PES UNIVERSITY ONLINE

SOCIAL NETWORK ANALYTICS

Properties of Networks

Path, Path length, Average path length, and Diameter

Prakash C O

Department of Computer Science and Engineering

SOCIAL NETWORK ANALYTICS

Properties of Networks

Path, Path length, Average path length, and Diameter

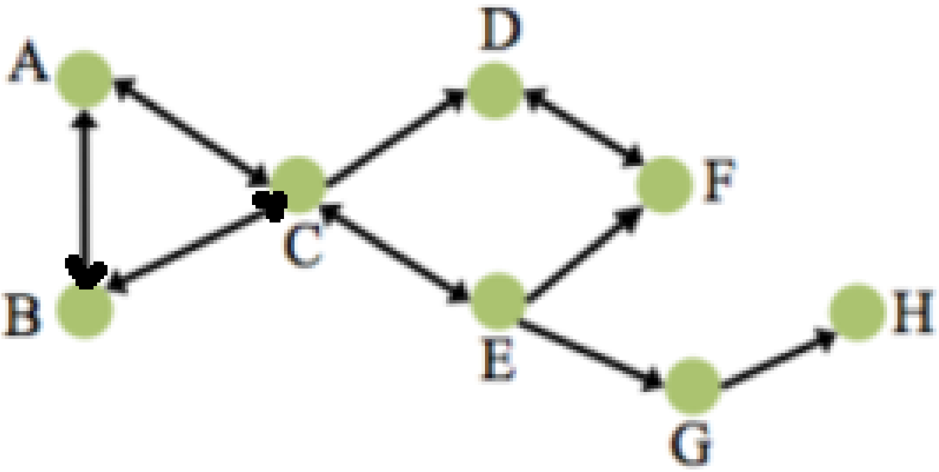
Prakash C O

Department of Computer Science and Engineering

Average Path Length

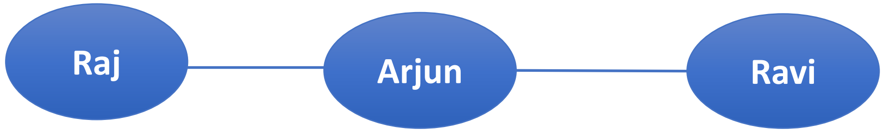
Path

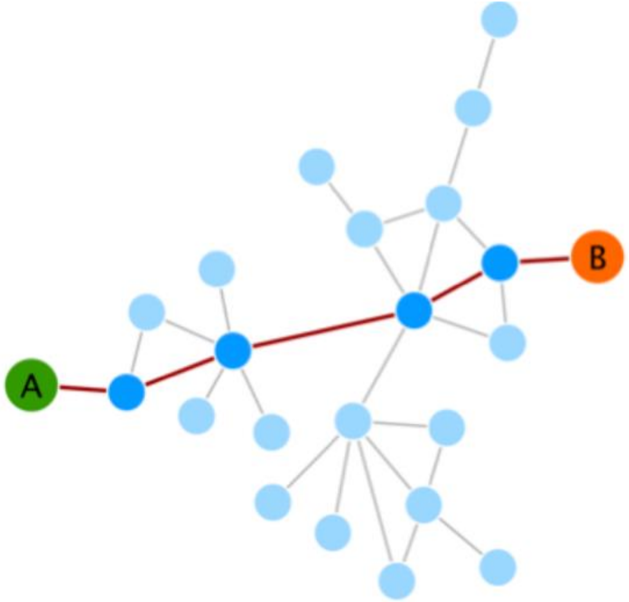
- In a network, a path is **a sequence of nodes** in which each node is connected by an edge to the next.
- For example, in the network below the paths between A and F are: ACDF, ACEF, ABCDF, and ABCEF.



Path Length

- Path length is simply **the distance between two nodes**, measured as the **number of edges between them**.
- If Raj is Arjun friend, and Arjun is Ravi’s friend, then the path length between Raj and Ravi is 2.





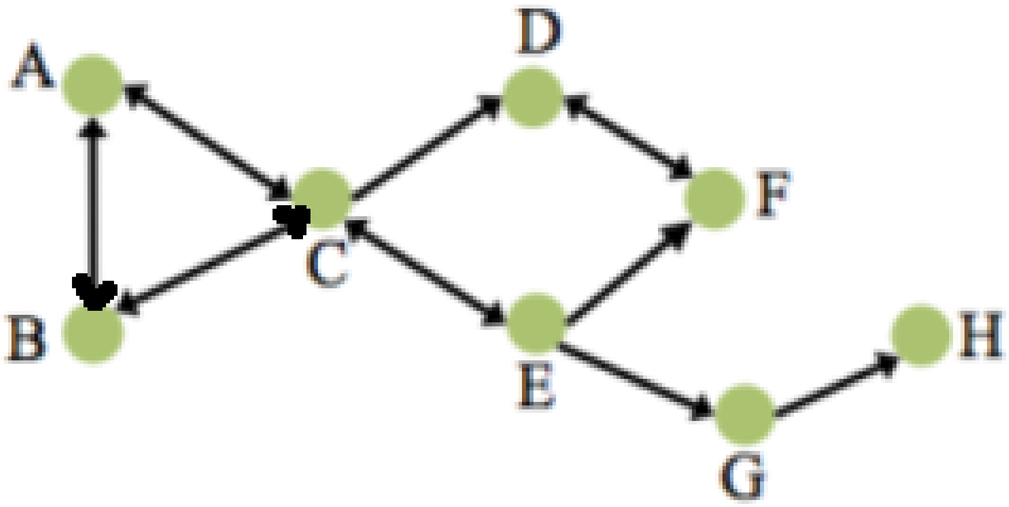
Average Path Length

Path Length

- For example, in the network below **the paths between A and F** are:

  ACDF, ACEF, ABCDF, ABCEF, with path lengths 3,3,4,4 respectively.

  The shortest paths are the first two.



Average Path Length

Path Length

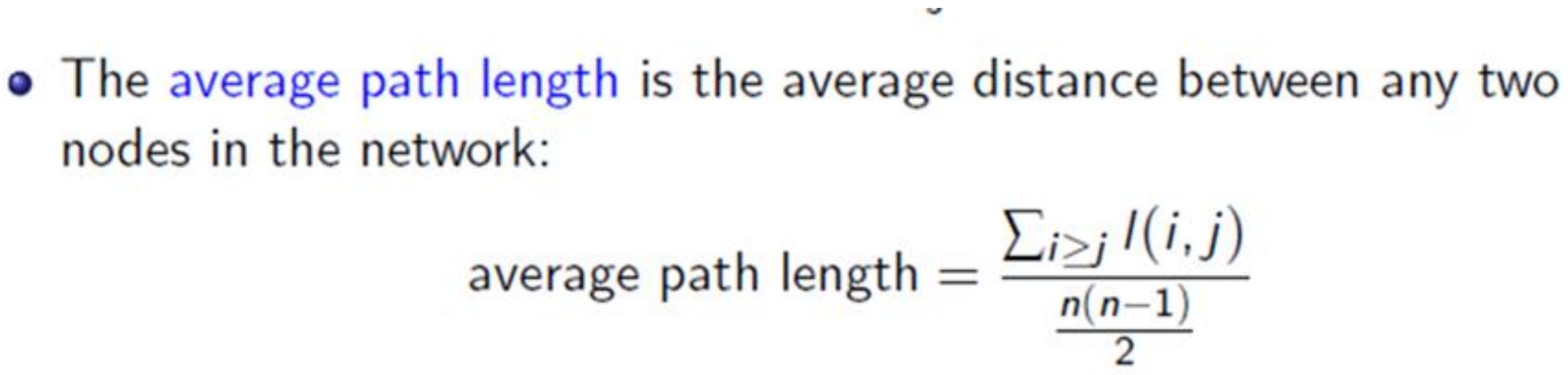
- **LinkedIn** is a good example of a social network that **uses paths** and **path lengths** to show how you might connect to other people.
- When you look at someone’s profile page, it will calculate the shortest path from you to them, and show you the first person in that path who might be able to introduce you.

Average Path Length

Why do we care about path?

- **Path is interesting for several reasons.**
  - **Path mean connectivity**.
  - **Path captures the indirect interactions in a network**, and individual nodes benefit (or suffer) from indirect relationships because friends might provide access to favors from their friends and information might spread through the links of a network.
  - **Path is closely related to small-world phenomenon**.
  - **Path is related to many centrality measures**.
  - …

Average Path Length



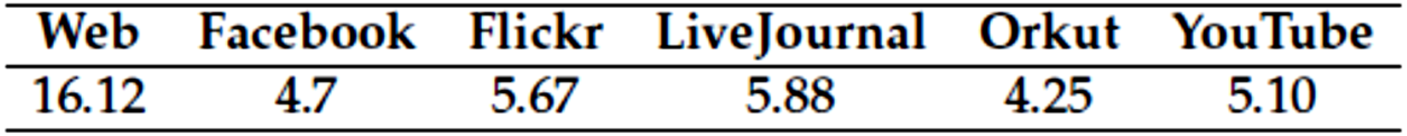
- The average path length is the average distance between any two nodes in the network:

$$\text{average path length} = \frac{\sum_{i \geq j} l(i,j)}{\frac{n(n-1)}{2}}$$

- Average path length(APL) is one of the three most robust measures of network topology.
- **Examples**:
  - The average number of clicks which will lead you from one website to another, or
  - The number of people you will have to communicate through, on an average, to contact a complete stranger.

Average Path Length

- In real-world networks, any two members of the network are usually connected via short paths. In other words, the average path length is small.
  - **Six degrees of separation:**
    - **Stanley Milgram** In the well-known small-world experiment conducted in the 1960’s conjectured that people around the world are connected to one another via a path of at most 6 individuals.
  - **Four degrees of separation:**
    - **Lars Backstrom et al.** in May 2011, the average path length between individuals in the Facebook graph was 4.7. (4.3 for individuals in the US)



Web	Facebook	Flickr	LiveJournal	Orkut	YouTube
16.12	4.7	5.67	5.88	4.25	5.10

Table 4.2 provides the average path length for real-world social networks and the web.

Average Path Length

Applications

- In a real network like the **Internet, a small average path length facilitates the quick transfer of information and reduces costs**.

- **Most real networks have a very small average path length leading to the concept of a small world where everyone is connected to everyone else through a very short path.**

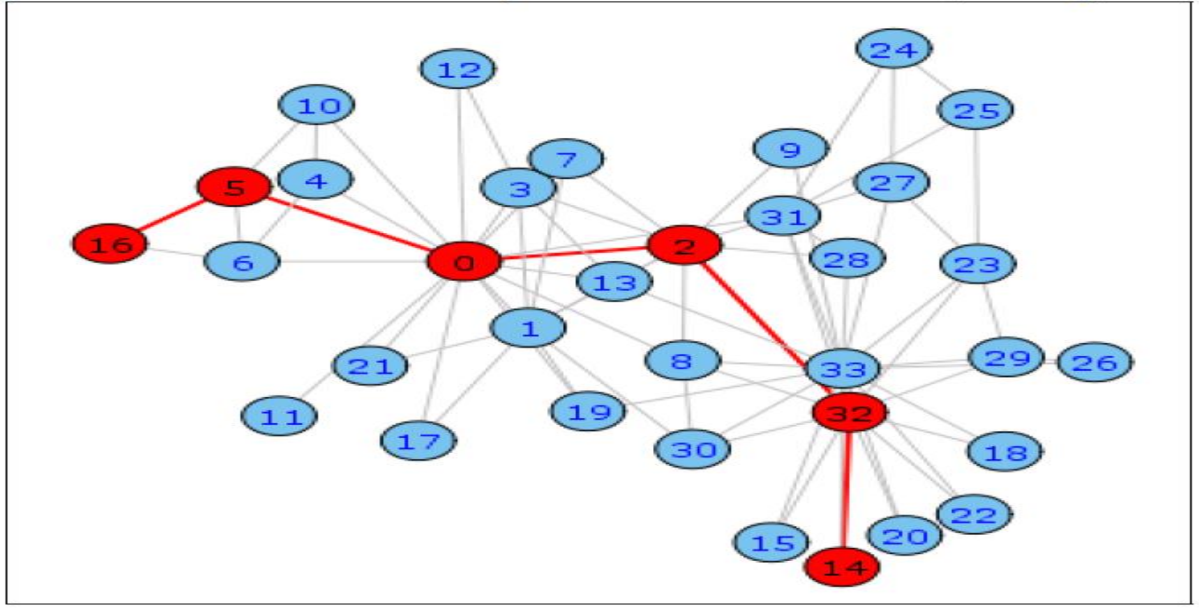
- A **power grid network** will have **fewer losses if its average path length is minimized.**

Network Diameter

- Let $l(i,j)$ denote the length of the shortest path (or geodesic) between node i and j (or the distance between i and j).
- The diameter of a network is the largest distance between any two nodes in the network:

$$\text{diameter} = \max_{i,j} l(i,j)$$

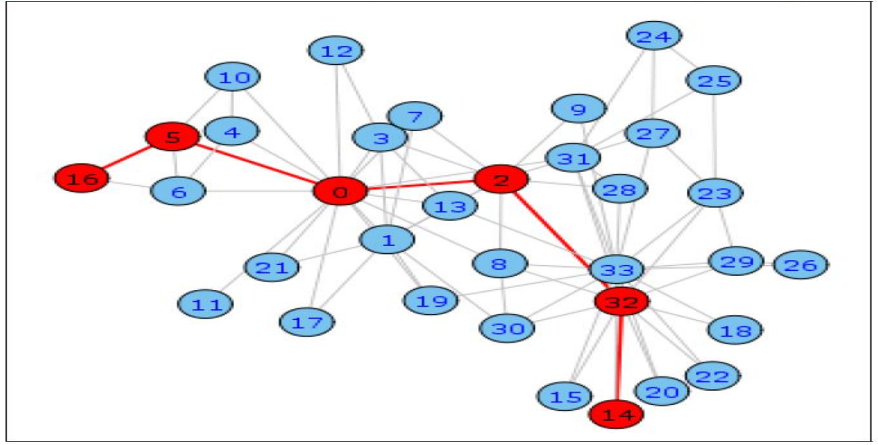
Figure 15: Diameter of Zachary's Karate Club Network (Zachary, 1977



Network Diameter

- The diameter of a network is **the longest of all the calculated shortest paths** in a network.
  Diameter is the **shortest distance between the two most distant nodes** in the network.
- The diameter is a representative of the linear size of a network.
- **Diameter is thus a signal about the ability for information or disease to diffuse on the network.**

Figure 15: Diameter of Zachary's Karate Club Network (Zachary, 1977



Average Path Length

Demo using NetworkX

- AvgPathLength_Diameter_CC_Measures.py
- https://networkx.github.io/documentation/stable/tutorial.html

SOCIAL NETWORK ANALYTICS

Properties of Networks

Reciprocity and Transitivity/Clustering coefficient

Prakash C O

Department of Computer Science
and Engineering

SOCIAL NETWORK ANALYTICS

Properties of Networks

Reciprocity

Prakash C O

Department of Computer Science and Engineering

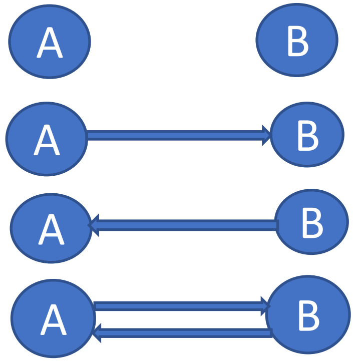
Linking behavior

- Often, we need to observe a specific behavior in a social media network. One such behavior is **linking behavior**.
- **Linking behavior determines how links (edges) are formed in a social graph/network.**
- For analyzing Linking behavior, we discuss two well-known quantitative measures:
  1. **Reciprocity** and
  2. **Transitivity**.

Both measures are commonly used in directed networks, and transitivity can also be applied to undirected networks.

Dyadic relationships

- With directed graph/network, **there are four possible dyadic relationships**:
  1. A and B are not connected,
  2. A sends to B,
  3. B sends to A, or
  4. A and B send to each other.

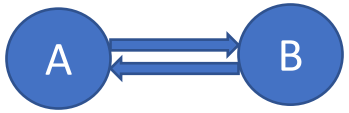


- **What is the prevalence of reciprocity in the given directed network?** One approach is to focus on the dyads, and ask what proportion of pairs have a reciprocated tie between them?

- In network science, reciprocity is **a measure of the likelihood of vertices in a directed network to be mutually linked.**

Reciprocity

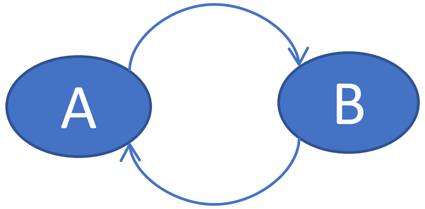
- In social media networks, the reciprocity is **the extent to which two actors reciprocate each other's friendship or other interaction.**
- In social media networks where users send messages to one another, the issue of reciprocity naturally arises:
  Does the communication between two users take place only in one direction, or is it reciprocated?
- Reciprocity refers to responding to a positive action with another positive action; it creates, maintains and strengthens various social bounds. It is the foundation of social order and is a major key to success.
- In real network problems, people are interested in determining the likelihood of occurring double links (with opposite directions) between vertex pairs.



Reciprocity

- Reciprocity considers closed loops of length 2, which can only happen in directed graphs.

  Formally, if node A is connected to node B, B by connecting to A exhibits reciprocity.



- **On microblogging site Tumblr**, for example, these nodes are known as "**mutual followers**."

  Informally, reciprocity is ***If you become my friend, I'll be yours.***

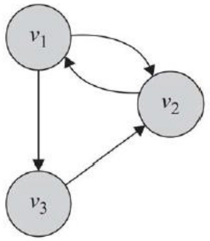


Figure : A Graph with Reciprocal Edges.

Reciprocity

- **Reciprocity computation counts the number of reciprocal pairs in the graph**. Any directed graph can have a maximum of |E|/2 pairs. This happens when all edges are reciprocal.
- **Reciprocity can be computed using the adjacency matrix A**:

$$R = \frac{\sum_{i,j,i<j} A_{i,j}A_{j,i}}{|E|/2} = \frac{2}{|E|}\sum_{i,j,i<j} A_{i,j}A_{j,i} = \frac{2}{|E|}\sum_{i,j,i<j} A_{i,j}A_{j,i} = \frac{2}{|E|} \times \frac{1}{2}\mathrm{Tr}(A^2)$$

$$= \frac{1}{|E|}\mathrm{Tr}(A^2) = \frac{1}{m}\mathrm{Tr}(A^2),$$

where $\mathrm{Tr}(A^2) = A_{1,1} + A_{2,2} + \cdots + A_{n,n} = \sum_{i=1}^{n} A_{i,i}$ and m is the number of edges in the network. Note that the maximum value for $\sum_{i,j} A_{i,j}A_{j,i}$ is m when all directed edges are reciprocated.

Reciprocity

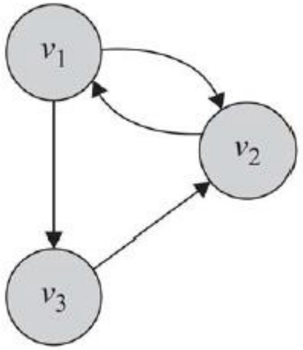


Figure : A Graph with Reciprocal Edges.

Example. *For the graph shown in Figure the adjacency matrix is*

$$A = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}.$$

Its reciprocity is

$$R = \frac{1}{m}\mathrm{Tr}(A^2) = \frac{1}{4}\mathrm{Tr}\left(\begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 0 \end{bmatrix}\right) = \frac{2}{4} = \frac{1}{2}.$$

The reciprocity value indicates what percentage of edges in the network are reciprocal.

Reciprocity - Assignment

Paper Reading

- **Predicting Reciprocity in Social Networks -** Justin Cheng, Daniel Romero, Brendan Meeder, Jon Kleinberg

In this paper we study the problem of reciprocity prediction: given the characteristics of two users, we wish to determine whether the communication between them is reciprocated or not.

SOCIAL NETWORK ANALYTICS

Properties of Networks

Transitivity/Clustering coefficient

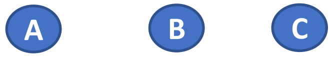
Prakash C O

Department of Computer Science and Engineering

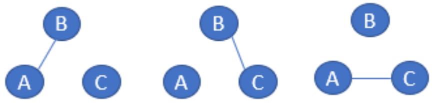
Triadic relationships

- **With undirected graph, there are 4-possible Triadic relationships:**

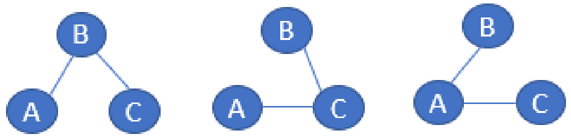
1. **no ties,**



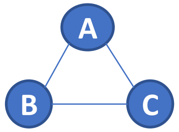
2. **one tie,**



3. **two ties (open triad – two edges with a shared vertex), or**



4. **all three ties(transitive/balanced triad/closed triad)**



- The counts of the relative prevalence of these **four types of relations across all possible triples** can give a good sense of the extent to which a population is characterized by "**isolation**," "**couples only**," "**structural holes**“ or "**clusters**.“
- **With directed graph, there are actually 16 possible types of relations among 3 actors**, including relationships that exhibit hierarchy, equality, and the formation of exclusive groups (e.g. where two actors connect, and exclude the third).

Network Transitivity

- In transitivity, **we analyze the linking behavior to determine whether a network demonstrates** a **transitive behavior**.
- In mathematics, for a transitive relation R, aRb ^ bRc → aRc.

Transitive Linking behavior(Network Transitivity)

- The **transitive linking behavior** can be described as follows.
  - **Let v1, v2, v3 denote three nodes. When edges (v1, v2) and (v2, v3) are formed, if (v3, v1) is also formed, then we have observed a transitive linking behavior (transitivity).** This is shown in Figure below

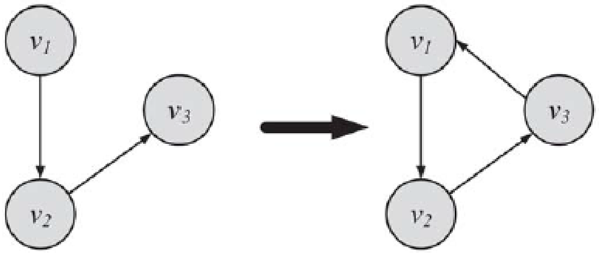


Figure : Transitive Linking.

  - In a less formal setting,

    Transitivity is when a friend of my friend is my friend.

Network Transitivity

Transitive Linking behavior(Network Transitivity)

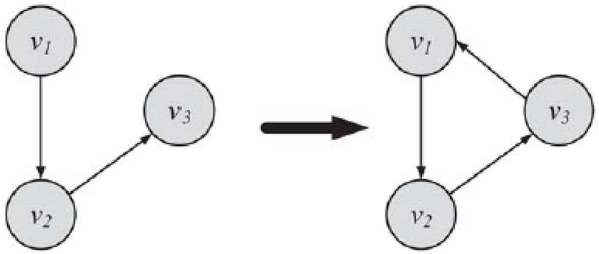


Figure : Transitive Linking.

Transitivity is when a friend of my friend is my friend.

- As shown in the definition, **a transitive linking behavior needs at least three edges. These three edges, along with the participating nodes, create a triangle.**
- **Higher transitivity in a graph results in a denser graph, which in turn is closer to a complete graph.**
  Thus, **we can determine how close graphs are to the complete graph by measuring transitivity**.
  This can be performed by measuring the [global] clustering coefficient and local clustering coefficient.

Clustering Coefficient

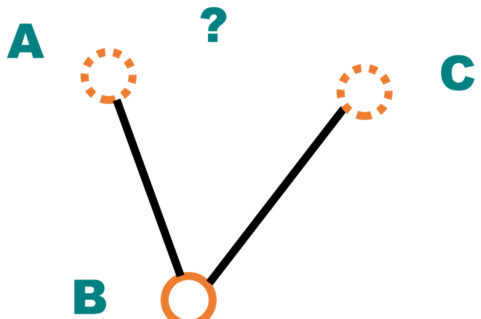
- **The clustering coefficient analyzes transitivity in an undirected graph.**
- **In graph theory**, a **clustering coefficient** is **a measure of the degree to which nodes in a graph tend to cluster together**

Clustering Coefficient

- A network is said to show clustering if the **probability of two vertices being connected by an edge is higher when the vertices in question have a common neighbor.**

- **Watts and Strogatz** measured this clustering by defining a **clustering coefficient C**.

- **In many real-world networks the clustering coefficient is found to have a high value**, anywhere from a few percent to 50 percent or even more.



Clustering Coefficient

- **Clustering is a typical property of social networks, where two individuals with a common friend are likely to know each other.**

- The clustering coefficient is a measure of an "**all-my-friends-know-each-other**" property.

- 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.
  We refer to this principle as **triadic closure.**

- **To measure the clustering in a social network, a common measure is the clustering coefficient.**

Clustering Coefficient

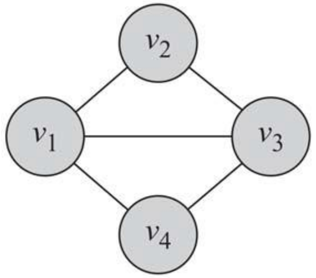
Two versions of this measure exist:

1. The **global clustering coefficient**(the overall clustering coefficient)
2. The **local clustering coefficient**(the individual clustering coefficient)

- **The global version** was designed to give **an overall indication of the clustering in the network**,
- **The local version** gives **an indication of the embeddedness of single nodes**

Clustering Coefficient (Overall)

- **The clustering coefficient analyzes transitivity in an undirected graph.**
- Transitivity in networks is observed when triangles are formed, we can measure transitivity by counting paths of length 2 (edges (v1, v2) and (v2, v3)) and checking whether the third edge (v3, v1) exists (i.e., the path is closed).
  Thus, clustering coefficient C is defined as

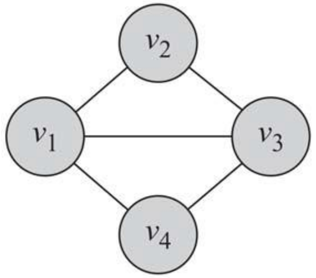


$$C = \frac{|\text{Closed Paths of Length 2}|}{|\text{Paths of Length 2}|}$$

$$\mathbf{C} = \frac{\text{Closed paths of length-2}=\{v_1v_2v_3,\ v_2v_3v_1,\ v_3v_1v_2,\ v_1v_3v_4,\ v_3v_4v_1,\ v_4v_1v_3\}}{\begin{array}{l}\text{Paths of length-2 } (v_1 \text{ centered}) =\{v_2v_1v_4,\ v_2v_1v_3,\ v_3v_1v_4\} \\ \text{Paths of length-2 } (v_3 \text{ centered}) =\{v_2v_3v_4,\ v_2v_3v_1,\ v_1v_3v_4\} \\ \text{Paths of length-2 } (v_2 \text{ centered}) =\{v_1v_2v_3\} \\ \text{Paths of length-2 } (v_4 \text{ centered}) =\{v_1v_4v_3\}\end{array}} = \mathbf{6/8}$$

Clustering Coefficient (Overall)

$$C = \frac{|\text{Closed Paths of Length 2}|}{|\text{Paths of Length 2}|}.$$



- Alternatively, we can count the number of triangles in network. Since **every triangle has three closed paths of length 2**, we can rewrite equation as

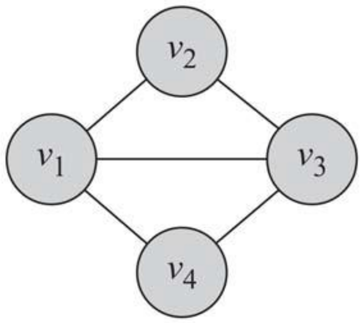
$$C = \frac{(\text{Number of Triangles}) \times 3}{\text{Number of Connected Triples of Nodes}}.$$

In this equation, a **triple** is **a set of three nodes, connected by two edges**

Two triples are different when their nodes are different, or their nodes are the same, but the triples are missing different edges.

Clustering Coefficient (Overall)

- **Example: For the graph in below Figure, the clustering coefficient is**

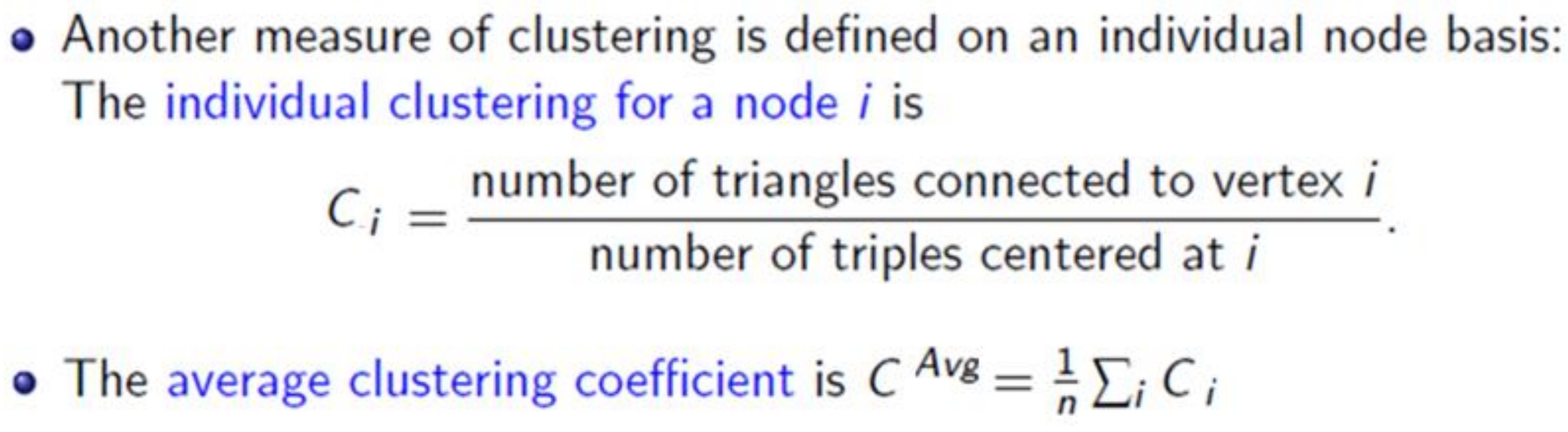


$$C = \frac{(Number\ of\ Triangles) \times 3}{Number\ of\ Connected\ Triples\ of\ Nodes}$$

$$= \frac{2 \times 3}{2 \times 3 + \underbrace{2}_{v_2v_1v_4,\, v_2v_3v_4}} = 0.75.$$

- Note:
  - **C** lies between 0 and 1.
  - In the above example, **C** is **overall clustering coefficient, that is overall indication of the clustering in the network**.
  - C **is the percentage of Transitive linking behavior exhibited by the network.**

Clustering Coefficient (Local/Individual)



- Another measure of clustering is defined on an individual node basis: The individual clustering for a node i is

$$C_i = \frac{\text{number of triangles connected to vertex } i}{\text{number of triples centered at } i}.$$

- The average clustering coefficient is $C^{Avg} = \frac{1}{n}\sum_i C_i$

- **Example:**

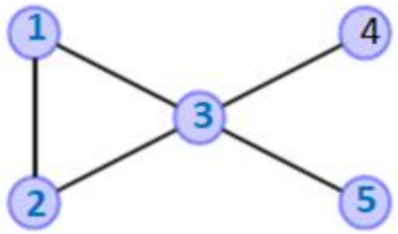


Figure: The overall clustering coefficient for this network is 3/8. The individual clustering for the nodes are 1, 1, 1/6, 0, and 0.

The clustering coefficient of a node *i* is

- **the fraction of pairs of i's friends that are connected to each other by edges, or**
- **What portion of i's neighbors are connected?**

Clustering Coefficient (Local/Individual)

Finding triples (simple approach) **in a network to compute Clustering Coefficient.**

➢ **In an undirected graph,**

- **For a node i having degree=1 , number of triples centered at node i = 0**
- **For a node i having degree=2 , number of triples centered at node i = 1**
- **For a node i having degree=3 , number of triples centered at node i =2+1=3**
- **For a node i having degree=4 , number of triples centered at node i =3+2+1=6**
- **In general, for a node i having degree=n_i, number of triples centered at node i = $[n_i*(n_i-1)]/2$**

$$C_i = \frac{\text{Number of triangles connected to node i}}{\text{Number of triples centered at node i}} = \frac{\text{Number of triangles connected to node i}}{[n_i*(n_i-1)]/2}$$

Where n_i is node i degree/connections

Clustering Coefficient (Local/Individual)

- **Individual Clustering Coefficient Computation** - Approach II

For a vertex i, let n_i be the number of neighbors of vertex i

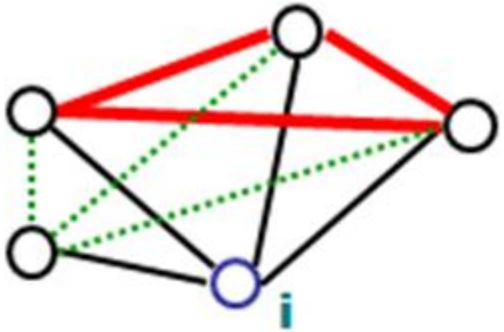
$$C_i = \frac{\text{Number of connections between i's neighbors}}{\text{Max. number of possible connections between i's neighbors}}$$

$$C_{i\ \textbf{directed}} = \frac{\text{Number of directed connections between i's neighbors}}{n_i * (n_i - 1)}$$

$$C_{i\ \textbf{undirected}} = \frac{\text{Number of undirected connections between i's neighbors}}{n_i * (n_i - 1)/2}$$

Clustering Coefficient (Local/Individual)

Example 1:



$n_i = 4$

max number of connections: 4*3/2 = 6

3 connections between i's neighbors

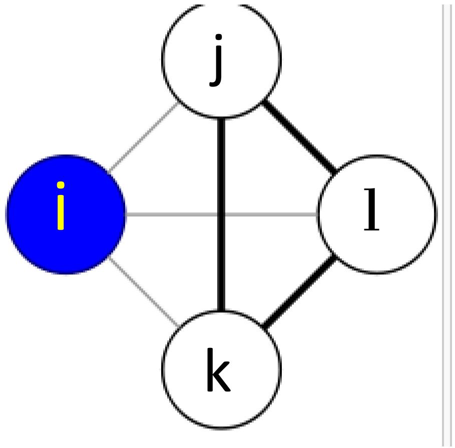
$C_i = 3/6 = 0.5$

━━ link present between i's neighbors

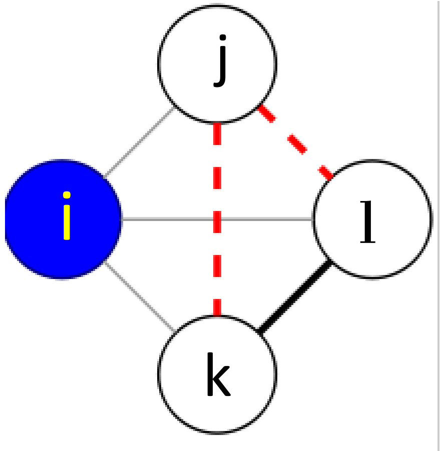
┈┈ link absent between i's neighbors

Clustering Coefficient (Local/Individual)

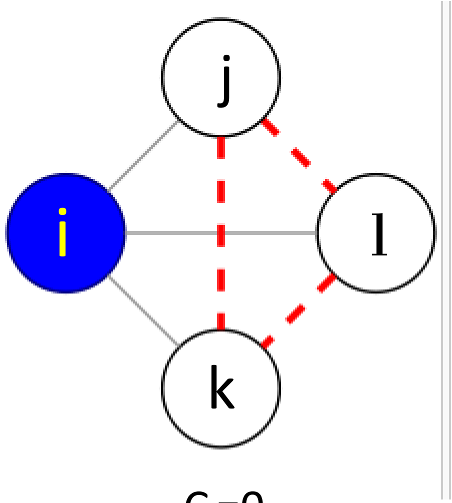
Example 2: Find clustering coefficient of vertex *i* in the following networks.



$C_i=1$
(Complete Network, i.e., all of his friends know each other)



$C_i=1/3$



$C_i=0$
(Star Network, i.e., None of his friends know each other)

Note: **Thick black line – Connections among his friends** and **dotted red lines - missing connections among his friends.**

Clustering Coefficient (Local/Individual)

- **In real-world social networks,**
  - **friendships are highly transitive**. In other words, friends of an individual are often friends with one another.
  - **Friendship-triads/transitive-friendships result in high average [local]clustering coefficients.**

- In May 2011, Facebook had an average clustering coefficient of 0.5 for individuals who had two friends; their degree was 2 .
  This indicates that for 50% of all users with two friends, their two friends were also friends with each other.

Clustering Coefficient (Local/Individual)

- Table 4.1 provides the average clustering coefficient for several real-world social networks and the web.

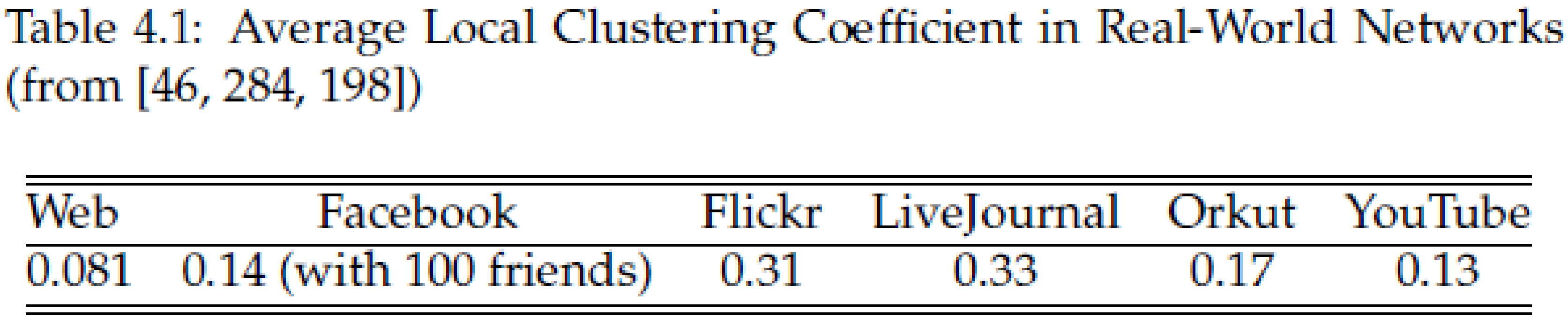


Table 4.1: Average Local Clustering Coefficient in Real-World Networks (from [46, 284, 198])

Web	Facebook	Flickr	LiveJournal	Orkut	YouTube
0.081	0.14 (with 100 friends)	0.31	0.33	0.17	0.13

Ref: **Social Media Mining**: *An Introduction* - Reza Zafarani

Clustering Coefficient

Why do we care about clustering coefficient?

Clustering is interesting for several reasons.

- **A clustering coefficient is a measure of the degree to which nodes in a graph tend to cluster together.**

- **Local clustering can be used as a probe for the existence of so-called structural holes in a network**, which are missing links between neighbors of a person.

Structural holes can be bad when we are interested in efficient spread of information or other traffic around a network because they reduce the number of alternative routes information can take through the network.

Clustering Coefficient

Why do we care about clustering coefficient?

Clustering is interesting for several reasons.

- **Structural holes can be good thing for the central vertex** whose friends lack connections because they give node i power over information flow between those friends.

- **The local clustering coefficient measures how influential node i is in this sense, taking lower values the more structural holes there are in the network around node i.**

- **Local clustering can be regarded as a type of centrality measure,** even though **one that takes small values for powerful individuals rather than large ones.**

Clustering Coefficient

Demo using NetworkX

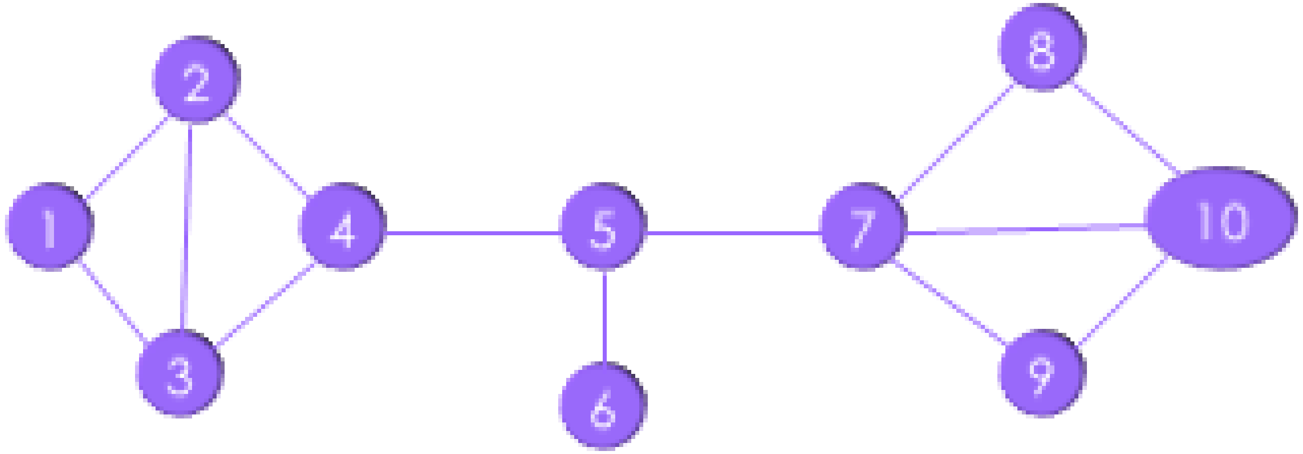
- AvgPathLength_Diameter_CC_Measures.py

Demo using Gephi

Clustering Coefficient

Exercise:

1. Find overall clustering coefficient C(g) and individual clustering coefficients for nodes 4 and 10 in the following network.



2. What is Triadic Closure? Give reasons for Triadic Closure.

Clustering Coefficient - Assignment

1. How do new friend suggestions appear on the notifications in Facebook?

2. Facebook's 'People You May Know' feature can be really creepy. How does it work?

Additional reading:

1. http://www.whatafuture.com/facebook-friend-suggestion-algorithm/

Link Prediction

- **Link Prediction is a process of recommending missing edges in a social network graph**.

- The networks are evolving over time, new users are joining, adding friends, new connections between old users, etc. Based on the current network we want to be able to predict the upcoming changes in the network and make recommendations accordingly.

- Social media sites such as Facebook provides a snapshot of its social network at time(say 't') and based on it, **we need to predict the future possible links**.

References

- Social Network Analysis: **Lada Adamic,** University of Michigan.
- Analyzing the Social Web- Jennifer Golbeck , Morgan Kaufmann
- Social Media Mining - Reza Zafarani
- Wikipedia – Current Literature

PES UNIVERSITY ONLINE

SOCIAL NETWORK ANALYTICS

Properties of Networks

Degree Distribution

Prakash C O

Department of Computer Science and Engineering

SOCIAL NETWORK ANALYTICS

Properties of Networks

Degree Distribution

Prakash C O

Department of Computer Science and Engineering

Degree Distribution

- To create a degree distribution, calculate the degree for each node in the network. Table 3.3 shows the degrees for each node in the graph shown in Figure 3.1.

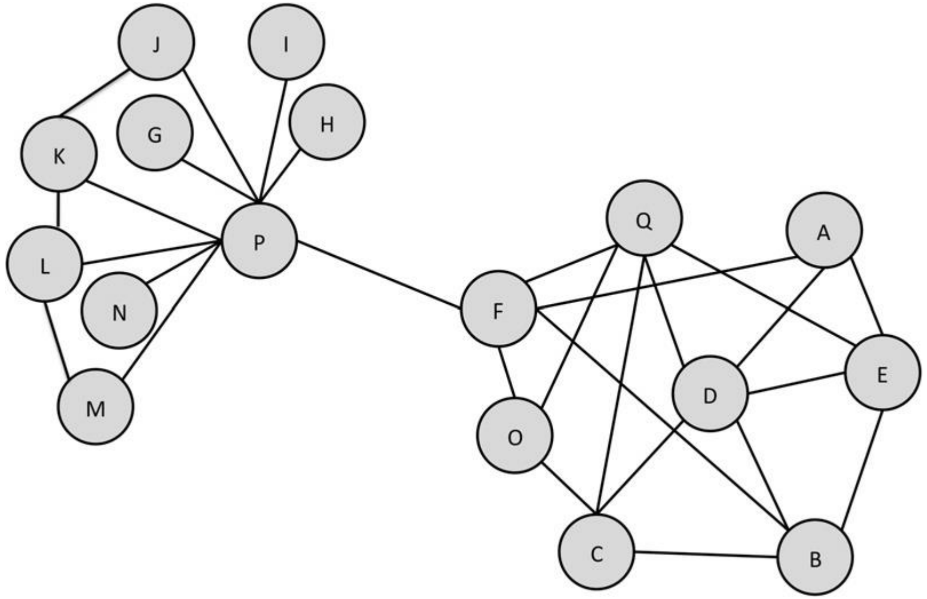


FIGURE 3.1 A sample undirected network.

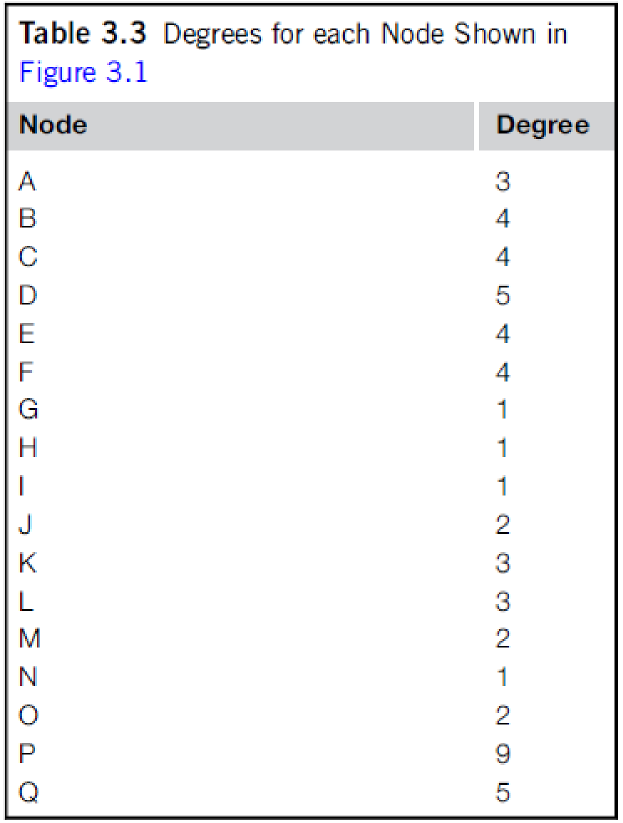


Table 3.3 Degrees for each Node Shown in Figure 3.1

Node	Degree
A	3
B	4
C	4
D	5
E	4
F	4
G	1
H	1
I	1
J	2
K	3
L	3
M	2
N	1
O	2
P	9
Q	5

Degree Distribution

- The next step is to count how many nodes have each degree. This is totaled for each degree, including those for which there are no nodes with that count.
- Table 3.4 shows the node count for each degree in this network.

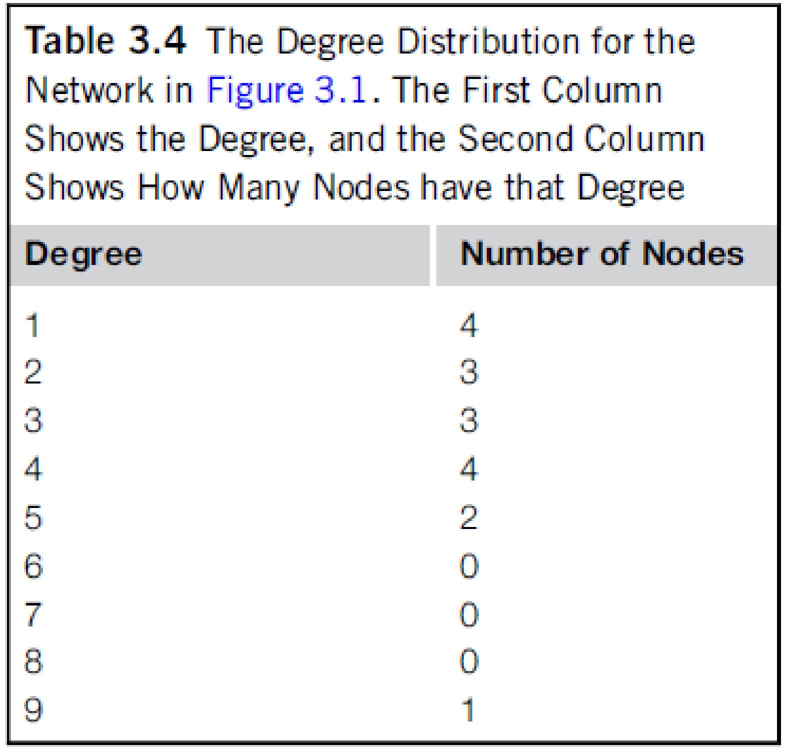


Table 3.4 The Degree Distribution for the Network in Figure 3.1. The First Column Shows the Degree, and the Second Column Shows How Many Nodes have that Degree

Degree	Number of Nodes
1	4
2	3
3	3
4	4
5	2
6	0
7	0
8	0
9	1

Degree Distribution

- The most common way to show a degree distribution is in a bar graph.
- The x-axis has the degrees in ascending order, and the Y-axis indicates how many nodes have a given-degree.
- For the data in Table 3.4, we would make a bar graph as shown in Figure 3.5.

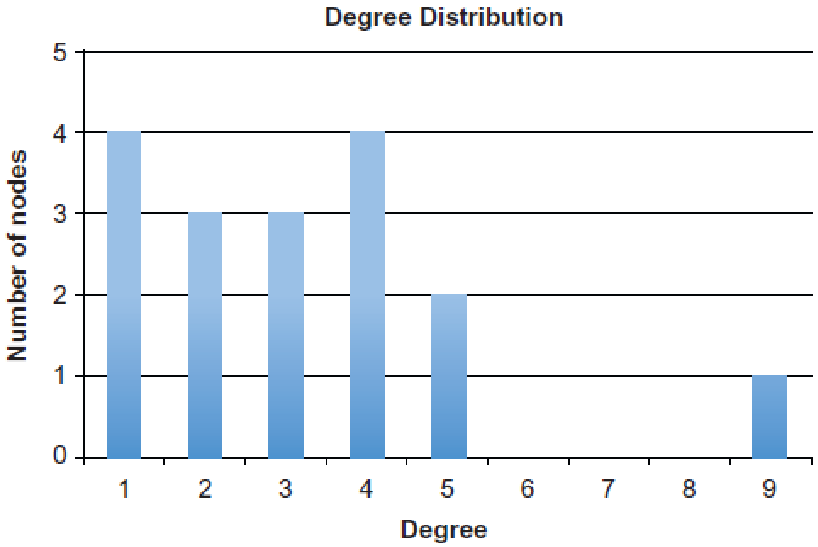


FIGURE 3.5

The degree distribution for the graph shown in Figure 3.1.

Degree Distribution

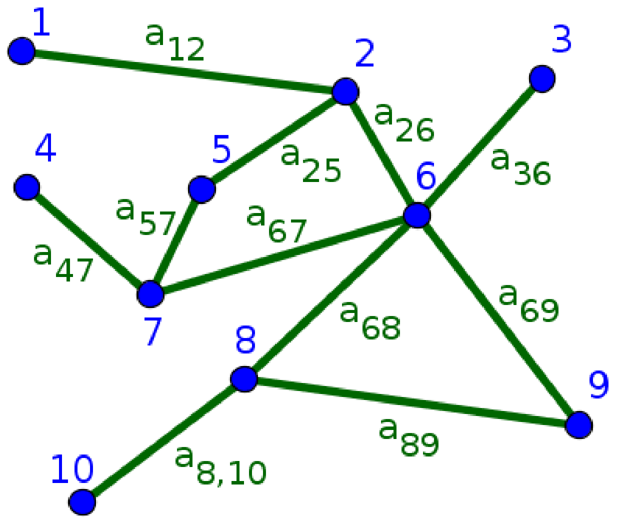
- In the study of graphs and networks,
  - the **degree of a node** in a network is **the number of connections it has to other nodes** and
  - the **degree distribution** is **the probability distribution of these degrees over the whole network.**

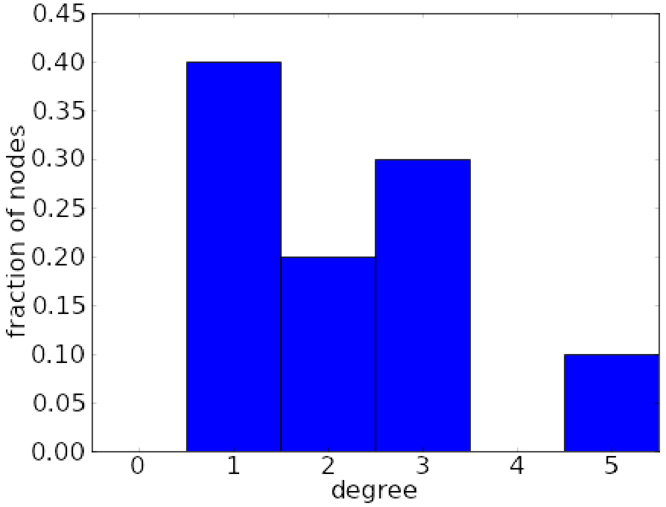
- By counting how many nodes have each degree, we form the degree distribution $P_{deg}(k)$, defined by

  $P_{deg}(k)$ = fraction of nodes in the graph with degree k.

Degree Distribution

Example:





- For this undirected network, the degrees are k_1=1, k_2=3, k_3=1, k_4=1, k_5=2, k_6=5, k_7=3, k_8=3, k_9=2, and k_{10}=1.
- Its degree distribution is $P_{deg}(1)$=2/5, $P_{deg}(2)$=1/5, $P_{deg}(3)$=3/10, $P_{deg}(5)$=1/10, and all other $P_{deg}(k)$=0.

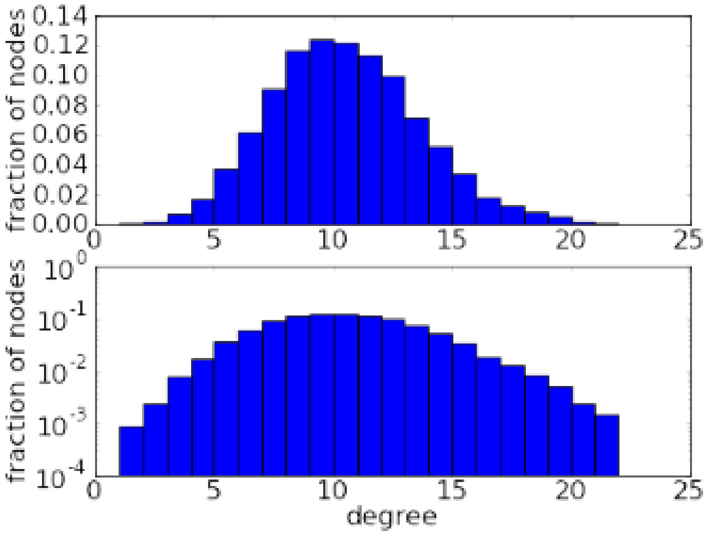
Degree Distribution

- The degree distribution clearly captures only a small amount of information about a network. But that information still gives important clues into structure of a network.

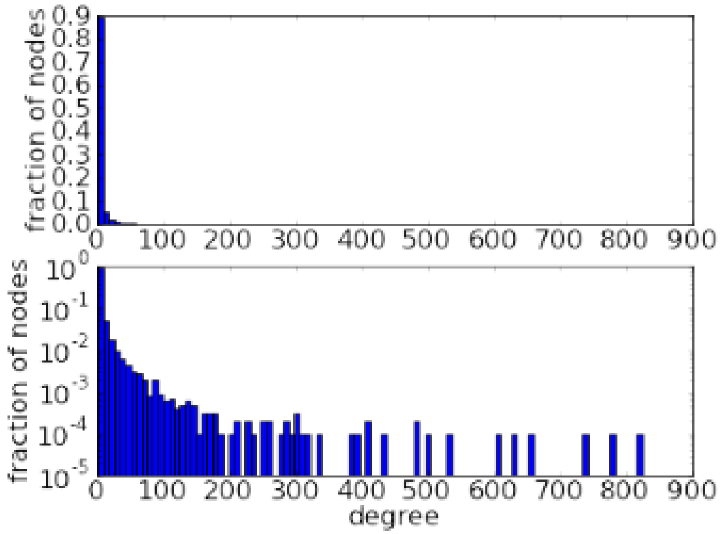
- In a real world network, most nodes have a relatively small degree, but a few nodes will have very large degree, being connected to many other nodes.

- In networks, **large-degree nodes are often referred to as *hubs***, in analogy to transportation network such as one connecting airports, where some very large hub airport have connections to many others.

Degree Distribution



A binomial degree distribution of a network with 10,000 nodes and average degree of 10. The top histogram is on a linear scale while the bottom shows the same data on a log scale.



A power law degree distribution of a network with 10,000 nodes and average degree of around 7. The top histogram is on a linear scale while the bottom shows the same data on a log scale.

References

- Social Network Analysis: **Lada Adamic,** University of Michigan.
- Analyzing the Social Web- Jennifer Golbeck , Morgan Kaufmann
- Social Media Mining - Reza Zafarani
- Wikipedia – Current Literature

THANK YOU

Prakash C O

Department of Computer Science and Engineering

coprakasha@pes.edu

+91 98 8059 1946

Properties of real world networks

- While a **small network** can be visualized directly by its graph (N, g), **larger networks** can be more difficult to visualize and describe.

  Therefore, we define a set of **quantitative performance measures to study complex networks**:

  - Diameter and average path length
  - Clustering Coefficient
  - Degree distributions

Clustering Coefficient

Average Clustering Coefficient

- Average - all n vertices individual Clustering Coefficients

$$C = \frac{1}{n} \sum_i C_i$$