Network Science: Principles and Applications

CS 695 - Spring 2019

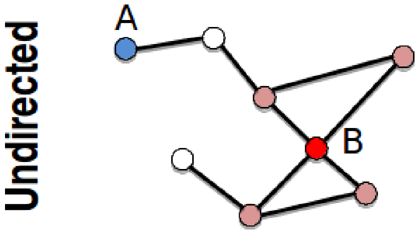
Amarda Shehu

[amarda](AT)gmu.edu
Department of Computer Science
George Mason University

1. Outline of Today's Class
2. Measures of Interest for Networks
   - Three Central Quantities in Network Science
   - Degree Distribution
     - Node Degree, Average Degree
   - Shortest-Path Lengths, Diameter, and Betweenness
   - A Fast Algorithm for Calculation of Betweenness Centrality
3. Central Quantities in Network Science
   - Clustering
   - Examples
   - Motifs
   - Community Structures
   - Graph Spectra
4. Topology of Real Networks

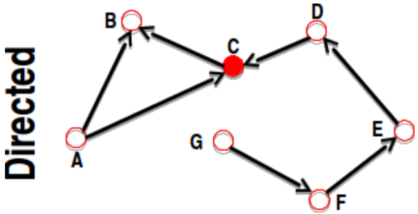
Three Central Quantities in Network Science

- Degree distribution p_k
- Average path length $\langle d \rangle$
- Clustering coefficient C



Node degree: nr. of links connected to node

$k_A = 1$ $\quad k_B = 4$



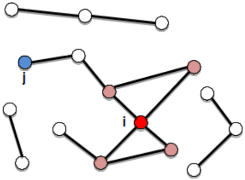
Node degree: sum of in- and out-degree

$k_C^{\mathrm{in}} = 2$ $\quad k_C^{\mathrm{out}} = 1$ $\quad k_C = 3$

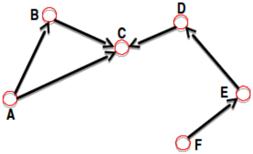
Source: node with 0 in-degree
Sink: node with 0 out-degree

Network Average Node Degrees

Undirected



Directed



Node degree: nr. of links connected to node

$$\langle k \rangle = \frac{1}{N} \sum_{i=1}^{N} k_i \qquad \langle k \rangle = \frac{2L}{N}$$

$$N = |V| \qquad L = |E|$$

Node degree: sum of in- and out-degree

$$\langle k^{\mathrm{in}} \rangle = \frac{1}{N} \sum_{i=1}^{N} k_i^{\mathrm{in}}$$

$$\langle k^{\mathrm{out}} \rangle = \frac{1}{N} \sum_{i=1}^{N} k_i^{\mathrm{out}}$$

$$\langle k^{\mathrm{in}} \rangle = \langle k^{\mathrm{out}} \rangle \qquad \langle k \rangle = \frac{L}{N}$$

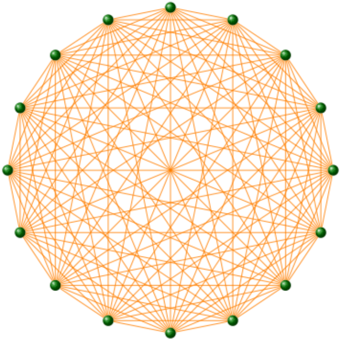
Source: node with 0 in-degree
Sink: node with 0 out-degree

Complete Graph

Maximum number of links in a network of N nodes: $L_{\max} = \binom{N}{2} = \frac{N \cdot (N-1)}{2}$

A graph with degree $L = L_{\max}$ is a **complete** graph

Its average degree is $\langle k \rangle = N - 1$



Real Networks are Sparse

Most networks observed in real systems are sparse

$$L << L_{\max}$$

or

$$\langle k \rangle << N - 1$$

WWW (ND Sample):	$N = 325,729$	$L = 1.4106$	$L_{\max} = 1012$	$\langle k \rangle = 4.51$
Protein (S. Cerevisiae):	$N = 1,870$	$L = 4,470$	$L_{\max} = 107$	$\langle k \rangle = 2.39$
Coauthorship (Math):	$N = 70,975$	$L = 2105$	$L_{\max} = 31010$	$\langle k \rangle = 3.9$
Movie Actors:	$N = 212,250$	$L = 6106$	$L_{\max} = 1.81013$	$\langle k \rangle = 28.78$

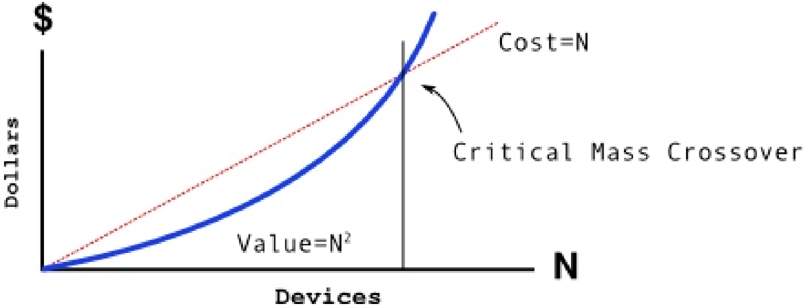
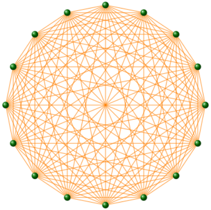


Figure: The value of a telecommunications network is proportional to the square of the number of connected users of the system.

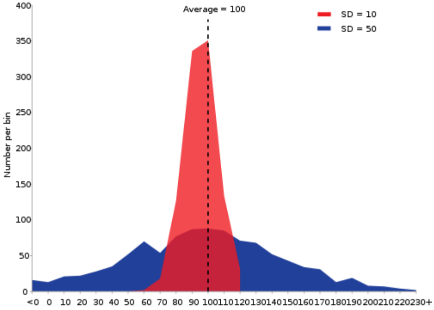
Maximum number of links a network of N nodes:
$L_{\max} = \binom{N}{2} = \frac{N \cdot (N-1)}{2}$



We have a sample of values $x_1, \ldots, x_N$

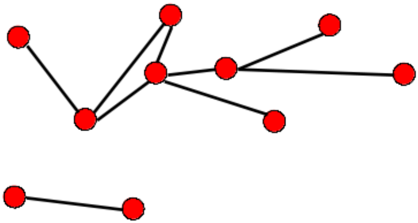
- Distribution of x (a.k.a. PDF): probability that a randomly chosen value is x
- $P(x) = (\#values x)/N$
- $\sum_i P(x_i) = 1$ always!

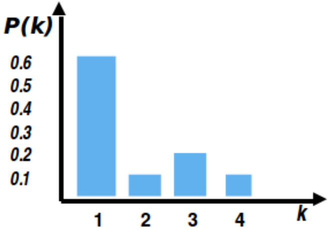
Histogram



We have a sample of values $x_1, \ldots, x_N$

- Degree distribution $P(k)$: probability that a randomly chosen vertex has degree k
- $N_k = \#nodes$ with degree k
- $P(k) = N_k/N$ plot





Degree Distribution

- **discrete representation:** p_k is the probability that a node has degree k
- **continuum description:** p_k is the pdf of the degrees, where $\int_{k_1}^{k_2} p_k dk$ represents the probability that a node's degree is between k_1 and k_2
- **Normalization condtion:** $\sum_0^\infty p_k = 1$ or $\int_0^\infty p_k dk = 1$
- $K_{\min}$ is the minimal degree in the network

Do Random Network Models Reproduce Deg. Distributions of Real Networks?

- Depends on the model

Do Random Network Models Reproduce Deg. Distributions of Real Networks?

- Depends on the model
- Depends when we are satisfied

Do Random Network Models Reproduce Deg. Distributions of Real Networks?

- Depends on the model
- Depends when we are satisfied
- More next lecture

Do Random Network Models Reproduce Deg. Distributions of Real Networks?

- Depends on the model
- Depends when we are satisfied
- More next lecture
- Let's derive the first momenta of the degree distribution in one of the earliest random models ... on the board

Do Random Network Models Reproduce Deg. Distributions of Real Networks?

- Depends on the model
- Depends when we are satisfied
- More next lecture
- Let's derive the first momenta of the degree distribution in one of the earliest random models ... on the board

Real Networks are Degree Correlated

The Erdos-Renyi model has a very narrow deviation, small σ_k, from $\langle k \rangle$, missing hubs.

Real Networks are Degree Correlated

The Erdos-Renyi model has a very narrow deviation, small σ_k, from $\langle k \rangle$, missing hubs.

Moreover, a real network is often degree correlated:

The probability that a node of degree k is connected to another node of degree $k^{'}$ **depends** on k.

Real Networks are Degree Correlated

The Erdos-Renyi model has a very narrow deviation, small σ_k, from $\langle k \rangle$, missing hubs.

Moreover, a real network is often degree correlated:

The probability that a node of degree k is connected to another node of degree $k^{'}$ **depends** on k.

- Necessary to introduce the **conditional** probability $P(k^{'}|k)$, defined as the probability that a link from a node of degree k points to a node of degree $k^{'}$.
- $P(k^{'}|k)$ satisfies the normalization $\sum_{k'} P(k^{'}|k) = 1$
- $P(k^{'}|k)$ satisfies the degree detailed balance condition $k \cdot P(k^{'}|k) \cdot P(k) = k^{'} \cdot P(k|k^{'}) \cdot P(k^{'})$
- For uncorrelated graphs, where $P(k^{'}|k)$ does not depend on k, the detailed balance condition and the normalization give $P(k^{'}|k) = k^{'} P(k^{'})/\langle k \rangle$.[a]

[a] S. Boccaletti et al. Physics Reports 424:175-308, 2006.

Real Networks are Correlated

Direct Evaluation of $P(k^{'}|k)$ is Noisy for Real Networks (Finite N)

- Can be overcome by defining **average nearest neighbors degree** of a node i:
  $$k_{nn,i} = \frac{1}{k_i}\sum_{j\in\mathcal{N}_i} k_j = \frac{1}{k_i}\sum_{j=1}^{N} a_{ij}\cdot k_j,$$
  where $\mathcal{N}_i$ refers to set of first neighbors of i.
- Then, average degree of nearest neighbors of nodes with degree k, $k_{nn}(k)$ can be expressed in terms of the conditional probability as:
  $$k_{nn}(k) = \sum_{k'} k^{'} P(k^{'}|k).$$
- In absence of corelations, $k_{nn}(k) = \langle k^2\rangle/\langle k\rangle$ (i.e., $k_{nn}(k)$ is independent of k.
- Correlated graphs are classified as:
  - **assortative** if $k_{nn}(k)$ is an increasing function of k (nodes tend to connect to their connectivity peers).
  - **disassortative** if $k_{nn}(k)$ is a decresing function of k (nodes with low degree are more likely connected with highly connected ones).
- Degree correlations are quantified by reporting:
  - slope of $k_{nn}(k)$ as a function of k.
  - Pearson correlation coefficient of degrees at either ends of a link.

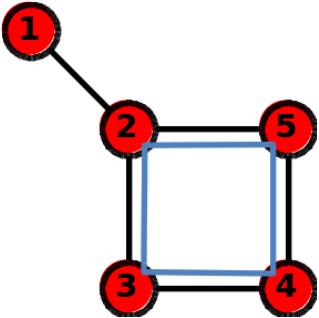
Three Central Quantities in Network Science

- Degree distribution p_k
- Average path length $\langle d \rangle$
- Clustering coefficient C

Paths in Measures

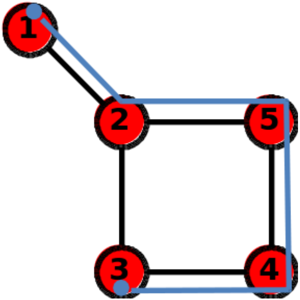
Concepts of a path connecting two nodes and shortest path connecting two nodes are central to various network measures. Let's see some paths first.

Cycle



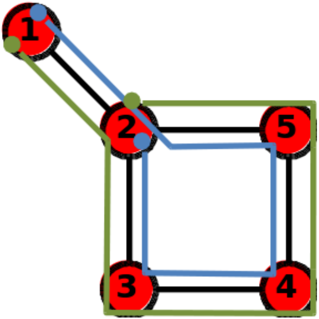
A path with the same start and end node.

Self-avoiding Path



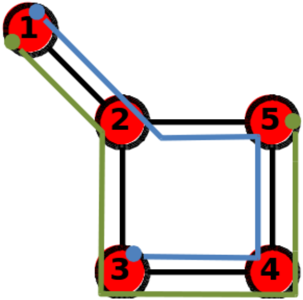
A path that does not intersect itself.

Eulerian Path



A path that traverses each link exactly once.

Hamiltonian Path



A path that visits each node exactly once.

Eulerian PATH or CIRCUIT: return to the starting point by traveling each link of the graph once and only once.
Eulerian graph has an eulerian path.

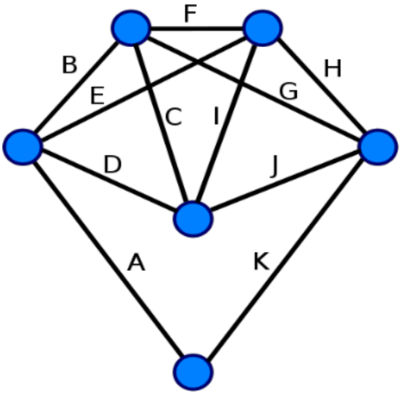
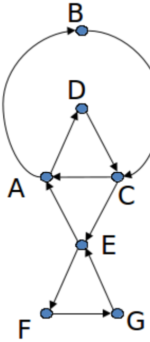


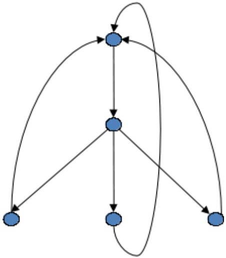
Figure: Every vertex of this graph has an even degree, therefore this is an Eulerian graph. Following the edges in alphabetical order gives an Eulerian circuit/cycle.

Eulerian Circuits in Directed Graphs

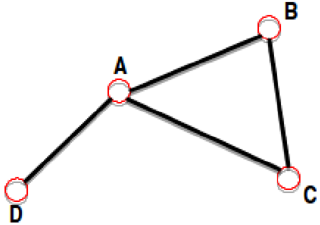


If a digraph is strongly connected and the in-degree of each node is equal to its out-degree, then there is an Eulerian circuit

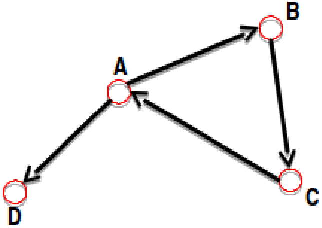
Otherwise there is no Eulerian circuit.
In a circuit we need to enter each node as many times as we leave it.



Undirected



Directed



distance (shortest path, geodesic path)

- between two nodes is defined as the number of edges along the shortest path connecting them.
- If the two nodes are disconnected, the distance is infinity.

distance (shortest path, geodesic path)

- In directed graphs each path needs to follow the direction of the arrows.
- In a digraph the distance from node A to B (on an AB path) is generally different from the distance from node B to A (on a BA path).

Shortest Paths

- Shortest paths play an important role in the transport and communication within a network.

Shortest Paths

- Shortest paths play an important role in the transport and communication within a network.
- Suppose one needs to send a data packet from one computer to another through the Internet: the geodesic provides an optimal path way, since one would achieve a fast transfer and save system resources.

Shortest Paths

- Shortest paths play an important role in the transport and communication within a network.
- Suppose one needs to send a data packet from one computer to another through the Internet: the geodesic provides an optimal path way, since one would achieve a fast transfer and save system resources.
- For such a reason, shortest paths have also played an important role in the characterization of the internal structure of a graph.

Shortest Paths

- Shortest paths play an important role in the transport and communication within a network.
- Suppose one needs to send a data packet from one computer to another through the Internet: the geodesic provides an optimal path way, since one would achieve a fast transfer and save system resources.
- For such a reason, shortest paths have also played an important role in the characterization of the internal structure of a graph.
- It is useful to represent all the shortest path lengths of a graph G as a matrix D in which the entry d_{ij} is the length of the geodesic from node i to node j.

Shortest Paths

- Shortest paths play an important role in the transport and communication within a network.
- Suppose one needs to send a data packet from one computer to another through the Internet: the geodesic provides an optimal path way, since one would achieve a fast transfer and save system resources.
- For such a reason, shortest paths have also played an important role in the characterization of the internal structure of a graph.
- It is useful to represent all the shortest path lengths of a graph G as a matrix D in which the entry d_{ij} is the length of the geodesic from node i to node j.
- How does one find the (shortest) distance between two nodes in a graph?

Shortest Paths

- Shortest paths play an important role in the transport and communication within a network.
- Suppose one needs to send a data packet from one computer to another through the Internet: the geodesic provides an optimal path way, since one would achieve a fast transfer and save system resources.
- For such a reason, shortest paths have also played an important role in the characterization of the internal structure of a graph.
- It is useful to represent all the shortest path lengths of a graph G as a matrix D in which the entry d_{ij} is the length of the geodesic from node i to node j.
- How does one find the (shortest) distance between two nodes in a graph?
- For unweighted graphs: BFS

Shortest Paths

- Shortest paths play an important role in the transport and communication within a network.
- Suppose one needs to send a data packet from one computer to another through the Internet: the geodesic provides an optimal path way, since one would achieve a fast transfer and save system resources.
- For such a reason, shortest paths have also played an important role in the characterization of the internal structure of a graph.
- It is useful to represent all the shortest path lengths of a graph G as a matrix D in which the entry d_{ij} is the length of the geodesic from node i to node j.
- How does one find the (shortest) distance between two nodes in a graph?
- For unweighted graphs: BFS
- For (non-negative) weighted graphs: Dijkstra, A*, D*, and variants.

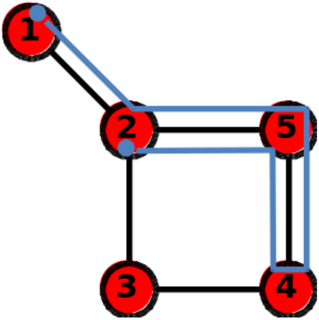
Shortest Paths

- Shortest paths play an important role in the transport and communication within a network.
- Suppose one needs to send a data packet from one computer to another through the Internet: the geodesic provides an optimal path way, since one would achieve a fast transfer and save system resources.
- For such a reason, shortest paths have also played an important role in the characterization of the internal structure of a graph.
- It is useful to represent all the shortest path lengths of a graph G as a matrix D in which the entry d_{ij} is the length of the geodesic from node i to node j.
- How does one find the (shortest) distance between two nodes in a graph?
- For unweighted graphs: BFS
- For (non-negative) weighted graphs: Dijkstra, A*, D*, and variants.
- How does one final all-pair shortest paths?

Shortest Paths

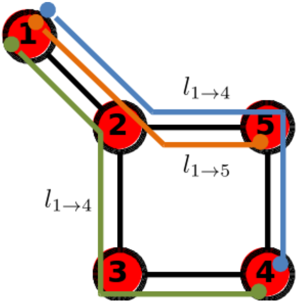
- Shortest paths play an important role in the transport and communication within a network.
- Suppose one needs to send a data packet from one computer to another through the Internet: the geodesic provides an optimal path way, since one would achieve a fast transfer and save system resources.
- For such a reason, shortest paths have also played an important role in the characterization of the internal structure of a graph.
- It is useful to represent all the shortest path lengths of a graph G as a matrix D in which the entry d_{ij} is the length of the geodesic from node i to node j.
- How does one find the (shortest) distance between two nodes in a graph?
- For unweighted graphs: BFS
- For (non-negative) weighted graphs: Dijkstra, A*, D*, and variants.
- How does one final all-pair shortest paths?
- Floyd-Warshall

Path



A sequence of nodes such that each node is connected to the next node along the path by a link.

Shortest Path



$l_{1\to4} = 3$

$l_{1\to5} = 2$

The path with the shortest length between two nodes (distance).

Network Diameter and Average Distance

Diameter $d_{\max}$: the maximum distance (shortest path length) between any two nodes in the graph ($\max\{d_{ij}\}$), often also denoted as $diam(G)$.

A measure of the typical separation between two nodes in a network is given by the **average shortest path length** also known as the **characteristic path length**, defined as the mean geodesic lengths over all pairs of nodes.

Network Diameter and Average Distance

Diameter $d_{\max}$: the maximum distance (shortest path length) between any two nodes in the graph ($\max\{d_{ij}\}$), often also denoted as $diam(G)$.

A measure of the typical separation between two nodes in a network is given by the **average shortest path length** also known as the **characteristic path length**, defined as the mean geodesic lengths over all pairs of nodes.

Average distance $\langle d \rangle$ for a connected graph:
$\langle d \rangle = \frac{1}{2L_{\max}} \sum_{i \neq j} d_{ij}$, where d_{ij} is the distance from node i to node j

Network Diameter and Average Distance

Diameter $d_{\max}$: the maximum distance (shortest path length) between any two nodes in the graph ($\max\{d_{ij}\}$), often also denoted as $diam(G)$.

A measure of the typical separation between two nodes in a network is given by the **average shortest path length** also known as the **characteristic path length**, defined as the mean geodesic lengths over all pairs of nodes.

Average distance $\langle d \rangle$ for a connected graph:
$\langle d \rangle = \frac{1}{2L_{\max}} \sum_{i \neq j} d_{ij}$, where d_{ij} is the distance from node i to node j

In undirected graph, $d_{ij} = d_{ji}$, so only counting once leads to:
$\langle d \rangle = \frac{1}{L_{\max}} \sum_{i \neq j} d_{ij}$

Network Diameter and Average Distance

Diameter $d_{\max}$: the maximum distance (shortest path length) between any two nodes in the graph ($\max\{d_{ij}\}$), often also denoted as $diam(G)$.

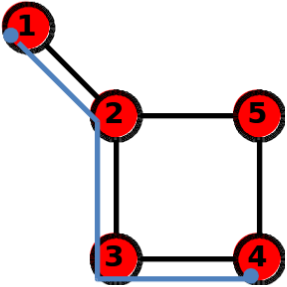
A measure of the typical separation between two nodes in a network is given by the **average shortest path length** also known as the **characteristic path length**, defined as the mean geodesic lengths over all pairs of nodes.

Average distance $\langle d \rangle$ for a connected graph:
$\langle d \rangle = \frac{1}{2L_{\max}} \sum_{i \neq j} d_{ij}$, where d_{ij} is the distance from node i to node j

In undirected graph, $d_{ij} = d_{ji}$, so only counting once leads to:
$\langle d \rangle = \frac{1}{L_{\max}} \sum_{i \neq j} d_{ij}$

Illustration

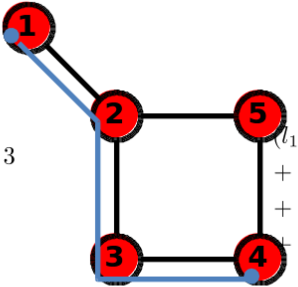
Diameter



$$l_{1\to4} = 3$$

The longest shortest path in a graph

Average Path Length



$$(l_{1\to2} + l_{1\to3} + l_{1\to4} + l_{1\to5} + l_{2\to3} + l_{2\to4} + l_{2\to5} + l_{3\to4} + l_{3\to5} + l_{4\to5})/10 = 1.6$$

The average of the shortest paths for all pairs of nodes.

Characteristic Path Length and Connectivity

Problem with $\langle d \rangle$ is that it diverges if there are disconnected components in the the graph.

How to address?

Characteristic Path Length and Connectivity

Problem with $\langle d \rangle$ is that it diverges if there are disconnected components in the the graph.

How to address?

First, let's define connectivity.

Characteristic Path Length and Connectivity

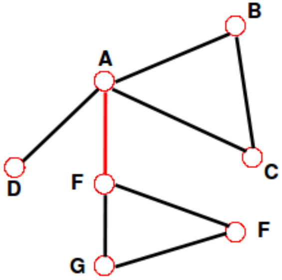
Problem with $\langle d \rangle$ is that it diverges if there are disconnected components in the the graph.

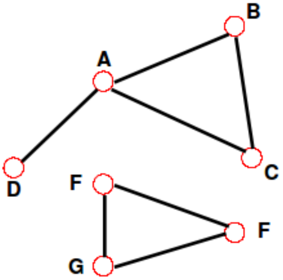
How to address?

First, let's define connectivity.

Connectivity of Undirected Graphs

- Connected (undirected) graph: any two vertices can be joined by a path.
- A disconnected graph is made up by two or more connected components.



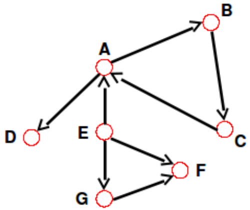


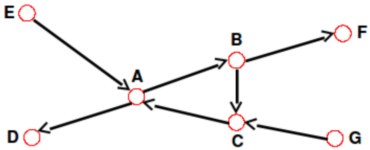
- Largest Component: **giant component**
- The rest: **isolates**
- Bridge: If we erase it, graph becomes disconnected

The adjacency matrix of a network with several components can be written in a block-diagonal form, so that nonzero elements are confined to squares, with all other elements being zero:

$$A = \begin{pmatrix} \blacksquare & 0 & \cdots \\ 0 & \blacksquare & \cdots \\ \vdots & \vdots & \ddots \end{pmatrix}$$

- **Strongly-connected directed** graph: has a path from each node to every other node and vice-versa
- **Weakly-connected directed** graph: connected if edge directions are disregarded.
- **Strongly-connected components (scc)** can be identified (via DFS-based algorithm), but not every node is part of a non-trivial scc.





- **In-component:** nodes that can reach the scc
- **Out-component:** nodes that can be reached from the scc

From Characteristic Path Length to Network Efficiency

Issue can be addressed by limiting formulation to largest connected component, or by considering the harmonic mean, so-called **efficiency** of G:
$E = \frac{1}{2L_{\max}} \sum_{i \neq j} \frac{1}{d_i j}$

Efficiency avoids divergence issue because any pair of nodes belonging to two different components yields a contribution of 0 the summation.

Issue can be addressed by limiting formulation to largest connected component, or by considering the harmonic mean, so-called **efficiency** of G:
$E = \frac{1}{2L_{\max}} \sum_{i \neq j} \frac{1}{d_i j}$

Efficiency avoids divergence issue because any pair of nodes belonging to two different components yields a contribution of 0 the summation.

Efficiency is an indicator of traffic capacity of a network.

From Characteristic Path Length to Network Efficiency

Issue can be addressed by limiting formulation to largest connected component, or by considering the harmonic mean, so-called **efficiency** of G:
$E = \frac{1}{2L_{\max}} \sum_{i \neq j} \frac{1}{d_i j}$

Efficiency avoids divergence issue because any pair of nodes belonging to two different components yields a contribution of 0 the summation.

Efficiency is an indicator of traffic capacity of a network.

Mathematical properties and extensions of **efficiency** have been studied by Criado et al. J Comput. Appl. Math. 2005 and Vragovic et al. Phys. Rev. E. 2005.

From Characteristic Path Length to Network Efficiency

Issue can be addressed by limiting formulation to largest connected component, or by considering the harmonic mean, so-called **efficiency** of G:

$E = \frac{1}{2L_{\max}} \sum_{i \neq j} \frac{1}{d_i j}$

Efficiency avoids divergence issue because any pair of nodes belonging to two different components yields a contribution of 0 the summation.

Efficiency is an indicator of traffic capacity of a network.

Mathematical properties and extensions of **efficiency** have been studied by Criado et al. J Comput. Appl. Math. 2005 and Vragovic et al. Phys. Rev. E. 2005.

Another useful measure is the **closeness** of a node, defined as the inverse of the average distance from all other nodes.

From Characteristic Path Length to Network Efficiency

Issue can be addressed by limiting formulation to largest connected component, or by considering the harmonic mean, so-called **efficiency** of G:
$E = \frac{1}{2L_{\max}} \sum_{i \neq j} \frac{1}{d_i j}$

Efficiency avoids divergence issue because any pair of nodes belonging to two different components yields a contribution of 0 the summation.

Efficiency is an indicator of traffic capacity of a network.

Mathematical properties and extensions of **efficiency** have been studied by Criado et al. J Comput. Appl. Math. 2005 and Vragovic et al. Phys. Rev. E. 2005.

Another useful measure is the **closeness** of a node, defined as the inverse of the average distance from all other nodes.

All above quantities aim to measure communication in a network.

From Characteristic Path Length to Network Efficiency

Issue can be addressed by limiting formulation to largest connected component, or by considering the harmonic mean, so-called **efficiency** of G:
$E = \frac{1}{2L_{\max}} \sum_{i \neq j} \frac{1}{d_i j}$

Efficiency avoids divergence issue because any pair of nodes belonging to two different components yields a contribution of 0 the summation.

Efficiency is an indicator of traffic capacity of a network.

Mathematical properties and extensions of **efficiency** have been studied by Criado et al. J Comput. Appl. Math. 2005 and Vragovic et al. Phys. Rev. E. 2005.

Another useful measure is the **closeness** of a node, defined as the inverse of the average distance from all other nodes.

All above quantities aim to measure communication in a network.

Communication in a Network: Node Betweenness

The communication of two non-adjacent nodes, j and k, depends on the nodes belonging to the paths connecting j and k.

A measure of the *relevance* of a given node can be obtained by counting the number of geodesics going through it, and defining the so-called **node betweenness**.

[1] Wasserman et al. Social Network Analysis, Cambridge University Press 1994

[2] see S. Boccaletti review for a list of references.

Communication in a Network: Node Betweenness

The communication of two non-adjacent nodes, j and k, depends on the nodes belonging to the paths connecting j and k.

A measure of the *relevance* of a given node can be obtained by counting the number of geodesics going through it, and defining the so-called **node betweenness**.

Like degree and closeness, **betweenness** is a standard measure of **node centrality**, originally introduced to quantify the *importance of an individual in a social network*[1].

[1] Wasserman et al. Social Network Analysis, Cambridge University Press 1994

[2] see S. Boccaletti review for a list of references.

Communication in a Network: Node Betweenness

The communication of two non-adjacent nodes, j and k, depends on the nodes belonging to the paths connecting j and k.

A measure of the *relevance* of a given node can be obtained by counting the number of geodesics going through it, and defining the so-called **node betweenness**.

Like degree and closeness, **betweenness** is a standard measure of **node centrality**, originally introduced to quantify the *importance of an individual in a social network*[1].

Betweenness b_i of a node i, sometimes referred to also as **load**, is defined as:

$$b_i = \sum_{j \neq k} \frac{n_{jk(i)}}{n_{jk}}$$

where n_{jk} is the number of shortest paths connecting j and k, and $n_{jk}(i)$ is the number of shortest paths connecting j and k that go through i.

[1] Wasserman et al. Social Network Analysis, Cambridge University Press 1994

[2] see S. Boccaletti review for a list of references.

Communication in a Network: Node Betweenness

The communication of two non-adjacent nodes, j and k, depends on the nodes belonging to the paths connecting j and k.

A measure of the *relevance* of a given node can be obtained by counting the number of geodesics going through it, and defining the so-called **node betweenness**.

Like degree and closeness, **betweenness** is a standard measure of **node centrality**, originally introduced to quantify the *importance of an individual in a social network*[1].

Betweenness b_i of a node i, sometimes referred to also as **load**, is defined as:

$$b_i = \sum_{j \neq k} \frac{n_{jk(i)}}{n_{jk}}$$

where n_{jk} is the number of shortest paths connecting j and k, and $n_{jk}(i)$ is the number of shortest paths connecting j and k that go through i.

Betweenness distributions, betweennes-betweennes correlations, and betweenness-degree correlationshave been investigated in many papers[2]

[1] Wasserman et al. Social Network Analysis, Cambridge University Press 1994

[2] see S. Boccaletti review for a list of references.

Communication in a Network: Node Betweenness

The communication of two non-adjacent nodes, j and k, depends on the nodes belonging to the paths connecting j and k.

A measure of the *relevance* of a given node can be obtained by counting the number of geodesics going through it, and defining the so-called **node betweenness**.

Like degree and closeness, **betweenness** is a standard measure of **node centrality**, originally introduced to quantify the *importance of an individual in a social network*[1].

Betweenness b_i of a node i, sometimes referred to also as **load**, is defined as:

$$b_i = \sum_{j \neq k} \frac{n_{jk(i)}}{n_{jk}}$$

where n_{jk} is the number of shortest paths connecting j and k, and $n_{jk}(i)$ is the number of shortest paths connecting j and k that go through i.

Betweenness distributions, betweennes-betweennes correlations, and betweenness-degree correlationshave been investigated in many papers[2]

Concept extends to edges, defining **edge betweenness** as the number of shortest paths utilizing an edge.

[1] Wasserman et al. Social Network Analysis, Cambridge University Press 1994

[2] see S. Boccaletti review for a list of references.

Communication in a Network: Node Betweenness

The communication of two non-adjacent nodes, j and k, depends on the nodes belonging to the paths connecting j and k.

A measure of the *relevance* of a given node can be obtained by counting the number of geodesics going through it, and defining the so-called **node betweenness**.

Like degree and closeness, **betweenness** is a standard measure of **node centrality**, originally introduced to quantify the *importance of an individual in a social network*[1].

Betweenness b_i of a node i, sometimes referred to also as **load**, is defined as:

$$b_i = \sum_{j \neq k} \frac{n_{jk(i)}}{n_{jk}}$$

where n_{jk} is the number of shortest paths connecting j and k, and $n_{jk}(i)$ is the number of shortest paths connecting j and k that go through i.

Betweenness distributions, betweennes-betweennes correlations, and betweenness-degree correlationshave been investigated in many papers[2]

Concept extends to edges, defining **edge betweenness** as the number of shortest paths utilizing an edge.

[1] Wasserman et al. Social Network Analysis, Cambridge University Press 1994

[2] see S. Boccaletti review for a list of references.

Some Simple Formulas: Number of Paths Between Two Nodes

Let N_{ij} be number of paths between nodes i and j:

- Length $n = 1$: If there is a link between i and j, then $A_{ij} = 1$ and A_{ij}=0 otherwise.
- Length $n = 2$: If there is a path of length two between i and j, then $A_{ik} \cdot A_{kj}$=1, and $A_{ik} \cdot A_{kj} = 0$ otherwise.
- **Number of paths of length 2:**
  $$N_{ij}^2 = \sum_{k=1}^{N} A_{ik} \cdot A_{kj} = [A^2]_{ij}$$
- Length n: In general, if there is a path of length n between i and j, then $A_{ik} \cdot \ldots \cdot A_{lj} = 1$ and $A_{ik} \cdot \ldots \cdot A_{lj} = 0$ otherwise.
- **Number of paths of length n between i and j:**[a]:
  $$N_{ij}^n = \sum_{k=1}^{N} A_{ik} \cdot A_{kj} = [A^n]_{ij}$$

[a] for both directed and undirected graphs

A Fast Algorithm for Calculation of Betweenness Centrality

Published by Ulrik Brandes in J Mathematical Sociology 2001.

Lemma 1 (Bellman Criterion): *A vertex* $v \in V$ *lies on a shortest path between vertices* $s, t \in V$ *iff* $d_G(s,t) = d_G(s,v) + d_G(v,t)$.

Given pairwise distances and shortest-path counts, the **pair-dependency** $\delta_{st}(v) = \frac{\sigma_{st}(v)}{\sigma_{st}}$ of a pair $s, t \in V$ on an intermediary $v \in V$, i.e. the ratio of shortest paths between s, t on which v lies, can be derived from:

$\sigma_{st}(v) = 0$ if $d_G(s,t) < d_G(s,v) + d_G(v,t)$ and $\sigma_{sv} \cdot \sigma_{vt}$ otherwise.

To obtain the betweenness-centrality index of a vertex v, we simply sum the pair-dependencies of all pairs on that vertex:
$C_B(v) = \sum_{s \neq v \neq t \in V} \delta_{st}(v)$.

A Fast Algorithm for Calculation of Betweenness Centrality

To compute betweenness-centrality, two steps are needed:

- compute length and number of shortest paths between all pairs
- sum all-pair dependencies.

Step 2. involves $\theta(n^3)$ summations and $\theta(n^2)$ storage of pair-dependencies.

Both BFS and Dijkstra's algorithm can be easily augmented to count the number of shortest paths:

- BFS can run in $O(m)$ time (unweighted graph).
- DFS can run in time $O(m + n \cdot logn)$ for weighted graphs if priority queue is implemented as a Fibonacci heap.

Corollary: *Given a source $s \in V$, both the length and number of all shortest paths to other vertices can be determined in time $O(m + nlogn)$ for weighted graphs and $O(m)$ for unweighted graphs.*

The explicit summation of all pair-dependencies can be avoided via a recursive formulation of the dependency of a vertex $\delta_{s*}(v) = \sum_{t \in V} \delta_{st}(v)$.

A Fast Algorithm for Calculation of Betweenness Centrality

Corollary: *Given the directed acyclic graph of shortest paths from $s \in V$ in G, the dependencies of s on all other vertices can be computed in $O(m)$ time and $O(n+m)$ space (details in Brandes paper).*

Idea: Traverse the vertices in non-increasing order of their distance from s and accumulate dependencies. Need to store a dependency per vertex, and lists of predecessors. There is at most one element per edge in any of these lists.

The betweenness centrality index can be determined by solving one single-source shortest-paths problem for each vertex. At the end of each iteration, the dependencies of the source on each other vertex are added to the centrality score of that vertex.

Theorem: *Betwennes centrality can be computed in $O(nm + n^2 logn)$ time and $n + m$ space for weighted graphs and in $O(nm)$ time for unweighted graphs.*

Brandes also shows how to compute other centrality measures via a similar efficient process.

Central Quantities in Network Science

- Degree distribution p_k
- Average path length $\langle d \rangle$
- Clustering coefficient C

Clustering, also known as **transitivity**, is a typical property of *acquaintance* networks, where two individuals with a common friend are likely to know each other[3].

Transitivity means the presence of a high number of triangles.

This can be quantified by defining the transitivity T of a graph G as the relative number of transitive triples, i.e. the fraction of connected triples of nodes (triads) which also form triangles[4].

$$T = \frac{3 \times \#\text{triangles in } G}{\#\text{connected components in } G}$$

The factor 3 compensates for the fact that each complete triangle of three nodes contributes three connected triples, one centered on each of the three nodes, and ensures that $0 \leq T \leq 1$ with $T = 1$ for K_N (a complete graph of N nodes).

An alternative measure is the **clustering coefficient**, introduced by Watts and Strogatz.

[3] Wasserman et al. Social Network Analysis, Cambridge University Press 1994

[4] Newman, SIAM Rev 2003

Clustering Coefficient of a Node

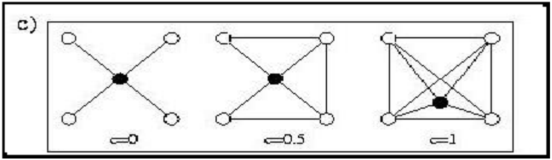
Clustering Coefficient of a Node:

- What portion of your neighbors are connected?
- Introduced by Watts and Strogatz in Nature 1998.
- **Local clustering coefficient** c_i of node i is introduced, expressing how likely $a_{jm} = 1$ for two neighbors j, m of node i.
- c_i of a node with degree k_i is obtained by counting actual number of edges e_i in subgraph G_i induced by neighbors of i normalizing by maximum possible number of edges in G_i:

$$c_i = \frac{2e_i}{k_i(k_i-1)} \qquad\qquad 0 \leq c_i \leq 1$$

- Fast algorithms to compute c_i are presented in Alon et al. Algorithmica 1997.

$$\boxed{C_i = \frac{2e_i}{k_i(k_i-1)}}$$



Clustering coefficient of a graph G is then the average of c_i over all nodes in G:

$$C = \langle c_i \rangle = \frac{1}{N} \sum_i c_i \qquad 0 \leq C \leq 1$$

Clustering coefficient of a connectivity class k, $c(k)$ is defined as the average of c_i taken over all nodes with a given degree k

Clustering Coefficient of a Graph G and Variants

Clustering coefficient of a graph G is then the average of c_i over all nodes in G:

$$C = \langle c_i \rangle = \frac{1}{N} \sum_i c_i \qquad\qquad 0 \leq C \leq 1$$

Clustering coefficient of a connectivity class k, $c(k)$ is defined as the average of c_i taken over all nodes with a given degree k

HIgher-order clustering coefficients have been proposed, such as the k-clustering coefficient that acccounts for k-neighbors, other measures based on internal structure of cycles of order 4, or on the number of cycles of a generic order.

Clustering Coefficient of a Graph G and Variants

Clustering coefficient of a graph G is then the average of c_i over all nodes in G:

$$C = \langle c_i \rangle = \frac{1}{N} \sum_i c_i \qquad 0 \leq C \leq 1$$

Clustering coefficient of a connectivity class k, $c(k)$ is defined as the average of c_i taken over all nodes with a given degree k

HIgher-order clustering coefficients have been proposed, such as the k-clustering coefficient that acccounts for k-neighbors, other measures based on internal structure of cycles of order 4, or on the number of cycles of a generic order.

An alternative of the clustering properties of G is the **local efficiency**:

$$E_{\mathrm{loc}} = \frac{1}{N} \sum_i E(G_i) \qquad E(G_i) \text{ is the efficiency of } G_i$$

Clustering Coefficient of a Graph G and Variants

Clustering coefficient of a graph G is then the average of c_i over all nodes in G:

$$C = \langle c_i \rangle = \frac{1}{N} \sum_i c_i \qquad\qquad 0 \leq C \leq 1$$

Clustering coefficient of a connectivity class k, $c(k)$ is defined as the average of c_i taken over all nodes with a given degree k

HIgher-order clustering coefficients have been proposed, such as the k-clustering coefficient that acccounts for k-neighbors, other measures based on internal structure of cycles of order 4, or on the number of cycles of a generic order.

An alternative of the clustering properties of G is the **local efficiency**:

$$E_{\mathrm{loc}} = \frac{1}{N} \sum_i E(G_i) \qquad\qquad E(G_i) \text{ is the efficiency of } G_i$$

Let's look at some simple examples.

Clustering Coefficient of a Graph G and Variants

Clustering coefficient of a graph G is then the average of c_i over all nodes in G:

$$C = \langle c_i \rangle = \frac{1}{N} \sum_i c_i \qquad\qquad 0 \leq C \leq 1$$

Clustering coefficient of a connectivity class k, $c(k)$ is defined as the average of c_i taken over all nodes with a given degree k

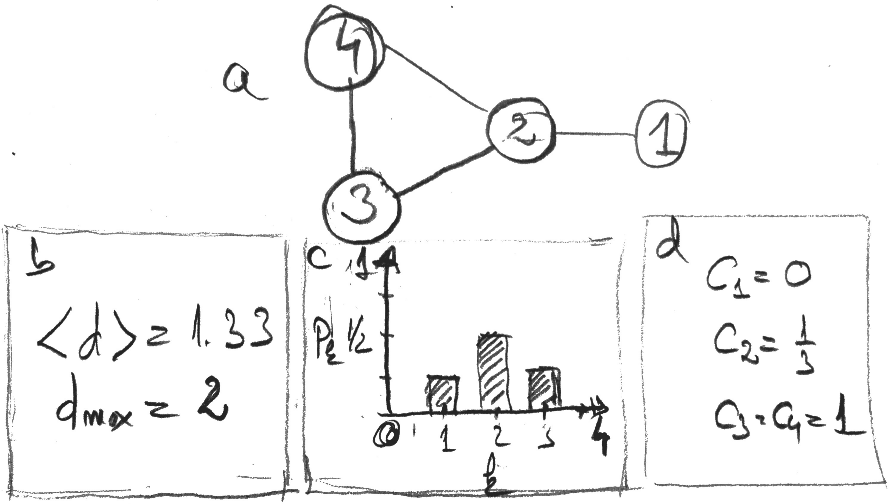
HIgher-order clustering coefficients have been proposed, such as the k-clustering coefficient that acccounts for k-neighbors, other measures based on internal structure of cycles of order 4, or on the number of cycles of a generic order.

An alternative of the clustering properties of G is the **local efficiency**:

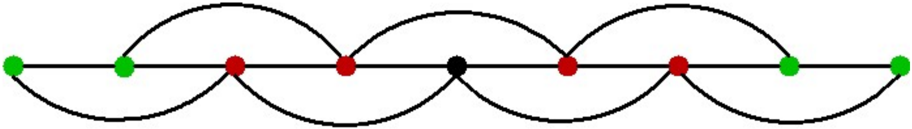
$$E_{\mathrm{loc}} = \frac{1}{N} \sum_i E(G_i) \qquad\qquad E(G_i) \text{ is the efficiency of } G_i$$

Let's look at some simple examples.

Example: Three Quantities

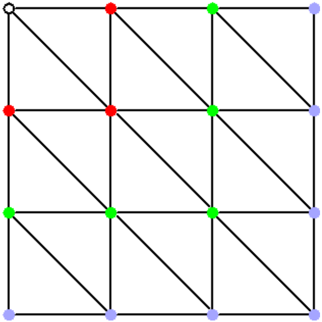


1D Lattice



- $P_k = \delta(k-4) \qquad k = 4$ for each node here
- $C = 1/2$ for each node if $N > 6$
- $1 + \sum_{l=1}^{l_{\max}} 4 \approx N \Rightarrow d_{\max} = \frac{N}{4} \qquad \langle d \rangle = \frac{4\sum_{d=1}^{d_{\max}} d}{N} \Rightarrow \langle d \rangle \approx \frac{N}{8}$
- The average path length varies as $\langle d \rangle \approx N$
- Constant degree
- Constant clustering coefficient

2D Lattice



- $P_k = \delta(k-6)$ $\qquad$ $k = 6$ for inside nodes
- $C = 6/15$ for inside nodes
- $1 + \sum_{l=1}^{l_{\max}} 6l \approx N \Rightarrow l_{\max} \propto \frac{N}{0.5}$ $\qquad$ $\langle l \rangle \approx L \approx N^{1/2}$
- In general, the average distance varies as $\langle l \rangle \approx L \approx N^{1/D}$ where D is the dimensionality of the lattice
- Constant degree (coordination number) $\qquad$ Constant clustering coefficient

A Motif M

- is a pattern of interconnections occurring either in a undirected or in a directed graph G at a number significantly higher than in randomized versions of the graph, i.e. in graphs with the same number of nodes, links and degree distribution as the original one, but where the links are distributed at random.
- As a pattern of interconnections, M is usually meant as a connected (undirected or directed) n-node graph which is a subgraph of G.

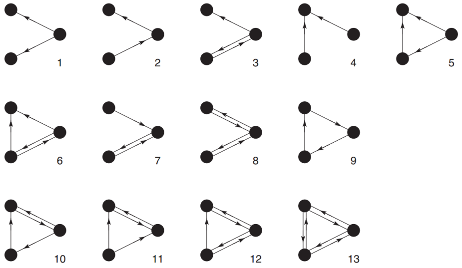


Figure: All possible 3-node connected directed graphs

Motifs

- The concept of motifs was introduced by Alon and co-workers, who studied small n motifs in biological networks and more.
- Significant motifs in a graph G are found by using matching algorithms that count the total number of occurrences of each n-node subgraph M in G and compare that to the count in randomized graphs.
- Statistical significance is determined by Z-score, defined as:

$$Z_m = \frac{n_M - \langle n_M^{\mathrm{rand}} \rangle}{\sigma_{n_M}^{\mathrm{rand}}}$$

  - where n_M is the number of times the subgraph M appears in G, and $\langle n_M^{\mathrm{rand}} \rangle$ and $\sigma_{n_M}^{\mathrm{rand}}$ are the average and standard deviations of the number of occurrences in a randomized network ensemble.

- Notion of **community** (or cluster, cohesive subgraph) first proposed in social sciences as a subgraph whose nodes are tighly conneced, i.e., cohesive.

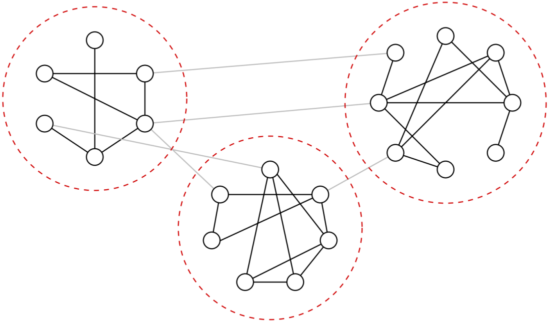


Figure: Communities can be defined as groups of nodes such that there is a higher density of edges within groups than between them. In the case shown in figure there are three communities, denoted by the dashed circles. ©2004 by the American Physical Society[6].

[5] Newman, Girvan Phys Rev E. 2004
[6] Newman, Girvan Phys Rev E. 2004

Structural Cohesion Measures

- Structural cohesion of the subgraph can be quantified in several ways, so there are different definitions of community structures.
- Strongest definition is that of a **clique**, a maximally-complete subgraph of three or more nodes.
- Weaker requirement uses **reachability**: an n-clique s a maximal subgraph in which the largest geodesic between any two nodes is no greater than n.
  - $n = 1$: this is just a clique.
  - $n = 2$: not all nodes are adjacent, but are reachable through at most one intermediate node.
  - $n = 3$: non-adjacent nodes are reachable through at most 2 intermediate nodes.
  - and so on.
- Alternative weakening involves reducing the number of nodes to which each node must be connected: a **k-plex** is a maximal subgraph containing n nodes, in which each node is adjacent to no fewer than $n - k$ nodes in the subgraph.

More Structural Cohesion Measures

- A different definition is based on the frequency of links; in this case communities are seen as groups of nodes within which connections are dense, and between which connections are sparse (previous figure was an example of this).
- Simplest formal definition has been proposed in Seidman, Social Network, 1983.
- Less stringent definition: $G^{'}$ is a community if the sum of degrees within $G^{'}$ is larger than the sum of all degrees towards $G - G^{'}$.
- Several other definitions are available, as in Wasserman et al. Social Network Analysis 1994.

The Spectrum of a Graph

- Is the set of eigenvalues of its adjacency matrix A
- a Graph $G_{N,K}$ (of N vertices and K edges) has N eigenvalues μ_i and N associated eigenvectors v_i.
- When G is a simpled undirected graph, $\mathcal{A}$ is real and symmetrix, so the eigenvalues are real, and the eigenvectors corresponding to distinct eigenvalues are orthogonal.
- When G is directed, the eigenvalues can have immaginary parts, as for instance in the tournament graph with 3 nodes; ordering and properties of eigenvalues and eigenvectors here is more complicated.

More on Graph Spectra

- The spectrum of the **normal** and **Laplacian** matrix of a graph G reveals important information regarding its connectivity.
- The **normal matrix** is defined as $\mathcal{N} = \mathcal{D}^{-1} \cdot \mathcal{A}$, where D is the diagonal matrix with $D_{ii} = \sum_j a_{ij} = k_i$.
- The **Laplacian matrix** Δ, also known as the **Kirchhoff matrix** is defined as $\Delta = \mathcal{D} - \mathcal{A}$.
- The multiplicity of the eigenvalue 0 of Δ equals the number of connected components in G.
- The second smallest eigenvalue λ_2 is important, too; several theorems from spectral graph theory prove that the larger λ_2, the more difficult it is to cut G into pieces.
- The spectrum of A and N have been used to discover cohesive subgroups and other local features of real networks, as we will cover later in this course.

Topology of Real Networks

- Despite inherent differences, most of the *real networks* are characterized by the same topological properties, such as:
  - relatively small characteristic path lengths
  - high clustering coefficients
  - fat tailed shapes in the degree distributions
  - degree correlations
  - presence of motifs and community structures.
- These features make real networks radically different from regular lattices and random graphs, the standard models studied in mathematical graph theory.
- This observation has led to a large attention towards:
  - *understanding of the evolution mechanisms that have shaped the topology of a network*
  - *design of new models retaining the most significant properties empirically observed*

Topology of Real Networks

Specifically, two properties observed about real networks are:

- Small-world property ($*$-degree separation)
- Scale-free degree distributions (power-law shaped degree distribution)

The focus of the next 2 lectures will be on random network models that can reproduce the topology of real networks in partially or fully.