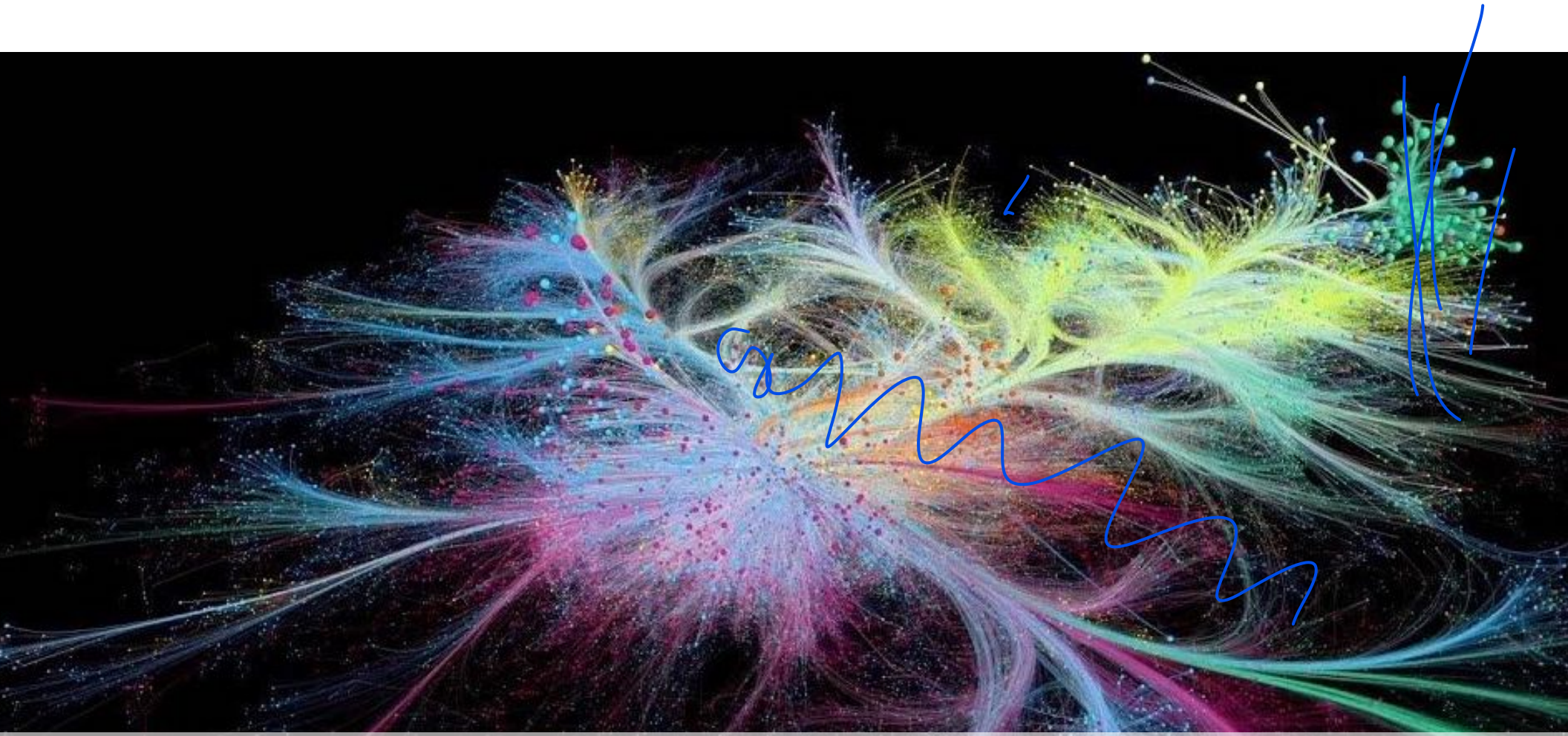


CPSC 572/672: Fundamentals of Social Network Analysis and Data Mining

Admin

Assessment: Project proposal Sept. 27th

Presentation

2 minutes (time limit will be enforced)
4 slides

Discuss:

- What are your nodes and links?
- Describe your dataset. How will you get it?
- Expected size of the network (number of nodes, number of links)
- What questions do you plan to ask? These may change during the course of the class.
- Why do we care about the network you plan to study?

Written component

1 page
Written summary of the details in your presentation.

Project Matchmaking: Week 1

CPSC 572/672 Project Matchmaking

File Edit View Insert Format Data Tools Add-ons Help Last edit was 2 minutes ago

Share

D8

	A	B	C	D	E	F	G	H	I
1	**Name**	**572/672**	**Skills I have**	**Skills I want to develop**	**Datasets or questions I am interested in**	**Partner**			
2									
3									
4									
5									
6									
7									
8									
9									
10									
11									
12									
13									
14									
15									

https://docs.google.com/spreadsheets/d/1vAOhbuaQooo_O5cRZru4QlqgwzTncoXp9URFmTqvfkM/edit?usp=sharing

Project Proposal: [My Social Network Project Title]

Author 1, Author 2, [Author 3 if three]

1. Nodes and Links

Describe in plain language what the node and links are in your social network. This only needs to be a sentence or two.

2. Dataset

Tell us about your dataset. What is the subject matter? Where are you getting it? Most important: How are you getting it and how long will that take you? Describe any expected challenges, need for permissions, which tools you will use, what format the data is in, any expected data cleaning. Note: You may not use data that is already in network format.

3. Expected size of the network

State the number of nodes and number of links. Tell us why – is this everything available? Are you taking a subset? If a subset, why, and how do you plan to define it?

4. Questions you plan to ask and why we care

This is the part to really practice your skills at something akin to a scientific abstract (minus the results). Motivate us – what is so important and/or interesting about your social network? What research questions do you plan to pursue? Do you have any preliminary ideas about how you want to pursue these questions (e.g. community detection)?

Page limit: 1 page
You can use this however you see fit, there is no limit per section. You will all present a variety of networks that require more or less info in each section. You also do not to fill the whole page – succinct is good, just as long as you cover each component as described above.

BUDDY GROUP CONTRACT **DATE: _____**

EXPECTATIONS FROM TEAM MEMBERS

(e.g., Attend all meetings – Bring cinnamon rolls after missing a meeting, Complete project task as promised – Kicked out of team if not completed 3 times, Be open to contributions and ideas from all team members, etc.)

Expectation	**Consequence if expectation not met**
We'll meet as a team every week at: **_____ for _____ hrs and at:** **_____ for _____ hrs**	(One to Two team meetings a week is recommended. Your group may choose more or less. Consequences for missing a few team meetings should be less severe than missing many team meetings.)
Be on time and prepared for team meetings. (Both in and out of class time)	
Follow through on commitments made to the team. (These are likely to be minimal as your projects are individual, but may include committing to giving thoughtful feedback on members' work.)	
Contribute to the team voluntarily. (These contributions may be ideas, questions, code, organizing meetings, managing code repository, creating charts for the report, etc.)	
Welcome and invite contributions by other team members.	*(How will you deal with a team member that consistently dismisses/discourages ideas from other team members or a single team member? It may be helpful to assign a single team member that manages discussions.)*
How you define the line between support and plagiarism. (helpful to play out a scenario, e.g. a buddy group member is stuck with a block of code)	
Each team member will talk for **at least _____ minutes and** **at most _____ minutes** **during a 1-hour meeting.**	*(Some team members have a tendency to dominate discussions and others hesitate to contribute. Set some guidelines on what is expected and how you, respectfully, will make team members aware that they need to contribute more/less.)*
(Think of any other team behaviours that hindered or helped the team performance and formalize this here.)	

If you've read and agreed with this contract, add your name here:

_____ _____

_____ _____

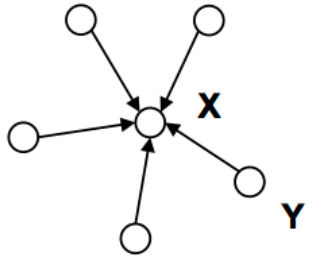
Questions?

Centrality measures

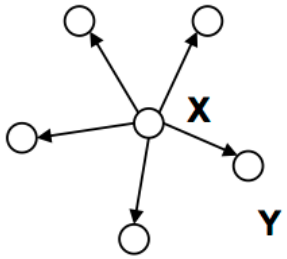
Centrality measures

"There is certainly no unanimity on exactly what centrality is or on its conceptual foundations, and there is little agreement on the proper procedure for its measurement."
- Freeman 1979

Centrality: who's important based on their network

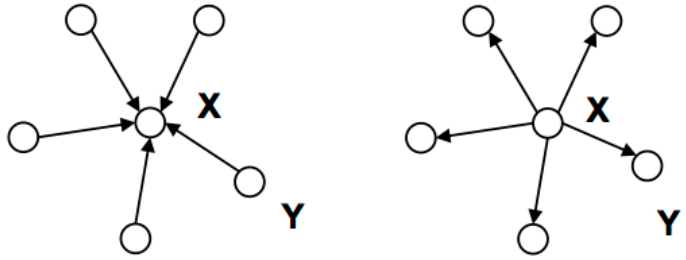


indegree



outdegree

Centrality: who's important based on their network



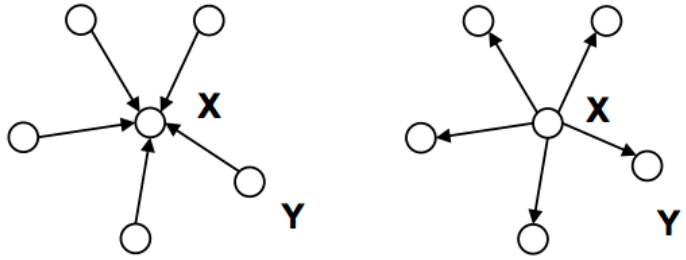
indegree outdegree

Best measure if importance means:

→ how popular you are

→ how many people you know

Centrality: who's important based on their network



indegree outdegree

Best measure if importance means:

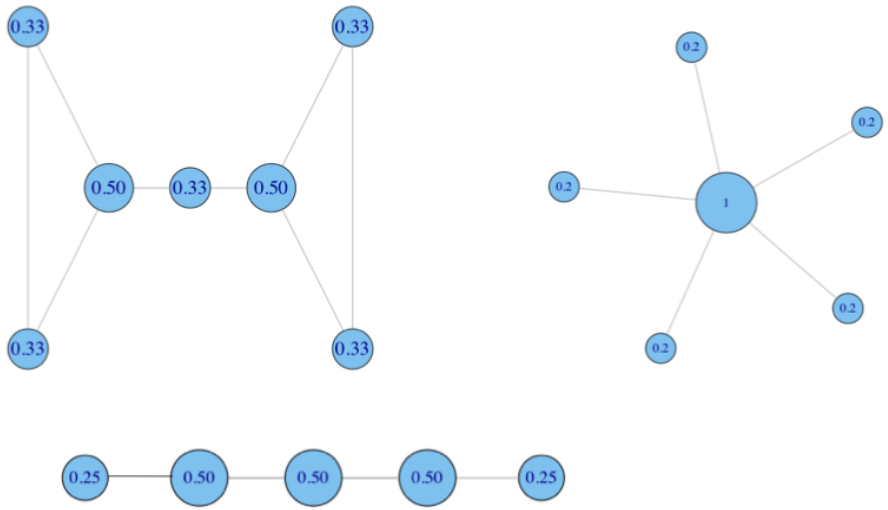
→ how popular you are

→ how many people you know

It is a local measure!

Degree centrality

$$C^D(i) = \frac{k_i}{N-1}$$

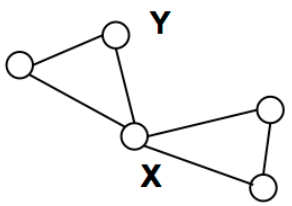


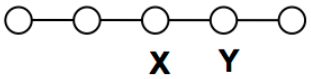
M.E.J. Newman. (2010). *Networks: An Introduction.* Oxford University Press.

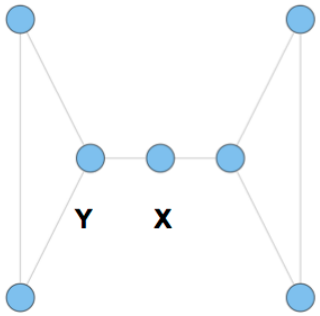
Degree is not everything

Who is more important in the following situations:

X or Y?







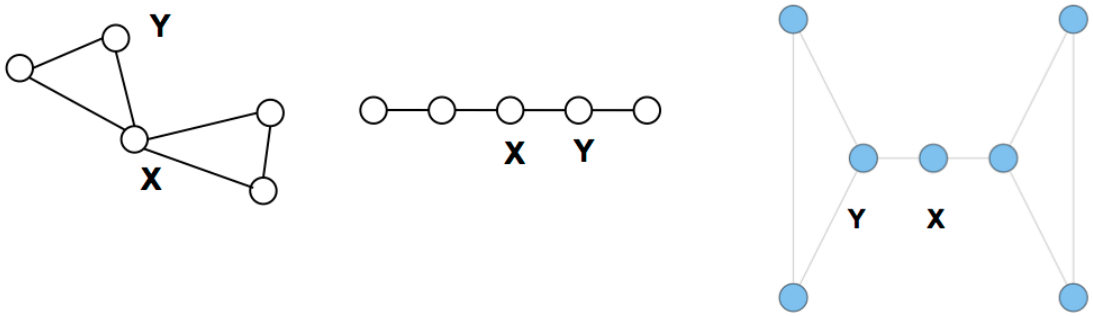
Y is going to be labeled as highest degree

X is a bridge node

Find diff between the two

Degree is not everything

Who is more important in the following situations:
X or Y?



We want to capture:

→ Ability to broker between groups

→ Likelihood that information originating anywhere in the network reaches you

Betweenness centrality

$$\tilde{C}^B(i) = \sum_{j<k} \frac{d_{jk}(i)}{d_{jk}}$$

d_{jk} # of shortest paths between j and k

$d_{jk}(i)$ # of shortest paths between j and k that go through i

cicle through all then

for all shortest path how many passes through node i

so if 3 shortest path and i gets passed 1 of those 3 : . : -> 1/3 Centrality

M.E.J. Newman. (2010). *Networks: An Introduction.* Oxford University Press.

Betweenness centrality

$$\tilde{C}^B(i) = \sum_{j<k} \frac{d_{jk}(i)}{d_{jk}}$$

d_{jk} # of shortest paths between j and k

$d_{jk}(i)$ # of shortest paths between j and k that go through i

Normalized

$$C^B(i) = \frac{\tilde{C}^B}{(N-1)(N-2)/2}$$

Number of pairs of vertices excluding i

M.E.J. Newman. (2010). *Networks: An Introduction.* Oxford University Press.

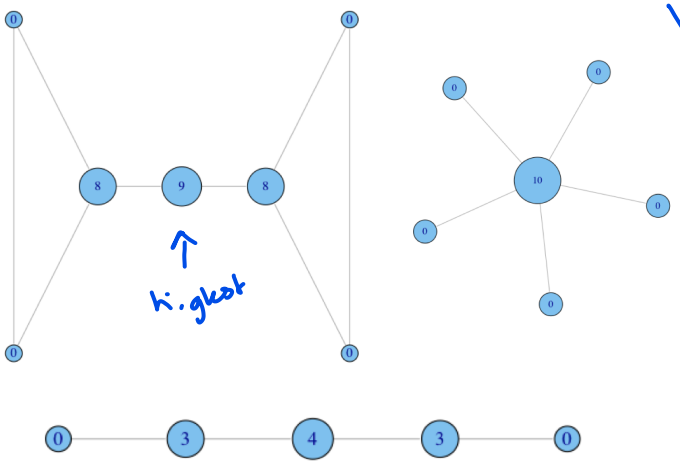
Betweenness centrality

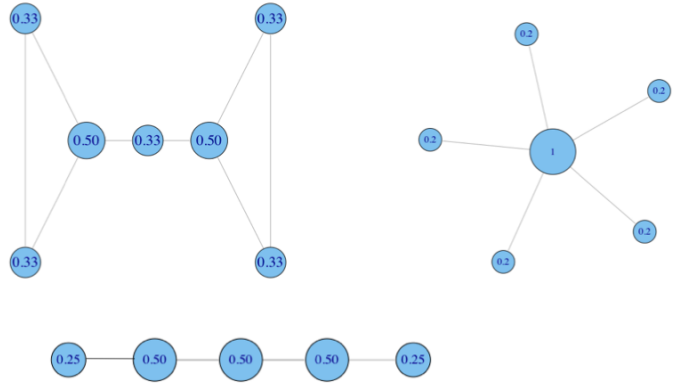
$$\tilde{C}^B(i) = \sum_{j<k} \frac{d_{jk}(i)}{d_{jk}}$$

Betweenness centrality
(not normalized)

Bridge isn't being counted correctly

Degree centrality



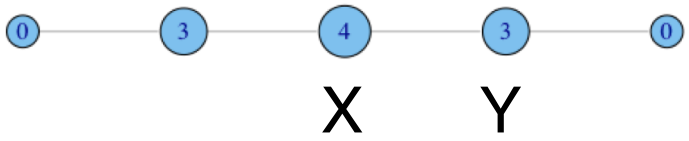


betweeness we wanted it to point out the middle node is important.

→ degree doesn't.

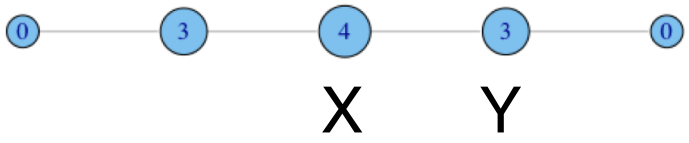
Betweenness is not everything

X and Y do not differ much for betweenness



Betweenness is not everything

X and Y do not differ much for betweenness

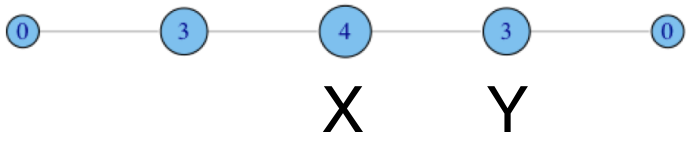


We want to capture:

⟶ Being close to all nodes

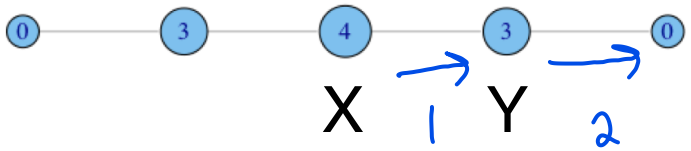
Closeness centrality

$$\tilde{C}^{C}(i) = \left[\sum_{j=1}^{N} d(i,j)\right]^{-1}$$



Closeness centrality

$$\tilde{C}^C(i) = \left[\sum_{j=1}^{N} d(i,j)\right]^{-1}$$



Normalized

$$C^C(i) = (N-1)\tilde{C}^C(i)$$

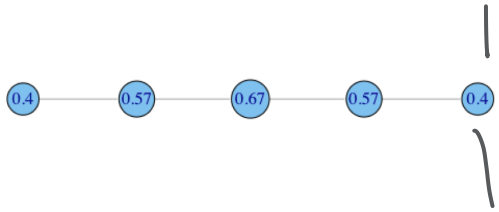
All other nodes in the network

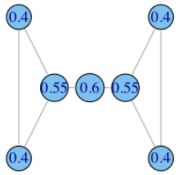
topological distance

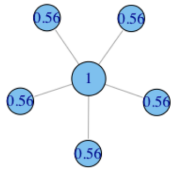
weights are equal

M.E.J. Newman. (2010). *Networks: An Introduction.* Oxford University Press.

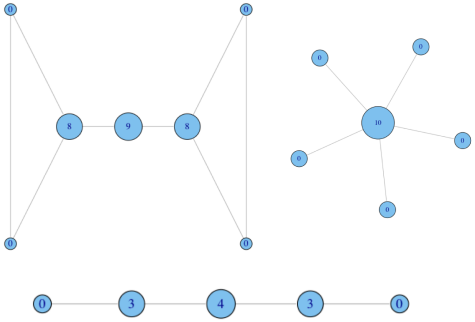
Closeness centrality



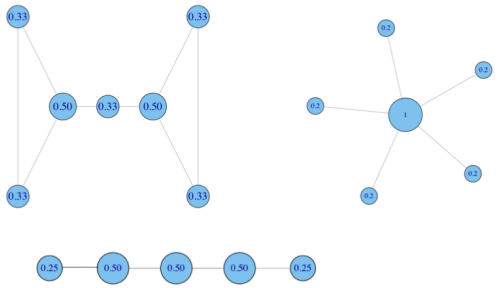


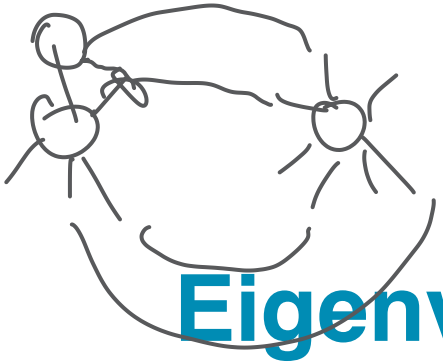


Betweenness centrality
(not normalized)



Degree centrality





Eigenvector centrality

A node is important if it is connected
to important nodes

Connected to a higher central node is worth more then connecting to a lower central node

Eigenvector centrality

A node is important if it is connected to important nodes

$$\forall i \in [1, N] \qquad X_i \propto \sum_{j \in \Lambda(i)} X_j$$

direct neighbour (node) of i

M.E.J. Newman. (2010). *Networks: An Introduction.* Oxford University Press.

Eigenvector centrality

A node is important if it is connected to important nodes

$$\forall i \in [1, N] \qquad X_i \sim \sum_{j \in \Lambda(i)} X_j \qquad X_i \sim \sum_{j=1}^{N} A_{ij} X_j$$

M.E.J. Newman. (2010). *Networks: An Introduction.* Oxford University Press.

Eigenvector centrality

A node is important if it is connected to important nodes

$$\forall i \in [1, N] \qquad X_i \sim \sum_{j \in \Lambda(i)} X_j \qquad X_i \sim \sum_{j=1}^{N} A_{ij} X_j \qquad \boxed{AX = \lambda X}$$

M.E.J. Newman. (2010). *Networks: An Introduction.* Oxford University Press.

Eigenvector centrality

A node is important if it is connected to important nodes

$$x_i = \frac{1}{\lambda} \sum_{j \in \Lambda(i)} x_j \qquad x_i = \frac{1}{\lambda} \sum_{j \in G} a_{ij} x_j \qquad \boxed{AX = \lambda X}$$

The solution (when exists) gives the node centrality. We take the highest λ

This concept is at the core of the ranking algorithm of Google

M.E.J. Newman. (2010). *Networks: An Introduction.* Oxford University Press.

Whatever page is connected to the most other pages

You can have a node that is central in all ways

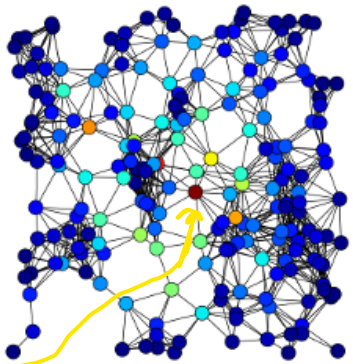
usually if central in one central in another

You can have nodes that contradict each other.

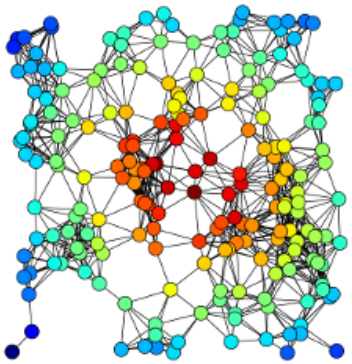
Summarizing

Bridge? → more central

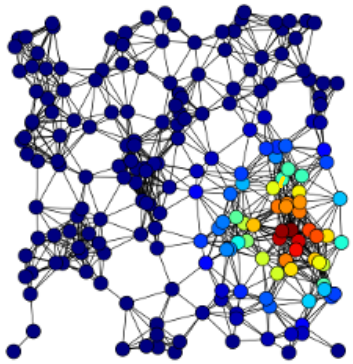
Centrality indices are answers to the question "**What characterizes an important node?**"
The word "importance" has a wide number of meanings, leading to many different definitions of centrality.



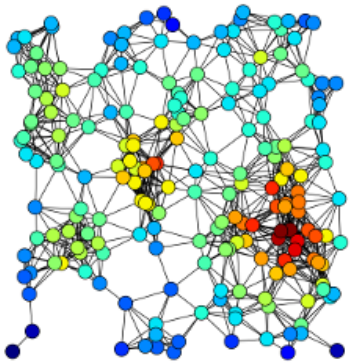
A Betweenness



B Closeness



C Eigenvector



D Degree

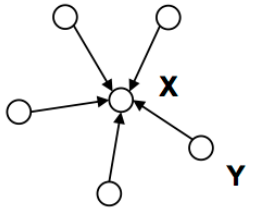
Source: http://en.wikipedia.org/wiki/Centrality

Which nodes are most "central"?

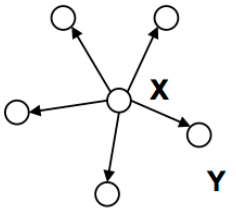
Definition of 'central' varies by context/purpose.

Local measure:

- degree



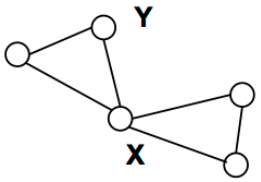
indegree



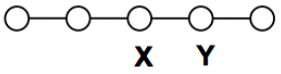
outdegree

Relative to the rest of network:

- betweenness
- closeness
- eigenvector



betweenness



closeness

high weight

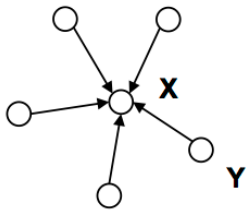
does not necessarily mean high distance.

A note on weights

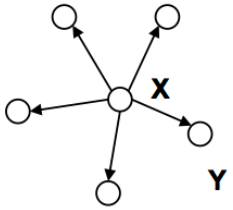
In your network, does a high weight increase centrality? Decrease it?

Betweenness and closeness rely on calculation of path length. Make sure your path weights make sense!

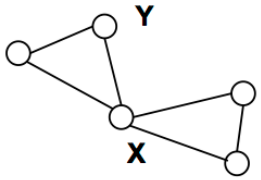
E.g. a strong friendship should mean individuals are *easier* to reach. Try inverting the weights.



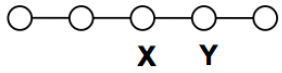
indegree



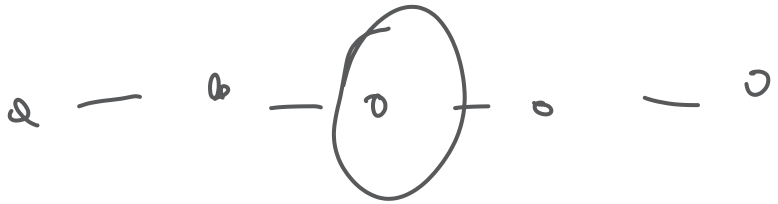
outdegree



betweenness



closeness



Many more definitions of centrality exist (Katz, random walk, eccentricity...). We have covered the most widely used metrics.

Can also extend to edges: edge betweenness etc.

Dynamics: some attempts

Exercise

Quite often in real datasets, you will find that the various network centrality measures are highly correlated with one another. To understand the differences between these measures, for each of the following cases, come up with an example network and a particular node in that network that has:

a) High closeness centrality but low degree centrality — management

b) High degree centrality but low closeness centrality → research paper

c) High betweenness centrality but low closeness centrality

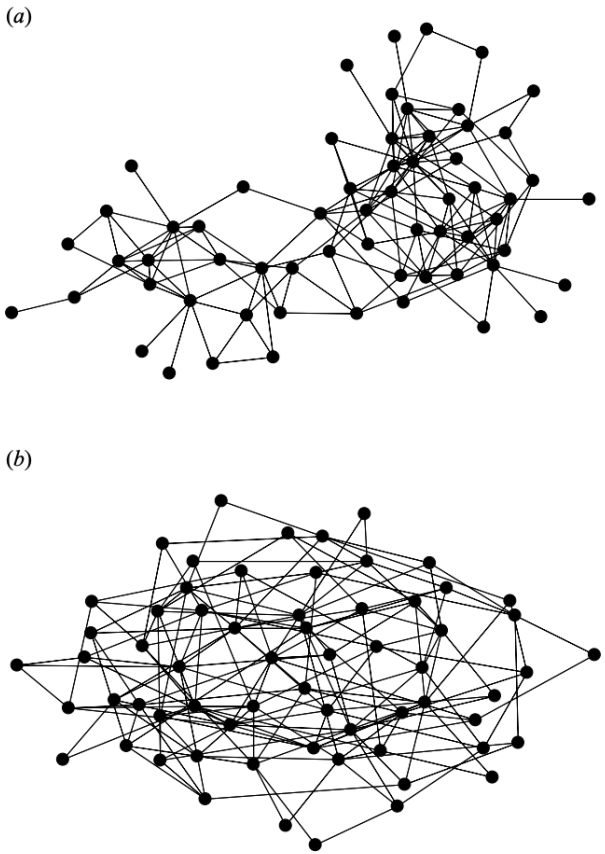
d) High closeness centrality but low betweenness centrality

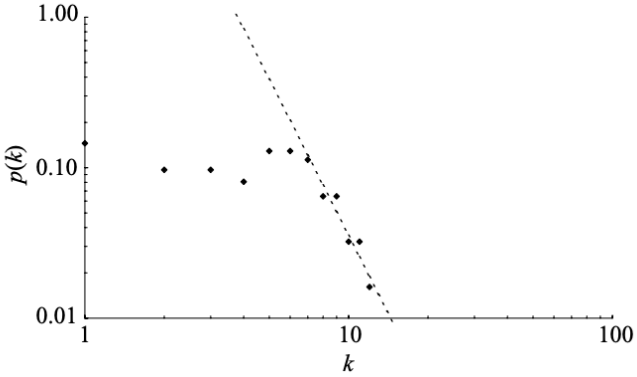
e) High degree centrality but low betweenness centrality

f) High betweenness centrality but low degree centrality

Examples

Dolphin social network



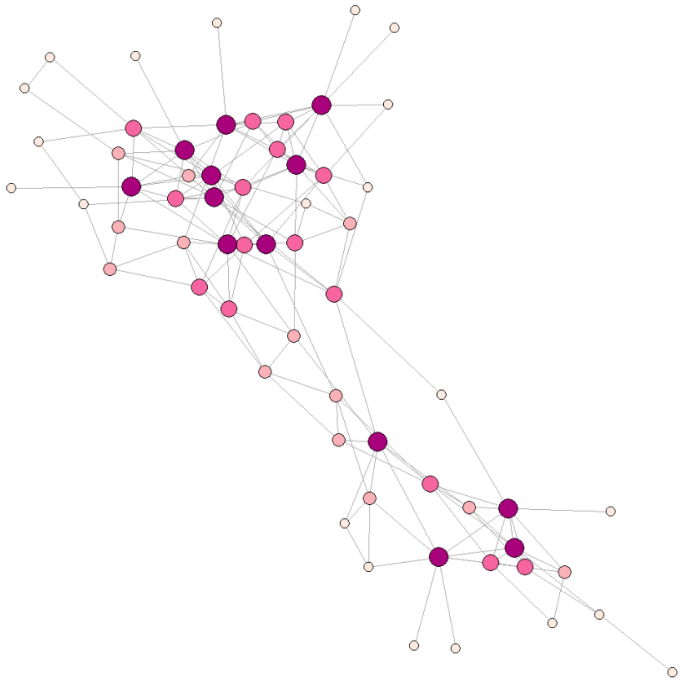


Lusseau 2003 RSoc Bio Letters
Source: https://rstudio-pubs-static.s3.amazonaws.com/341956_3febd99edad34d9a93e3bfe24c3fb54e.html

Dolphin social network

By degree

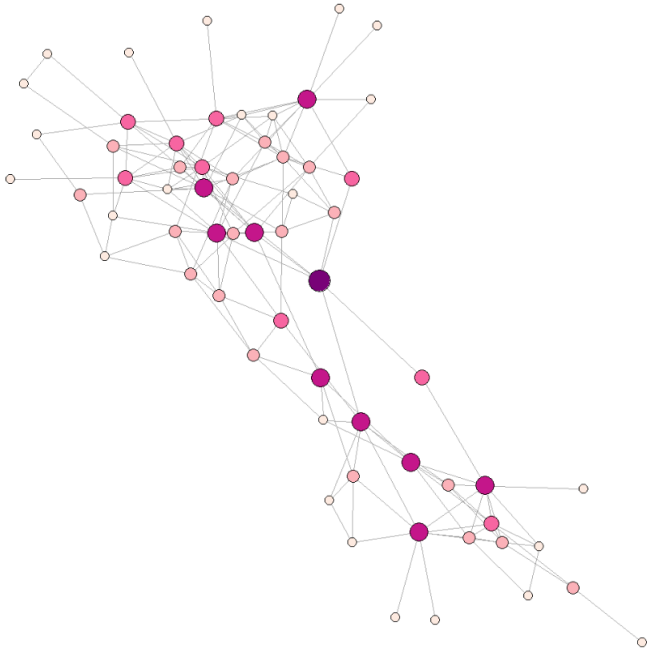
Node <chr>	Name <chr>	dCent <dbl>	bCent <dbl>	cCent <dbl>
14	Grin	12	113.408769	0.006172840
37	SN4	11	253.582713	0.006535948
45	Topless	11	74.426906	0.005681818
33	Scabs	10	104.614585	0.005988024
51	Trigger	10	154.959376	0.005405405
17	Jet	9	209.169298	0.005076142
20	Kringel	9	187.841704	0.006410256
29	Patchback	9	119.918587	0.005291005
57	Web	9	154.094571	0.004950495
1	Beescratch	8	390.383717	0.006097561



Dolphin social network

By betweenness

Node <chr>	Name <chr>	dCent <dbl>	bCent <dbl>	cCent <dbl>
36	SN100	7	454.274069	0.006849315
1	Beescratch	8	390.383717	0.006097561
40	SN9	8	261.963619	0.006622517
37	SN4	11	253.582713	0.006535948
7	DN63	5	216.376673	0.005988024
17	Jet	9	209.169298	0.005076142
20	Kringel	9	187.841704	0.006410256
54	Upbang	7	181.392614	0.005319149
51	Trigger	10	154.959376	0.005405405
57	Web	9	154.094571	0.004950495



Dolphin social network

By degree

Node <chr>	Name <chr>	dCent <dbl>	bCent <dbl>	cCent <dbl>
14	Grin	12	113.408769	0.006172840
37	SN4	11	253.582713	0.006535948
45	Topless	11	74.426906	0.005681818
33	Scabs	10	104.614585	0.005988024
51	Trigger	10	154.959376	0.005405405
17	Jet	9	209.169298	0.005076142
20	Kringel	9	187.841704	0.006410256
29	Patchback	9	119.918587	0.005291005
57	Web	9	154.094571	0.004950495
1	Beescratch	8	390.383717	0.006097561

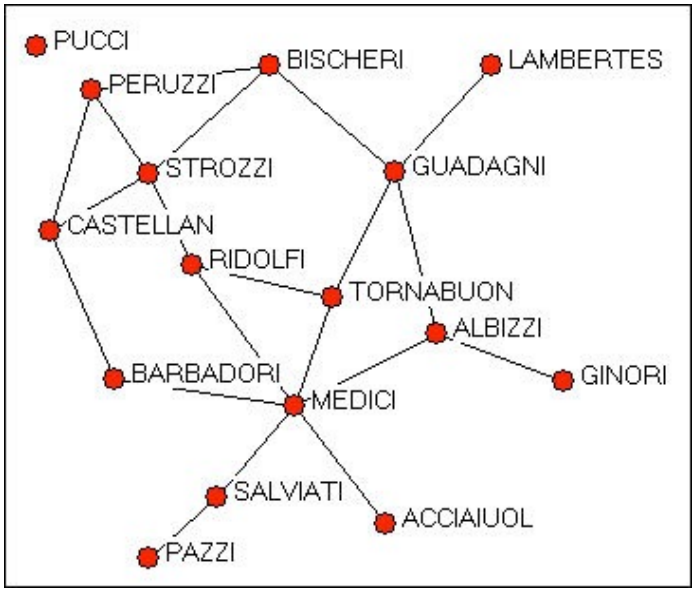
By betweenness

Node <chr>	Name <chr>	dCent <dbl>	bCent <dbl>	cCent <dbl>
36	SN100	7	454.274069	0.006849315
1	Beescratch	8	390.383717	0.006097561
40	SN9	8	261.963619	0.006622517
37	SN4	11	253.582713	0.006535948
7	DN63	5	216.376673	0.005988024
17	Jet	9	209.169298	0.005076142
20	Kringel	9	187.841704	0.006410256
54	Upbang	7	181.392614	0.005319149
51	Trigger	10	154.959376	0.005405405
57	Web	9	154.094571	0.004950495

By closeness

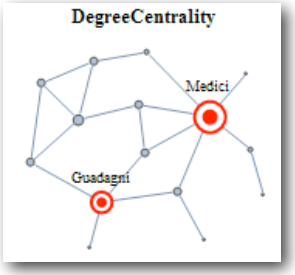
Node <chr>	Name <chr>	dCent <dbl>	bCent <dbl>	cCent <dbl>
36	SN100	7	454.274069	0.006849315
40	SN9	8	261.963619	0.006622517
37	SN4	11	253.582713	0.006535948
20	Kringel	9	187.841704	0.006410256
14	Grin	12	113.408769	0.006172840
1	Beescratch	8	390.383717	0.006097561
7	DN63	5	216.376673	0.005988024
28	Oscar	5	122.165227	0.005988024
33	Scabs	10	104.614585	0.005988024
8	Double	6	40.929300	0.005952381

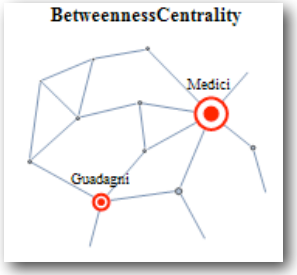
Prestige of florentine families during Renaissance

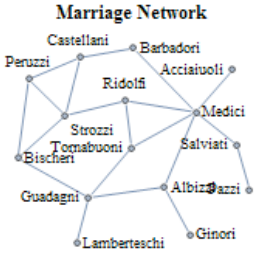


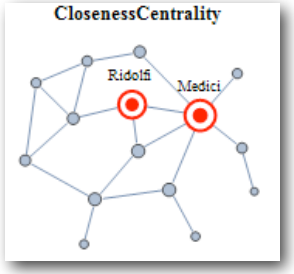
Robust action and the rise of the Medici, 1400-1434. Padgett and Ansell 1993

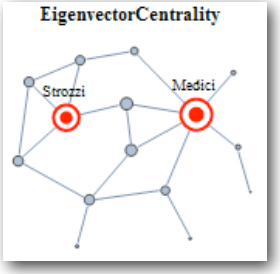
Prestige of florentine families during Renaissance











Robust action and the rise of the Medici, 1400-1434. Padgett and Ansell 1993

Next week installations

Contact

Mailing list
Issue tracker
Source

Releases

Stable (notes)

2.8.6 — August 2022
download | doc | pdf

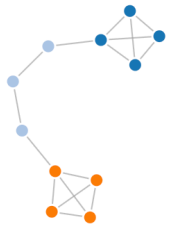
Latest (notes)

3.0 development
github | doc | pdf

Archive

NetworkX
Network Analysis in Python

NetworkX is a Python package for the creation, manipulation, and study of the structure, dynamics, and functions of complex networks.



Software for complex networks

- Data structures for graphs, digraphs, and multigraphs
- Many standard graph algorithms
- Network structure and analysis measures
- Generators for classic graphs, random graphs, and synthetic networks
- Nodes can be "anything" (e.g., text, images, XML records)
- Edges can hold arbitrary data (e.g., weights, time-series)
- Open source 3-clause BSD license
- Well tested with over 90% code coverage
- Additional benefits from Python include fast prototyping, easy to teach, and multi-platform

Recommend Anaconda

We will use Jupyter notebook

Python 3

Gephi

Gephi
makes graphs handy

Download | Blog | Wiki | Forum | Support | Bug tracker

Home | Features | Learn | Develop | Plugins | Services | Consortium

The Open Graph Viz Platform

Gephi is the leading visualization and exploration software for all kinds of graphs and networks. Gephi is open-source and free.

Runs on Windows, Mac OS X and Linux.

Learn More on Gephi Platform »

Download FREE
Gephi 0.9.1

Release Notes | System Requirements

- Features
- Quick start
- Screenshots
- Videos