BA Network Analytics
Lecture 3

Kalyan Talluri

Imperial College Business School, London

kalyan.talluri@imperial.ac.uk

Jan-Feb 2021

Three phenomena observed in social networks

Phenomenon observed in many large social networks

- Small-world phenomenon: The length between a random pair of nodes is small
- Homophily
- Strength-of-weak ties

We see now what they mean, and how they were discovered.

Small-World phenomenon[16]

The Milgram experiment (origin of "Six-degrees of separation"):

- Random people from Nebraska were asked to send a letter (via intermediaries) to a stock broker in Boston
- S/he could only send to someone with whom they were on a first-name basis
- Result: Median length was 6 — Six-degrees of Separation

Experiment frequently repeated on facebook yields a lower average number

[16] Reference: EK: Section 20.1; Barabási Section 3.8

Homophily[17]

"Birds of a Feather Flock Together" Philemon Holland (1600 "As commonly birds of a feather will flye together")

- Age, race, gender, religion, profession ...
  - Only 8% of people have any people of another race that they "discuss important matters" (Marsden '87)
  - Interracial marriages U.S.: 1% of white marriages, 5% of black marriages, 14% of Asian marriages (Fryer '07)
  - In middle school, less than 10% of "expected" cross-race friendships exist (Shrum et al. '88)
  - Closest friend: 10% of men name a woman, 32% of women name a man (Verbrugge '77)

[1] Reference: EK §4.1; creates "endogeneity" problems in the econometrics

Strength of weak-ties[18]

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 (Why?...)

- Links differ in terms of strength; e.g.: Friends vs. Acquaintance
  - Amount of contact time, affection, trust
  - Edges are labeled *strong* or *weak*
- Mark Granovetter (1974):Interviewed 54 people who found jobs. Job-seekers obtain useful job info through personal contacts
  - 16.7% via strong tie (at least 2 interactions/week)
  - 55.7% via medium tie (at least 1 interactions/week)
  - 27.6% via weak tie ($<$ 1 interactions/year)
- Best job leads come more often from acquaintances than from close friends
- Why? your close friends likely don't have any more information than you do

Motivation for the measure of *clustering coefficient of a node*: Fraction of your neighbors who are friends with each other. Real-world networks tend to have local clusters with high clustering coefficient, connected by bridges, with hubs, surprisingly, having lower clustering coefficient.

[18] Reference: EK: Section 3.1

Centrality Measures of a graph.

Their main use is to identify "important" nodes. If nodes represent actors in a social network, the interpretation could be that they are the most powerful, or influential, or important.

The measures can also be used for ranking. Order the nodes by decreasing values of the measure.

There is no single or dominant measure in use. Many have been proposed, and they all have some intuition. We cover the main ones.

Example 1 from Organizational Studies: Network Analysis of an organization[19]

Nodes colored by organizational position

(d) The links of the largest hub (red) and those two links away from this hub (orange), demonstrate that a significant fraction of employees are at most two links from this hub. But who is this hub? He is the employee in charge of safety and environmental issues. Hence he regularly visits each location and talks with the employees. He is connected to everyone except the top management. With little knowledge of the true intentions of the management, he passes on information that he collects along his trail, effectively running a gossip center.

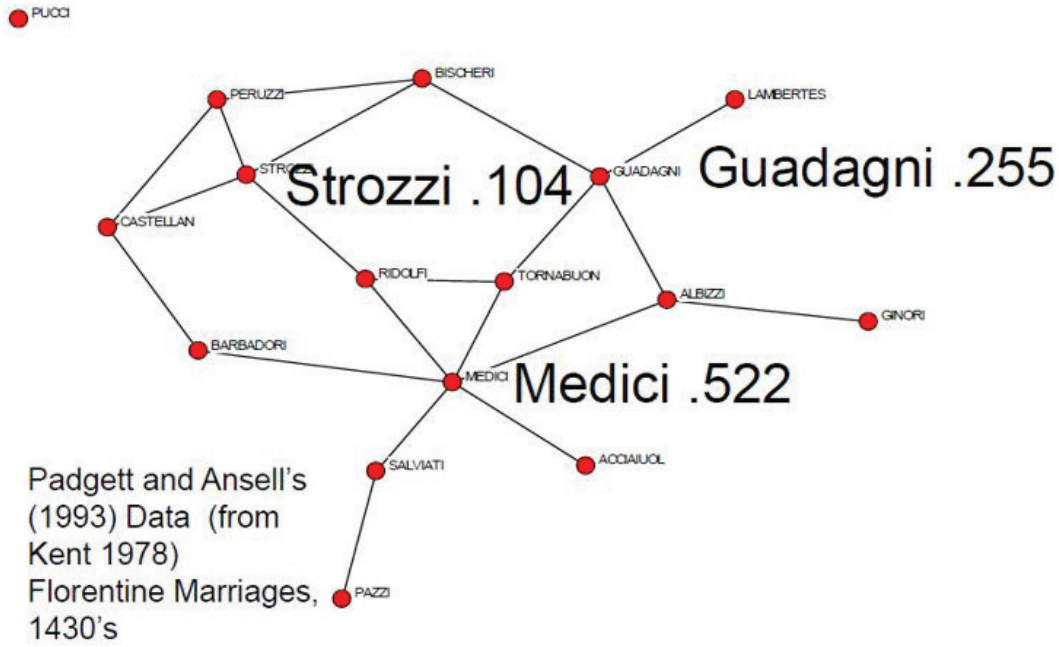
Should they fire or promote the biggest hub? What is the best solution to this problem?

`http://networksciencebook.com/chapter/1#societal-impact`

[19] Example from Barabasi, Figure 1.7; be aware that story is from his consulting company and engagement

Example 2 from Economic History: Understand power via network connections

The following is a network of families in 15th century Florence, interlinked by marriage.[20]



Similarly the 2009 financial crises, other financial networks . . . talk
Takeaway: Centrality measures give a number to each actor, i.e. quantify visual notions of centrality and importance. Once calculated, we have to find a meaning as to why

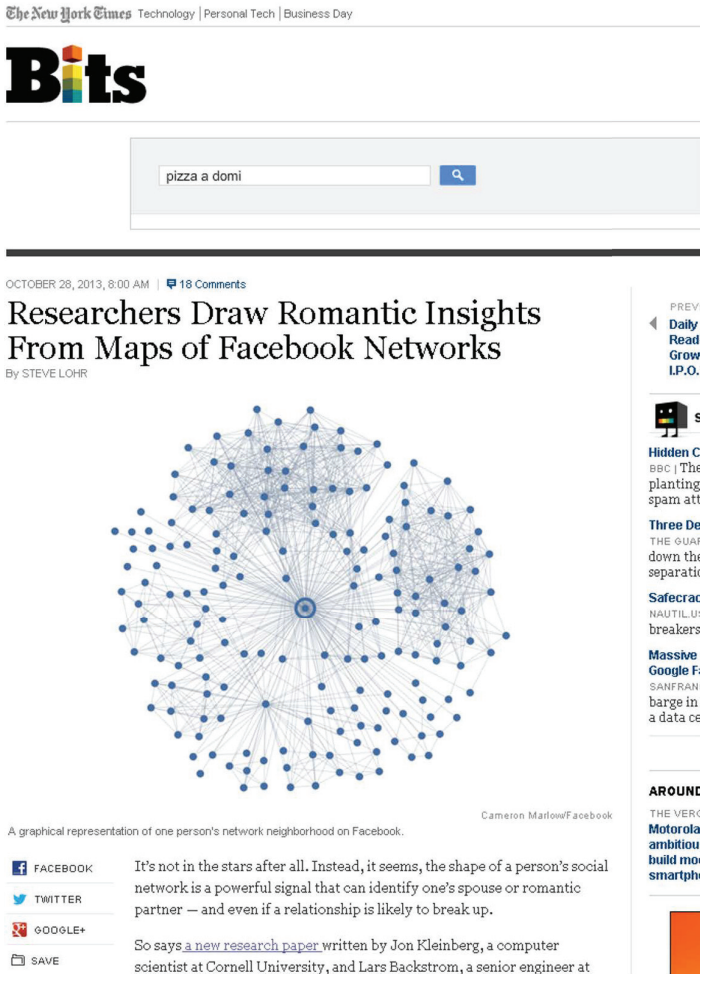
[20] I took the example from Jackson *Social and Economic Networks*

Example 3 from Social Networks for ranking and predictin: Relationship prediction based on a centrality measure prediction based on some network measures (knowing only the network topology)[21]

Discussion question: How do they use centrality measure to understand an unobservable phenomenon? (*we will come to this later*)

[21] Takeaway: we can come up with some task-specific measures also that might work better

Casestudy: Relationship prediction based on a centrality measure[22]



The New York Times Technology | Personal Tech | Business Day

Bits

pizza a domi

OCTOBER 28, 2013, 8:00 AM | 18 Comments

Researchers Draw Romantic Insights From Maps of Facebook Networks

By STEVE LOHR

Cameron Marlow/Facebook

A graphical representation of one person's network neighborhood on Facebook.

FACEBOOK
TWITTER
GOOGLE+
SAVE

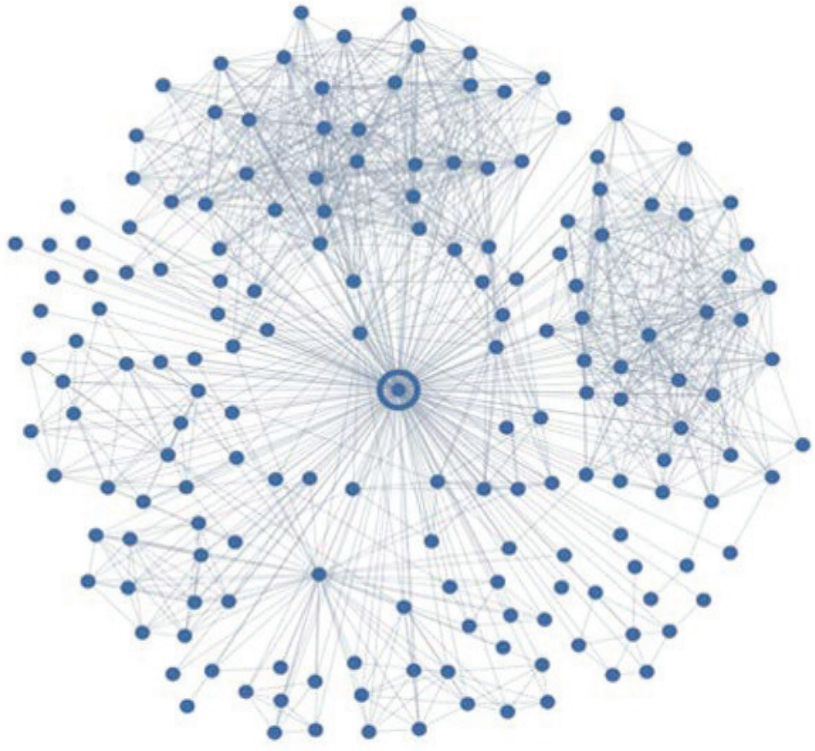
It's not in the stars after all. Instead, it seems, the shape of a person's social network is a powerful signal that can identify one's spouse or romantic partner — and even if a relationship is likely to break up.

So says a new research paper written by Jon Kleinberg, a computer scientist at Cornell University, and Lars Backstrom, a senior engineer at

[1] *Romantic Partnerships and the Dispersion of Social Ties: A Network Analysis of Relationship Status on Facebook*, Backstrom and Kleinberg. Authors have unique access to the ground-truth!

Example: Relationship prediction

The person declares on Facebook he/she is in a relationship. Prediction task: With whom?



Data: The person's ego network (the induced subgraph of their neighbors)
Performance measure: How often can you predict correctly over 1.3 million such users
Information: Network structure only, or [+ Features (messages, photos, photos with both appearing, . . .)]

Centrality measures[23]

insight + algebra + algorithms $\rightarrow$ condensed and valuable view/meaning on individual actors (the nodes) of large networks

[23] Note: A new centrality measure comes up practically every other week. They are all intuitive and have a reason. Many are difficult (time-consuming) to compute. We concentrate on the bread-and-butter ones and not the exotic new ones. NetworkX has around 20 coded up.

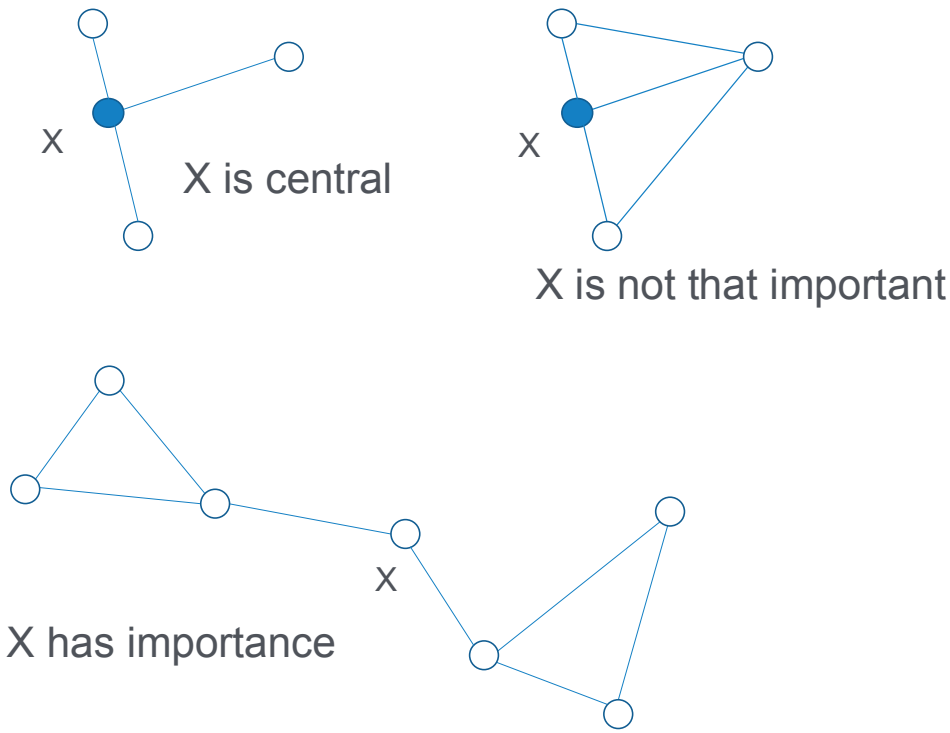
Centrality measures

Usually proxies for Importance of actors (also power, prestige, influence . . . all hard to define precisely)
There is no single measure that captures what we are looking for (to begin with, what exactly do we mean *importance*?).

- Modus operandi in analytics: calculate multiple measures, dig deeper . . . create a plausible story on why.
- We go from simple measures to more global (and mathematically deeper) measures that often give superior insight and known to work well (like PageRank).

Importance measures

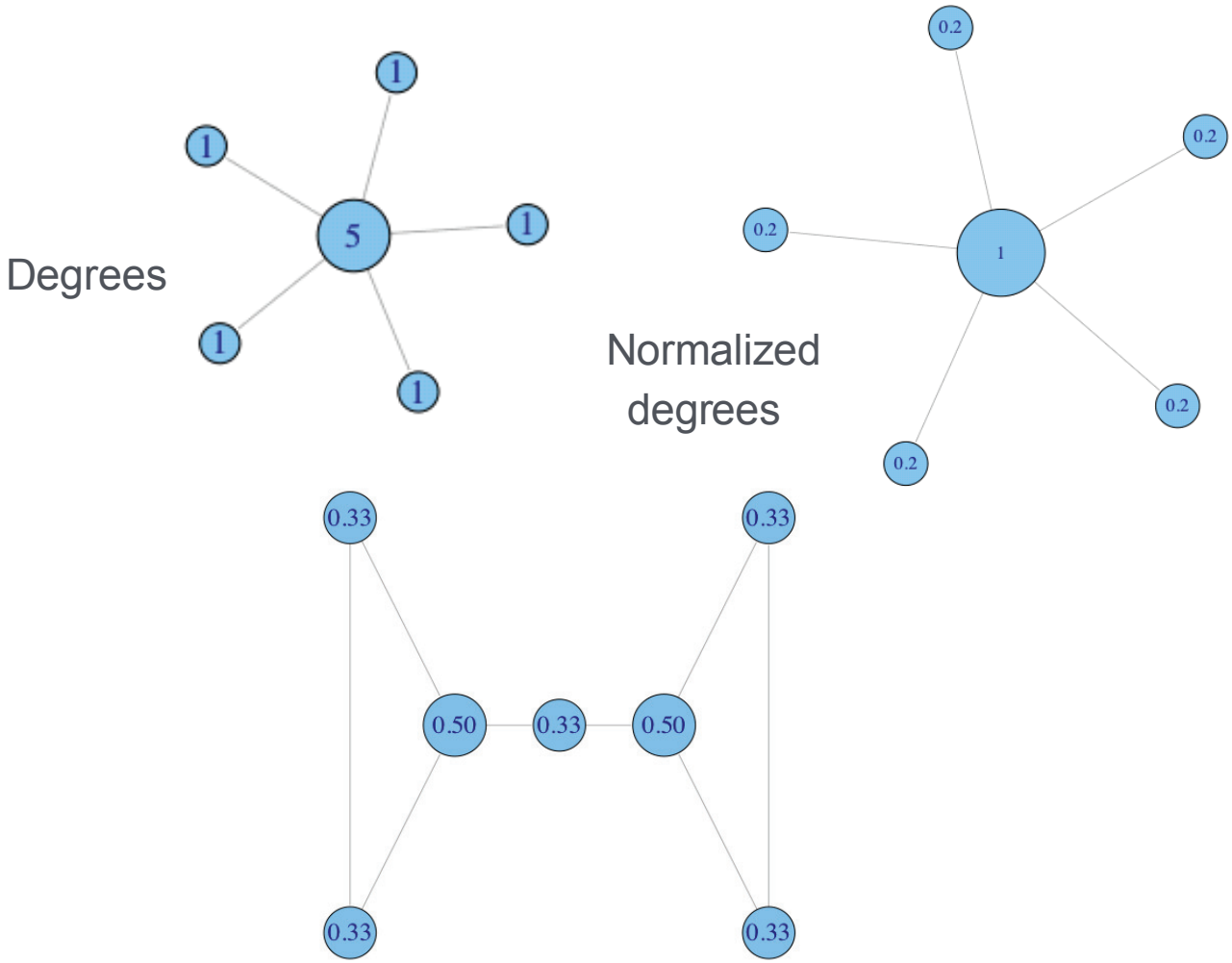
Nodes as actors, edges represent a relationship between the actors. The structure of the graph leads to notions of importance.



Intuitive notions of importance of a node, but can be for different reasons. We try to capture this with a metric—typically a single number for each node.

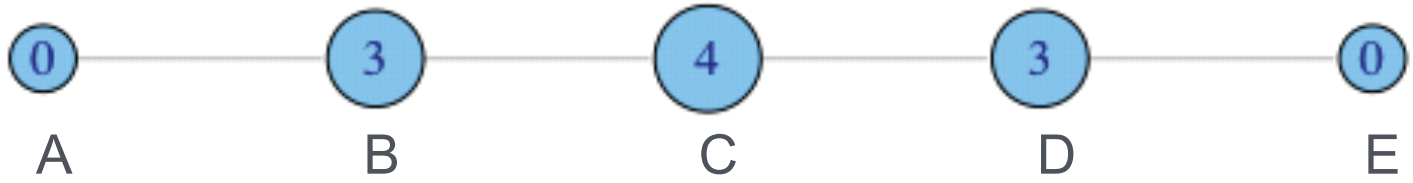
Degree centrality

Degree is the number of edges incident on a node. Higher degree means more important.
Normalized centrality of a node =degree/$N-1$ where N is the number of nodes.

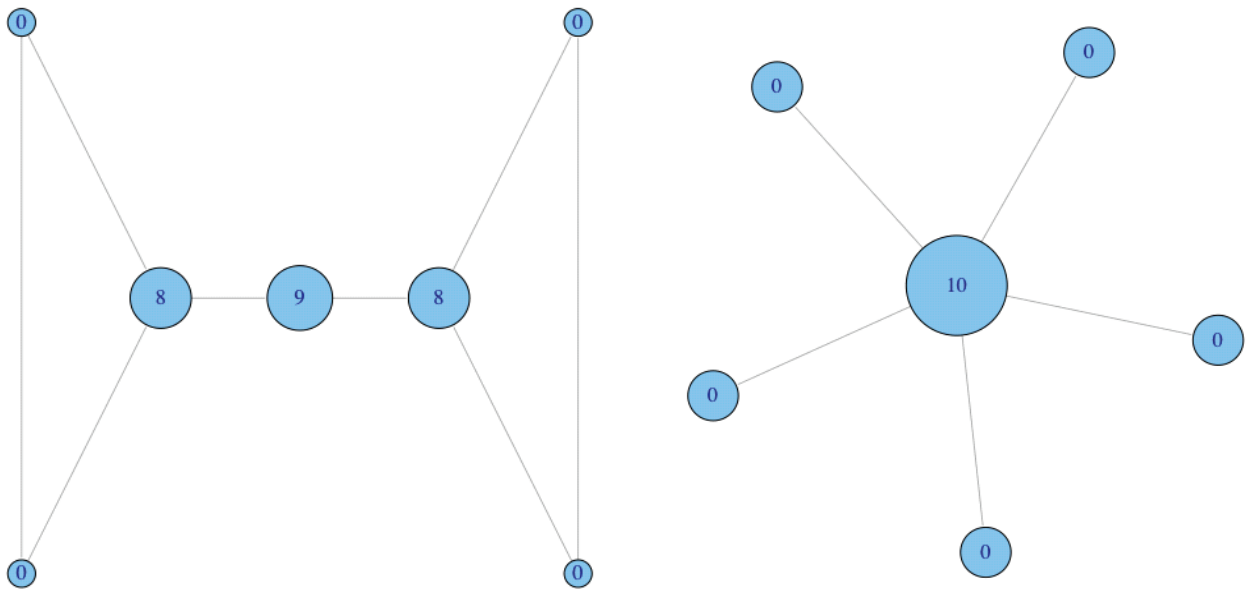


Betweenness centrality

Betweenness of a node x: How many pairs of nodes have to go through x in a shortest path between the pair



B lies between A and C, D, E; C lies on shortest paths between $(A, D), (A, E), (B, D), (B, E)$



Betweenness centrality formulas

Betweenness of node i:

$$C_B(i) = \sum_{j<k} \frac{g_{jk}(i)}{g_{jk}}$$

where g_{jk} is the number of shortest-paths connecting j and k and $g_{jk}(i)$ is the number of these shortest paths passing through node i.
Normalized betweenness of node i

$$C'_B(i) = \frac{C_B(i)}{\frac{(n-1)(n-2)}{2}}$$

i.e., divide by the number of pairs of vertices, excluding the vertex i itself

Flow centrality

Do people communicate only by shortest paths?

- For a pair of nodes j, k what proportion of *total number of paths* pass through i?
- Defines the importance of i in the communication between j and k

Many variations on these ideas: Give longer paths less weight etc..
How to calculate these? Max-flow problem (later)

List of common centrality measures[24]

Local (many variations on the basic theme with different names; edge versions of each)

Centrality	Formula	Description	Comments
Degree	$d(i)$	Degree	too simplistic
Betweenness	$C_B(i) = \sum_{j<k} \frac{g_{jk}(i)}{g_{jk}}$	where $g_{jk}(i)$ is the number of shortest-paths connecting j and k that go through i; g_{jk} is the number of shortest paths between j and k	
Closeness	$C(i) = \frac{1}{\sum_j d(j,i)}$	Reciprocal of sum of shortest distances from all other nodes	

Global: Eigenvector, Katz, PageRank (all use eigenvalues and eigenvectors)

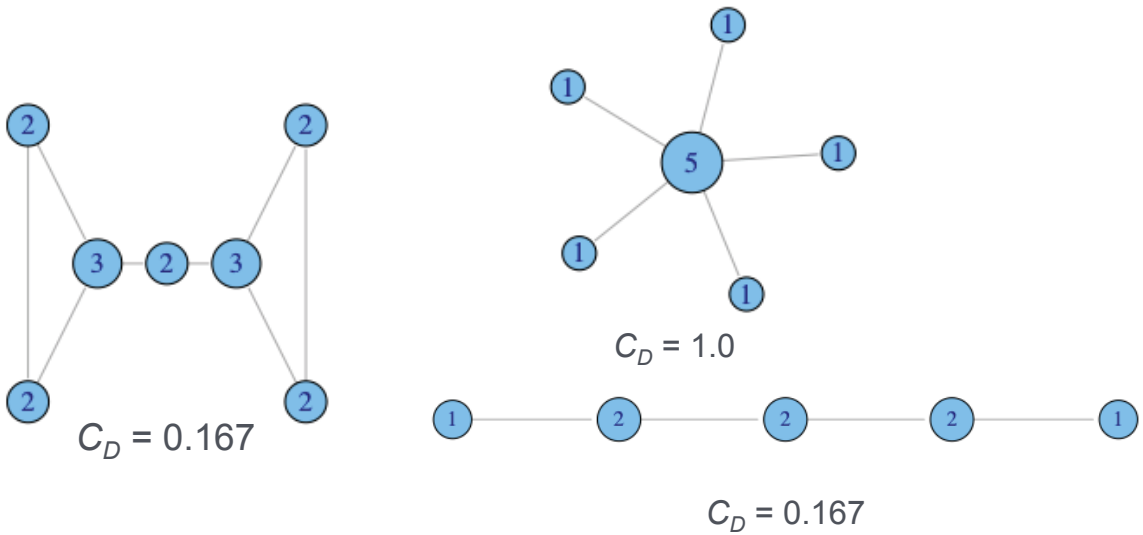
[24] Most sociological measures used are described here (for reference): http://faculty.ucr.edu/ hanneman/nettext/C10_Centrality.html

Characterizing a Network by its *dispersion* of centrality

How equitable or balanced is the network? Measure of the variance in the centralities of the nodes

$$C_D = \frac{\sum_{i=1}^{N}[C_D(n^*) - C_D(i)]}{(N-1)(N-2)}$$

where $C_D(\cdot)$ is some measure of centrality, n^* is the node that achieves the maximum value according to this measure.



Spectral measures of centrality

A basic insight (attributed to Bonacich, called eigenvector centrality): Importance is not just a function of how many friends, but also the importance of the friends

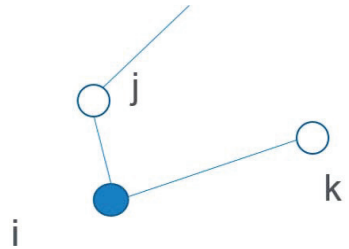
- You have 500 friends
- Warren Buffett has 500 friends
- What's the difference? . . . Amongst his 500 are Bill Gates, Obama, Clinton, Zuckerberg . . .

So not just number but also who they are counts. How to capture this based only on network topology?[25]

[25] You can of course gain more insight on their influence by analyzing what they say and how what they say spreads in the network. But often we don't have this information.

Eigenvector centrality

Importance of a node is proportional (we use the symbol $\sim$; interpreted as some constant times this quantity) to the Importance of its neighbours



- Say nodes j and k are neighbors of i
- Let x_i be a number measuring importance of i
- $x_i \propto x_j + x_k$; i.e. i's importance is proportional to that of its neighbors' importance

But this is recursive, as we define the importance of j the same way . . . and we know how to represent this with a matrix, the adjacency matrix!

(we hope) some sort of equilibrium is reached with the equation $A\boldsymbol{x} = \boldsymbol{x}$. i.e., start with $\boldsymbol{x}^0$

$$\boldsymbol{x}^1 = A\boldsymbol{x}^0, \quad \boldsymbol{x}^2 = A\boldsymbol{x}^1, \ldots$$

do we get $A\boldsymbol{x} = \boldsymbol{x}$ eventually for some $\boldsymbol{x}$?
In Linear Algebra terminology: $\boldsymbol{x}$ is the eigenvector of the matrix A corresponding to an eigenvalue of 1. So it should be clear that Eigenvalues and Eigenvectors from Linear Algebra are the critical concepts here. They answer questions like: Why should such an $\boldsymbol{x}$ exist? Does it always exist?

Eigenvalues and eigenvectors refresher

("eigen" comes from German $\rightarrow$ "own" vector)

Eigenvalues and eigenvectors

A *nonzero* vector $\boldsymbol{x}$ is an eigenvector of a *square* matrix A if there exists a scalar λ such that $A\boldsymbol{x} = \lambda\boldsymbol{x}$. Then λ is an eigenvalue of A. $\boldsymbol{x}$ is called an eigenvector corresponding to λ

Computation is in general *very* tedious. In high school you would have done this: $A\boldsymbol{x} = \lambda\boldsymbol{x}$ or $(A - \lambda I)\boldsymbol{x} = 0$, $\boldsymbol{x} \neq 0$. Therefore, $(A - \lambda I)$ does not have full rank, so we should be having $det(A - \lambda I) = 0$ (a polynomial equation to be solved for λ).
Example:

$$A = \begin{pmatrix} 1 & 6 \\ 5 & 2 \end{pmatrix}$$

Solve then

$$det \begin{pmatrix} 1 - \lambda & 6 \\ 5 & 2 - \lambda \end{pmatrix} = 0$$

$$(1 - \lambda)(2 - \lambda) - 30 = 0$$
$$\lambda^2 - 3\lambda - 28 = 0$$
$$(\lambda - 7)(\lambda + 4) = 0$$
$$\lambda = 7, \lambda = -4$$

Fortunately modern software does this for you (R, Matlab, Mathematica, Python packages) using better methods.

A few important truths of eigenvalues and eigenvectors (just accept them)

The following facts (that you should just accept as basic Math facts):

1. In general eigenvalues may be complex numbers. Eg:

$$A = \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix}$$

   The characteristic polynomial to find eigenvalues is $\lambda^2 - 2\lambda + 2 = 0$ which has only complex roots

2. However, if A is a symmetric square matrix, all the eigenvalues are real (Note: the adjacency matrix of an undirected graph is symmetric)
3. (Perron-Froebenius) If A is a positive matrix (all entries ≥ 0, not necessarily symmetric), the largest eigenvalue $\lambda_{\max}$ is positive, and the corresponding eigenvector has all non-negative real elements.
4. If A is a *stochastic* matrix (non-negative elements and sum of all elements of each column (or each row) sum to 1), the largest eigenvalue $\lambda_{\max} = 1$
5. (the following is used in dimensionality-reduction and community detection algorithms) The Rayleigh Quotient interpretation: If $\lambda_{\max} \geq \ldots \geq \lambda_n$ are the eigenvalues of a symmetric matrix B, the λ_1 is the value of

$$\lambda_{\max} = \max_{\boldsymbol{x} \neq 0} \frac{\boldsymbol{x}^T A \boldsymbol{x}}{\boldsymbol{x}^T \boldsymbol{x}}$$

   and the solution is the corresponding eigenvector. Similarly for the second highest value over an orthogonal space, etc. and eventually $\lambda_n = \min_{\boldsymbol{x} \neq 0} \frac{\boldsymbol{x}^T A \boldsymbol{x}}{\boldsymbol{x}^T \boldsymbol{x}}$