Social Network Analysis

Prof. Vrijendra Singh
Dept of IT, IIIT Allahabad
vrij@iiita.ac.in
7376581984

Fair Use Disclaimer

The presentation contains a very few information/material/images from the web resources, journal papers, books etc. under fair use policy. The materials have been taken for explaining the relevant issues related to the Social Network Analysis to the scientific community. If you wish to use any material from this presentation for the purpose other than the “fair use”, you must obtain the explicit permission from the copyright owner.

TABLE 2.1 Examples of Social Media and Pre-digital Media Systems Organized by the Size of Producer and Consumer Populations

Size of Consumer Population	Size of Producer Population		
	Small	Medium	Large
Small	Instant messaging Personal messaging (e.g., within Facebook) Video chat Phone call Face-to-face office meeting	Committee report to a decision maker Online survey Social networking friend feed Twitter homepage showing tweets of people you follow	Professional services reports for decision makers Personalized suggestions based on recommender systems
Medium	"Social" or family blog Profile page on community site or social network Departmental email list Tweet sent to followers Wall post on Facebook	Group blog on niche topic Internet relay chat room Internal department wiki Facebook group Niche YouTube channels Local markets (e.g., Craigslist)	Professional report for specialty group NASA clickworkers[1] Idea-generation sites (e.g., IdeaConnection[2])
Large	Popular blog, podcast, or webcast Message to large forum or email list Popular Twitter user's tweet Popular YouTube video Company web site Novel	News rating site (e.g., Digg) Wikipedia page Television program Popular discussion forum Online user-generated databases (e.g., IMDB) or marketplace (e.g., Threadless)	Large online marketplaces (e.g., eBay) Wikipedia encyclopedia YouTube video sharing Flickr photo sharing Popular massively multiplayer game

[1] *www.scienceofcollaboratories.org/resources/collab.php?317*
[2] *www.ideaconnection.com/crowdsourcing/procter-gamble-00007.html*

TABLE 2.2 Examples of Social Media Categorized by the Pace of Interaction and the Granularity of Control over Content

Pace of Interaction	Granularity of Control		
	Fine: Users can directly control smallest units of content (characters, pixels, bytes)	Medium: Users control medium-sized blocks of content (objects, attributes, tracks, players) that they can only indirectly alter or that can be altered by other users	Coarse: Users control large block of content (documents, messages, blog posts, photos), rarely edited or modified by others
Synchronous	Real-time shared canvas	Virtual worlds, multiplayer games, real-time networked musical jamming	Chat, instant messaging, texting, Twitter
Asynchronous	Shared documents (e.g., Google Docs), source code, Wikipedia	Contribution to collected works like an album, anthology, report section, discussion group, or photosets	Email; blog posts and comments; sharing of links, photos, videos, and documents; turn-based games

TABLE 2.3 Types of Social Media Listed with Example Services

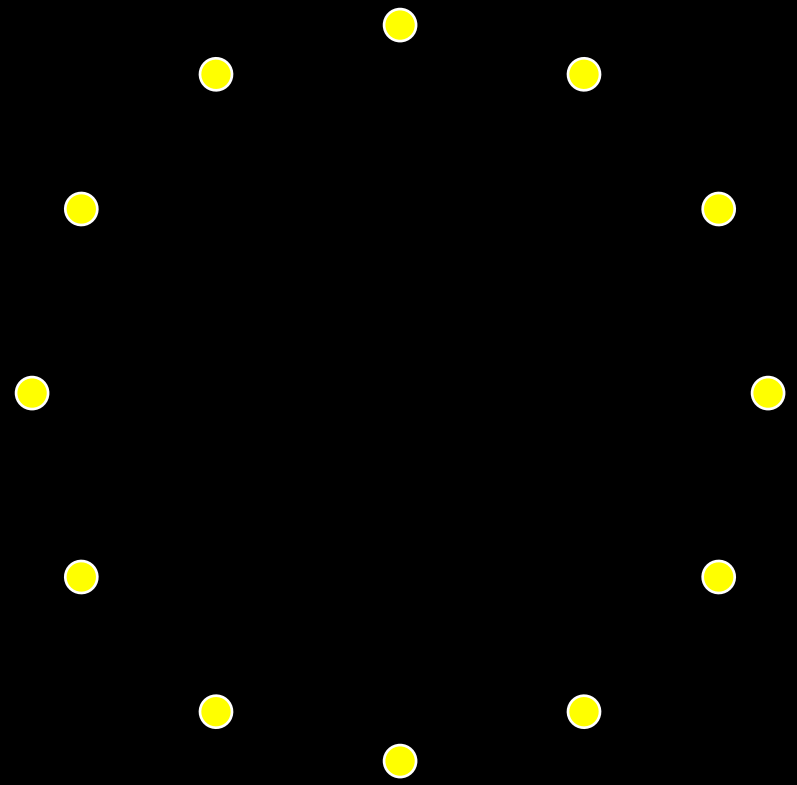
Social Media Type	Examples
ASYNCHRONOUS THREADED CONVERSATION	
Email	Gmail, Hotmail, AIM Mail, Yahoo! Mail, MS Outlook
BBS, discussion forums, Usenet newsgroups, email lists	Slashdot, Google groups, Yahoo! Groups, Yahoo! Answers, Listserv
SYNCHRONOUS CONVERSATIONS	
Chat, instant messaging, texting	UNIX Talk, IRC, Yahoo! Messenger, MSN Messenger, AIM, Google Talk, ChaCha
Audio and videoconferencing	Skype, Gizmo, iChat, Window's Live
WORLD WIDE WEB	**TRADITIONAL WEB SITES, HOMEPAGES, AND DOCUMENTS**
Corporate, organizational, and government websites and documents	Ford.com, UMD.edu, Prevent.org, Serve.gov; Data.gov
Homepages	Faculty member websites, artists' portfolio websites, family history websites
COLLABORATIVE AUTHORING	
Wiki	Wikipedia, Wikia (Lostpedia), pbwiki, wetpaint
Shared documents	Google Docs, Zoho, Etherpad

BLOGS AND PODCASTS	
Blogs	LiveJournal, Blogger, WordPress
Microblogs and activity streams	Twitter, Yammer, Buzz, Activity Streams
Multimedia blogs and podcasts	Vlogs (video blogs such as Qik), photo blogs (Fotolog, FAILblog.org), moblog (mobile blogging such as moblog.net), podcasts (iTunes, NPR)
SOCIAL SHARING	
Video and TV	YouTube, Hulu, Netflix, Vimeo, Chatroulette
Photo and art	Flickr, Picasso, deviantART
Music	Last.Fm; imeem; Sonic Garden
Bookmarks, news, and books	Delicious, Digg, Reddit, StumbleUpon, Goodreads, LibraryThing, citeulike
SOCIAL NETWORKING SERVICES	
Social and dating	Facebook, MySpace, BlackPlanet, Tagged, eHarmony, Match
Professional	LinkedIn, Plaxo, XING
Niche networks	Ning (e.g., classroom 2.0), Ravelry, Grou.ps
ONLINE MARKETS AND PRODUCTION	
Financial transaction	eBay, Amazon, craigslist, Kiva
User-generated products	Instructables, Threadless, TopCoder, Sourceforge, Codeplex
Review sites	ePinions, Amazon, Angie's List, Yelp
IDEA GENERATION	
Idea generation, selection, and challenge sites	IdeaConnection, Chaordix, IdeaScale, Imaginatik

VIRTUAL WORLDS	
Virtual reality worlds	Second Life, Club Penguin, Webkinz, Habbo
Massively multiplayer games	World of Warcraft, Lord of the Rings Online, Aion
MOBILE-BASED SERVICES	
Location sharing, annotation, and games	Foursquare, Gowalla, Loopt, MapMyRun, Geocaching, Letterboxing, SCVNGR

Society as a Graph

People are represented as *nodes/actors/vertices/points.*



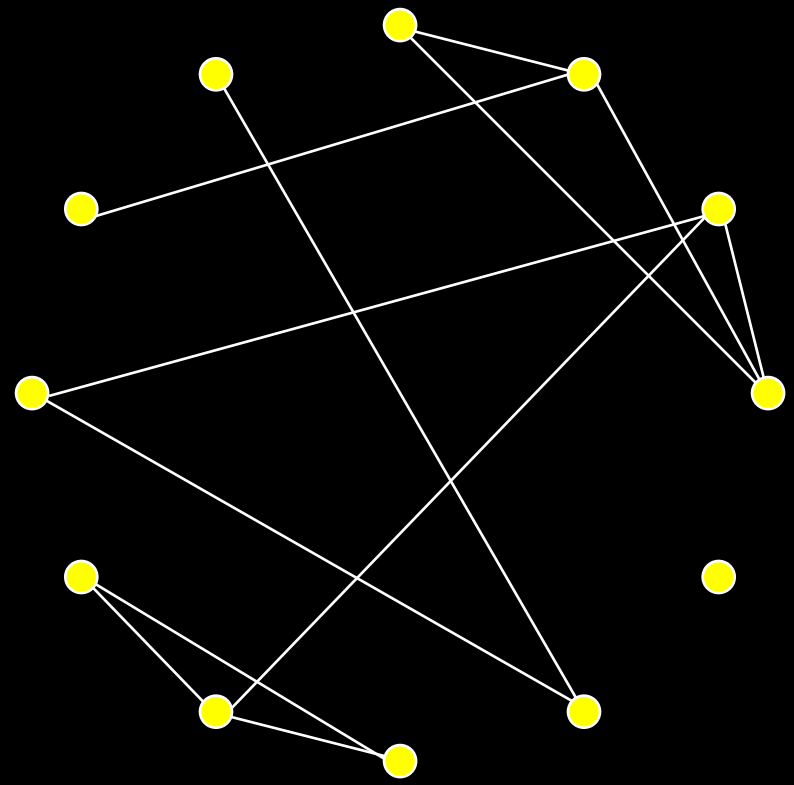
Actors / nodes / vertices / points
Computers / Telephones
Persons / Employees
Companies / Business Units
Articles / Books
Can have properties (attributes)

Society as a Graph

People are represented as *nodes/actors/vertices/points.*

Relationships are represented as *edges/ties/arcs/lines/links.*

(Relationships may be acquaintanceship, friendship, co-authorship, etc.)



Society as a Graph

Actors / nodes / vertices / points

Ties / edges / arcs / lines / links

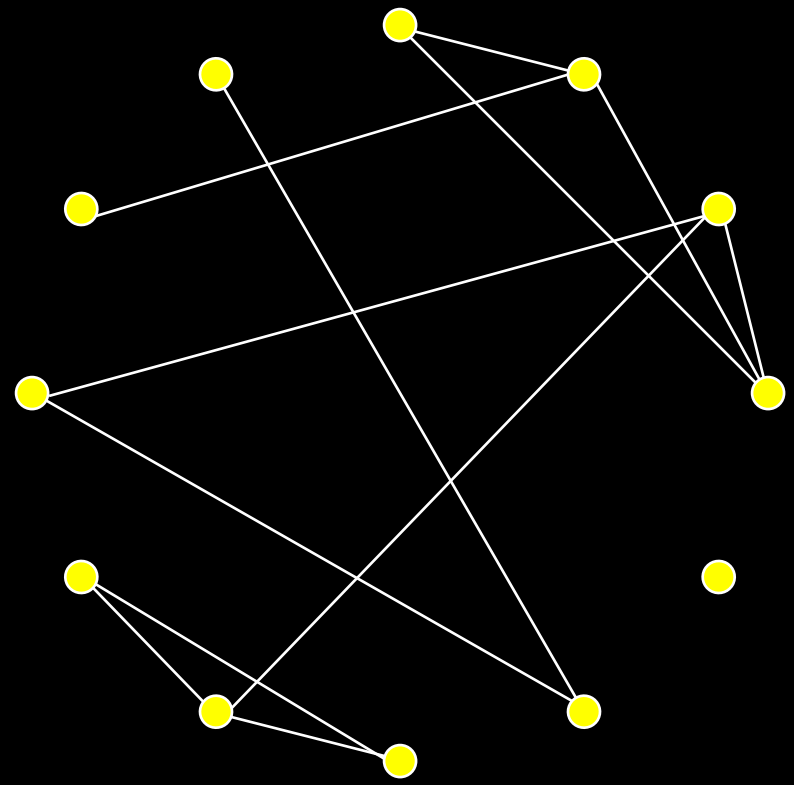
connect pair of actors

types of social relations

- friendship
- acquaintance
- kinship
- advice
- hindrance
- sex

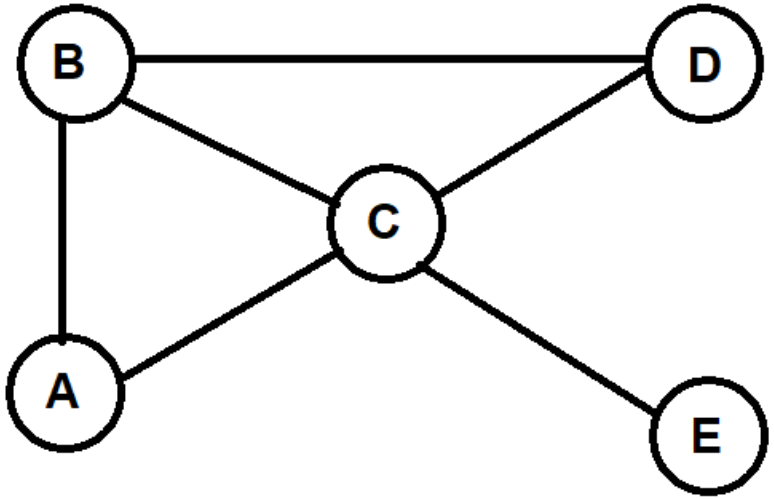
allow different kind of flows

- messages
- money
- diseases

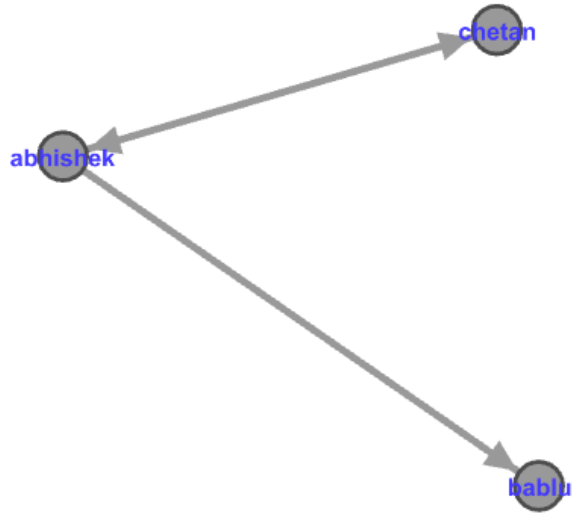


Edge Types

- **Directed Edge-** One way connection – Twitter, Instagram, ResearchGate etc.- you can follow someone without them following you back.



- **Undirected Edge-** two – way or mutual Connection- Facebook, LinkedIn friend connection



Data Management: Adjacency Matrix

- Symmetric binary: Who knows who?

	A	B	C
A	0	1	0
B	1	0	1
C	0	1	0

- Symmetric weighted: distance between places

	Allahabad	Lucknow	Varanasi
Allahabad	0	205	120
Lucknow	205	0	315
Varanasi	120	315	0

Data Management: Adjacency Matrix

- Asymmetric binary: choose 3 friends to sit with in the class

	A	B	C	D	E	F
A	-	1	1	1	0	0
B	1	-	0	0	1	1
C	0	1	-	1	0	1
D	0	0	1	-	1	1
E	1	1	1	0	-	0
F	1	0	1	0	1	-

- Asymmetric weighted: number of emails sent to friends

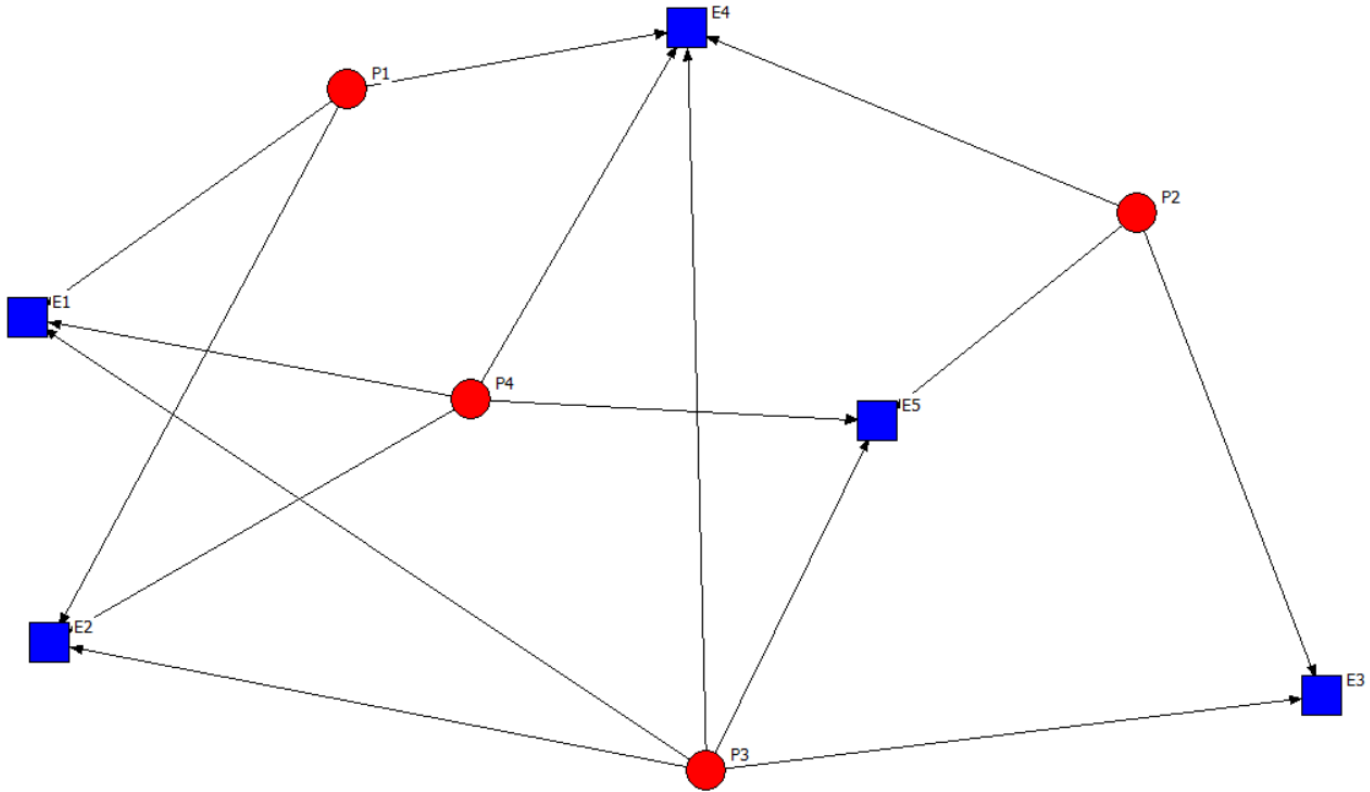
	A	B	C
A	0	2	7
B	3	0	5
C	1	4	0

One mode vs two mode

- Direct ties between actors: Adjacency Matrix **– one mode**
- Ties between different set of actors: Affiliation Matrix **- two mode**

	A	B	C	D	E	F
A	-	1	1	1	0	0
B	1	-	0	0	1	1
C	0	1	-	1	0	1
D	0	0	1	-	1	1
E	1	1	1	0	-	0
F	1	0	1	0	1	-

	E 1	E 2	E 3	E 4	E 5
	-	-	-	-	-
P1	1	1	0	1	0
P2	0	0	1	1	1
P3	1	1	1	1	1
P4	1	1	0	1	1



Adjacency Matrix vs Attribute Matrix

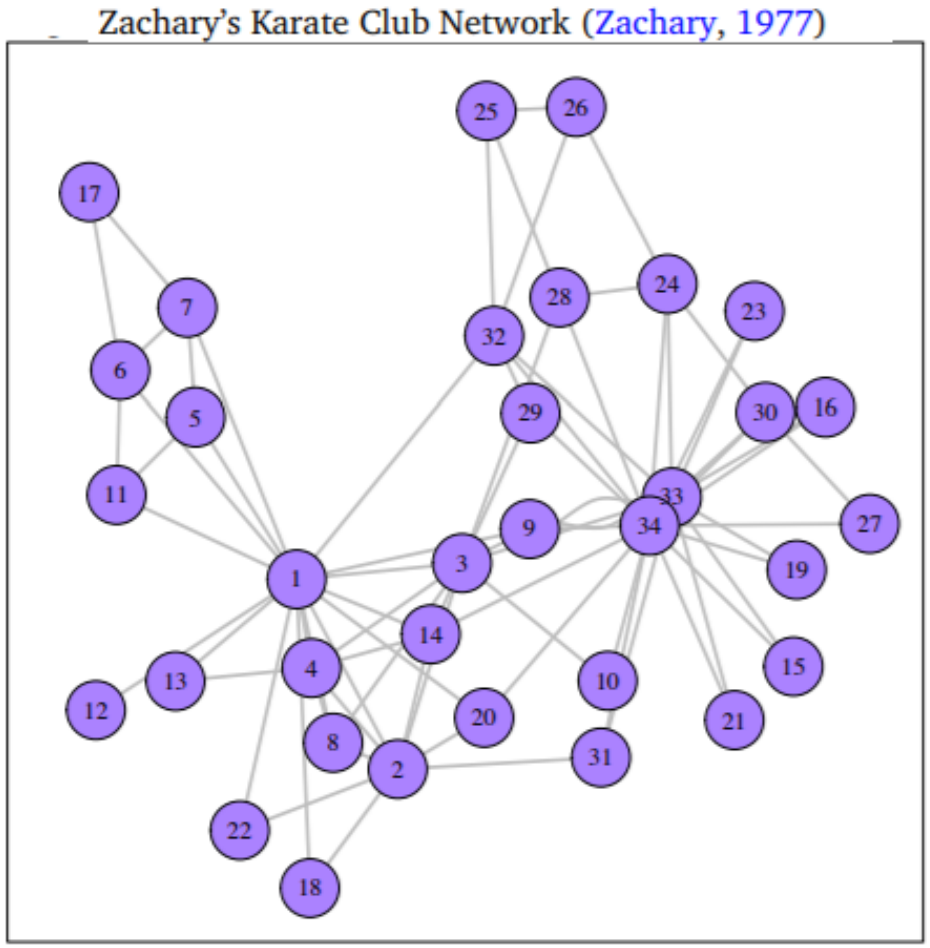
- Sociomatrix or Adjacency matrix only records ties between nodes
- Each column is different attribute of the nodes. (nodelist)

Node id/Name	Node lebel	Node type
Name	**Gender**	**Species**
Ram	M	Human
Sita	F	Human
Pavanhans	-	Machine

Zachary's Karate Club Network Edgelist Representation (Zachary, 1977)

1 32	1 4	3 10	6 7	20 34	26 32
1 22	1 3	3 33	7 17	21 34	27 34
1 20	1 2	3 29	9 34	21 33	27 30
1 18	2 31	3 28	9 33	23 34	28 34
1 14	2 22	3 8	9 33	23 33	29 34
1 13	2 20	3 4	10 34	24 30	29 32
1 12	2 18	4 14	14 34	24 34	30 34
1 11	2 14	4 13	15 34	24 33	30 33
1 9	2 8	4 8	15 33	24 28	31 34
1 8	2 4	5 11	16 34	24 26	31 33
1 7	2 3	5 7	16 33	25 32	32 34
1 6	3 14	6 17	19 34	25 28	32 33
1 5	3 9	6 11	19 33	25 26	33 34

Edgelist



Edgelist

Source	Target	Weight	Type	Relations
1	2	1	undirected	friend
1	3	1	undirected	friend
2	3	1	undirected	friend
2	4	1	undirected	family
1	4	1	undirected	friend
3	4	1	undirected	friend
1	5	1	undirected	friend
2	5	1	undirected	family
3	5	1	undirected	friend
4	5	1	undirected	family
1	6	1	undirected	friend
2	6	1	undirected	family
3	6	1	undirected	friend
4	6	1	undirected	family
5	6	1	undirected	family
1	9	1	undirected	friend
2	9	1	undirected	friend
3	9	1	undirected	friend
3	10	1	undirected	friend
2	10	1	undirected	friend
1	10	1	undirected	friend
4	10	1	undirected	friend

Characteristics

- Order
- Degree
- Clustering
- Distance to other node
- Centrality, influence, power…

Order and Size

- The **order** of a graph is the number of nodes in the graph
- The **cardinality** of the set is the number of edges in the graph
- The **size** of the graph is defined as how many edges there could possibly be if everyone had a relationship with everyone else in the network. It is the maximum possible number of ties, and is calculated as:
  - [n(n-1)]/2 for an undirected network, and
  - [n(n-1)] for a directed network.

Degree/ Degree Centrality/Popularity Measure

- Degree/Degree Centrality is the measure of the total number of edges attached/ connected to a particular node/vertex.

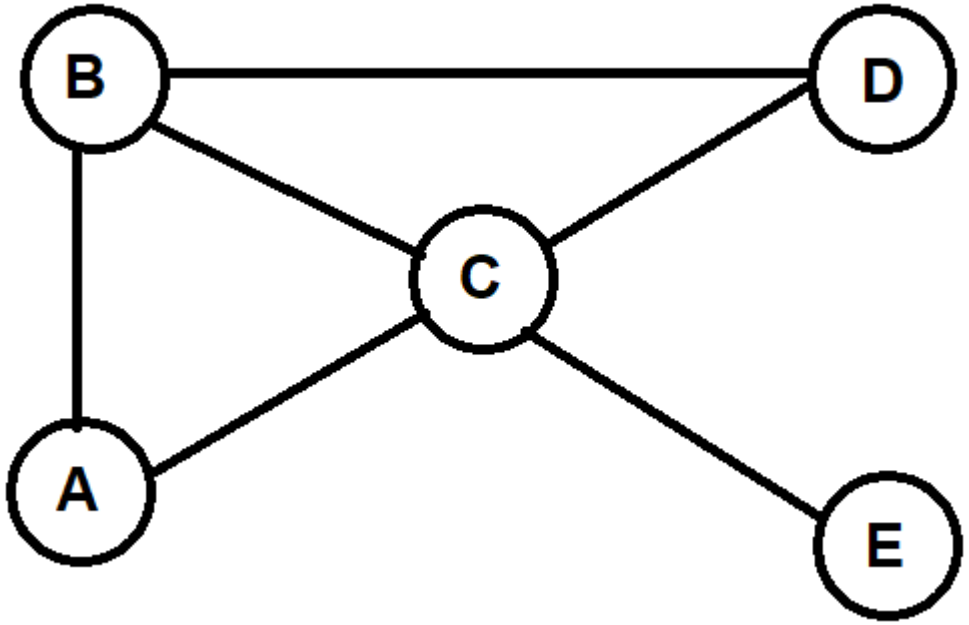
- Indicate most connected node in the network

- For Directed Network: two measures of degree
  - **In-Degree:** number of connections that point inward at node/vertex
  - **Out-Degree:** number of connections that originate at a vertex and point outward to other vertices

Degree/ Degree Centrality/Popularity Measure

- A simplest measure of node connectivity
- Degree centrality **assigns an importance score based simply on the number of links held by each node**
- A high degree centrality score simply means **that a node has a larger than average number of connections for that graph**.

Degree/ Degree Centrality/Popularity Measure



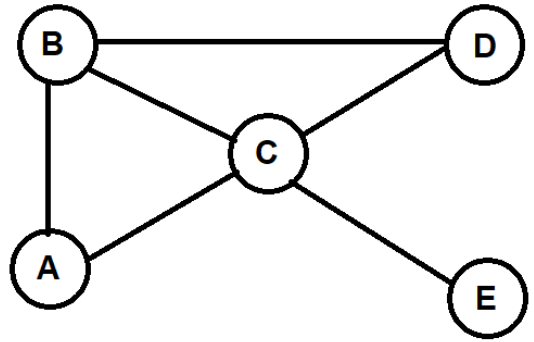
- Deg(A) = 2
- Deg(B) = 3
- Deg(C) = 4
- Deg(D) = 2
- Deg(E) = 1

- Average Degree=Total Edges/Total Node
- =6/5=1.2

- LOOP
- Deg(F) = 4

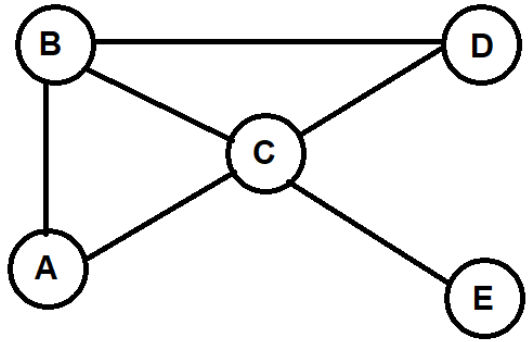
Degree Centrality Score

Node (n)	Score (s)	Standardised Score = s/(n-1)
A	2	2/4 = 0.5
B	3	¾ = 0.75
C	4	4/4 = 1
D	2	2/4 = 0.5
E	1	¼ = 0.25



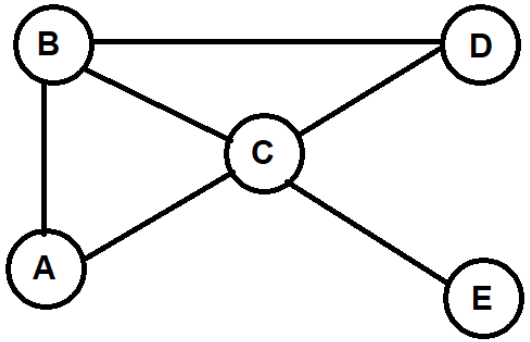
- Standardised Score is a normalised form. The most well connected node in the network is “C”.

Closeness Centrality



- How close a node is to the rest of the nodes of the network.
- It is the ability of the node to reach the other nodes in the network.
- It is calculated as the inverse of the sum of the distance between a node and other nodes in the network.

Closeness Centrality Score



	A	B	C	D	E	Total Distance (td)
A	0	1	1	2	2	6
B	1	0	1	1	2	5
C	1	1	0	1	1	4
D	2	1	1	0	2	6
E	2	2	1	2	0	7

Node (n)	Closeness Score (cs = 1/td)	Standardised Score = (n-1)*cs
A	1/6	4*1/6= 2/3=0.67
B	1/5	4*1/5= 4/5=0.80
C	¼	4*1/4 = 1= 1.00
D	1/6	4*1/6= 2/3=0.67
E	1/7	4*1/7= 4/7= 0.57

- The most closest/ central node in the network is “C” with highest closeness score of 1.

Betweenness Centrality

- Freeman (1977) gave the first formal definition of betweenness centrality
- It is a measure of how often a node appears in the shortest path connecting two other nodes.

Definition

The betweenness centrality of a node v is given by the expression:

$$g(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

where σ_{st} is the total number of shortest paths from node s to node t and $\sigma_{st}(v)$ is the number of those paths that pass through v (not where v is an end point).[2]

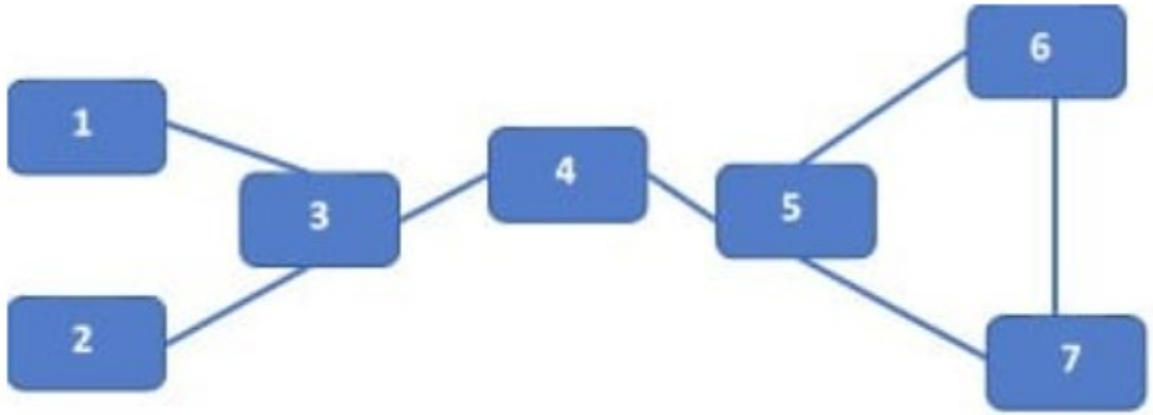
Note that the betweenness centrality of a node scales with the number of pairs of nodes as suggested by the summation indices. Therefore, the calculation may be rescaled by dividing through by the number of pairs of nodes not including v, so that $g \in [0, 1]$. The division is done by $(N-1)(N-2)$ for directed graphs and $(N-1)(N-2)/2$ for undirected graphs, where N is the number of nodes in the giant component. Note that this scales for the highest possible value, where one node is crossed by every single shortest path. This is often not the case, and a normalization can be performed without a loss of precision

$$\mathrm{normal}(g(v)) = \frac{g(v) - \min(g)}{\max(g) - \min(g)}$$

which results in:

$$\max(normal) = 1$$
$$\min(normal) = 0$$

Note that this will always be a scaling from a smaller range into a larger range, so no precision is lost.



Node	Score	Standardized Score
1	0	0
2	0	0
3	16/3	16/45
4	13/3	13/45
5	13/3	13/45
6	0	0
7	0	0

Node pairs	Path value of Node 4
1,5	½
1,6	1/3
1,7	1/3
2,5	½
2,6	1/3
2,7	1/3
3,5	1
3,6	½
3,7	½
Total Score for Node 4	13/3

Nodes with high betweenness centrality are critical in controlling and maintaining flow in the network

Discussion

Ques. Suppose the above network refers to friendship network. Each node represents a person, and each edge represents friendship between the persons at ends. If you are interested in finding **the most poplar person** in the network, which centrality measure is the most appropriate? Give the answer with reasons why it is the most appropriate.

- **The most popular person should have the highest number of friends. Thus, degree centrality is the most appropriate measure.**

- **Ques.** Suppose the above network refers to information flow network of an organization. Each node represents a section in the organization, and each edge represents a possible information exchange between the sections at ends. If you are interested in finding **the section that can most efficiently obtain information from every other section**, which centrality measure is the most appropriate? Give the answer with reasons why it is the most appropriate.

- **To obtain information, one should be near from everyone. In this sense, the node in the nearest position on average can most efficiently obtain information. Thus, closeness centrality is the most appropriate.**

- **Ques.** Again, suppose the above network refers to information flow network of an organization. If you are interested in finding **the section that can most frequently control information flow in the network**, which centrality measure is the most appropriate? Give the answer with reasons why it is the most appropriate

- **To control information flow, a node should be between other nodes because the node can interrupt information flow between them. Thus, betweenness centrality is the most appropriate measure.**

Thank You!