ST5225: Statistical Analysis of Networks
Lecture 3: Descriptive Statistics

WANG Wanjie
staww@nus.edu.sg

Department of Statistics and Applied Probability
National University of Singapore (NUS)

Sunday 28$^{\text{th}}$ January, 2018

Outline

- Review
- Degree Distribution
- Centrality
  - Closeness
  - Betweenness
  - Eigenvector
- Cohesion
  - Cliques, k-cores
  - Connectivity
  - Local Density

Relevant Chapter

Statistical Analysis of Network Data, Chapter 4.1– 4.3

Review

- Graph Sampling
  - Induced-subgraph sampling
  - Incident-subgraph sampling
  - Snowball sampling
  - Respondent-driven sampling
  - Trace route sampling
- Bias of estimates
  - Horvitz-Thompson estimator
- Degree

Degree

Recall:

- Degree of a node i:

$$d_i = \text{the number of edges incident on the node } i$$

- For directed graphs,

  $d_i^{in} = \#\text{edges pointing in towards } i, \quad d_i^{out} = \#\text{edges pointing out from } i.$

  We use $d_i^{tot} = d_i^{in} + d_i^{out}$ to denote the number of all the edges incident to i for directed graphs.

- Degree Sequence/Degree Vector: A vector containing the degrees of each node

If we look into the degree statistic...

- Degree of each node is a good summary statistic, which is also called *degree centrality*
- For each node, it shows how important this node is.
- For the whole network, we are more interested in the *degree distribution*

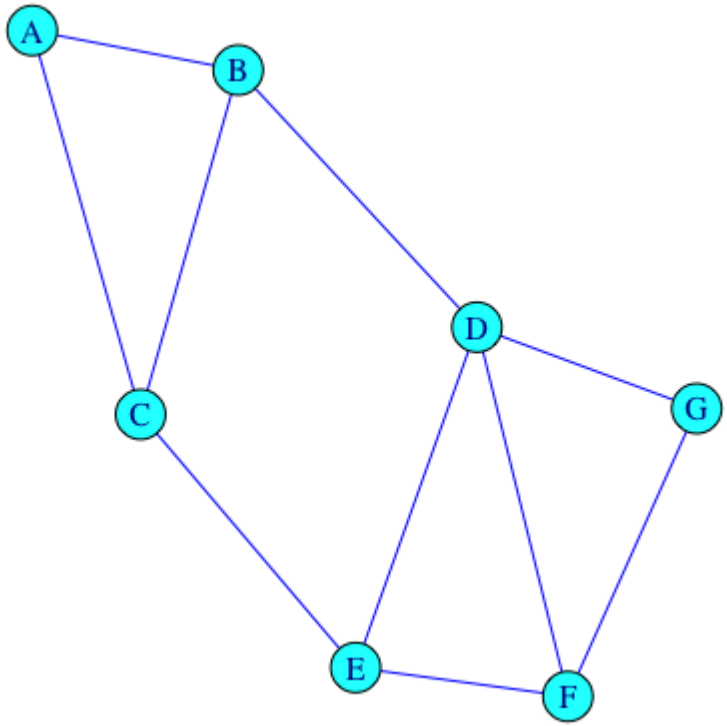
Degree Distribution

Given a network graph $G = (V, E)$, define f_d to be the fraction of vertices $v \in V$ with degree $d_v = d$. The collection $\{f_d\}_{d \geq 0}$ is called the *degree distribution* of G, which is simply the distribution from the degree vector.

- Just as other dist., we can learn its mean, median, standard deviation, quantiles, etc.
- The shape of the distribution also give some information
- If we have the population network, the degree dist. is the empirical distribution from the network. If we only have a sampling network, the degree dist. needs to be estimated.

Degree Distribution (Examples)

A toy example

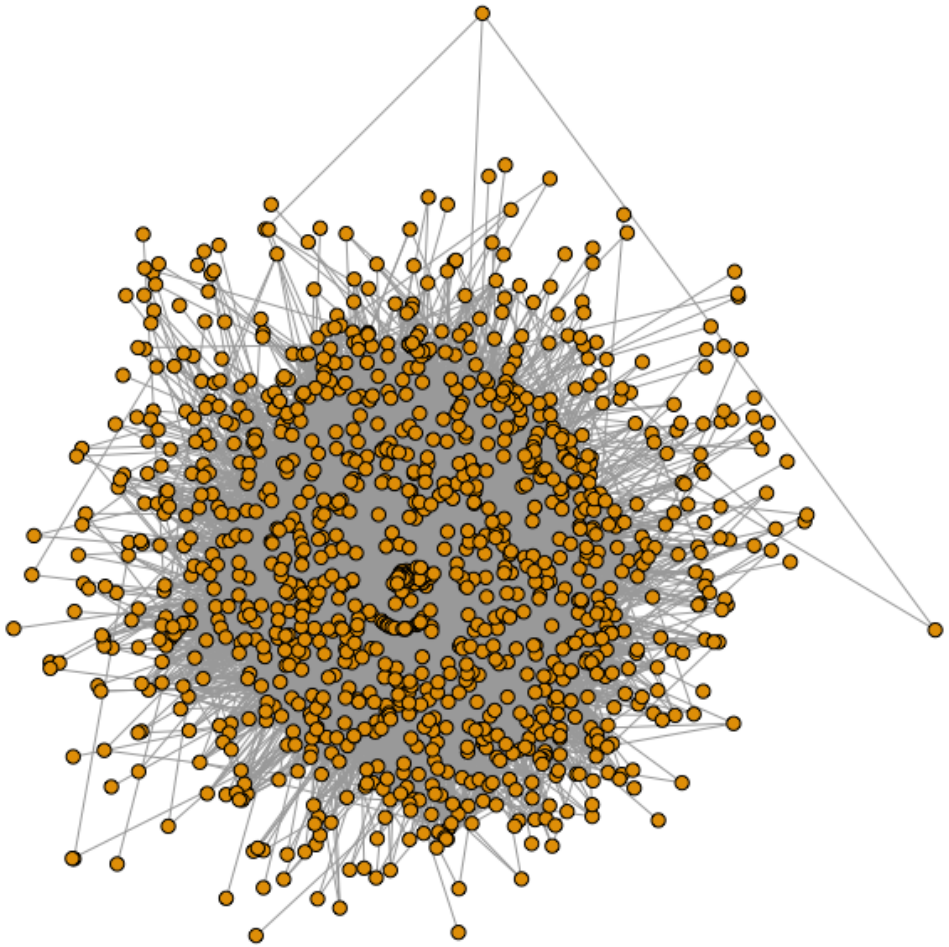


For this graph, the degree sequence is $(2, 3, 3, 4, 3, 3, 2)$. Therefore, the degree distribution is

$$f_2 = 2/7, \quad f_3 = 4/7, \quad f_4 = 1/7.$$

Degree Distribution (Examples)

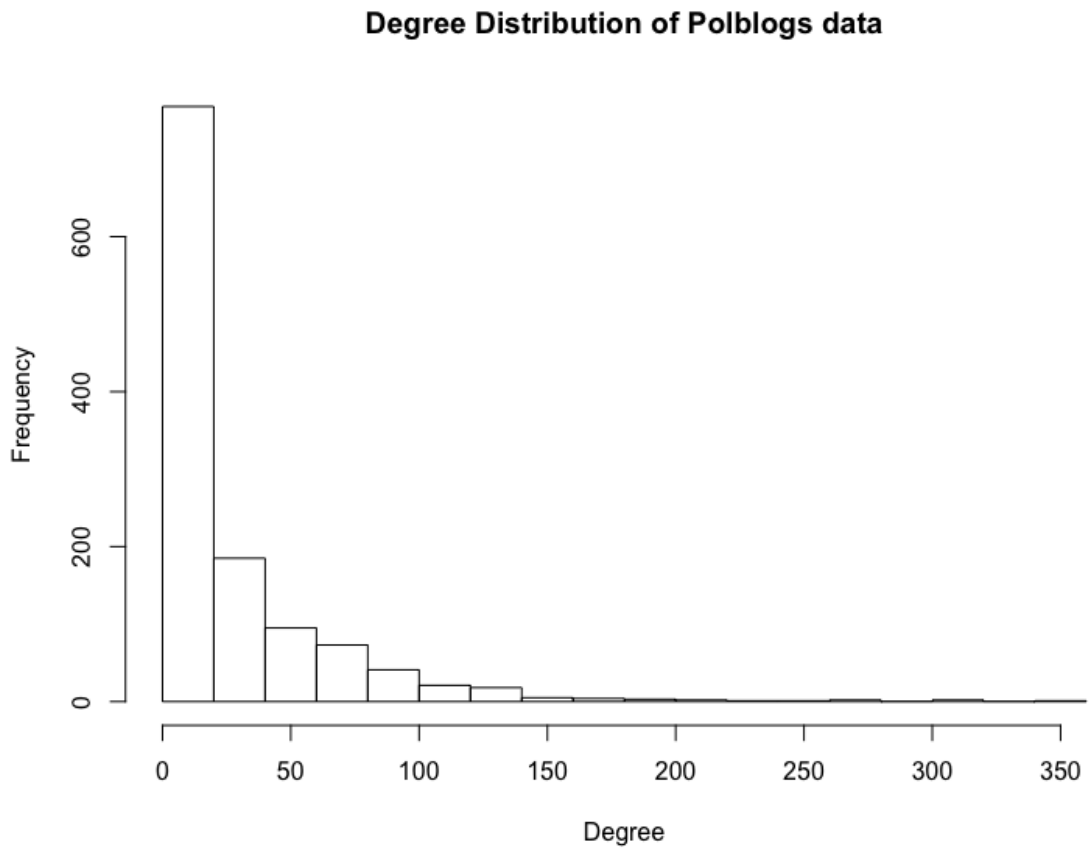
Politics Blogs Data: Blogs are labeled according to the political stand: liberal (0) or conservative (1). Links between blogs were automatically extracted from a crawl of the front page of the blog.



Adamic and Glance (2005), "The political blogosphere and the 2004 US Election"

Degree Distribution (Examples)

Degree Distribution for Politics Blogs Data:



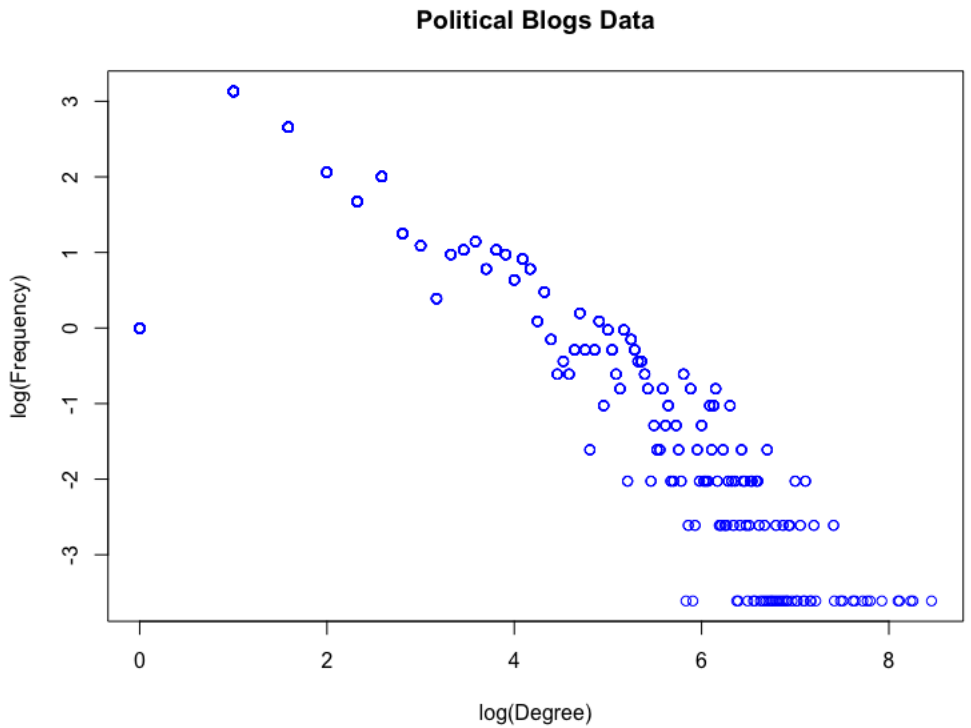
- It is right-skewed and heavy-tailed.
- It's quite normal for real data sets (but not required!!!). If not so, please double check your data.
- We need a better scale.

Degree Distribution (Examples)

Log-log scale of Degree Distribution:

- y-axis: $\log_2$(Frequency)
- x-axis: $\log_2$(Degree)

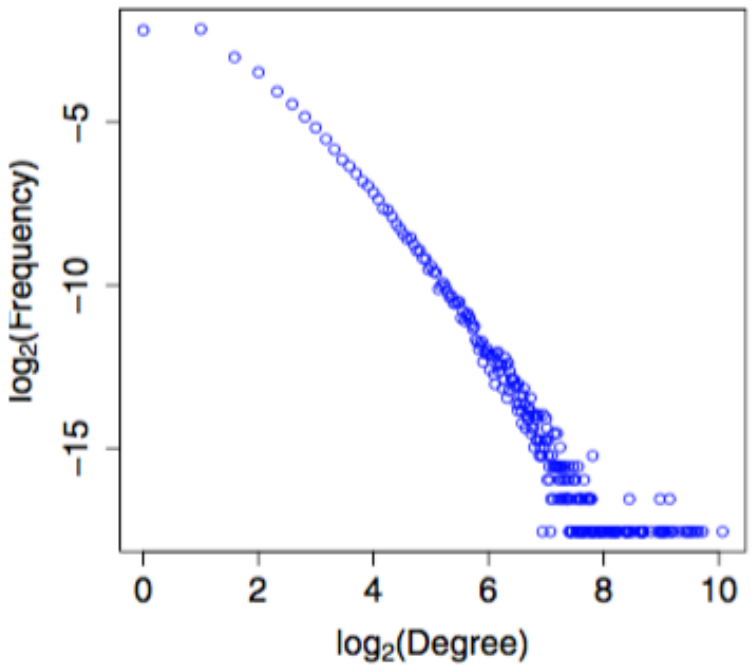
Political Blogs Data:

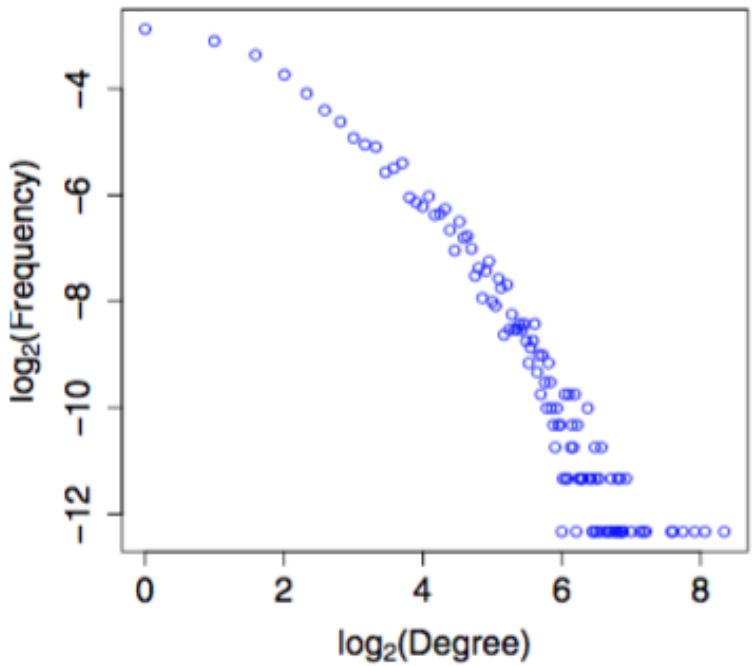


Remark. Very close to a *linear* relationship with *negative* coefficient!

Degree Distribution (Examples)

Two more examples from Textbook (Page 82):





Remark.

- Still quite *linear* !
- Say that $\log_2(f_d) \approx -\alpha \log_2(d) + C = \log_2(2^C d^{-\alpha})$, so

$$f_d \propto d^{-\alpha} \text{ Power-law degree distributions}$$

Degree Distribution

An approximation of degree distribution:

$$f_d \propto d^{-\alpha}$$

- The decreasing speed is $\log f_d \approx -\alpha \log(d)$. Compare to other dist:
  - Gaussian dist. $d \sim N(0, \sigma^2)$:

$$f_d = C \exp^{-d^2/2\sigma^2} \Longrightarrow \log f_d \approx -Cd^2 < -\alpha \log(d)$$

  - Exponential dist. $d \sim Exp(\lambda)$:

$$f_d = \lambda e^{-\lambda d} \Longrightarrow \log f_d \approx -Cd < -\alpha \log(d)$$

  - In all, f_d decreases slower than the exponential tail and Gaussian tail, which means heavy-tail
- The parameter α is an important quantity to evaluate the network.
  - Larger α means faster decreasing, which means a network with fewer "Hot nodes"

How to figure out α?

- Fitting directly with, say, least squares regression.
  Problem. The noise at the high degrees will cause much trouble
- Fitting α with cumulative densities $F(d) = P(\text{degree} \leq d)$. Given the definition, the probability for nodes with high degrees won't be affected much by the noise.
  The tail probabilities have the form

$$1 - F(d) \sim d^{-(\alpha-1)}.$$

  Consider a linear regression to estimate α.
- Instead of linear regression, use the estimates in other forms.

Recall:

- *Distance between two nodes* i, j:

$$d(i,j) = \text{the length of the shortest path between } i \text{ and } j.$$

- If i and j are unconnected, define $d(i,j) = \infty$

New:

- Average distance between nodes (in the same component)

$$\bar{d} = \frac{1}{n(n-1)} \sum_{i,j} d(i,j).$$

- Diameter of a graph.

Diameter

Given a graph $G = (V, E)$, the *diameter* of G is defined as

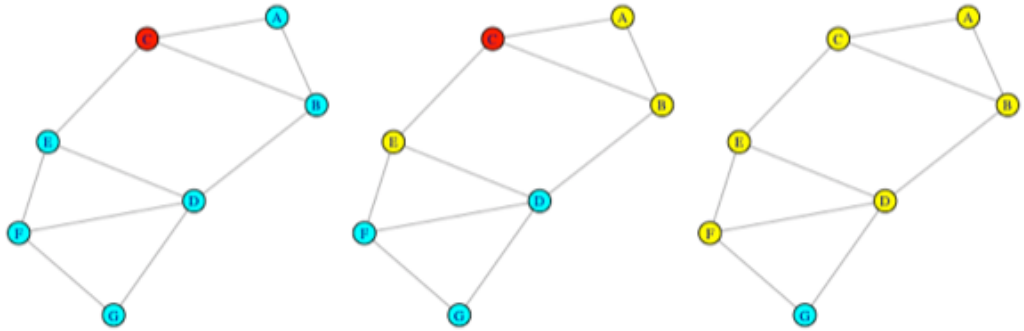
$$diam(G) = \max_{i,j} \min d(i,j),$$

which is the *maximum* geodesic distance between two nodes.

Diameter

Typically, the diameter of a graph is low. The intuition is as follows:

- Given an arbitrary starting node i, let $N(i, j)$ be the number of nodes which are reachable in a j-step path.



- Let $\bar{d}$ be the average degree $\Rightarrow \bar{d}$ is the average of number of neighbors of a node.
- $N(i, 1) \approx \bar{d}$, $N(i, 2) \approx \bar{d}(\bar{d} - 1)$, $\cdots$, $N(i, r) \approx \bar{d}(\bar{d} - 1)^{r-1} \approx \bar{d}^r$.
- Assume at step r, all the n nodes are covered, where $n \approx \bar{d}^r$. Then $r \approx Diam(G)$.
- Therefore, $Diam(G) \approx \log(n)/\log(\bar{d})$, which is at the rate of $\log(n)$.
- $\log(n)$ is quite small compared to n.

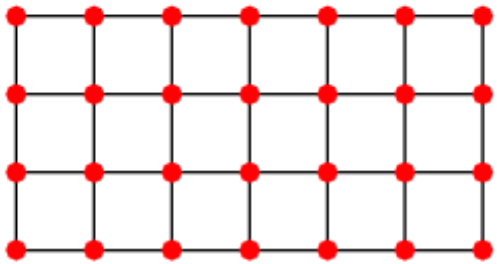
Diameter

- For most real data networks, the diameter is small.
  - A famous example is *Six-degree separation (small-world network)*. Assuming the average degree (connection one person has) is 40, then the diameter is approximately

$$\frac{\log(\text{whole world population})}{\log(\bar{d})} = \frac{\log(3,490,333,715)}{\log(40)} = 5.96.$$

- However, there are counter-examples
  - Recall the two-dimensional lattice network:

This network has 28 nodes, with average degree approximately 3. However, the diameter is 9, which is much larger than $\log(28)/\log(3) = 3$.



  - Usually, a p-dim lattices have diameters at $O(n^{1/p})$
- The real-world network can be thought of as in a *high-dimensional* space

Summary

When you have a real data network,

- Check whether this network is simple, connected, directed/undirected
- Get a summary of the number of nodes, edges, and the corresponding properties
- Examine the degree distribution. Have a plot of the degree distribution, and get the corresponding stats
- Take a look at the diameter of the graph.

All of these show the properties of the whole graph. However, sometimes we are interested in the role of *a node* in the graph.

Centrality

- We are interested in how *important* a node is in the network.
  - Airport network: which airport is the most important?
  - Facebook network: which person is most welcomed?
  - We call it as "*centrality*"
- Obviously, degree is one measurement for the centrality of nodes
- Sometimes we are more interested in the topological properties:
  - If we remove the node, some nodes lose connection with the other nodes, or
  - If we remove the node, it is more difficult for some nodes to connect with others
- We need other kind of *centrality*
- Unlike other statistical notions, there are various measurement for the *centrality* of a node, and many of them are widely used.

Closeness Centrality

Given a graph $G = (V, E)$ with $|V| = n$. Let $d(i, j)$ denote the distance between two nodes i and j. The *closeness centrality* of node i is defined as

$$C(i) = \frac{1}{\frac{1}{n-1}\sum_{j \neq i} d(i, j)},$$

where the denominator is the average distance between node i and all the other nodes in the network.

- If node i is close to the other nodes, then the average distance is small, and $C(i)$ is large; if node i is far from the other nodes, then $C(i)$ is small.
- If there are multiple components, there are two solutions
  - Use only the component of node i (and also the relative number of nodes n in this component)
  - Use the harmonic mean, which is

$$C'(i) = \frac{1}{n-1} \sum_{j \neq i} \frac{1}{d(i, j)}.$$

Closeness Centrality: Example

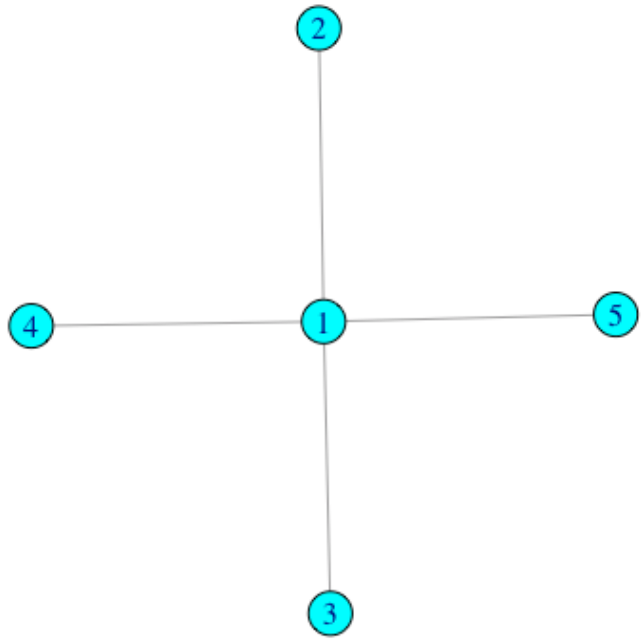
- For this graph, $n = 5$,
  $d(1, j) = 1$, $j = 2, 3, 4, 5$.
  $d(i, j) = 2$, $i \neq 1$ and $j \neq 1$.
- The average distance for node 1 is

$$\frac{1}{4} \sum_{j=2,3,4,5} d(1, j) = 1.$$

  The centrality for node 1 is
  $C(1) = \frac{1}{1} = 1$.
- The average distance for node 2 is

$$\frac{1}{4}[d(2, 1) + \sum_{j=3,4,5} d(2, j)] = 7/$$

  The centrality for node 2 is
  $C(2) = \frac{1}{7/4} = 4/7$.

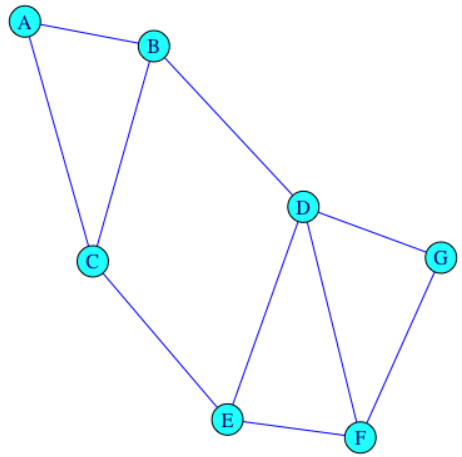


Shortest Path

Given a graph $G = (V, E)$. For two nodes i and j, the *shortest path* is a path $g(i, j)$, where the sum of the weights of its constituent edges is minimized. We also denote the shortest path between i and j which passes the node k as $g(i, j|k)$.

- $d(i, j)$ is the length of the shortest path between i and j.
- It is impossible for loops in the shortest paths.
- There can be multiple shortest paths between a pair of nodes. It is not unique!

For the node B and E, the paths $B--C--E$ and $B--D--E$ both the shortest paths.

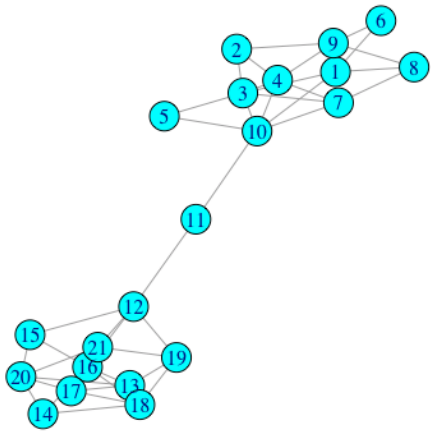


Betweenness Centrality

Given a graph $G = (V, E)$. Let $\sigma(u, v)$ denote the number of shortest paths between u and v, and $\sigma(u, v|i)$ denote the number of shortest paths between u and v which go through i. The *betweenness centrality* of node i is defined as

$$B(i) = \sum_{(u,v) \in E \times E, u \neq i, v \neq i, u \neq v} \frac{\sigma(u, v)}{\sigma(u, v|i)}.$$

- The betweenness of a node i is a measure of *how many shortest paths go through it.*
- An example of *high betweenness, low degree*



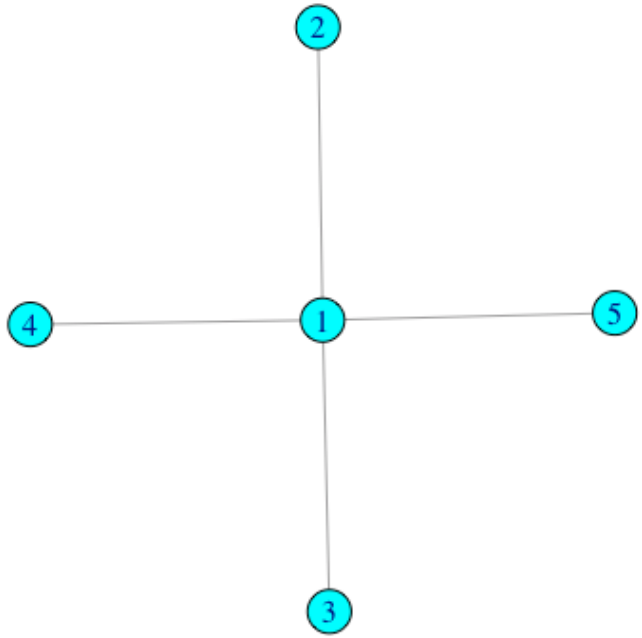
Betweenness Centrality: Example

- For this graph, $n = 5$,
  Every shortest paths go through node 1.
  No shortest path go through node 2, 3, 4, 5.
- The betweenness centrality for node 1 is

$$B(1) = \sum_{(u,v) \in E \times E} 1 = 6.$$

- The average distance for node 2 is

$$B(2) = \sum_{(u,v) \in E \times E} 0 = 0.$$



Eigenvector Centrality

- If you know more people who are "centers" of the network, then you are more likely to be a "center"
- Define the "importance" of a node according to its neighbors. Hope it satisfies

$$Cv(i) = \sum_{j \in \{\text{neighbors of } i\}} v(j)$$

- Recall the adjacency matrix A, where $A_{ij} = 1$ if there is $j \in \{\text{neighbors of } i\}$, and $A_{ij} = 0$ if not. So we rewrite the formula as

$$Cv(i) = \sum_j A_{ij} v(j).$$

Obviously, this indicates an eigenvector of A.

Eigenvector Centrality

Rewrite the formula as

$$v = \alpha A v.$$

According to The Perron-Frobenius theorem, when A has non-negative entries:

- The largest eigenvalue is positive
- The eigenvector corresponding to the largest eigenvalue is non-negative
- There is one such eigenvector for each connected component.

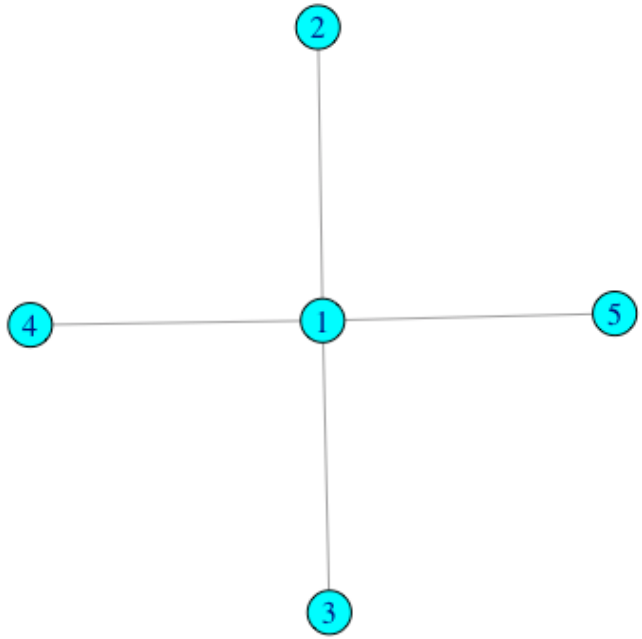
- The eigenvector is called the *eigenvector centrality* for the nodes.
- Variates of it is largely used in search engines (google...)

Eigenvector Centrality

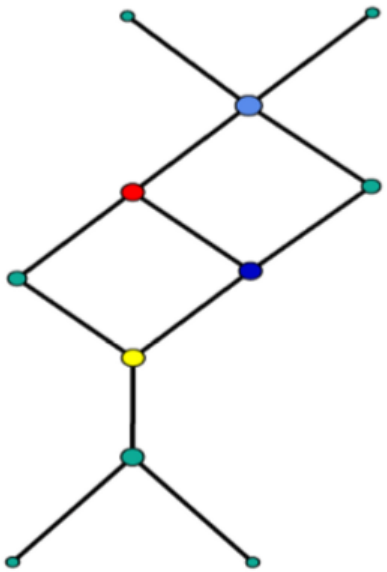
The adjacency matrix for this graph is

$$A = \begin{pmatrix} 0 & 1 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \end{pmatrix}$$

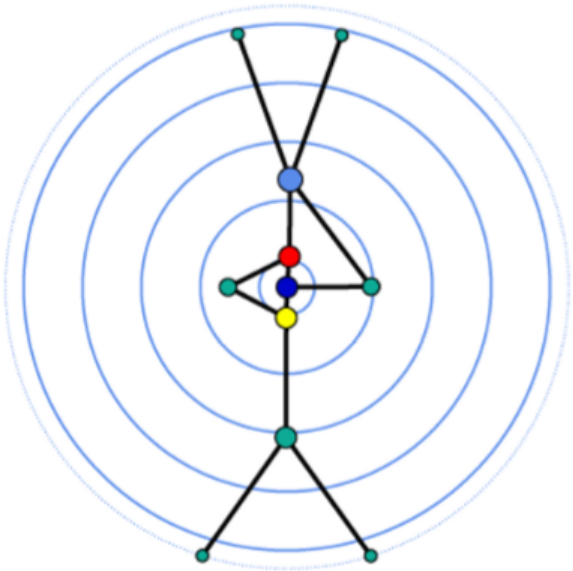
The eigenvector is $(1, 0.5, 0.5, 0.5, 0.5)^T$



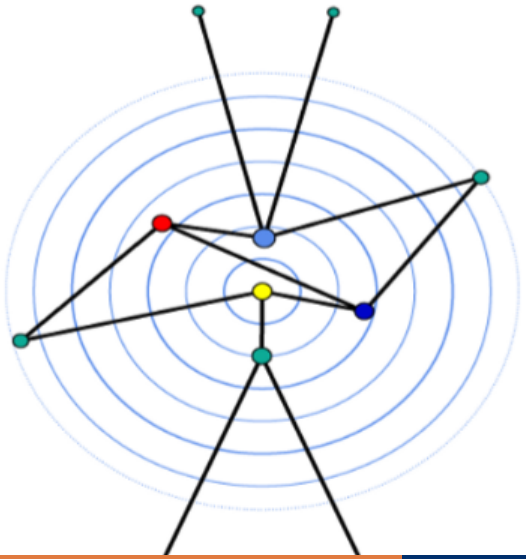
3 Types of Centrality

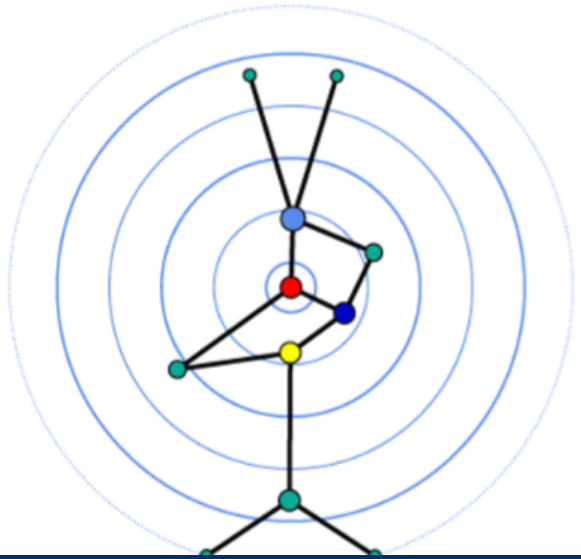


(a)



(b)





Additional Topics

- For directed graphs,
  - the betweenness centrality and closeness centrality can be defined in the same way
  - the eigenvector centrality can be adjusted in two ways: find the eigenvector of $M_{hub} = AA^T$, or of $M_{auth} = A^T A$. It is called the "Hubs and Authorities" algorithms.

- All the notions can be generalized to the *edge centrality*.
  - Generate the dual graph of G, say $G' = (V', E')$, where each node in V' is an edge in G, and each edge in E' is a node in G
  - The edge centrality for E is defined as the vertex centrality for V', correspondingly.

Network Cohesion

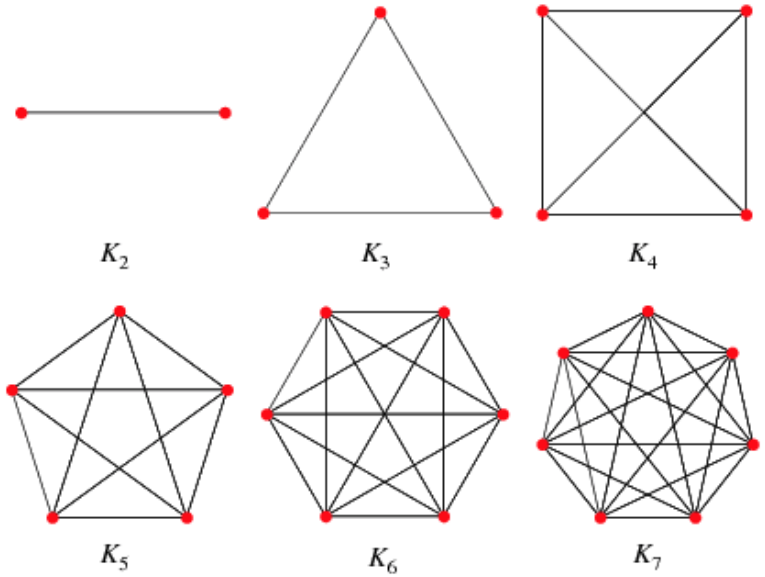
- Interested in *a subset of nodes*: whether these nodes are cohesive
- Interested in the whole network: *How to evaluate the cohesive parts in this network?*
- Examples:
  - If both B and C are friends of A, are B and C friends?
  - Does the structure of the internet pages tend to separate, with respect to distinct types of content?

- We need some measures for the network cohesion. Again, to describe it, there are multiple measurements, including
  - Densities (cliques, cores, local density)
  - Hierarchical structure (clusters)

Cliques

Recall:

- Cliques: A subgraph which is complete
  - Cliques are *fully cohesive*. Distance between each two nodes is 1.
- Examples of Cliques:

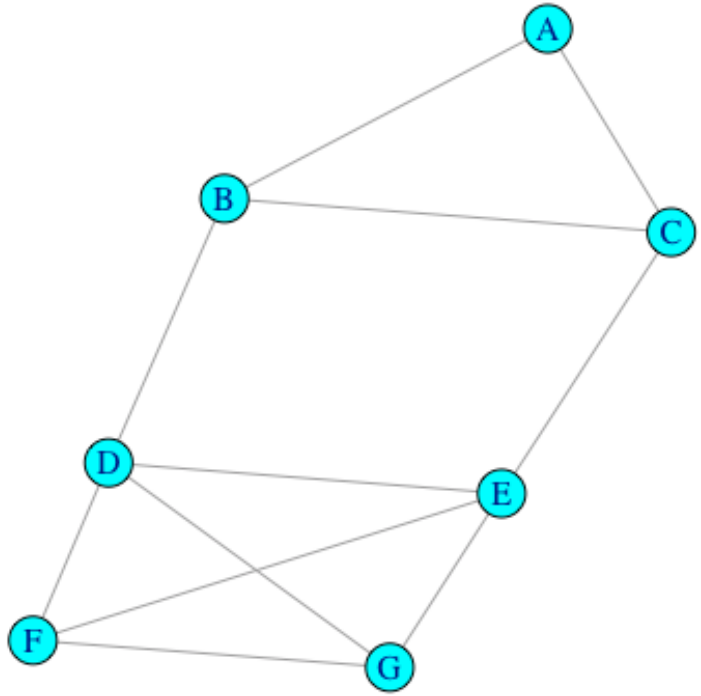


- If $H \subset G$ is a clique, then any induced subgraph of H is a clique.
- Maximal clique: a clique that no larger clique contains it

Cliques

What are the maximal cliques in the following network?
Solution: Find the maximal clique for each node

- For A, the maximal clique containing A is the subgraph formed by $\{A, B, C\}$
- For B, there are two maximal cliques, the one formed by $\{A, B, C\}$ and the one formed by $\{B, D\}$
- For C, there are also two maximal cliques, $\{A, B, C\}$ and $\{C, E\}$
- For D, the one formed by $\{B, D\}$ and the one formed by $\{D, E, F, G\}$
- For E, similarly we have $\{C, E\}$ and $\{D, E, F, G\}$
- For F and G, the maximal clique is only the one formed by $\{D, E, F, G\}$.



Cliques

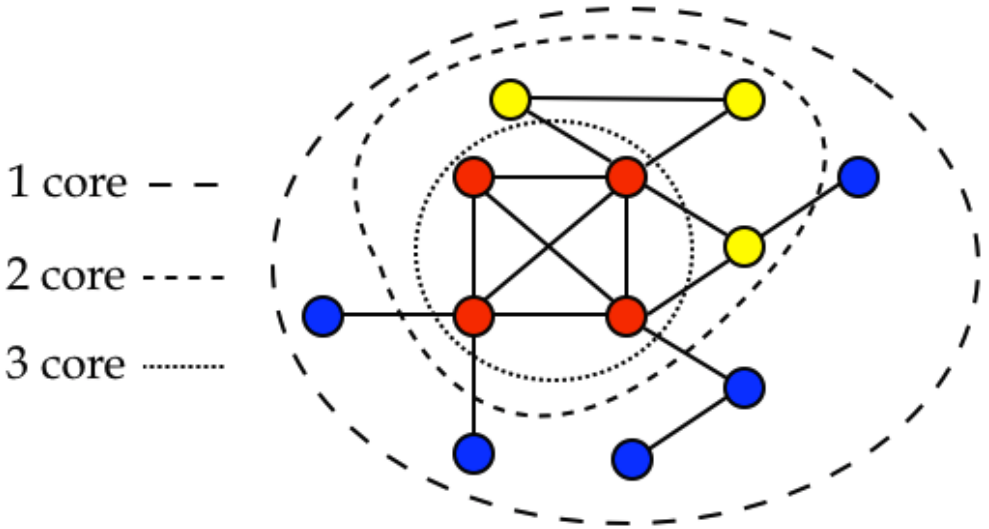
- The nodes in the same clique can be seen as a small "community". They connect with each other in this community.
- If a network has a large clique, then this network is more "cohesive"
- If there are a lot of edges ($|E| > (|V|^2/2)\frac{n-2}{n-1}$), there must be a clique (with size n).
- Computation cost:
  - Given a clique, whether it is maximal or not can be done in $O(|V| + |E|)$ time
  - Whether a graph has a maximal clique of at least size n is NP-complete (nondeterministic polynomial time)
  - NP-complete: e.g. the computation cost is $2^{|V|}$. For large networks where $|V|$ and $|E|$ are large, it is impossible to realize

- The clique requires connection between every pair of nodes – very strict
- Relaxation: every node has a high degree.

k-core

A subset of nodes is called a *k-core*, if in the induced subgraph, all nodes have degree at least k.

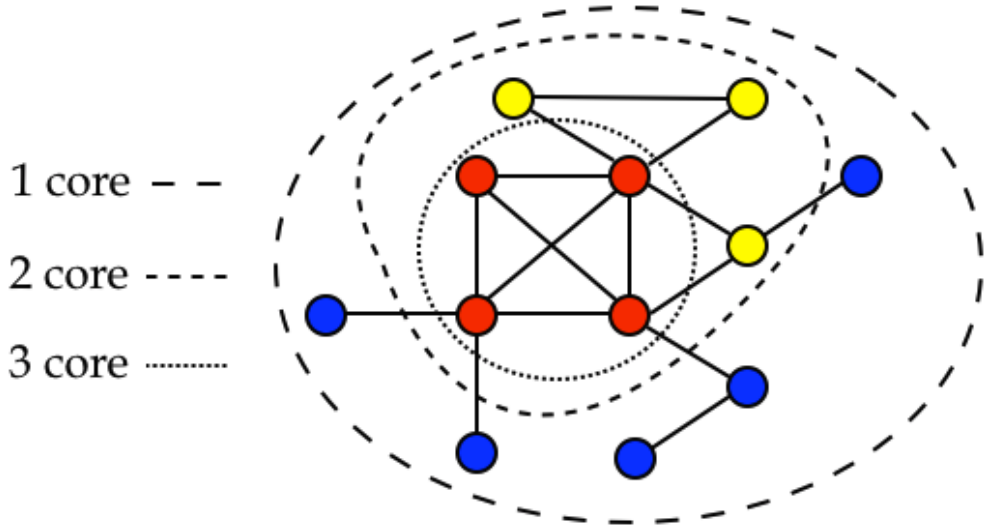
- Example[1]:



[1] https://chaoslikehome.wordpress.com/tag/k-core/

Maximal k-core

The *(maximal) k-core* is a k-core that cannot be enlarged to a larger k-core.



- If a node is in the k-core, it is also in the $k-1$-core.
- We can also define the *coreness of a vertex i*:

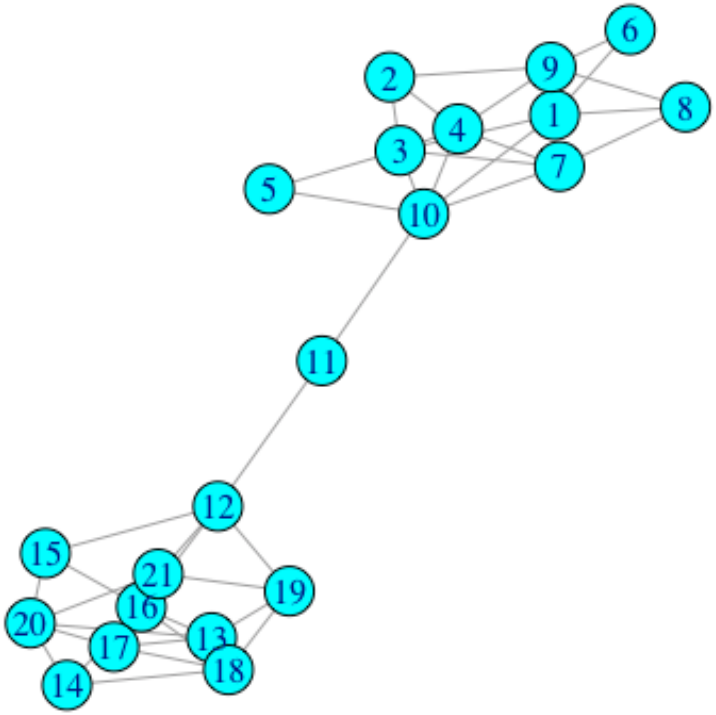
$$\text{Coreness}(i) = \max\{k : i \in \ k\text{-core } \}$$

- k-core can be found by deleting the node with smallest degrees

Connectivity

Recall:

- The network can be decomposed into several connected components
- The nodes in each component are connected. Nodes in different components are disconnected.
- It is possible that if several nodes/edges are deleted, the originally connected components are separated



Vertex-Connectivity

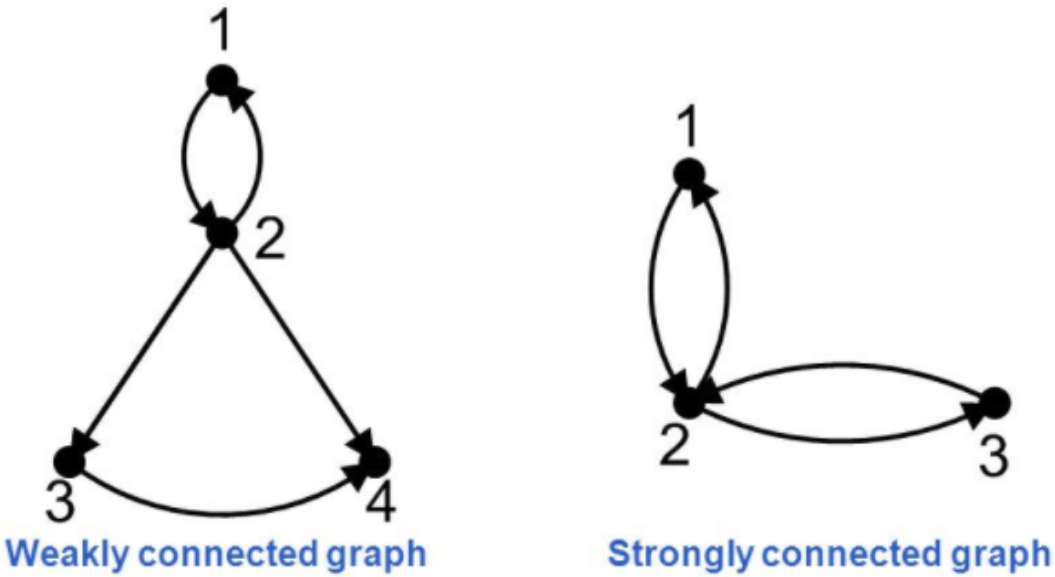
A graph G is called *k-vertex-connected* if

1. the number of nodes $|V| > k$;
2. the removal of any subset of vertices $X \subset V$ of cardinality $|X| < k$ leaves a subgraph $G - X$ that is connected.

- The smallest number of nodes you have to remove to disconnect the graph
- The previous graph is 1-vertex-connected
- $k < \min_i d(i)$, otherwise you can isolate the node with smallest degree
- Similarly, we can define *k-edge-connected* graphs.
  1. the number of edges $|E| > k$;
  2. the removal of any subset of edges $X \subset E$ of cardinality $|X| < k$ leaves a subgraph $G - X$ that is connected.

Connectivity

- The connectivity can be expanded to the directed graphs straightforwardly.
  - *Weakly connected.* If the underlying graph (the undirected graph where the labels 'tail' and 'head' are removed from G) is connected, G is called *weakly connected.*
  - *Strongly connected.* If every node v is reachable from every other node u by a *directed* walk, G is called *strongly connected.*
- Example[2]



- Vertex/Edge-connectivity can be extended analogously.

[2] http://slideplayer.com/slide/2433835/

Connectivity

- For the directed graph, there is a new characterization, a 'bowtie'.

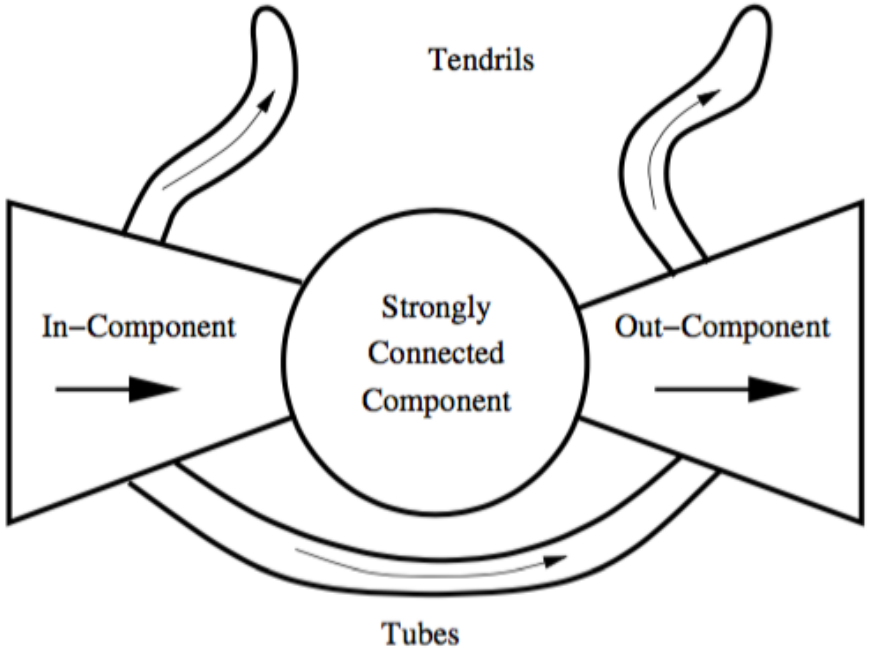


Fig. 4.5 'Bowtie' structure of a directed network graph. Adapted from Broder et al. [67].

- Strongly connected component (SCC)
- *in-component*, nodes can reach SCC, but cannot be reached from SCC
- *out-component*, nodes can be reached from the SCC but cannot reach SCC
- *tubes*, nodes between the in- and out- components, not SCC
- *tendrils*, nodes that can neither be reach nor be reached from the SCC

Connectivity

- The in- and out-components are in some sense 'upstream' and 'downstream' from the SCC
- First discovered when studying the WWW graph
  - Crawling on webpages, and record the hyperlinks in the pages
  - The hyperlinks are record as edges, while the webpages are the nodes
  - Several papers found this nature, and also studied on the reason
- Also found in other real data sets

Local Density

- For an undirected graph $G = (V, E)$, the possible number of edges is $\binom{|V|}{2} = \frac{|V|(|V|-1)}{2}$
- The density is the proportion of the truly observed edges

Local Density

For a graph G, the *density* of G is

$$den(G) = \frac{|E|}{|V|(|V|-1)/2}.$$

- Prob(two randomly picked nodes are connected by an edge)
- Recall the average degree $\bar{d}(G) = \frac{2|E|}{|V|}$, the density is just a rescaling of $\bar{d}(G)$ of G, where

$$den(G) = \frac{|E|}{|V|(|V|-1)/2} = \frac{\bar{d}(G)}{|V|-1}$$

- The definition can also be applied to an induced subgraph $H = (V_H, E_H) \subset G$, where

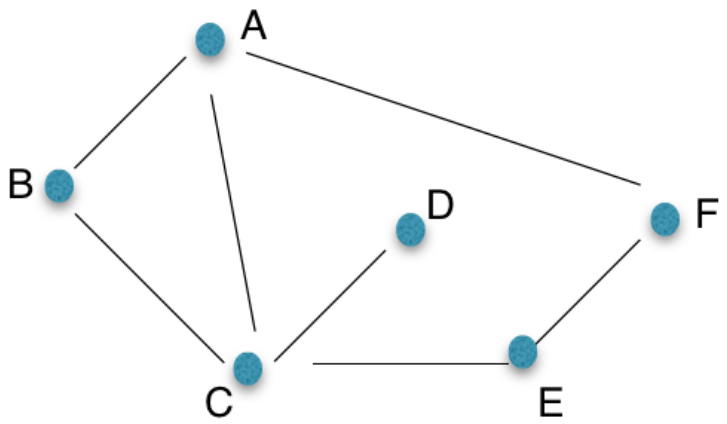
$$den(H) = \frac{|E_H|}{|V_H|(|V_H| - 1)/2}.$$

  It can be seen as the local density at subgraph H.

- Especially, for node v, take the subgraph $H = H_v$, which contains the nodes $\{v$ and neighbors of $v\}$ and the edges between them.
- Define the *density of node v* as

$$den(v) = den(H_v)$$

Example.



- Density of the whole graph: $\frac{7}{6*5/2} = 7/15 = 0.47$
- $Den(H_A) = \frac{4}{4*3/2} = 2/3$; $Den(H_B) = \frac{3}{3*2/2} = 1$;
  $Den(H_C) = \frac{5}{5*4/2} = 0.5$; $Den(H_F) = \frac{2}{3*2/2} = 2/3$.

- Obviously, $0 \leq den(G) \leq 1$, as it is a proportion
- $Den(G) = 1 \iff G$ is a clique

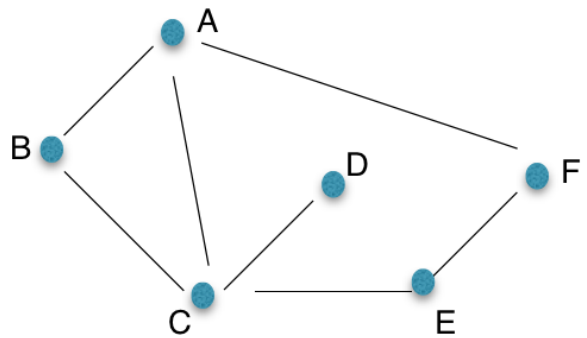
Local Density

- Call the 3-node complete graph as a *triangle*
- Call a 2-star graph as *connected triple*
- Note: The connected triple is one edge less than the triangle

- For a node v with $d(v) \geq 2$, define

$$cl(v) = \frac{\#\text{triangles } v \text{ falls into}}{\#\text{connected triples that both edges are incident to } v}$$

- Example.



$$cl(A) = 1/3, \quad cl(B) = 1, \quad cl(C) = 1/6, \quad cl(E) = 0, \quad cl(F) = 0$$

Local Density

- If a group contains more triangles, the group is more "dense"
- For a graph $G = (V, E)$, consider $V' \subset V$, which contains all the nodes with degrees ≥ 2. Each of these nodes have a
- Define the *clustering coefficient* for the graph as

$$cl(G) = \frac{1}{V'} \sum_{v \in V'} cl(v)$$

- However, this is not quite informative, so a weighted one is more generally used, which is

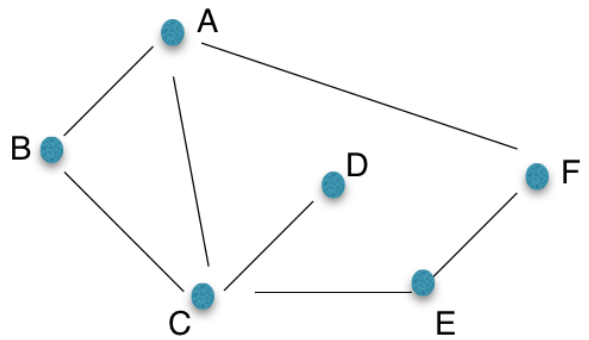
$$cl(G) = \frac{\sum_{v \in V'} \tau_3(v) cl(v)}{\sum_{v \in V'} \tau_3(v)} = \frac{3\tau_\triangle(G)}{\tau_3(G)},$$

where $\tau_3(v) = \#$ connected triples that two edges are incident with v; $\tau_3(G) = \#$ connected triples in G; and $\tau_\triangle(G) = \#$ triangles in G.

- Show the equality:

$$
\begin{aligned}
cl(G) &= \frac{\sum_{v \in V'} \tau_3(v) cl(v)}{\sum_{v \in V'} \tau_3(v)} \\
&= \frac{\sum_{v \in V'} \tau_3(v) \times \tau_\triangle(v)/\tau_3(v)}{\sum_{v \in V'} \tau_3(v)} \\
&= \frac{\sum_{v \in V'} \tau_\triangle(v)}{\tau_3(G)} \\
&= \frac{3\tau_\triangle(G)}{\tau_3(G)}
\end{aligned}
$$

Example

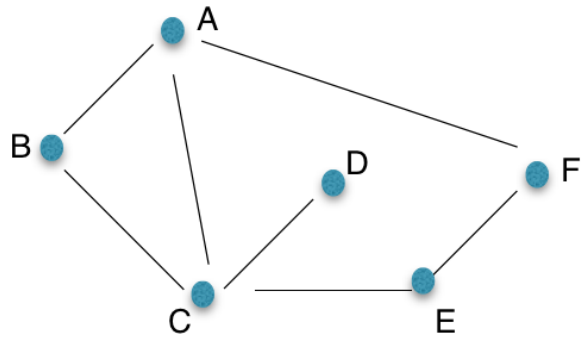


$\tau_3(G) = 12, \quad \tau_\triangle(G) = 1, \quad cl(G) = 0.25$

- Show the equality:

$$
\begin{aligned}
cl(G) &= \frac{\sum_{v \in V'} \tau_3(v) cl(v)}{\sum_{v \in V'} \tau_3(v)} \\
&= \frac{\sum_{v \in V'} \tau_3(v) \times \tau_\triangle(v)/\tau_3(v)}{\sum_{v \in V'} \tau_3(v)} \\
&= \frac{\sum_{v \in V'} \tau_\triangle(v)}{\tau_3(G)} \\
&= \frac{3\tau_\triangle(G)}{\tau_3(G)}
\end{aligned}
$$

Example



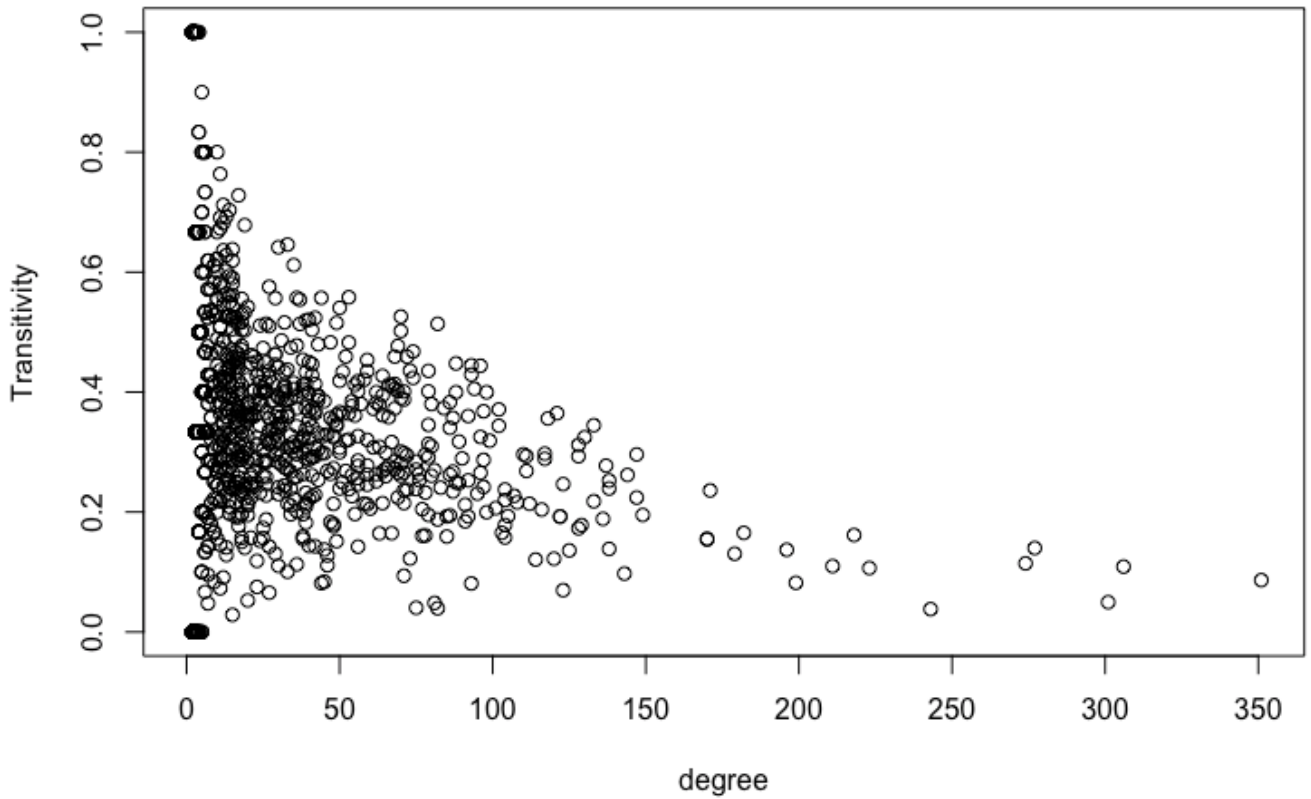
$$\tau_3(G) = 12, \quad \tau_\triangle(G) = 1, \quad cl(G) = 0.25$$

Local Density

- Clustering coefficient is also called *transitivity*. For one node v, it means if v knows u and v knows w, what is the chance that u knows w.
- For the whole network, it is the conditional prob. of three nodes knowing each other, given that one knows the two others
- Transitivity for some typical graphs:
  - Transitivity for k-stars are always 0.
  - Transitivity for k-rings are always 0.
  - Transitivity for a complete graph is always 1.
- For large-scale real networks, most of the times (of course not all!)
  - $cl(v)$ varies *inversely* with vertex degree

Local Density

Recall: Political Blogs, 1490 nodes, 16715 edges
The Degree-Clustering Coefficient figure is as following



For the whole network, the transitivity is 0.226.