Web Mining and Recommender Systems

Triadic closure; strong & weak ties

Triangles

So far we've seen (a little about) how networks can be characterized by their connectivity patterns

What more can we learn by looking at higher-order properties, such as relationships between **triplets** of nodes?

Motivation

Q: Last time you found a job, was it through:

- A complete stranger?
- A close friend?
- An acquaintance?

A: Surprisingly, people often find jobs through **acquaintances** rather than through close friends (Granovetter, 1973)

Motivation

- Your friends (hopefully) would seem to have the greatest motivation to help you
- But! Your closest friends have limited information that you don't already know about
- Alternately, acquaintances act as a "bridge" to a different part of the social network, and expose you to new information

This phenomenon is known as **the strength of weak ties**

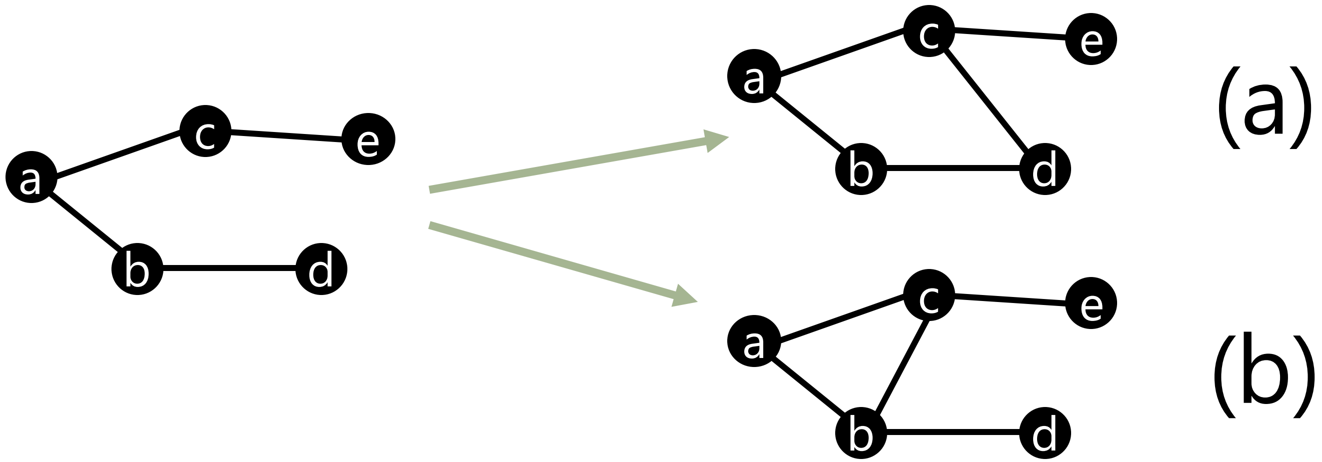
Motivation

- To make this concrete, we'd like to come up with some notion of "tie strength" in networks
- To do this, we need to go beyond just looking at edges in isolation, and looking at how an edge connects one part of a network to another

Refs:
"The Strength of Weak Ties", Granovetter (1973): http://goo.gl/wVJVlN
"Getting a Job", Granovetter (1974)

Triadic closure

Q: Which edge is most likely to form **next** in this (social) network?



A: (b), because it creates a **triad** in the network

Triangles

"If two people in a social network have a friend in common, then there is an increased likelihood that they will become friends themselves at some point in the future" (Ropoport, 1953)

Three reasons (from Heider, 1958; see Easley & Kleinberg):

- Every mutual friend *a* between *bik* and *camila* gives them an **opportunity** to meet
- If *bik* is friends with *aliyah*, then knowing that *camila* is friends with *aliyah* gives *bik* a reason to **trust** *camila*
- If *camila* and *bik* don't become friends, this causes stress for *aliyah* (having two friends who don't like each other), so there is an **incentive** for them to connect

Triangles

The extent to which this is true is measured by the (local) **clustering coefficient:**

- The clustering coefficient of a node *i* is the probability that two of *i*'s friends will be friends with each other:

neighbours of i

pairs of neighbours that are edges

$$C_i = \frac{\sum_{j,k \in \Gamma(i)} \delta((j,k) \in E)}{k_i(k_i - 1)}$$

(edges (j,k) and (k,j) are both counted for undirected graphs)

degree of node i

- This ranges between 0 (none of my friends are friends with each other) and 1 (all of my friends are friends with each other)

Triangles

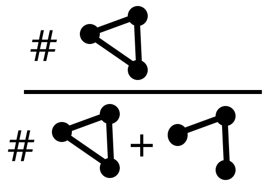
The extent to which this is true is measured by the (local) **clustering coefficient:**

- The clustering coefficient of the **graph** is usually defined as the average of local clustering coefficients

$$\bar{C} = \frac{1}{n} \sum_{i=1}^{n} C_i$$

- Alternately it can be defined as the fraction of connected triplets in the graph that are closed (these do not evaluate to the same thing!):

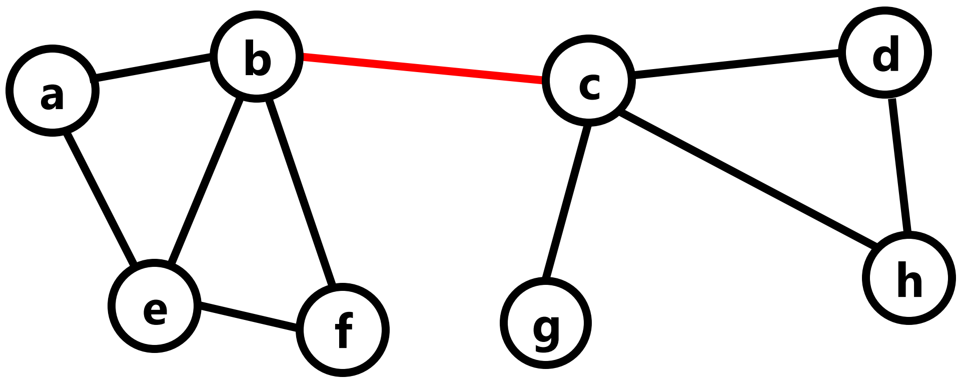
$$C = \frac{\#\ \text{of closed triplets}}{\#\ \text{of connected triplets}}$$



Bridges

Next, we can talk about the role of edges in relation to the rest of the network, starting with a few more definitions

1. Bridge edge

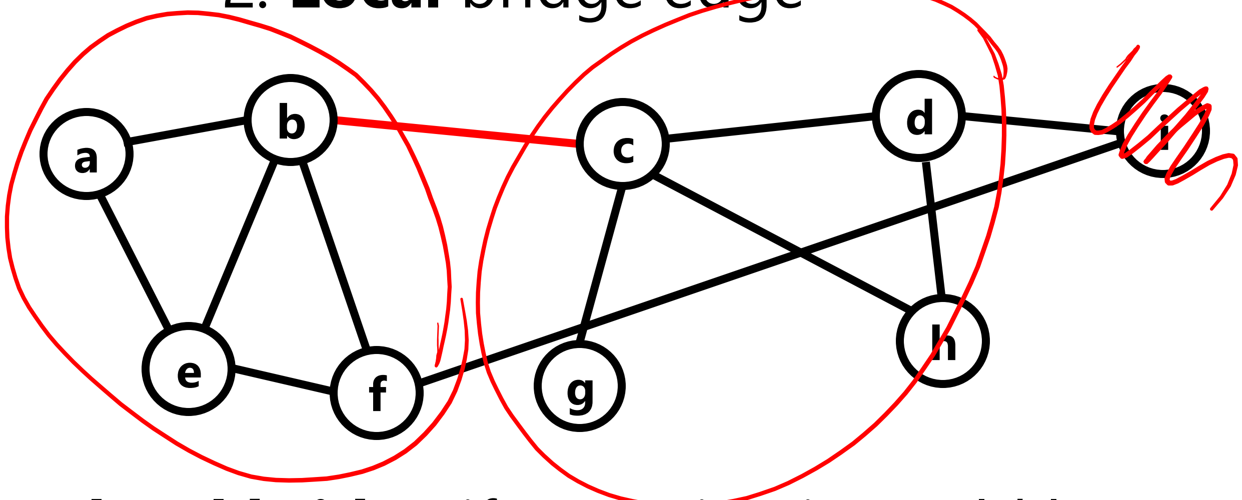


An edge (b,c) is a **bridge** edge if removing it would leave no path between b and c in the resulting network

Bridges

In practice, “bridges” aren’t a very useful definition, since there will be very few edges that completely isolate two parts of the graph

2. **Local** bridge edge

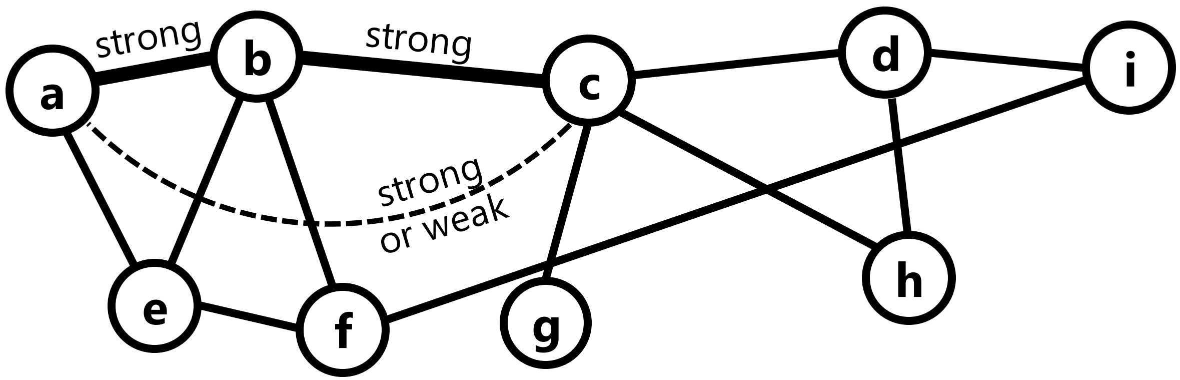


An edge (b,c) is a **local bridge** if removing it would leave no edge between b’s friends and c’s friends (though there could be more distant connections)

Strong & weak ties

We can now define the concept of "strong" and "weak" ties (which roughly correspond to notions of "friends" and "acquaintances"

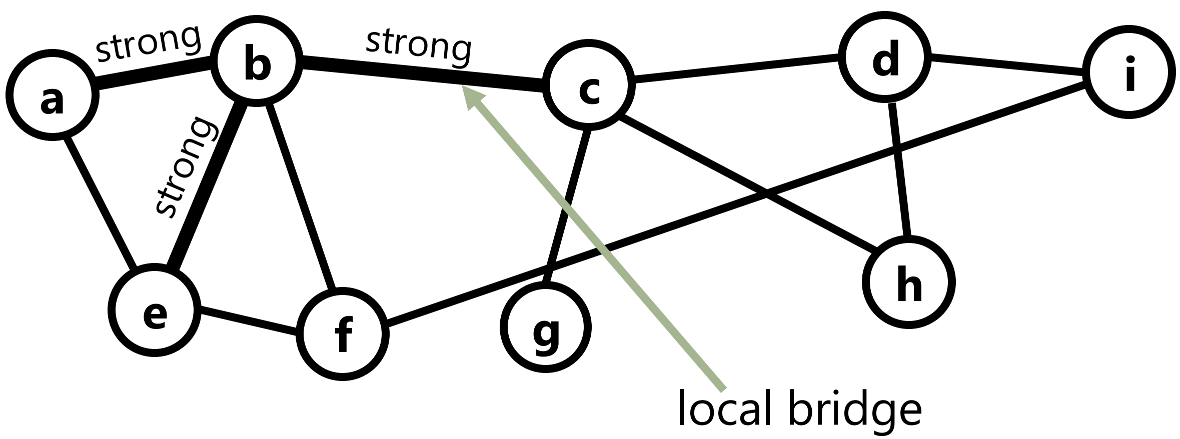
3. Strong triadic closure property



If (a,b) and (b,c) are connected by **strong** ties, there must be at least a **weak** tie between a and c

Strong & weak ties

Granovetter's theorem: if the strong triadic closure property is satisfied for a node, and that node is involved in two strong ties, then any incident local bridge must be a **weak tie**



Proof (by contradiction): (1) b has two strong ties (to a and e); (2) suppose it has a **strong** tie to c via a local bridge; (3) but now a tie must exist between c and a (or c and e) due to strong triadic closure; (4) so b → c cannot be a bridge

Strong & weak ties

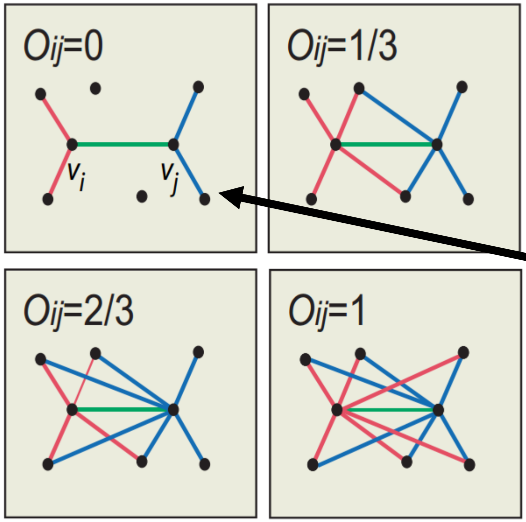
Granovetter's theorem: so, if we're receiving information from distant parts of the network (i.e., via "local bridges") then we must be receiving it via **weak ties**

Q: How to test this theorem empirically on real data?
A: Onnela et al. 2007 studied networks of mobile phone calls

Defn. 1: Define the "overlap" between two nodes to be the Jaccard similarity between their connections

$$O_{i,j} = \frac{\Gamma(i)\cap\Gamma(j)}{\Gamma(i)\cup\Gamma(j)}$$

neighbours of i



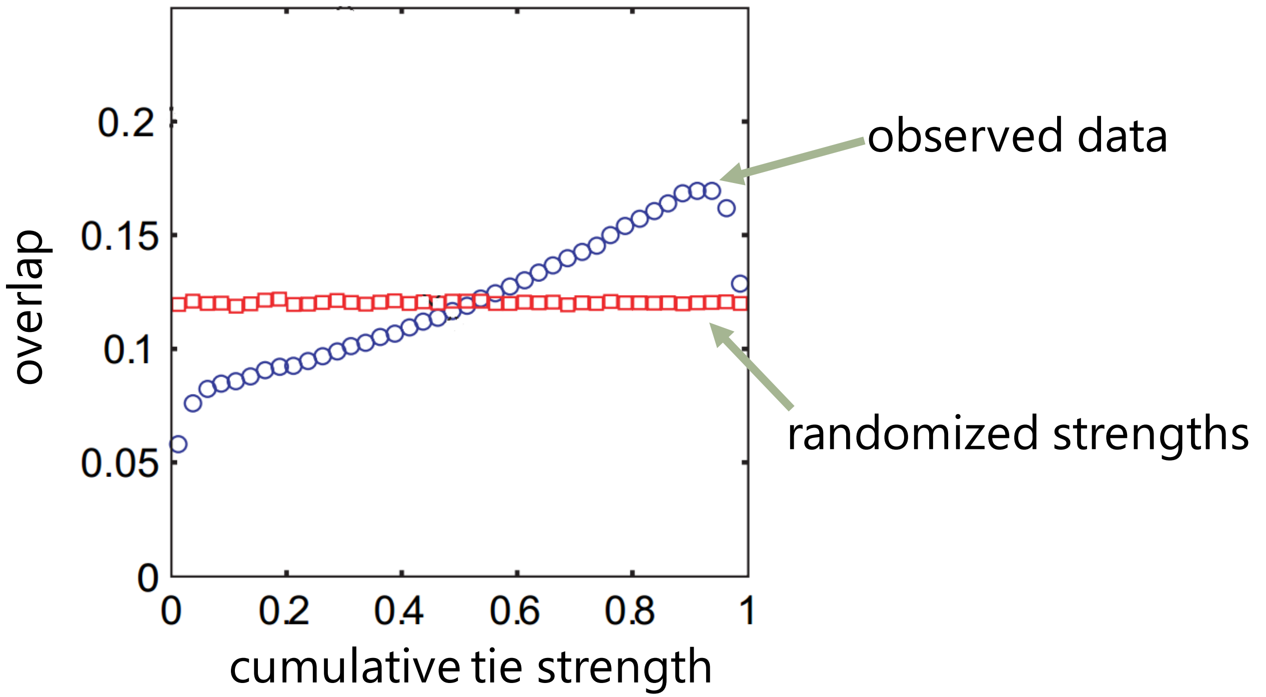
"local bridges" have overlap 0

(picture from Onnela et al., 2007)

Strong & weak ties

Secondly, define the "strength" of a tie in terms of the number of phone calls between *i* and *j*

finding: the "stronger" our tie, the more likely there are to be additional ties between our mutual friends



(picture from Onnela et al., 2007)

Strong & weak ties

Another case study (Ugander et al., 2012)

Suppose a user receives four e-mail invites to join facebook from users who are already on facebook. Under what conditions are we most likely to accept the invite (and join facebook)?

1. If those four invites are from four close friends?
2. If our invites are from found acquiantances?
3. If the invites are from a combination of friends, acquaintances, work colleagues, and family members?

hypothesis: the invitations are most likely to be adopted if they come from **distinct groups** of people in the network

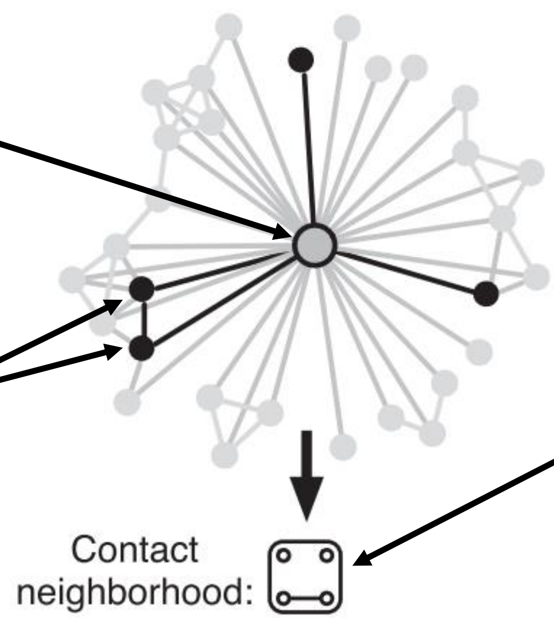
Strong & weak ties

Another case study (Ugander et al., 2012)

Let's consider the connectivity patterns amongst the people who tried to recruit us

user being recruited

users recruiting



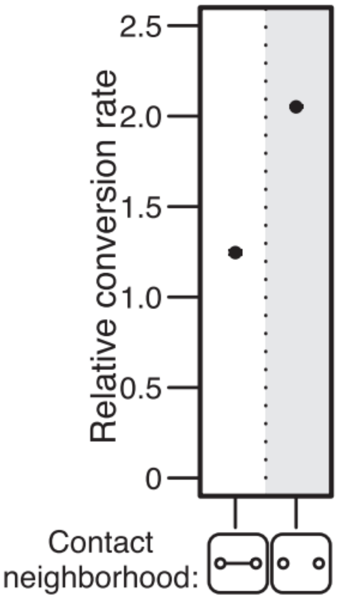
reachability between users attempting to recruit

(picture from Ugander et al., 2012)

Strong & weak ties

Another case study (Ugander et al., 2012)

Let's consider the connectivity patterns amongst the people who tried to recruit us



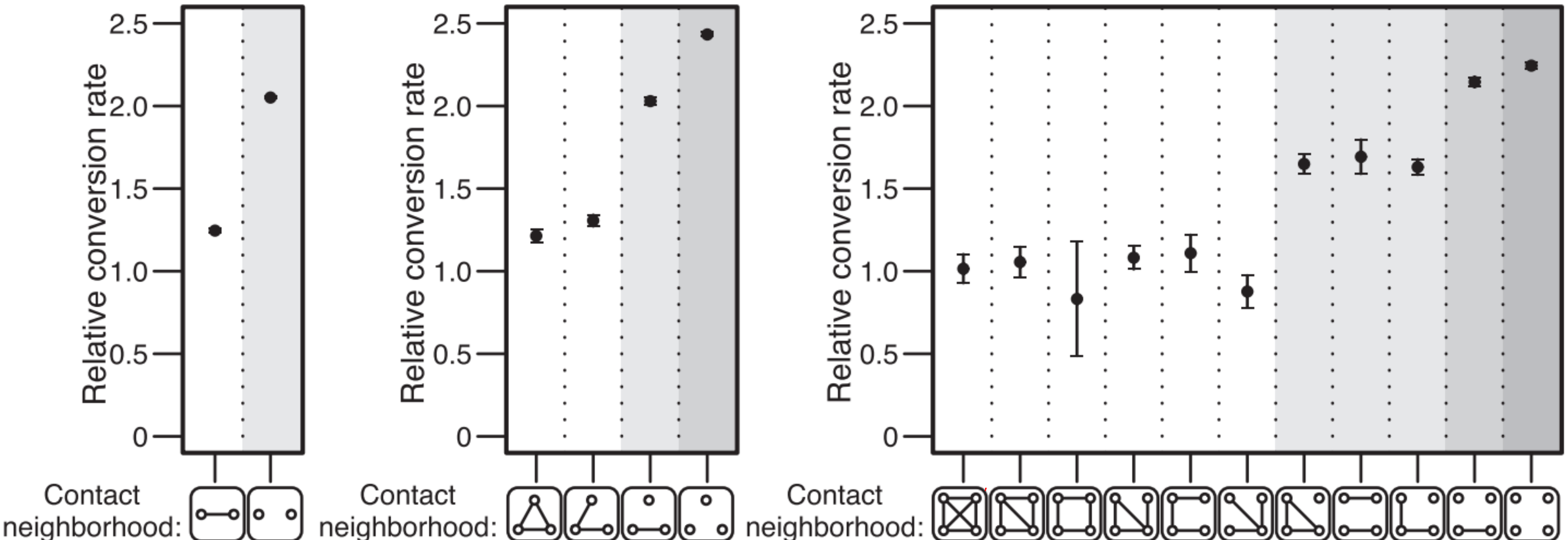
- **Case 1:** two users attempted to recruit
- **y-axis:** relative to recruitment by a single user
- **finding:** recruitments are **more likely to succeed** if they come from friends who are **not connected to each other**

(picture from Ugander et al., 2012)

Strong & weak ties

Another case study (Ugander et al., 2012)

Let's consider the connectivity patterns amongst the people who tried to recruit us



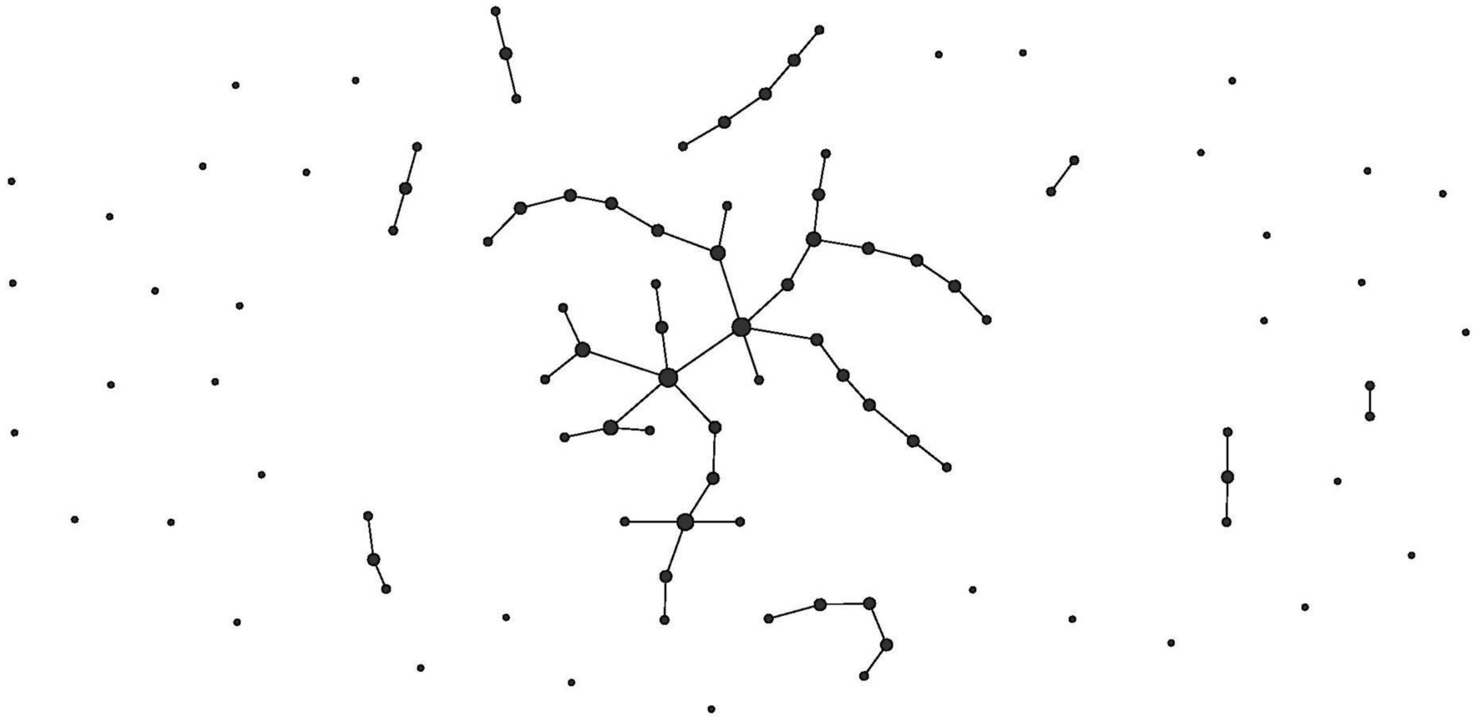
(picture from Ugander et al., 2012)

Web Mining and Recommender Systems

Social & Information Networks

Monday…

Random models of networks: Erdos Renyi random graphs



(picture from Wikipedia http://en.wikipedia.org/wiki/Erd%C5%91s%E2%80%93R%C3%A9nyi_model)

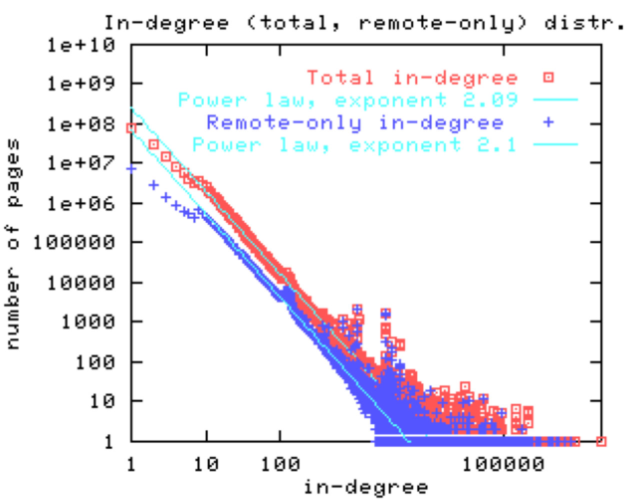
Preferential attachment models of network formation

Consider the following process to generate a network (e.g. a web graph):

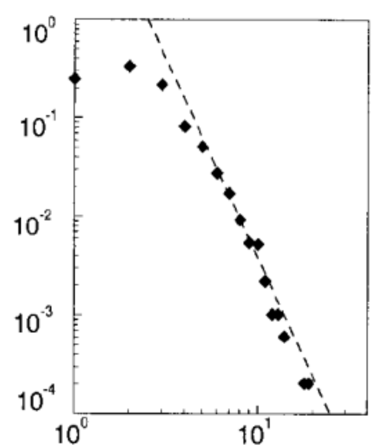
1. Order all of the N pages 1,2,3,…,N and repeat the following process for each page *j*:
2. Use the following rule to generate a link to another page:
   a. With probability *p*, link to a random page $i < j$
   b. Otherwise, choose a random page *i* and link to the page ***i links to***

Monday – power laws

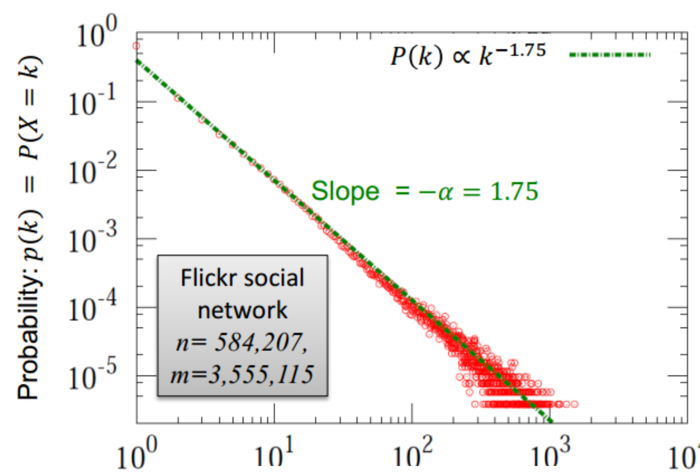
- Social and information networks often follow **power laws**, meaning that a few nodes have **many** of the edges, and many nodes have **a few** edges



e.g. web graph (Broder et al.)



e.g. power grid (Barabasi-Albert)

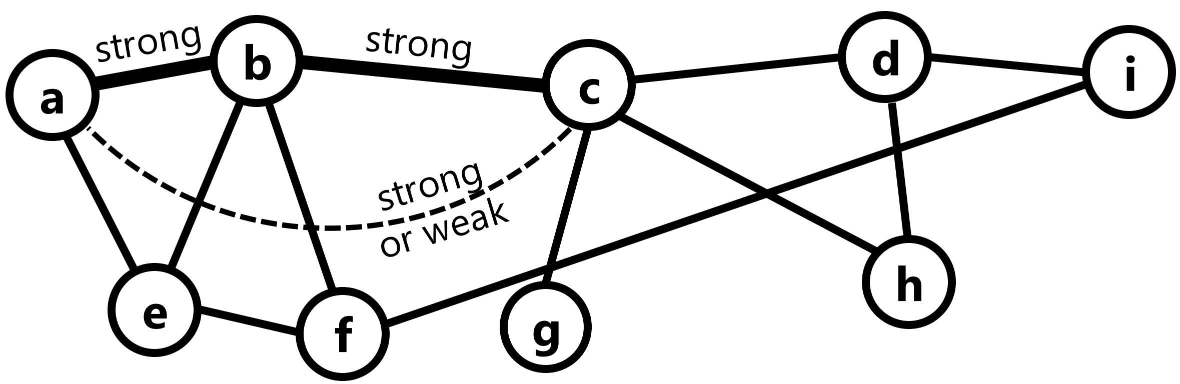


e.g. Flickr (Leskovec)

Monday - Strong & weak ties

We defined the concept of “strong” and “weak” ties (which roughly correspond to notions of “friends” and “acquaintances”)

3. Strong triadic closure property



If (a,b) and (b,c) are connected by **strong** ties, there must be at least a **weak** tie between a and c

Today

How can we **characterize, model**, and **reason about** the structure of social networks?

1. Models of network structure
2. Power-laws and scale-free networks, “rich-get-richer” phenomena
3. Triadic closure and “the strength of weak ties”
4. Small-world phenomena
5. Hubs & Authorities; PageRank

Web Mining and Recommender Systems

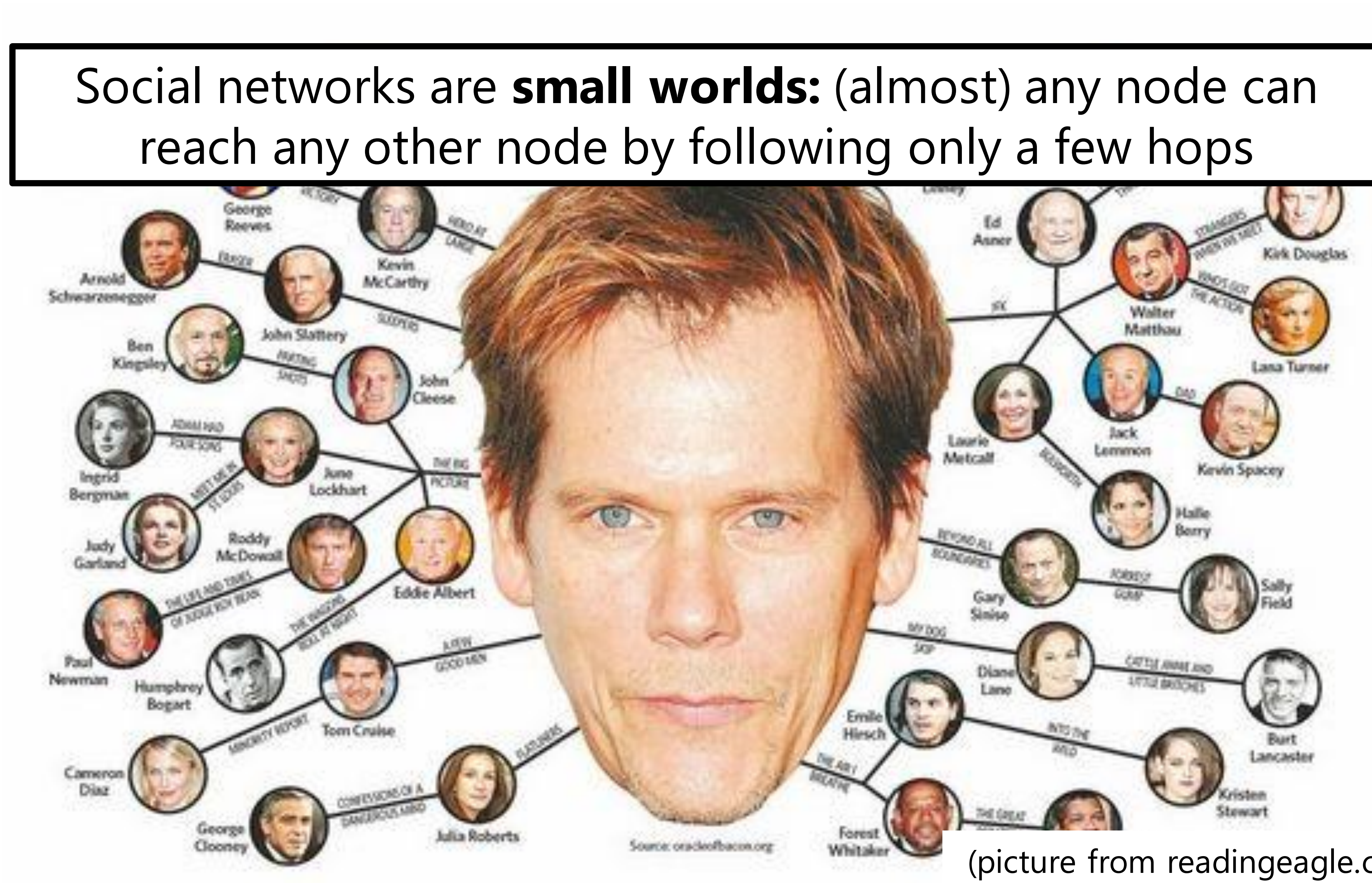
Small-world phenomena

Small worlds

- We’ve seen random graph models that reproduce the **power-law** behaviour of real-world networks
- But what about other types of network behaviour, e.g. can we develop a random graph model that reproduces small-world phenomena? Or which have the correct ratio of closed to open triangles?

Small worlds

Social networks are **small worlds:** (almost) any node can reach any other node by following only a few hops



(picture from readingeagle.com)

Six degrees of separation

Another famous study…

- Stanley Milgram wanted to test the (already popular) hypothesis that people in social networks are separated by only a small number of "hops"
- He conducted the following experiment:

1. "Random" pairs of users were chosen, with start points in Omaha & Wichita, and endpoints in Boston
2. Users at the start point were sent a letter describing the study: they were to get the letter to the endpoint, but only by contacting somebody with whom they had a direct connection
3. So, either they sent the letter directly, or they wrote their name on it and passed it on to somebody they believed had a high likelihood of knowing the target (they also mailed the researchers so that they could track the progress of the letters)

Six degrees of separation

Another famous study…

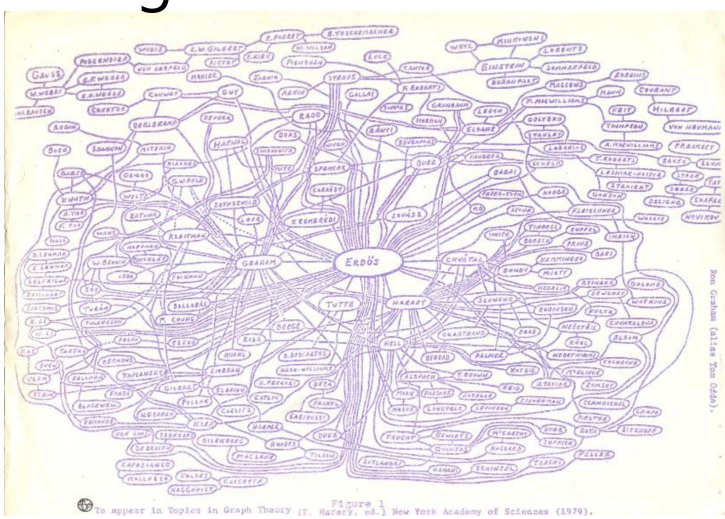
Of those letters that reached their destination, the average path length was between 5.5 and 6 (thus the origin of the expression). At least two facts about this study are somewhat remarkable:

- First, that short paths appear to be abundant in the network
- Second, that people are capable of discovering them in a "decentralized" fashion, i.e., they're somehow good at "guessing" which links will be closer to the target

Six degrees of separation

Such small-world phenomena turn out to be abundant in a variety of network settings

e.g. Erdos numbers:



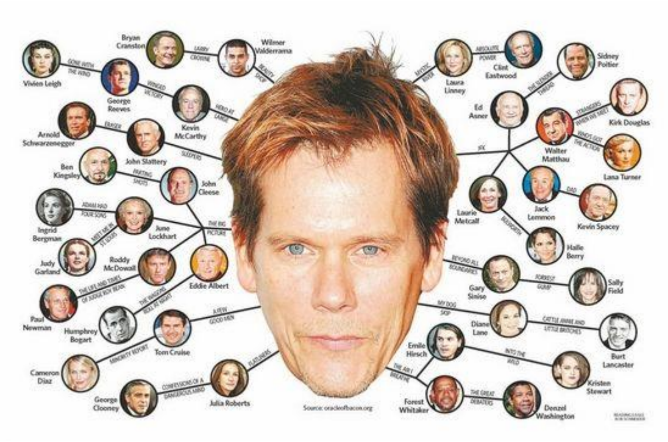
Erdös # 0	-	1	person
Erdös # 1	-	504	people
Erdös # 2	-	6593	people
Erdös # 3	-	33605	people
Erdös # 4	-	83642	people
Erdös # 5	-	87760	people
Erdös # 6	-	40014	people
Erdös # 7	-	11591	people
Erdös # 8	-	3146	people
Erdös # 9	-	819	people
Erdös #10	-	244	people
Erdös #11	-	68	people
Erdös #12	-	23	people
Erdös #13	-	5	people

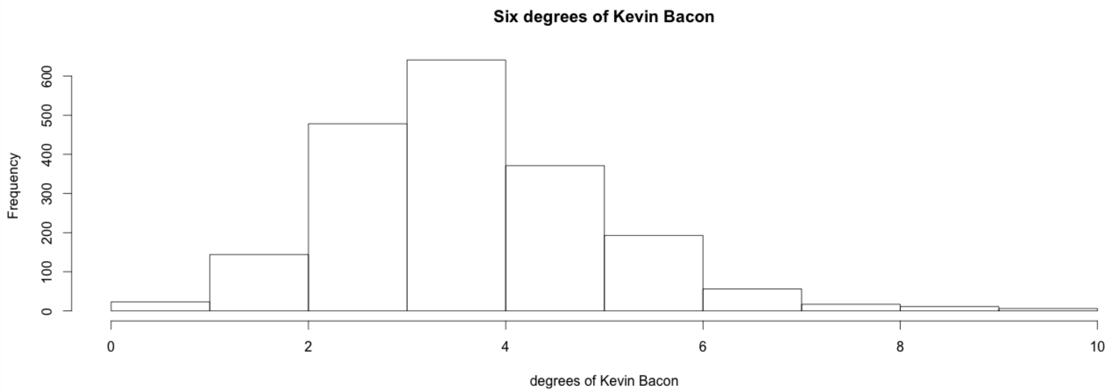
http://www.oakland.edu/enp/trivia/

Six degrees of separation

Such small-world phenomena turn out to be abundant in a variety of network settings

e.g. Bacon numbers:





linkedscience.org & readingeagle.com

Six degrees of separation

Such small-world phenomena turn out to be abundant in a variety of network settings

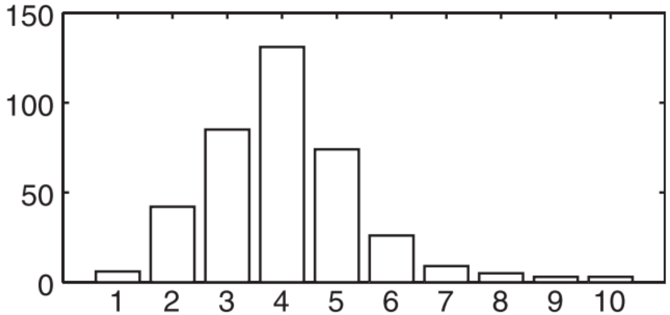
Bacon/Erdos numbers:

Kevin Bacon→Sarah Michelle Gellar→**Natalie Portman**→Abigail Baird→Michael Gazzaniga→J. Victor→Joseph Gillis→Paul Erdos

Six degrees of separation

Dodds, Muhamed, & Watts repeated Milgram's experiments using e-mail

- 18 "targets" in 13 countries
- 60,000+ participants across 24,133 chains
- Only 384 (!) reached their targets



Histogram of (completed) chain lengths – average is just **4.01!**

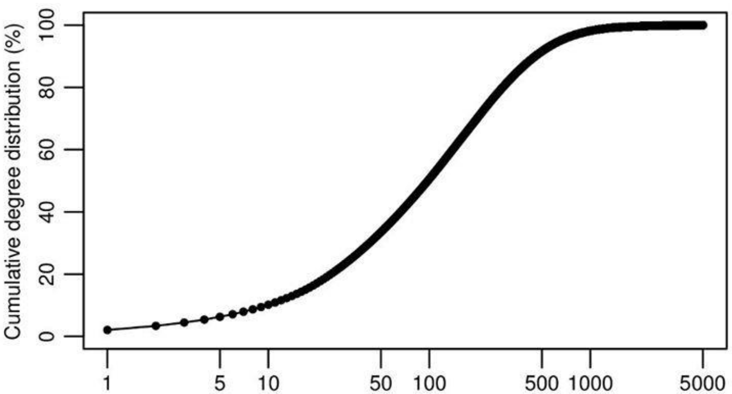
L	*N*	Location	Travel	Family	Work	Education	Friends	Cooperative	Other
1	19,718	33	16	11	16	3	9	9	3
2	7,414	40	11	11	19	4	6	7	2
3	2,834	37	8	10	26	6	6	4	3
4	1,014	33	6	7	31	8	5	5	5
5	349	27	3	6	38	12	6	3	5
6	117	21	3	5	42	15	4	5	5
7	37	16	3	3	46	19	8	5	0

Reasons for choosing the next recipient at each point in the chain

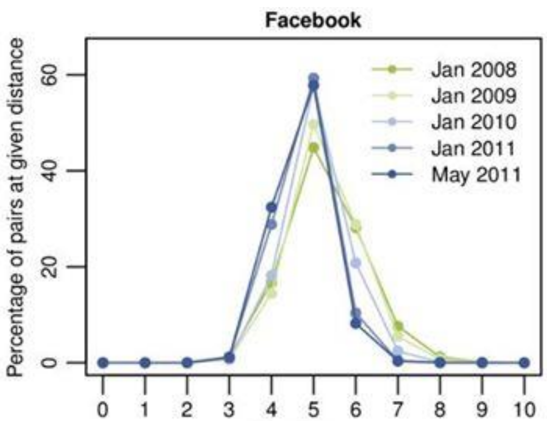
from http://www.cis.upenn.edu/~mkearns/teaching/NetworkedLife/columbia.pdf

Six degrees of separation

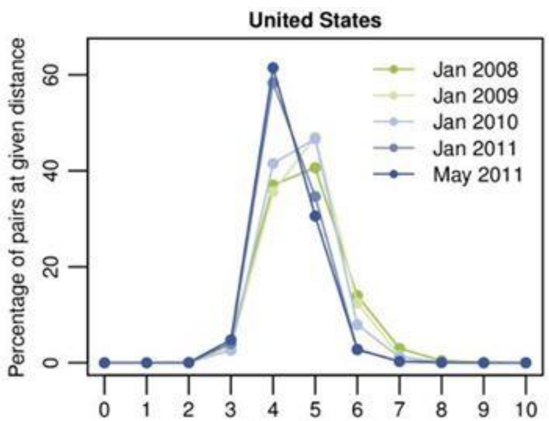
Actual shortest-path distances are similar to those in Dodds' experiment:



Cumulative degree distribution (# of friends) of Facebook users



Hop distance between Facebook users



Hop distance between users in the US

This suggests that people choose a reasonably good heuristic when choosing shortest paths in a decentralized fashion (assuming that FB is a good proxy for "real" social networks)

from "the anatomy of facebook": http://goo.gl/H0bkWY

Six degrees of separation

Q: is this result surprising?

- **Maybe not:** We have ~100 friends on Facebook, so 100^2 friends-of-friends, 10^6 at length three, 10^8 at length four, **everyone** at length 5

- **But:** Due to our previous argument that people close triads, the **vast majority** of new links will be between friends of friends (i.e., we're increasing the **density** of our local network, rather than making distant links more reachable)

- In fact **92%** of new connections on Facebook are to a friend of a friend (Backstrom & Leskovec, 2011)

Six degrees of separation

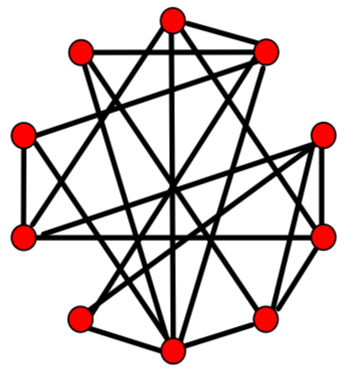
Definition: Network diameter

- A network's diameter is the length of its **longest shortest path**
- **Note:** iterating over all pairs of nodes *i* and *j* and then running a shortest-paths algorithm is going to be prohibitively slow
- Instead, the "all pairs shortest paths" algorithm computes all shortest paths simultaneously, and is more efficient (O(N^2logN) to O(N^3), depending on the graph structure)
- In practice, one doesn't **really** care about the diameter, but rather the distribution of shortest path lengths, e.g., what is the average/90^{th} percentile shortest-path distance
- This latter quantity can computed just by randomly sampling pairs of nodes and computing their distance
- When we say that a network exhibits the "small world phenomenon", we are really saying this latter quantity is small

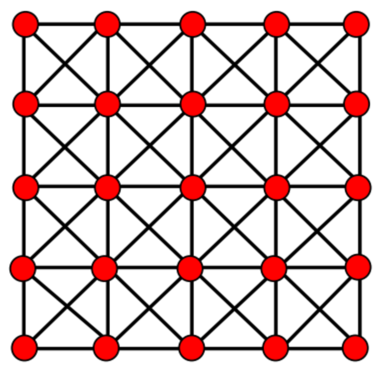
Six degrees of separation

Q: is this a contradiction?

- How can we have a network made up of **dense communities** that is simultaneously a **small world?**
- The shortest paths we could possibly have are O(log n) (assuming nodes have constant degree)



random connectivity – low diameter, low clustering coefficient



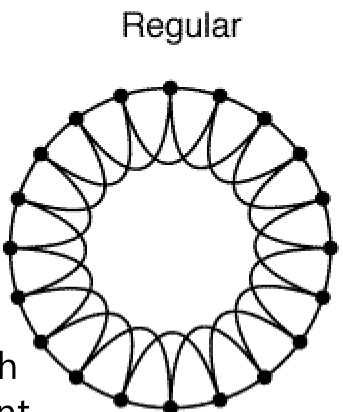
regular lattice – high clustering coefficient, high diameter

picture from http://cs224w.Stanford.edu

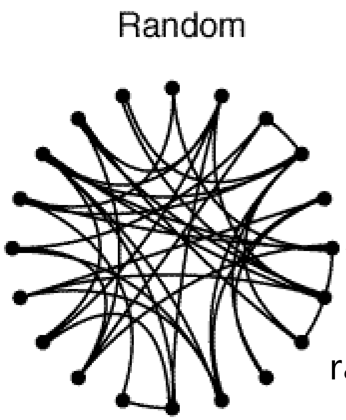
Six degrees of separation

We'd like a model that reproduces small-world phenomena

regular lattic – high clustering coefficient, high diameter



We'd like something "in between" that exhibits both of the desired properties (high cc, low diameter)



random connectivity – low diameter, low clustering coefficient

from http://www.nature.com/nature/journal/v393/n6684/abs/393440a0.html

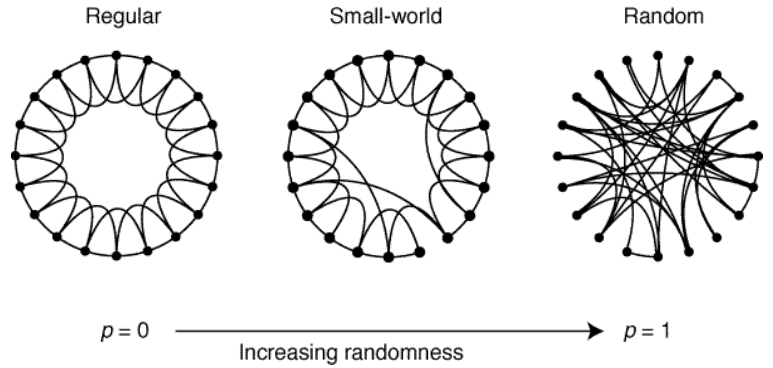
Six degrees of separation

The following model was proposed by Watts & Strogatz (1998)

1. Start with a regular lattice graph (which we know to have high clustering coefficient)

Next – introduce some randomness into the graph

2. For each edge, with prob. *p*, reconnect one of its endpoints



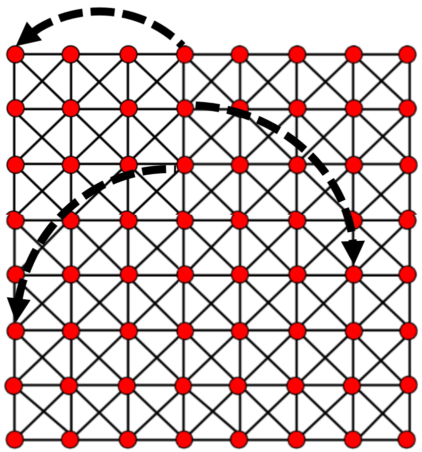
as we increase *p*, this becomes more like a random graph

from http://www.nature.com/nature/journal/v393/n6684/abs/393440a0.html

Six degrees of separation

Slightly simpler (to reason about formulation) with the same properties

1. Start with a regular lattice graph (which we know to have high clustering coefficient)
2. From each node, add an additional random link



etc.

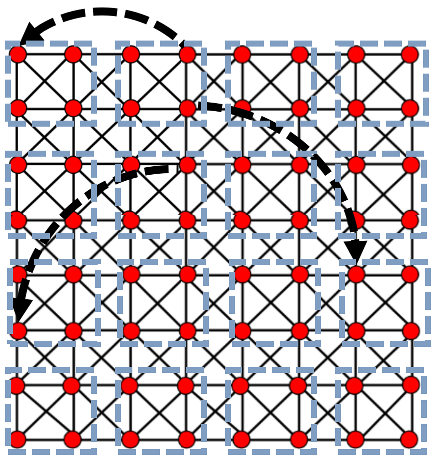
Six degrees of separation

Slightly simpler (to reason about formulation) with the same properties

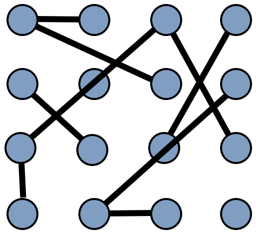
Conceptually, if we combine groups of adjacent nodes into “supernodes”, then what we have formed is a **4-regular** random graph

(very handwavy) proof:

- The clustering coefficient is still high (each node is incident to 12 triangles)
- **4-regular** random graphs have diameter O(log n) (Bollobas, 2001), so the whole graph has diameter O(log n)



connections between supernodes:



(should be a 4-regular random graph, I didn’t finish drawing the edges)

The Watts-Strogatz model

- Helps us to understand the relationship between dense clustering and the small-world phenomenon
- Reproduces the small-world structure of realistic networks
- Does **not** lead to the correct degree distribution (no power laws)

(see Klemm, 2002: "Growing scale-free networks with small-world behavior" http://ifisc.uib-csic.es/victor/Nets/sw.pdf)

Six degrees of separation

So far...

- Real networks exhibit **small-world** phenomena: the average distance between nodes grows only logarithmically with the size of the network
- Many experiments have demonstrated this to be true, in mail networks, e-mail networks, and on Facebook etc.
- But we know that social networks are highly **clustered** which is somehow inconsistent with the notion of having low diameter
- To explain this apparent contradiction, we can model networks as some combination of highly-clustered nodes, plus some fraction of “random” connections

Questions?

Further reading:

- Easley & Kleinberg, Chapter 20
- Milgram's paper

"An experimental study of the small world problem"
http://www.uvm.edu/~pdodds/files/papers/others/1969/travers1969.pdf

- Dodds et al.'s small worlds paper

http://www.cis.upenn.edu/~mkearns/teaching/NetworkedLife/columbia.pdf

- Facebook's small worlds paper

http://arxiv.org/abs/1111.4503

- Watts & Strogatz small worlds model

"Collective dynamics of 'small world' networks"
file:///C:/Users/julian/Downloads/w_s_NATURE_0.pdf

- More about random graphs

"Random Graphs" (Bollobas, 2001), Cambridge University Press

Web Mining and Recommender Systems

Hubs and Authorities; PageRank

Trust in networks

We already know that there's considerable variation in the connectivity structure of nodes in networks

So how can we find nodes that are in some sense "important" or "authoritative"?

- In links?
- Out links?
- Quality of content?
- Quality of linking pages?
- etc.

1. The "HITS" algorithm
Two important notions:

Hubs:

We might consider a node to be of "high quality" if it links to many high-quality nodes. E.g. a high-quality page might be a "hub" for good content
(e.g. Wikipedia lists)

Authorities:

We might consider a node to be of high quality if many high-quality nodes link to it
(e.g. the homepage of a popular newspaper)

Trust in networks

This “self-reinforcing” notion is the idea behind the HITS algorithm

- Each node *i* has a “hub” score h_i
- Each node *i* has an “authority” score a_i

- The hub score of a page is the sum of the authority scores of pages it links to
- The authority score of a page is the sum of hub scores of pages that link to it

Trust in networks

This "self-reinforcing" notion is the idea behind the HITS algorithm

Algorithm:

$$a_i^{(0)} = \frac{1}{\sqrt{n}} \qquad h_i^{(0)} = \frac{1}{\sqrt{n}}$$

iterate until convergence:

$$\forall_i a_i^{(t+1)} = \sum_{j \to i} h_j^{(t)}$$

pages that link to i

$$\forall_i h_i^{(t+1)} = \sum_{i \to j} a_j^{(t)}$$

pages that i links to

normalize:

$$\|a^{(t+1)}\|_2^2 = 1 \qquad \|h^{(t+1)}\|_2^2 = 1$$

Trust in networks

This “self-reinforcing” notion is the idea behind the HITS algorithm

This can be re-written in terms of the adjacency matrix (*A*)

$A =$

$$a_i^{(0)} = \frac{1}{\sqrt{n}} \qquad h_i^{(0)} = \frac{1}{\sqrt{n}}$$

iterate until convergence:

$$a^{(t+1)} = A^T h^{(t)}$$

$$h^{(t+1)} = A a^{(t)}$$

skipping a step:

$$a^{(t+2)} = (A^T A)^t a^{(t)}$$

$$h^{(t+2)} = (A A^T)^t h^{(t)}$$

normalize:

$$\|a^{(t+1)}\|_2^2 = 1 \qquad \|h^{(t+1)}\|_2^2 = 1$$

Trust in networks

This "self-reinforcing" notion is the idea behind the HITS algorithm

So at convergence we seek stationary points such that

$$A^T A a = c' \cdot a$$

$$A A^T h = c'' \cdot h$$

(constants don't matter since we're normalizing)

- This can only be true if the authority/hub scores are **eigenvectors** of A^TA and AA^T
- In fact this will converge to the eigenvector with the largest eigenvalue (see: Perron-Frobenius theorem)

Trust in networks

The idea behind PageRank is very similar:

- Every page gets to "vote" on other pages
- Each page's votes are proportional to that page's importance
- If a page of importance *x* has *n* outgoing links, then each of its votes is worth *x/n*
- Similar to the previous algorithm, but with only a single a term to be updated (the rank r_i of a page *i*)

$$\forall_i r_i^{(t+1)} = \sum_{j \to i} \frac{r_j^{(t)}}{|\Gamma(j)|}$$

rank of linking pages

# of links from linking pages

Trust in networks

The idea behind PageRank is very similar:

Matrix formulation:
each **column** describes the out-links of one page, e.g.:

$$M = \left.\overbrace{\begin{pmatrix} \frac{1}{3} & 0 & \frac{1}{4} & 1 \\ 0 & 0 & \frac{1}{4} & 0 \\ \frac{1}{3} & \frac{1}{2} & \frac{1}{4} & 0 \\ \frac{1}{3} & \frac{1}{2} & \frac{1}{4} & 0 \end{pmatrix}}^{\text{pages}}\right\}\text{pages}$$

this out-link gets 1/3 votes since this page has three out-links

column-stochastic matrix (columns add to 1)

Trust in networks

The idea behind PageRank is very similar:

Then the update equations become:

$$r^{(t+1)} = Mr^{(t)}$$

And as before the stationary point is given by the eigenvector of M with the highest eigenvalue

Summary

The level of "authoritativeness" of a node in a network should somehow be defined in terms of the pages that link to (it or the pages it links from), and *their* level of authoritativeness

- Both the HITS algorithm and PageRank are based on this type of "self-reinforcing" notion
- We can then measure the centrality of nodes by some iterative update scheme which converges to a stationary point of this recursive definition
- In both cases, a solution was found by taking the principal eigenvector of some matrix encoding the link structure

This week

- We’ve seen how to characterize networks by their degree distribution (degree distributions in many real-world networks follow power laws)
- We’re seen some random graph models that try to mimic the degree distributions of real networks
- We’ve discussed the notion of “tie strength” in networks, and shown that edges are likely to form in “open” triads
- We’ve seen that real-world networks often have small diameter, and exhibit “small-world” phenomena
- We’ve seen (very quickly) two algorithms for measuring the “trustworthiness” or “authoritativeness” of nodes in networks

Questions?

Further reading:

- Easley & Kleinberg, Chapter 14
- The "HITS" algorithm (aka "Hubs and Authorities")

"Hubs, authorities, and communities" (Kleinberg, 1999)

http://cs.brown.edu/memex/ACM_HypertextTestbed/papers/10.html

Web Mining and Recommender Systems

Assignment 1 solutions

Strong read prediction

HW3: $f(u,b) = [1,\ pop(b),\ \max_{b'} Jac(b,b'),$

min

mean

max $\quad Jac(u,u'),$

~~cat(u,b)~~ $]$

use classifier for ranking

for each u, top 50% of (u,b) → 1

bottom 50% → 0

$$\sigma\left(\gamma_{u}\cdot\gamma_{i}-\gamma_{u}\cdot\gamma_{j}\right)$$

rating pred

- $\underbrace{\alpha + \beta_i + \beta_u}_{\lambda_1} + \underbrace{\gamma_u \cdot \gamma_i}_{\lambda_2}$

- init. of $\alpha\ \beta\ \gamma$

- $\alpha\ \beta$: output of HW

- γ : normal