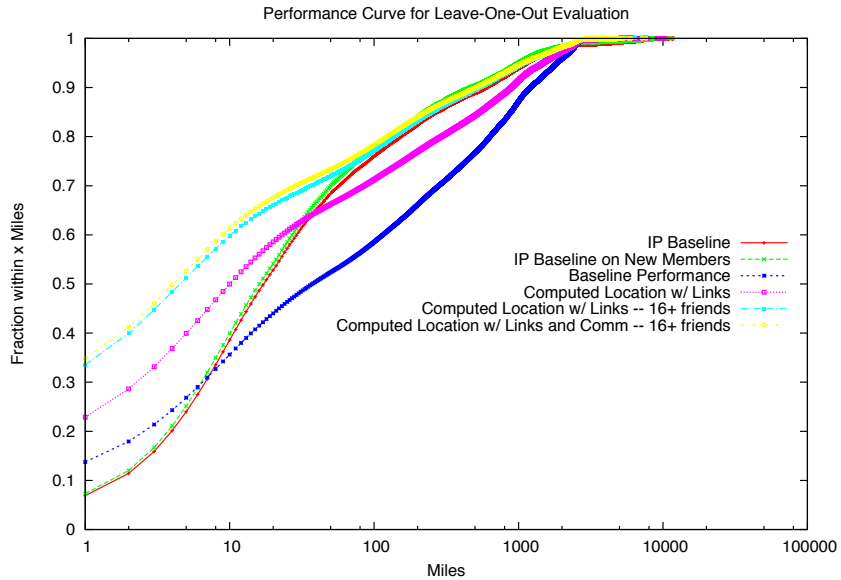


Figure 11: Location Prediction Performance. This figure compares external predictions from an IP geolocation service, the same service constrained to users who have recently updated their address, a baseline of randomly choosing the location of a friend, along with three predictions: our algorithm with all links, for users with 16+ friends, and finally for users with 16+ friends constraining to only those with whom they have communicated recently.

[Figure 11 from Backstrom et al]

From geographic distance to social distance

- What if there is no (reliable) distance information in a social network?

From geographic distance to social distance

- What if there is no (reliable) distance information in a social network?

- It is, of course, natural that we tend to have more common interests with people who live closer to us (e.g. based on ethnicity, economic status, etc), but clearly there are other notions of social distance that should be considered.

From geographic distance to social distance

- What if there is no (reliable) distance information in a social network?
- It is, of course, natural that we tend to have more common interests with people who live closer to us (e.g. based on ethnicity, economic status, etc), but clearly there are other notions of social distance that should be considered.
- Early in the course we considered social foci (clubs, shared interests, language, etc.) we tend to share a number of focal interests with the same person.
- But, of course, belonging to a small group of people in a course, is different than attending the same University, and speaking Mandarin is different than being interested in Esperanto.

From geographic distance to social distance

- What if there is no (reliable) distance information in a social network?
- It is, of course, natural that we tend to have more common interests with people who live closer to us (e.g. based on ethnicity, economic status, etc), but clearly there are other notions of social distance that should be considered.
- Early in the course we considered social foci (clubs, shared interests, language, etc.) we tend to share a number of focal interests with the same person.
- But, of course, belonging to a small group of people in a course, is different than attending the same University, and speaking Mandarin is different than being interested in Esperanto.
- So the suggestion is made that we define social distance $s(v, w)$ between individuals v, w to be the minimum size of a common focus.

Smallest size shared focus as a distance measure

- Kleinberg (2001) gives theoretical results indicating that when friendships follow a distribution proportional to $1/s(v, w)$ then the resulting social network will support efficient decentralized search.

- This is somewhat verified in a study (by Adamic and Adar) of 'who talks to whom' friendship data (based on frequency of email exchanges) amongst a small group of HP employees.

- The focal groups are defined by the organizational hierarchy of the company.

- The Adamic and Adar 2005 study shows that the distribution for this friendship relationship is proportional to the inverse of $s(v, w)^{-3/4}$ so that it doesn't match as closely with the previous geographical rank based results but still observes a power law relation governing how social ties decrease with "distance".

Probability of email exchanges vs distance in the organizational hierarchy

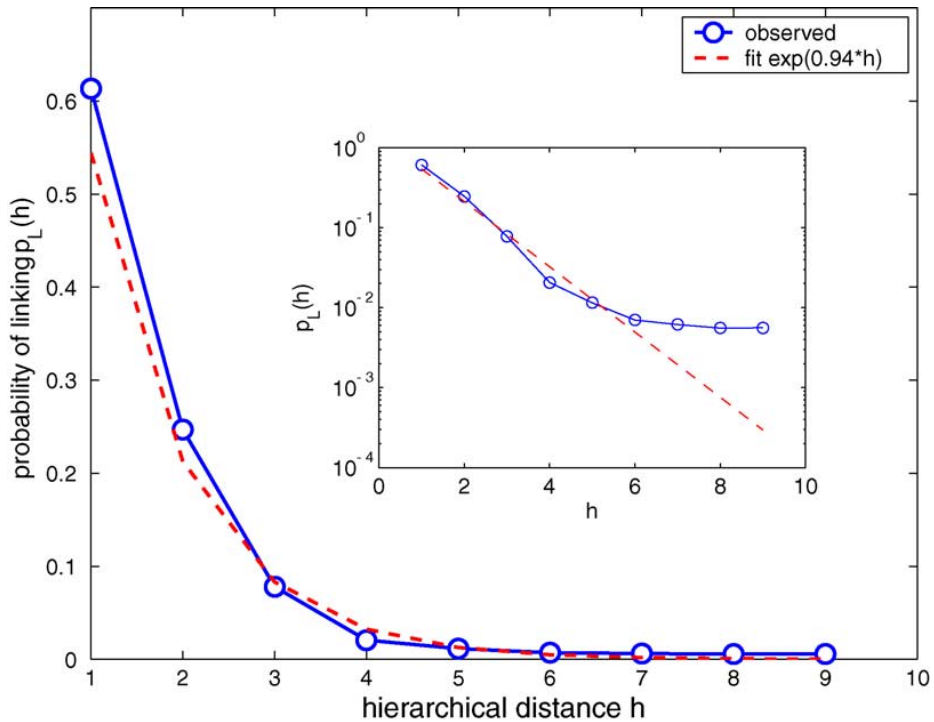


Fig. 4. Probability of linking as a function of the separation in the organizational hierarchy. The exponential parameter $\alpha = 0.94$, is in the searchable range of the Watts model (Watts et al., 2002).

[Figure 4 from Adamic and Adar]

Probability of email exchanges vs size of smallest common organizational unit

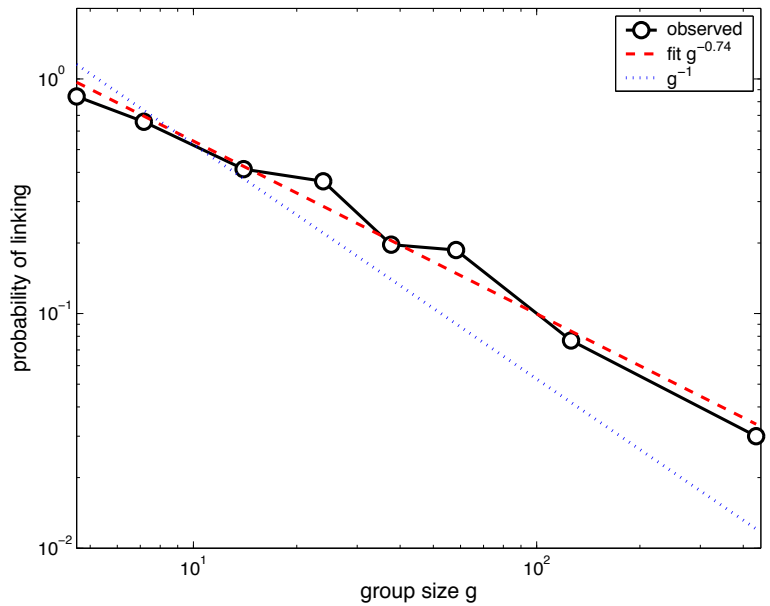


Figure 5: Probability of two individuals corresponding by email as a function of the size of the smallest organizational unit they both belong to. The optimum relationship derived in [7]is $p \sim g^{-1}$, g being the group size. The observed relationship is $p \sim g^{-3/4}$.

[Figure 5 from Adamic and Adar]

Final observations in chapter 20 of EK text

- The EK text suggests viewing the Milgram experiment as an example of decentralized problem solving (in this case solving a shortest path problem). An advertisement for distributed systems course.

- The EK text asks what other problem solving tasks might be amenable to such decentralized problem solving and how to analyze what can be done especially in large online networks.

- Finally the EK text briefly suggests the role of social status in determining the effectiveness of reaching a given target.
  - An email forwarding Milgram type 2003 study by Dodds et al shows that completion rates to all targets were low but were highest for "high status" targets and particularly small for "low status" targets.

- In section 12.6, the EK text speculates on structural reasons for the impact of status. This discussion leaves me with the sense that we are far from having any comprehensive understanding of such phenomena.