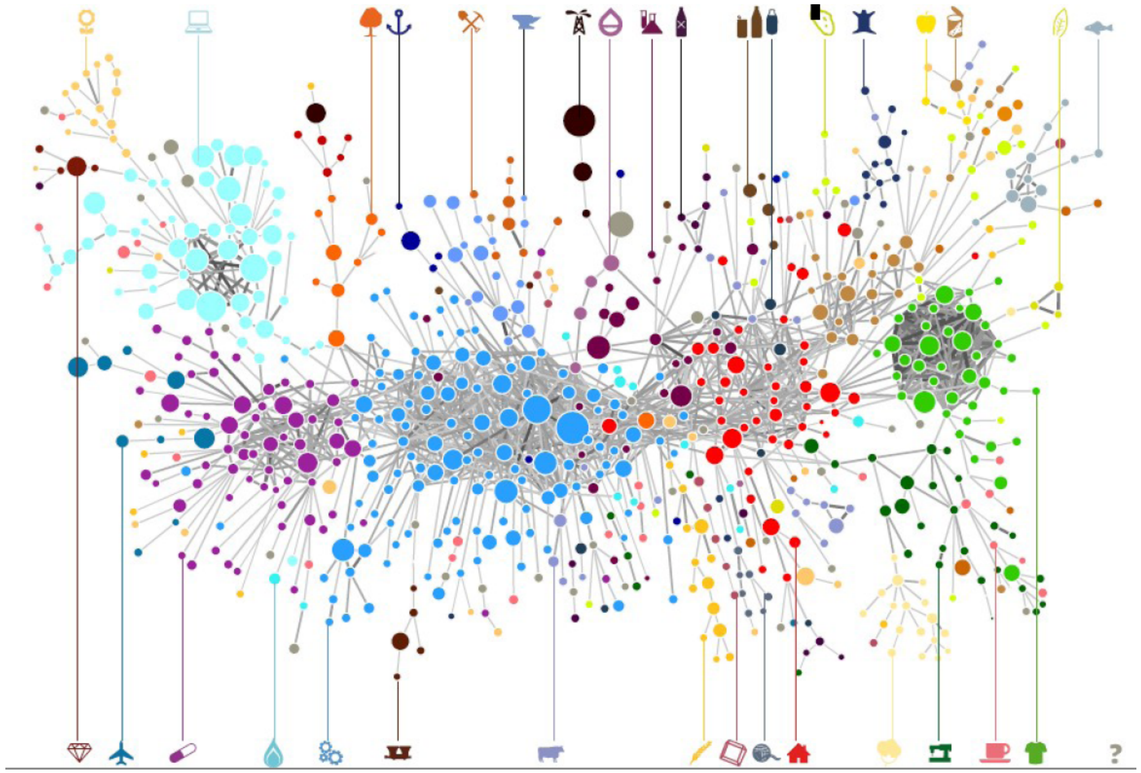
Social Network Analysis - Lecture 12

2162-F23

2023-03-15



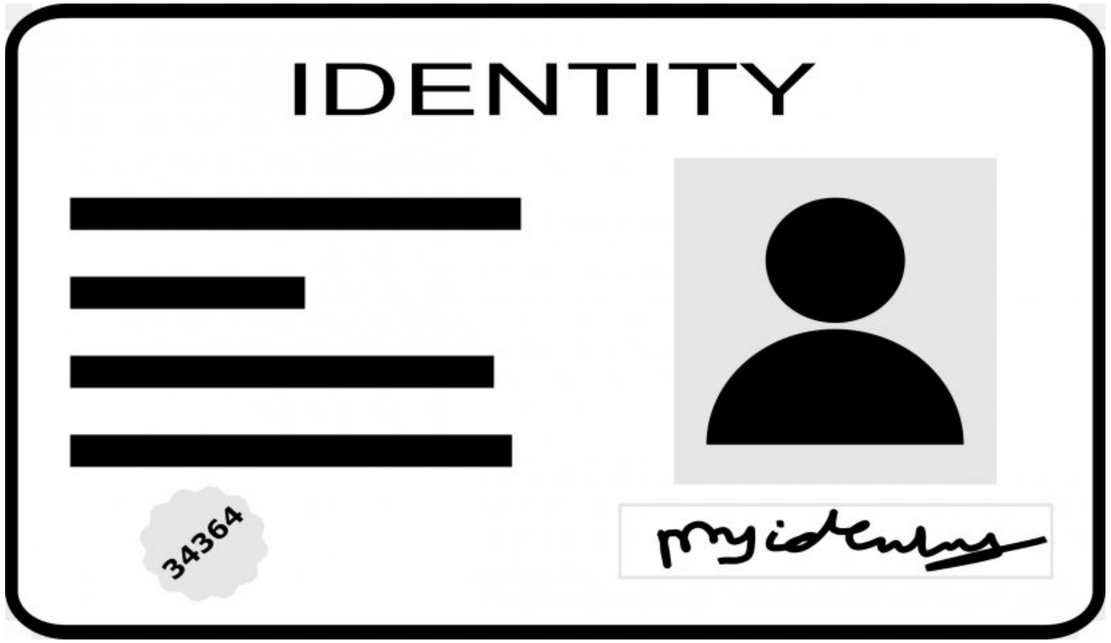
Product Space, from the Atlas of Economic Complexity

What do we do today?

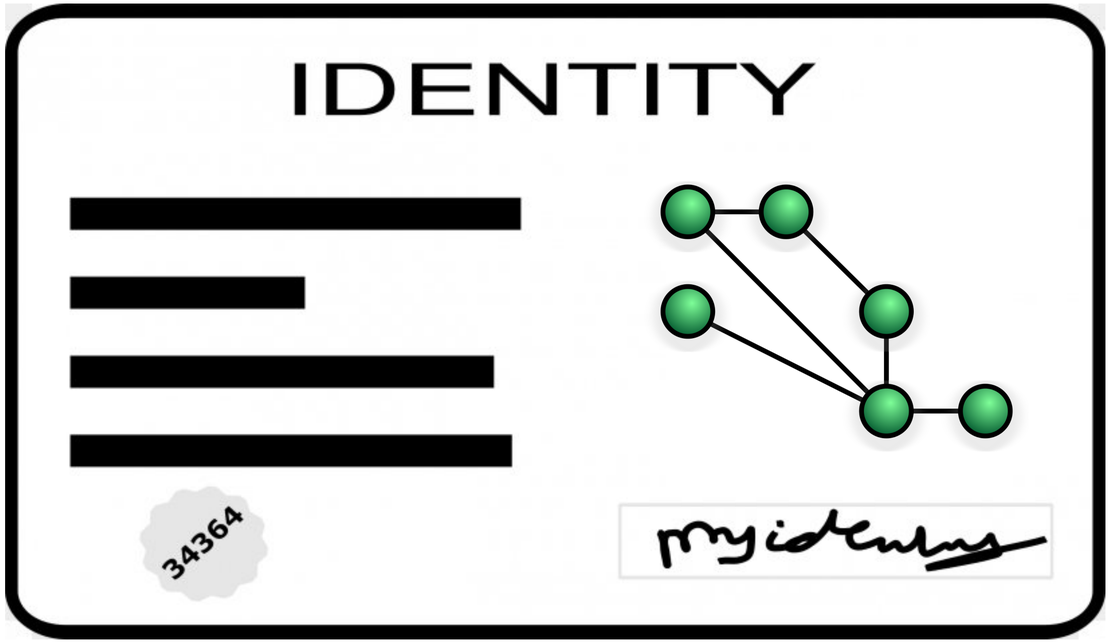
Basics of graph theory and graph measures

- Network identity card Part 1
- Network identity card Part 2
- Centralities

Identity card of a graph



Identity card of a graph



Nodes $N = 6$

Links $L = 6$

Clustering coefficient $< C >= ...$

Average path length $< d >= ...$

Average degree $< k >$

Density D

Diameter d_{max}

Why do we need these numbers?

These numbers alone mean very little

To give them a meaning, we need:

- To understand how they are constructed
- To have an expectation for these numbers
- To put these numbers in context and provide an expectation

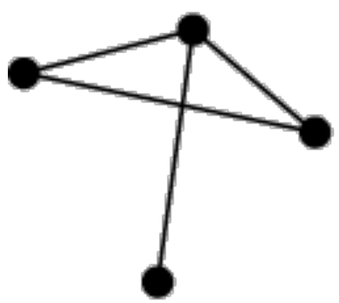
Identity card of a graph

Network	Nodes	Links	Directed / Undirected	N	L	‹K›
Internet	Routers	Internet connections	Undirected	192,244	609,066	.34
WWW	Webpages	Links	Directed	325,729	1,497,134	.60
Power Grid	Power plants, transformers	Cables	Undirected	4,941	6,594	.67
Mobile-Phone Calls	Subscribers	Calls	Directed	36,595	91,826	.51
Email	Email addresses	Emails	Directed	57,194	103,731	.81
Science Collaboration	Scientists	Co-authorships	Undirected	23,133	93,437	.08
Actor Network	Actors	Co-acting	Undirected	702,388	29,397,908	3.71
Citation Network	Papers	Citations	Directed	449,673	4,689,479	0.43
E. Coli Metabolism	Metabolites	Chemical reactions	Directed	1,039	5,802	.58
Protein Interactions	Proteins	Binding interaction	Undirected	2,018	2,930	.90

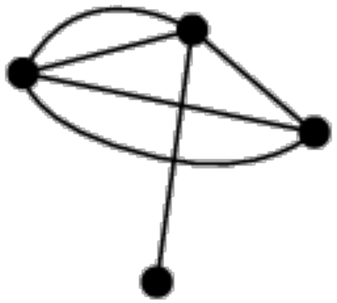
http://networksciencebook.com/chapter/2#networks-graphs

Graph types

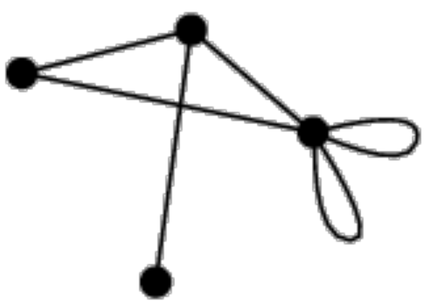
1. What is a simple graph? What is the difference with a multigraph? Draw examples for both.



simple graph



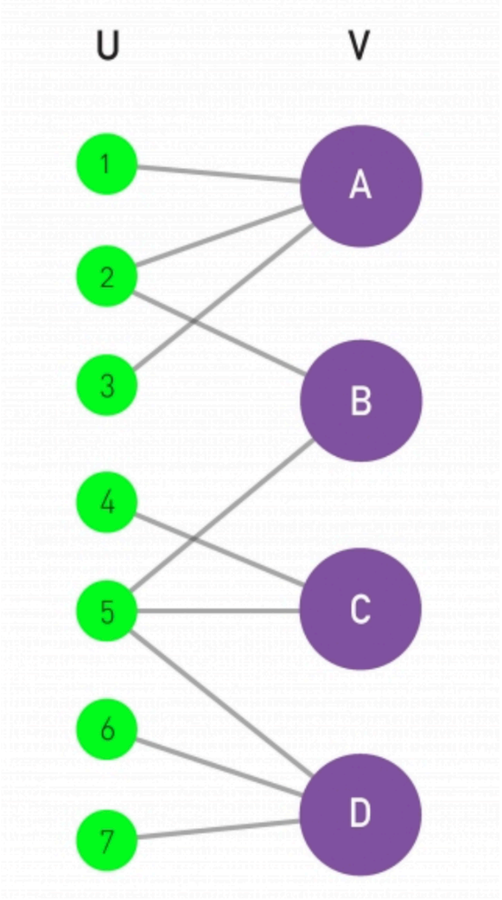
nonsimple graph with multiple edges



nonsimple graph with loops

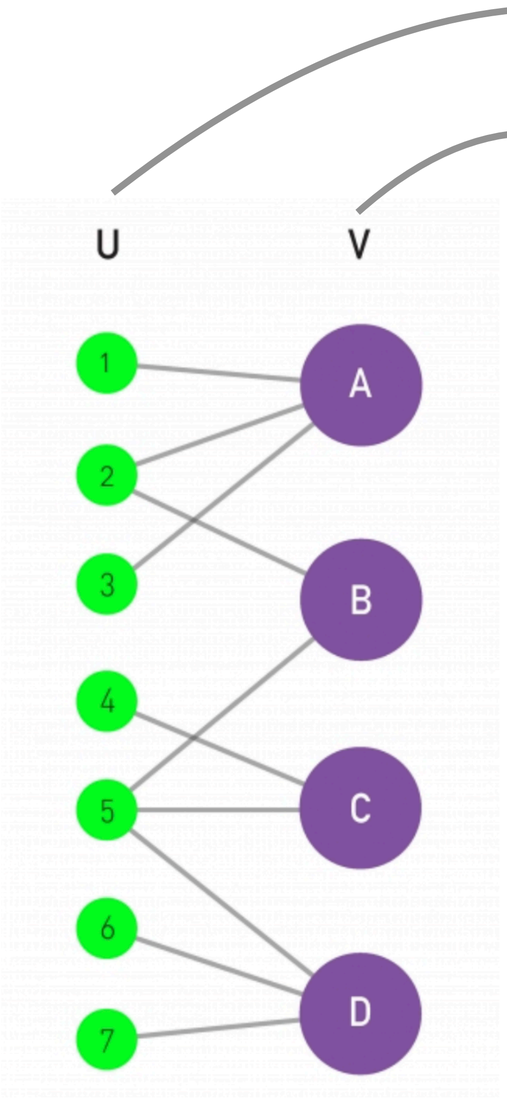
https://mathworld.wolfram.com/SimpleGraph.html

Unipartite - Bipartite



- Two node types
- Edges exclusively between unlike nodes
- Widely used, less studied

Unipartite - Bipartite

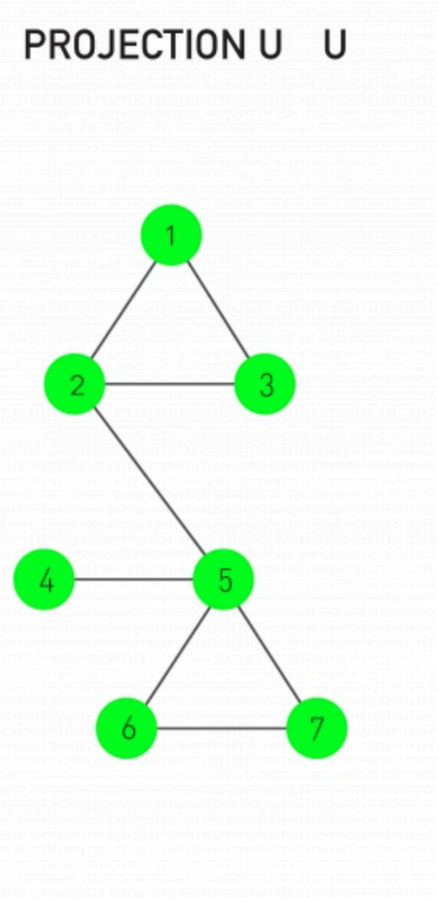


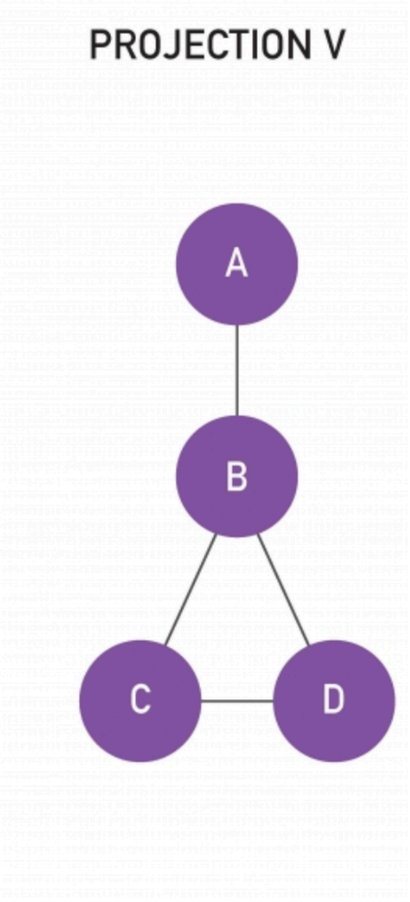
Readers	Books
Countries	Exports
Actor	Movie
Viewer	Video
Listener	Band

- Two node types
- Edges exclusively between unlike nodes
- Widely used, less studied

Unipartite - Bipartite

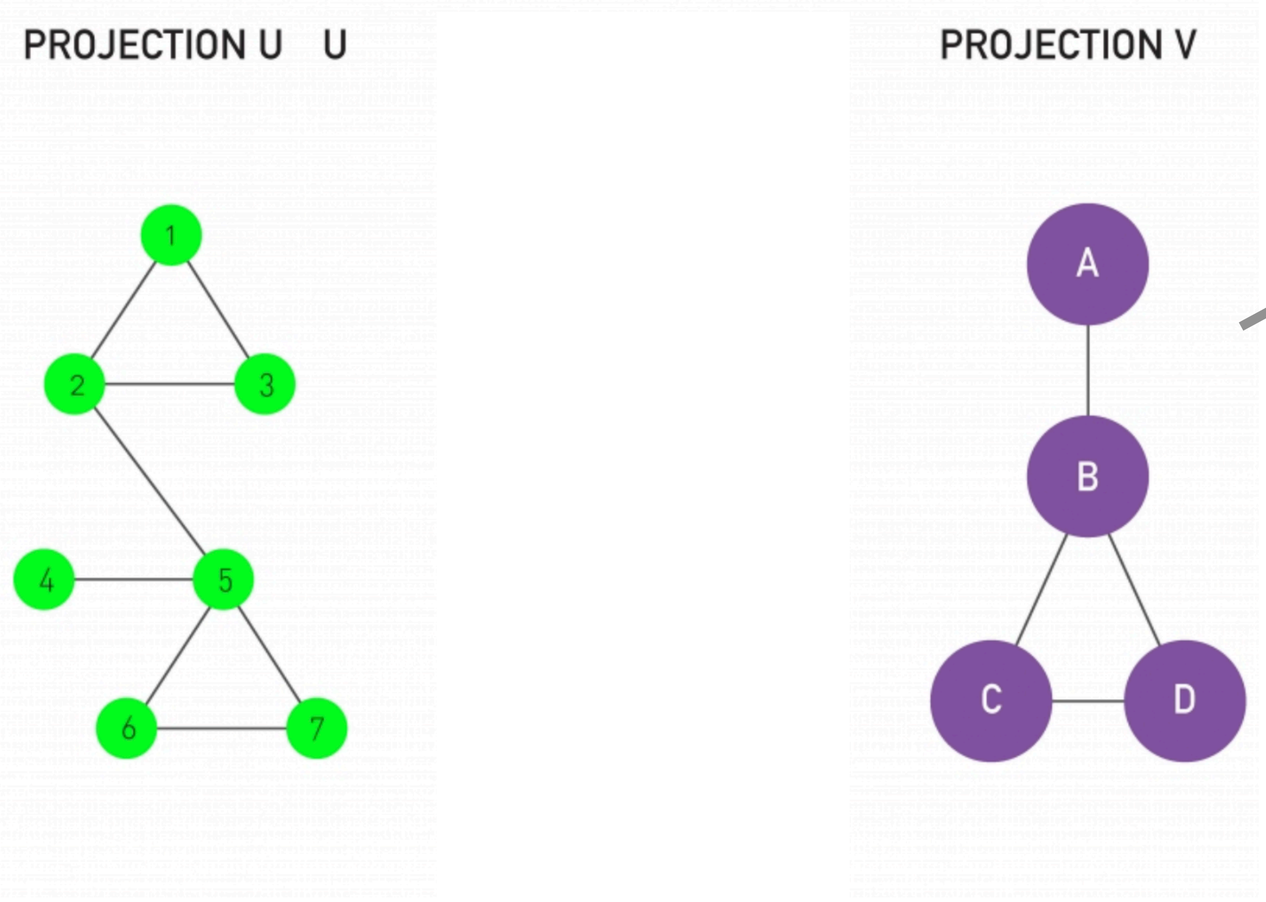
The nodes have their own projection

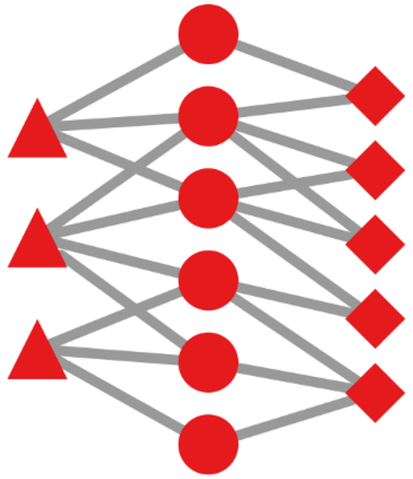




Unipartite - Bipartite

The nodes have their own projection

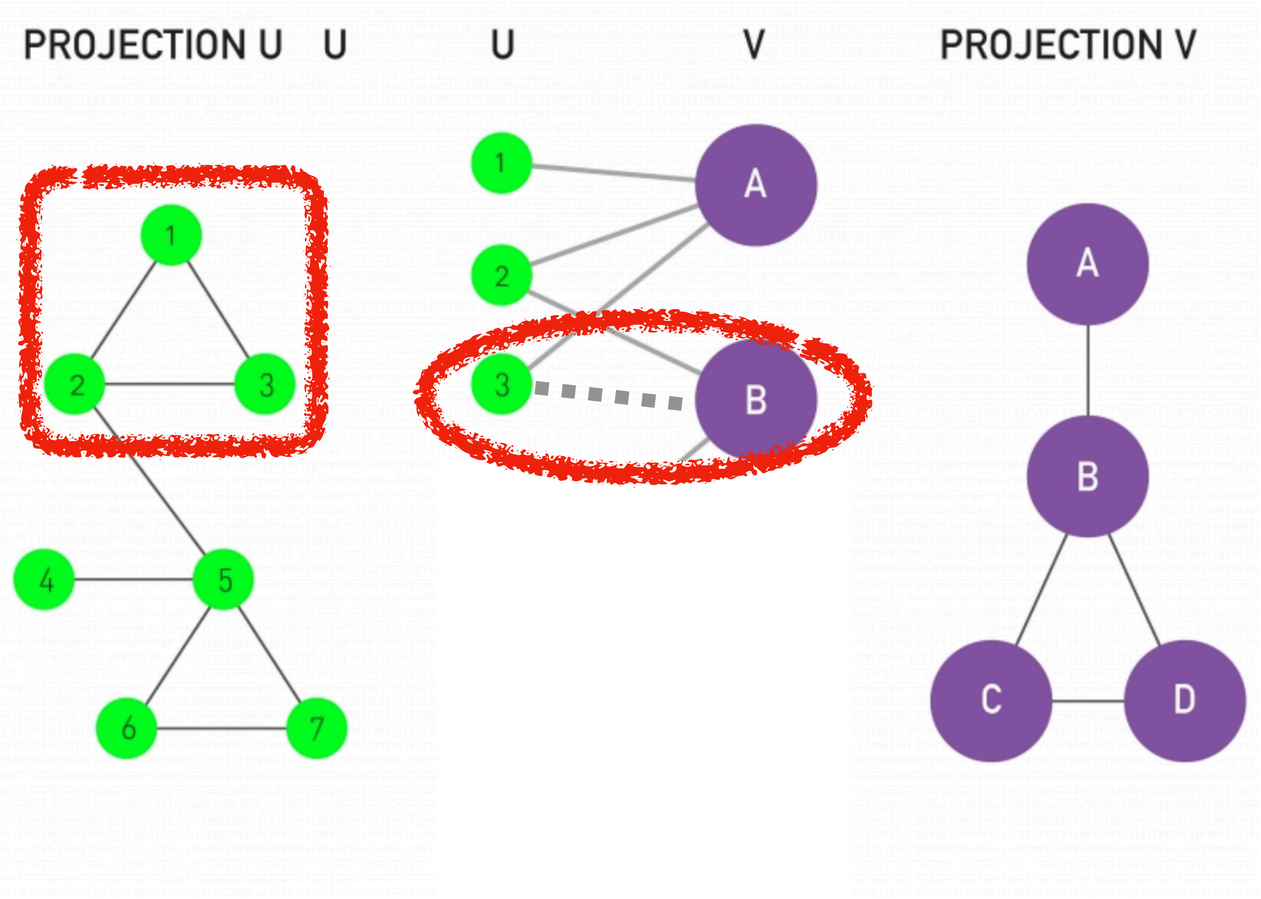




One can go n-partite, but marginal gains start building up at the cost of increasing complexity.

Unipartite - Bipartite

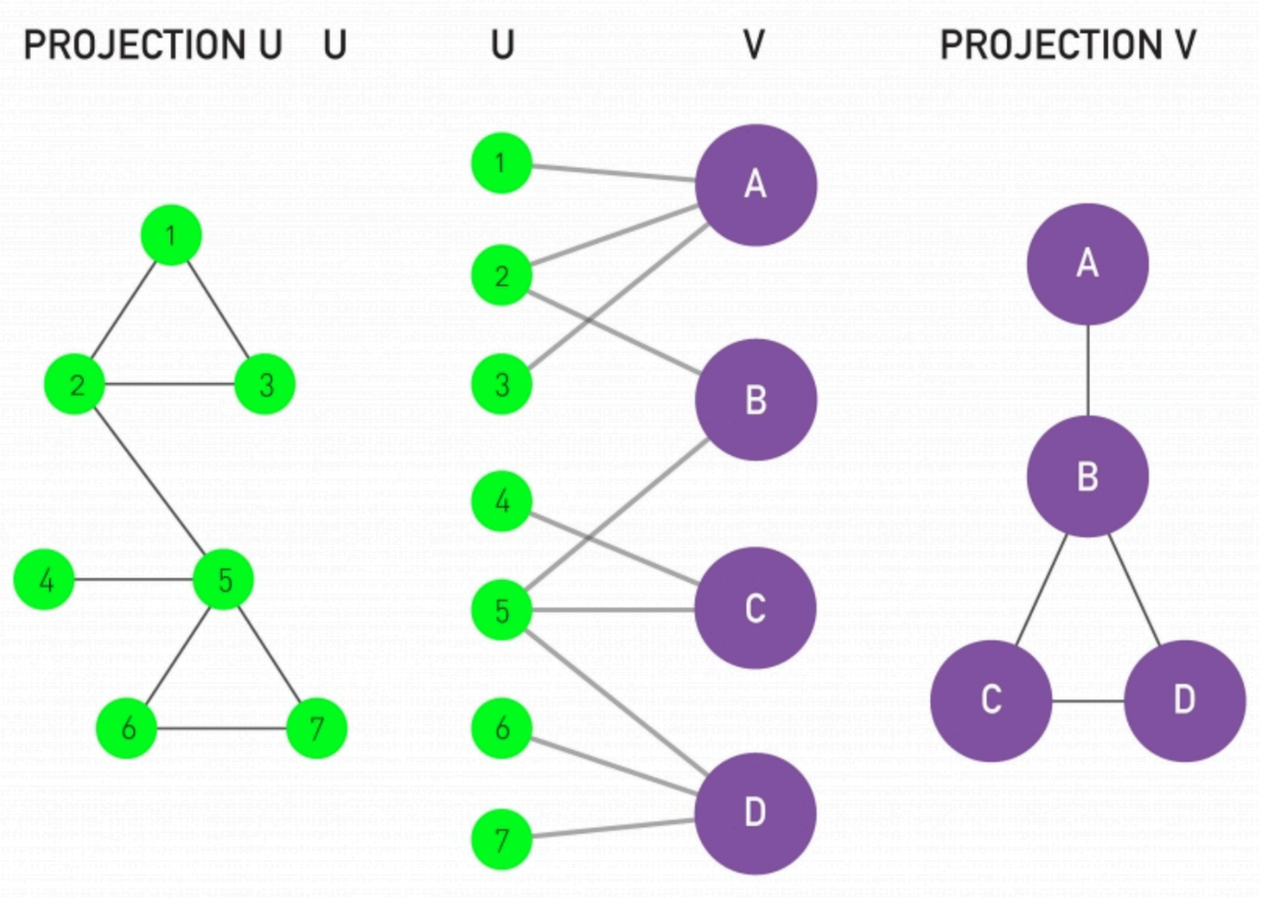
The nodes have their own projection



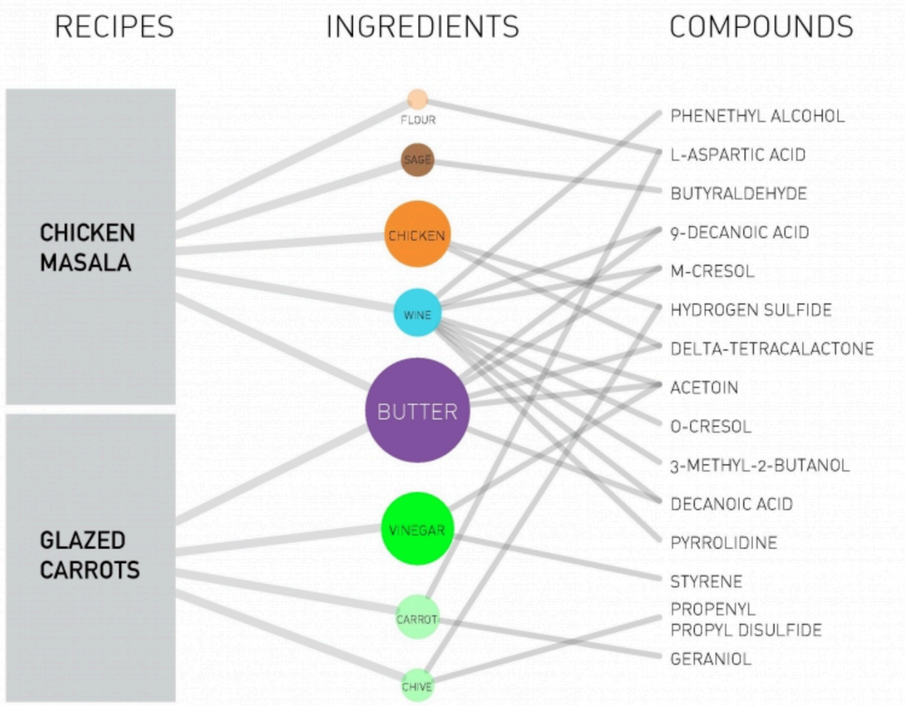
Recommendation system

Next time you see Netflix recommending stuff to watch, remember this!

Unipartite - Bipartite

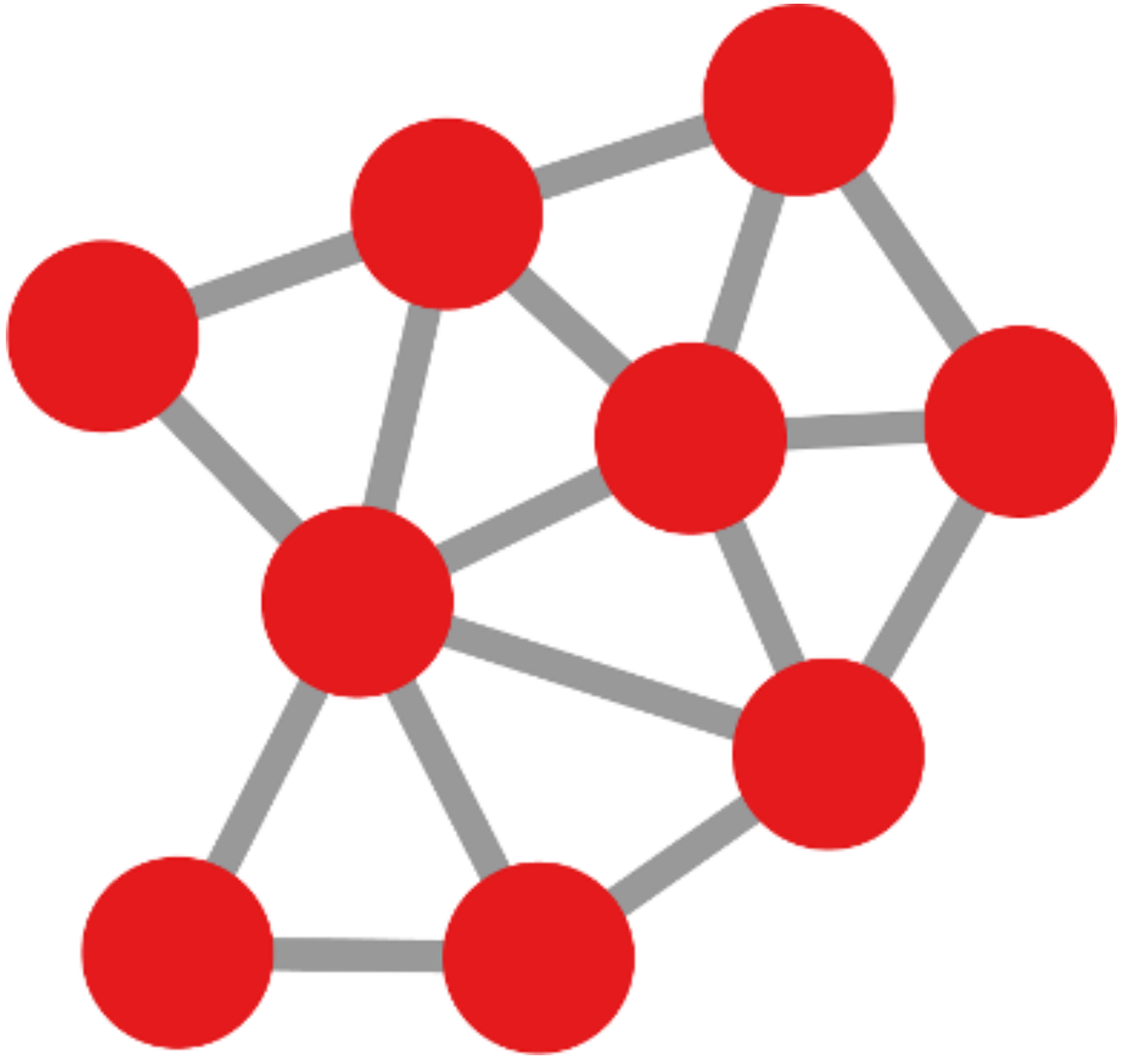


Recipe discovery



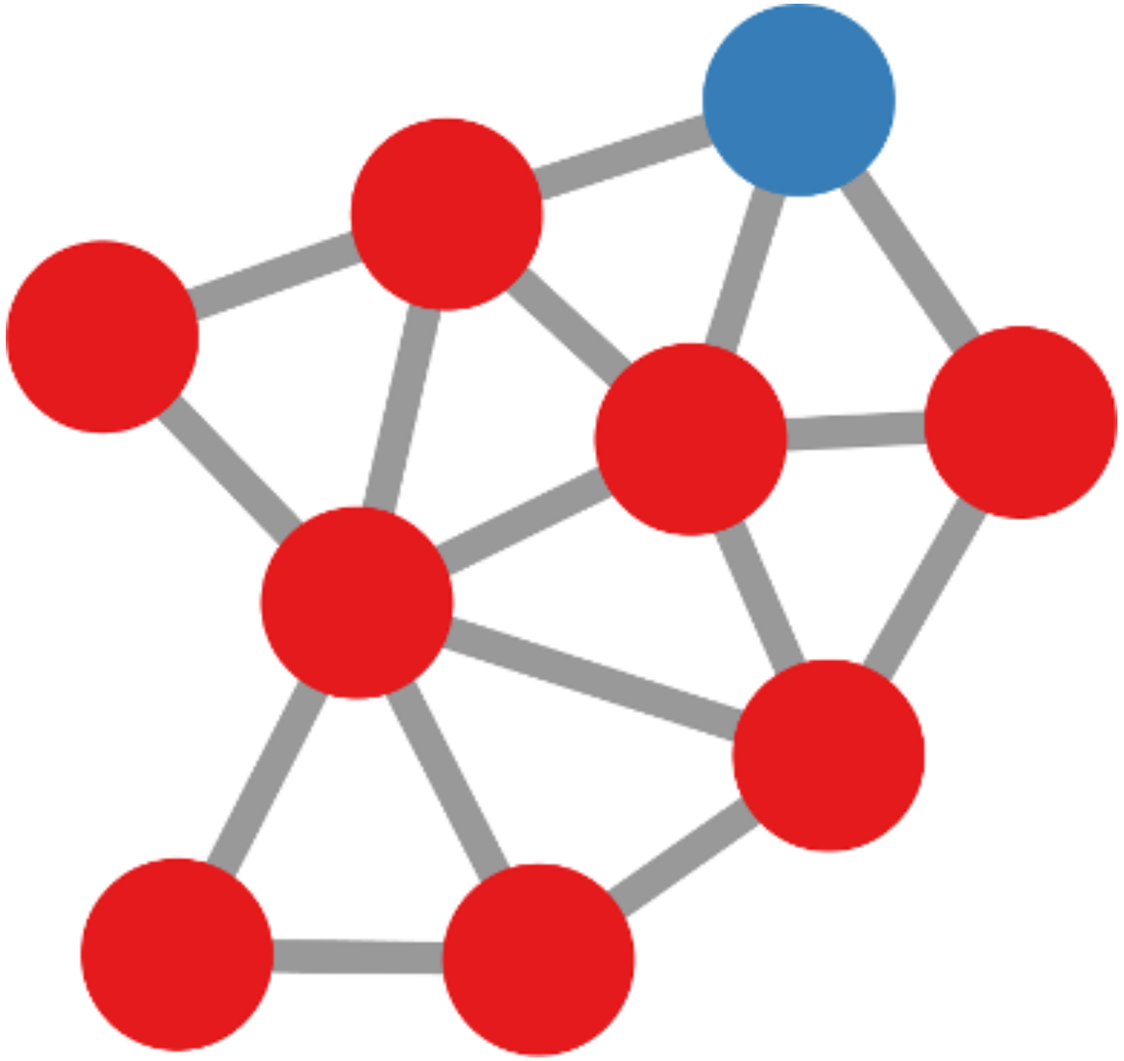
Shortest path detection

- Given a graph G



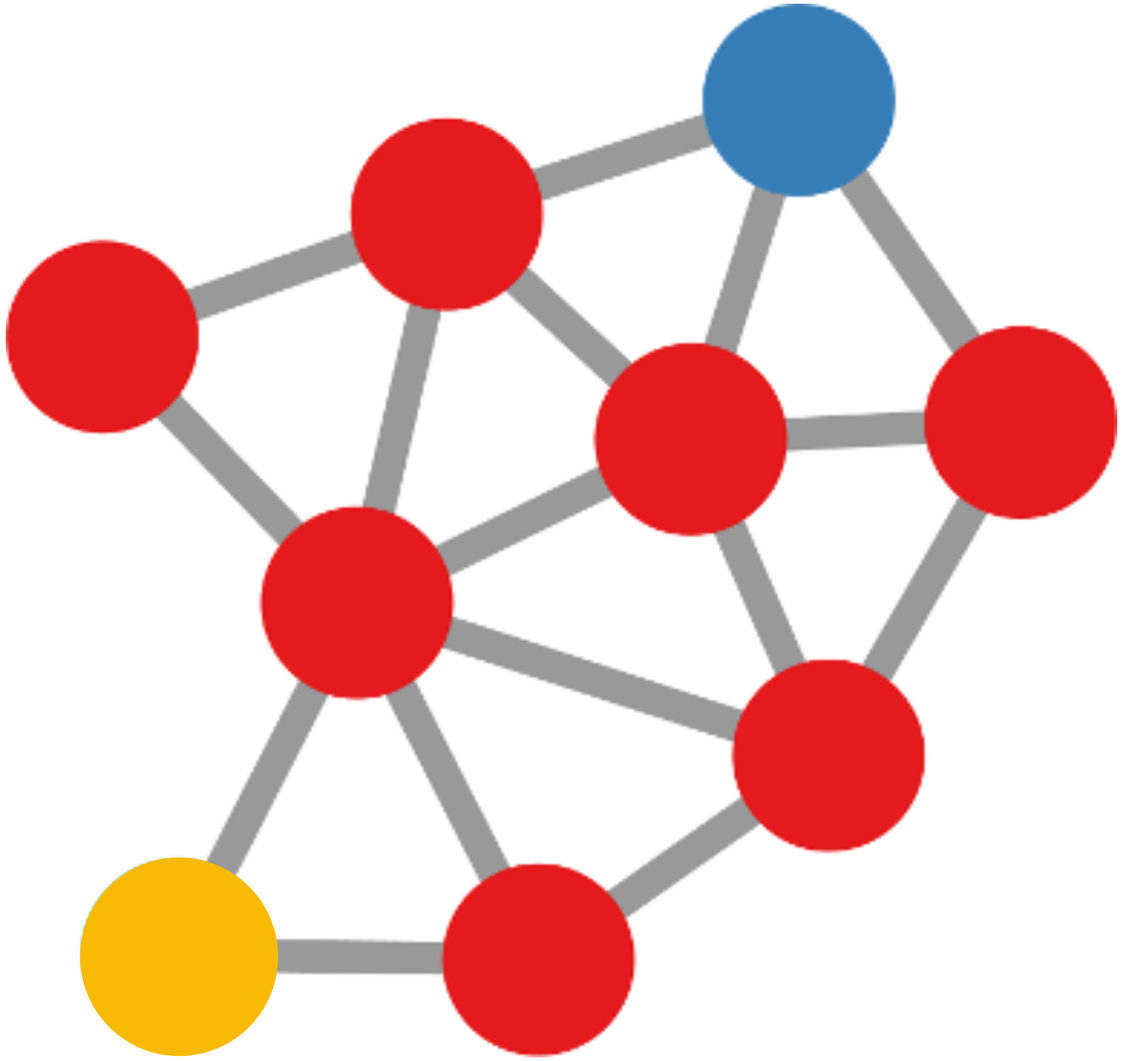
Shortest path detection

- Given a graph G
- A start at node i



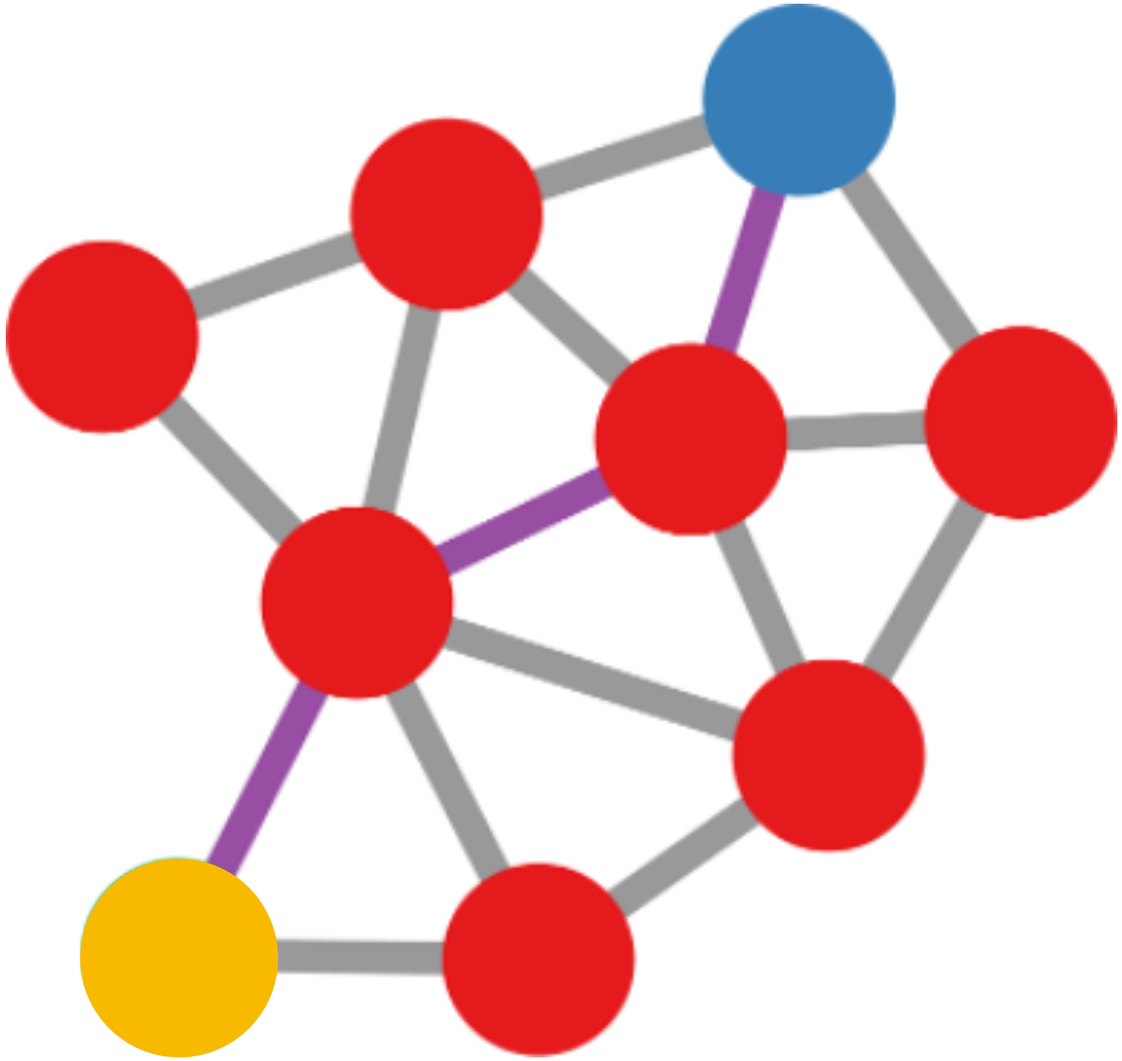
Shortest path detection

- Given a graph G
- A start at node i
- An end at node j



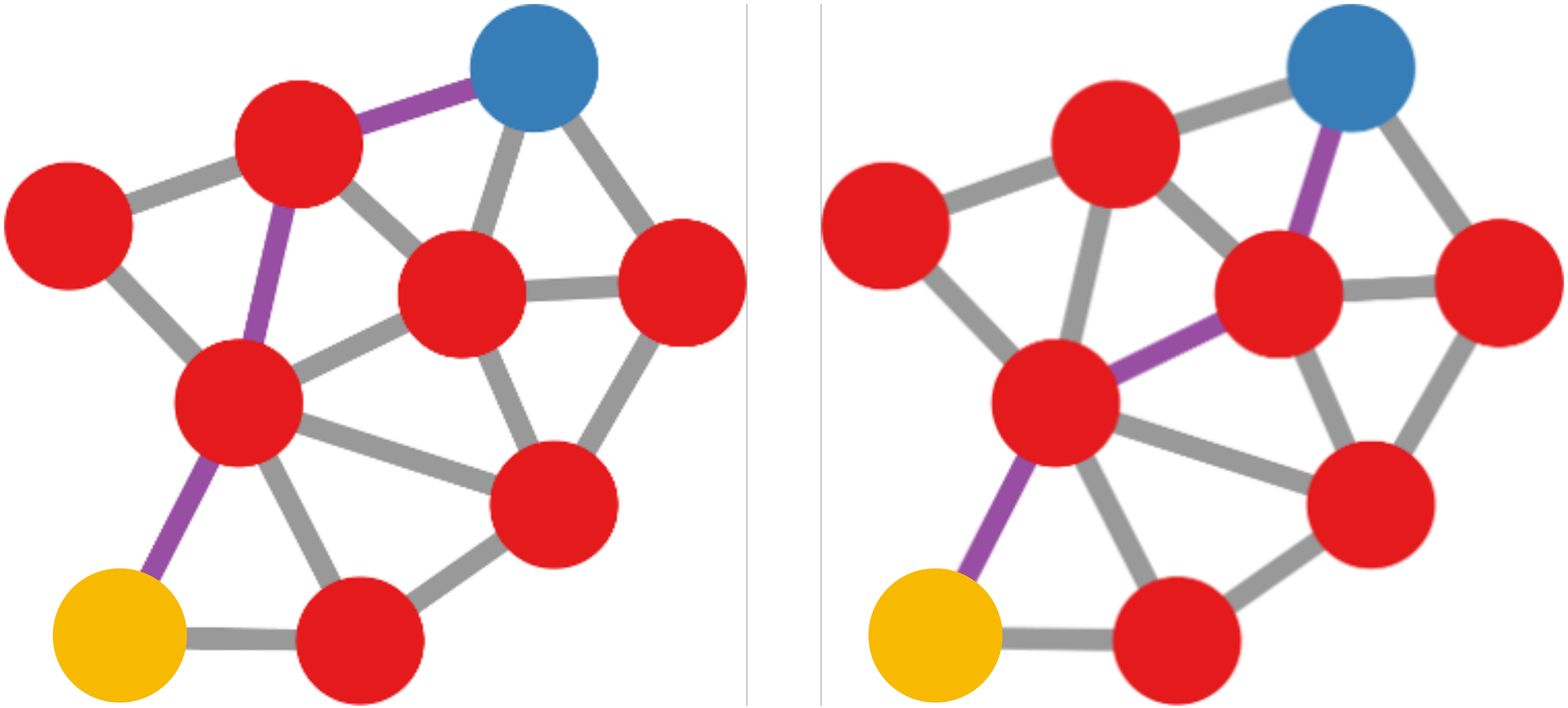
Shortest path detection

- Given a graph G
- A start at node i
- An end at node j
- Find the path crossing the least number of edges



Shortest path detection

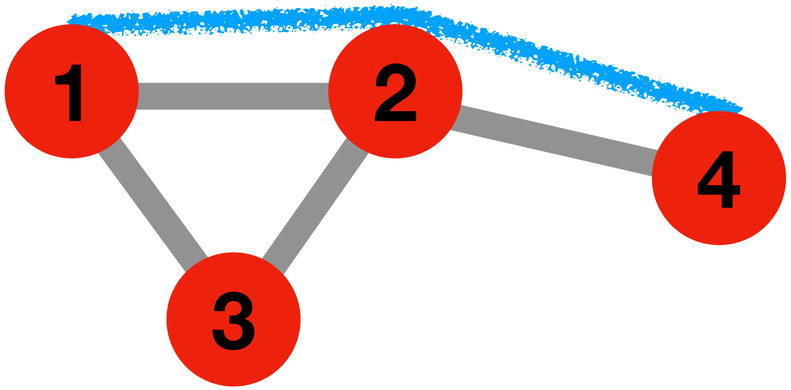
Not Necessarily Unique



Shortest path detection

Let's work on a small example…

$$d_{ij} = d_{14}(\text{blue}) = 2 \quad \{(1,2),(2,4)\}$$



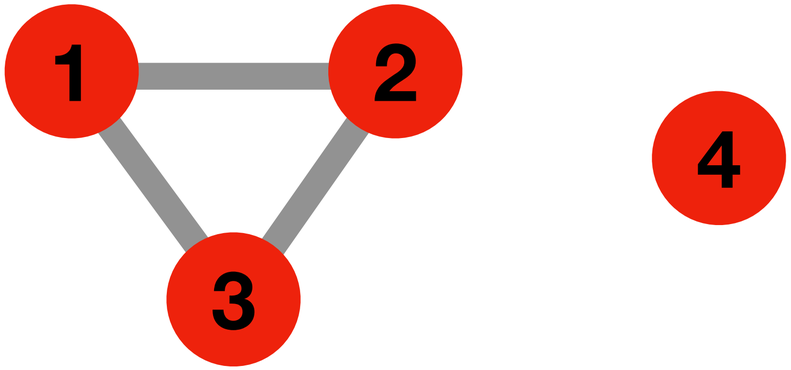
Shortest path & Connectedness

A node ***i*** is reachable from a node ***j*** if there exists a path connecting ***i*** to ***j***.

$d_{ij} = \infty$ if there's no path connecting ***i*** to ***j***.

A graph is connected if every node is reachable from every other node.

An isolated node is a node with 0 connections (node **4**)

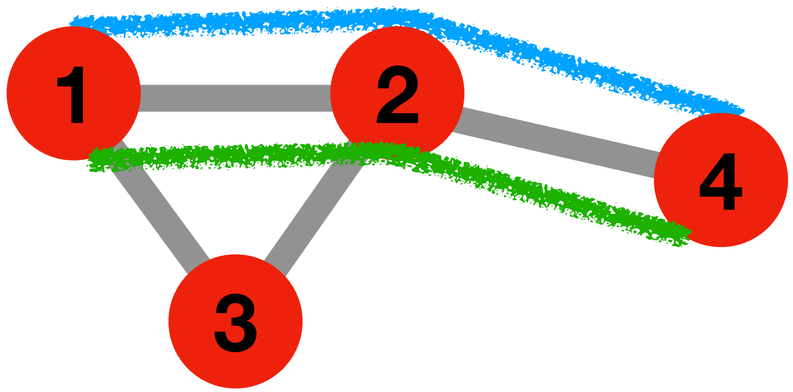


Shortest path & Diameter

Going through all shortest paths…

… tells the **Diameter** of the network.

$$d_{max} = max(d_{ij}) \text{ (green)} = 2$$



Diameter

Measure of max separation

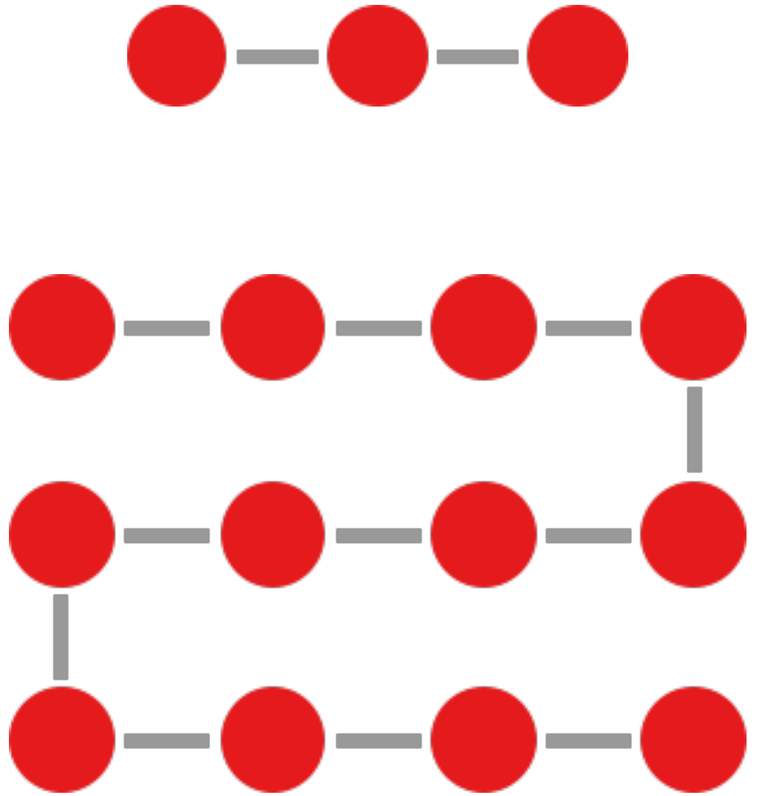
Small Diameter → Everybody is reachable

Diameter

Measure of max separation

Small Diameter → Everybody is reachable

Large Diameter → Full traversal might be impossible



Diameter

- Diameter = 1
- Diameter = 2
- Diameter = 3

- You know everyone
- Your friends know everyone
- Your friends know someone who knows everyone

Diameter of real networks

What is the difference between the average shortest path and the diameter?

Network	N	L	‹k›	‹d›	d_{max}
Internet	192,244	609,066	6.34	6.98	26
WWW	325,729	1,497,134	4.60	11.27	93
Power Grid	4,941	6,594	2.67	18.99	46
Mobile-Phone Calls	36,595	91,826	2.51	11.72	39
Email	57,194	103,731	1.81	5.88	18
Science Collaboration	23,133	93,437	8.08	5.35	15
Actor Network	702,388	29,397,908	83.71	3.91	14
Citation Network	449,673	4,707,958	10.43	11.21	42
E. Coli Metabolism	1,039	5,802	5.58	2.98	8
Protein Interactions	2,018	2,930	2.90	5.61	14

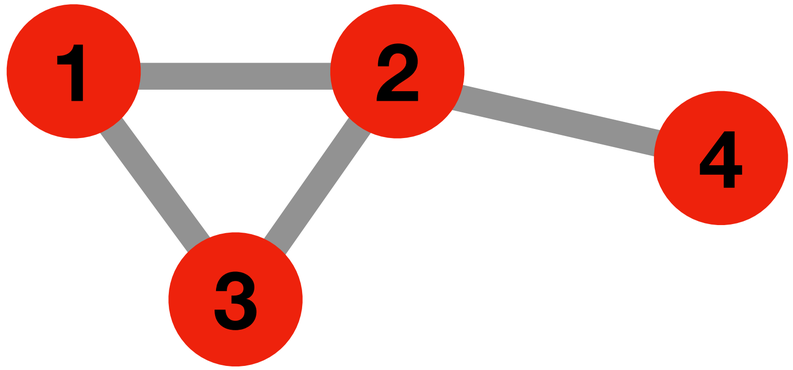
http://networksciencebook.com/chapter/3#small-worlds

Diameter & Average path length

In the worst-case scenario:

Diameter (d_{max}) = $< d >$ (average path length)

In practice, we calculate $< d >$



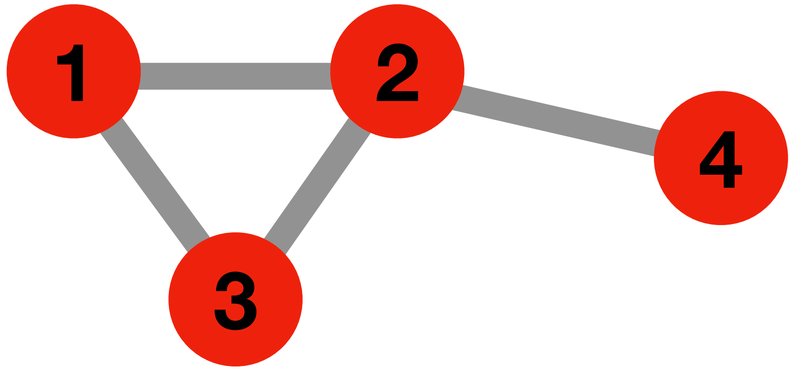
Average path length

In practice:

$$d_{1j} = 1+1+2$$

...

$$d_{4j} = 2+1+2$$



Average path length

The characteristic (or average) path length is the mean graph distance over all pairs of nodes.

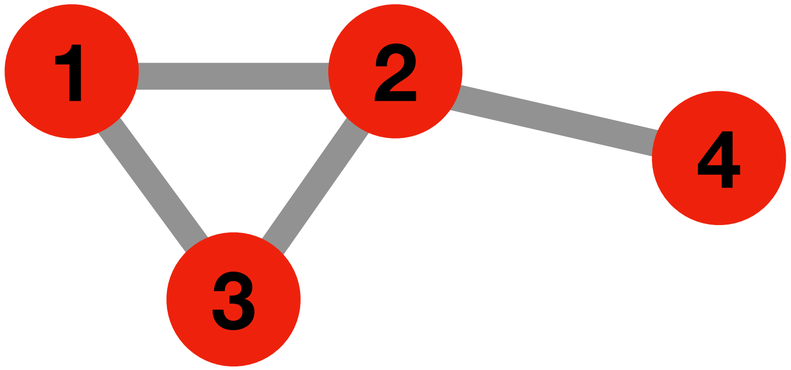
$$< d > = \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j=1}^{N} d_{i,j}$$

In practice:

$$d_{1j} = 1+1+2$$

...

$$d_{4j} = 2+1+2$$



Average path length

The characteristic (or average) path length is the mean graph distance over all pairs of nodes.

$$< d > = \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j=1}^{N} d_{i,j}$$

In practice:

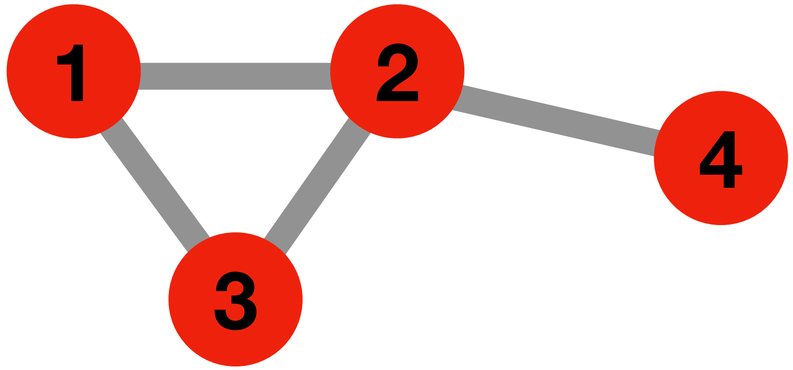
$$d_{1j} = 1+1+2$$

...

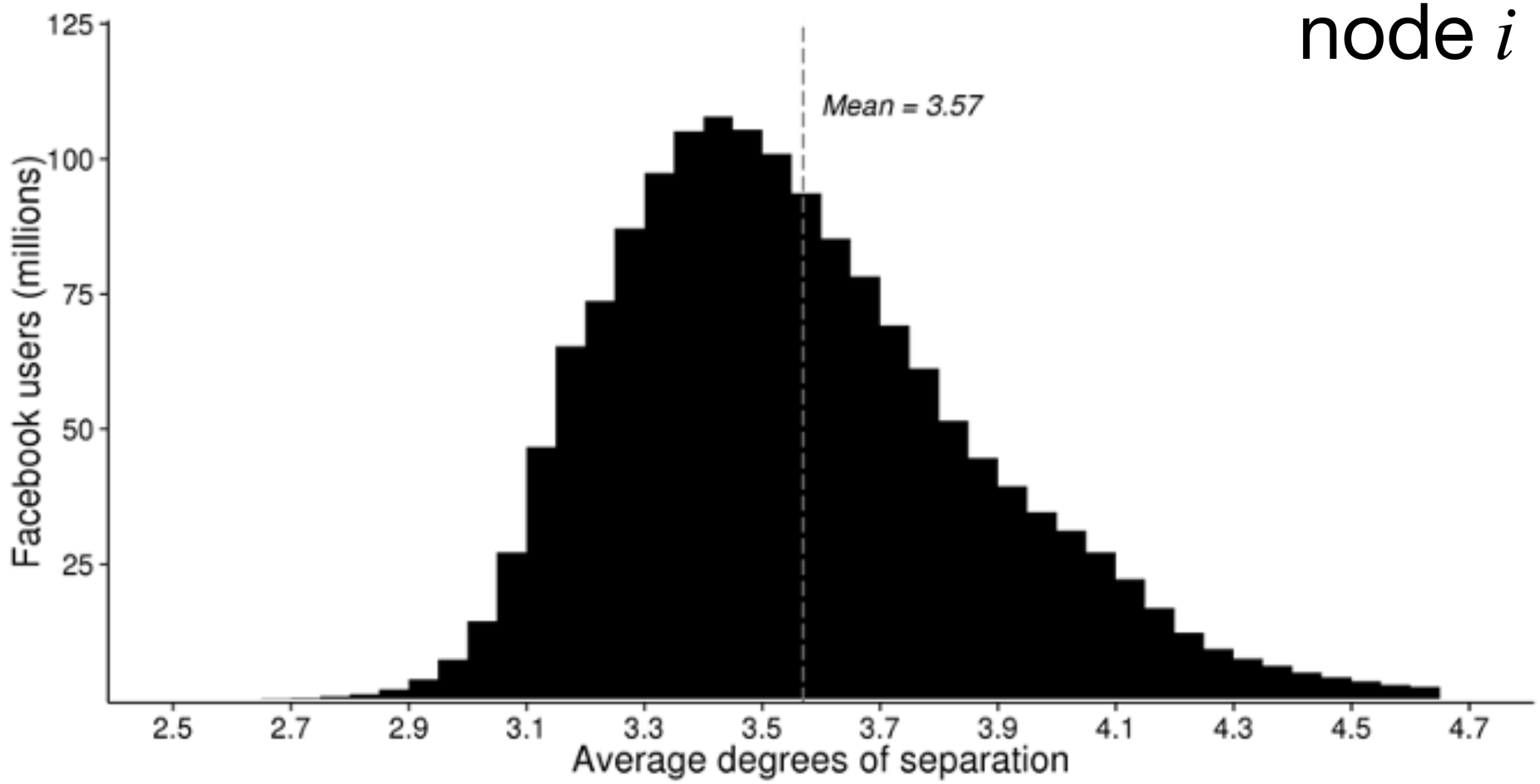
$$d_{4j} = 2+1+2$$

Example:

$$< d > = \frac{4+3+4+5}{4(4-1)} = 1.33$$



Average path length



If we fix a node i

Avg. path length of real networks

What is the difference between the average shortest path and the diameter?

Network	N	L	‹k›	‹d›	d_{max}
Internet	192,244	609,066	6.34	6.98	26
WWW	325,729	1,497,134	4.60	11.27	93
Power Grid	4,941	6,594	2.67	18.99	46
Mobile-Phone Calls	36,595	91,826	2.51	11.72	39
Email	57,194	103,731	1.81	5.88	18
Science Collaboration	23,133	93,437	8.08	5.35	15
Actor Network	702,388	29,397,908	83.71	3.91	14
Citation Network	449,673	4,707,958	10.43	11.21	42
E. Coli Metabolism	1,039	5,802	5.58	2.98	8
Protein Interactions	2,018	2,930	2.90	5.61	14

http://networksciencebook.com/chapter/3#small-worlds

Density

Density is the ratio of the number of edges $\mathrm{L} = |\mathrm{E}|$ to the maximum possible edges.

... maximum possible edges?

Density

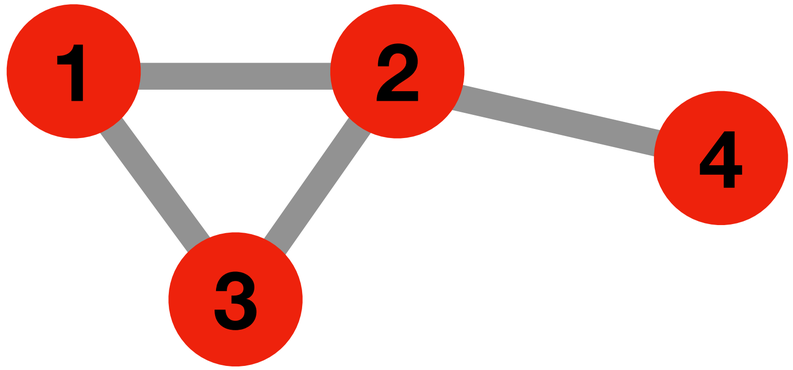
Density is the ratio of the number of edges $\mathrm{L} = |\mathrm{E}|$ to the maximum possible edges.

$$\binom{|V|}{2} = \frac{|V|(|V|-1)}{2}$$

where $\mathrm{N} = |\mathrm{V}|$ nodes.

Binomial (combine all nodes)

... maximum possible edges?



Density

Density is the ratio of the number of edges $\mathrm{L} = |\mathrm{E}|$ to the maximum possible edges.

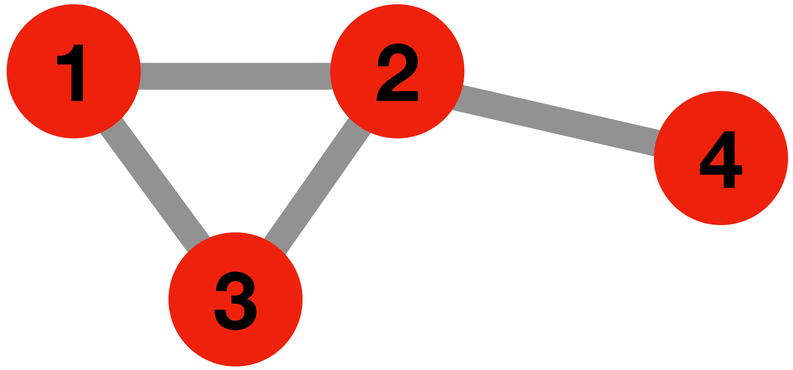
$$\frac{4 * (4 - 1)}{2} = \frac{12}{2} = 6 \quad \frac{4_{current}}{6_{possible}} = 0.66$$

... maximum possible edges?

$$\binom{|V|}{2} = \frac{|V|(|V| - 1)}{2}$$

Binomial (combine all nodes)

where $\mathrm{N} = |\mathrm{V}|$ nodes.



Density

Density is the ratio of the number of edges $\mathrm{L} = |\mathrm{E}|$ to the maximum possible edges.

For undirected graphs:

$$D = \frac{|E|}{\binom{|V|}{2}} = \frac{2|E|}{|V|(|V|-1)}$$

For directed graphs:

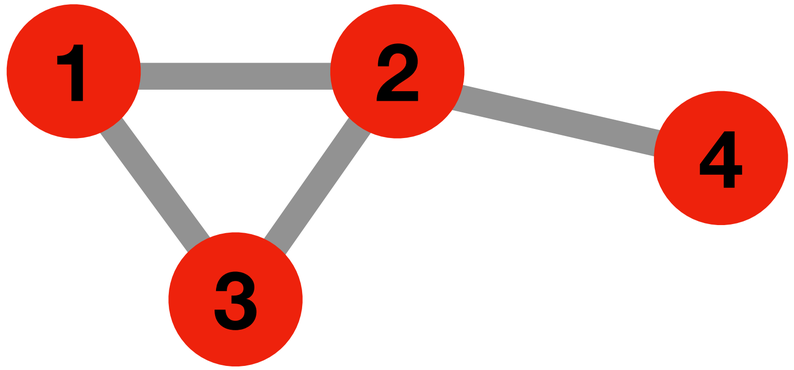
$$D = \frac{|E|}{2\binom{|V|}{2}} = \frac{|E|}{|V|(|V|-1)}$$

$$\binom{|V|}{2} = \frac{|V|(|V|-1)}{2}$$

Binomial (combine all nodes)

where $\mathrm{N} = |\mathrm{V}|$ nodes.

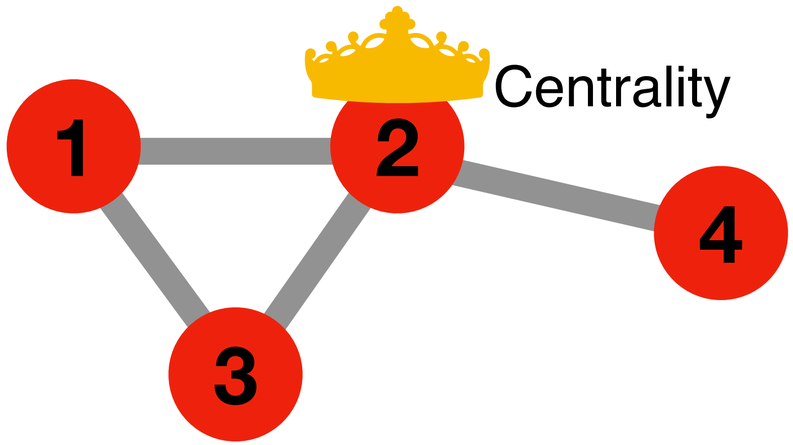
... maximum possible edges?



Degree

Degree (of a node ***i***) is the number of incident links

k_i

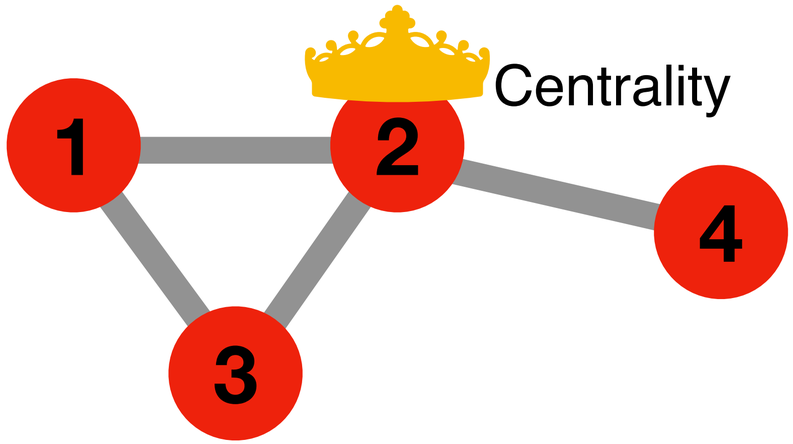


Degree

Degree (of a node ***i***) is the number of incident links k_i

In this example:

$$k_1 = 2 \qquad k_3 = 2$$
$$k_2 = 3 \qquad k_4 = 1$$



Degree

Degree (of a node i) is the number of incident links k_i

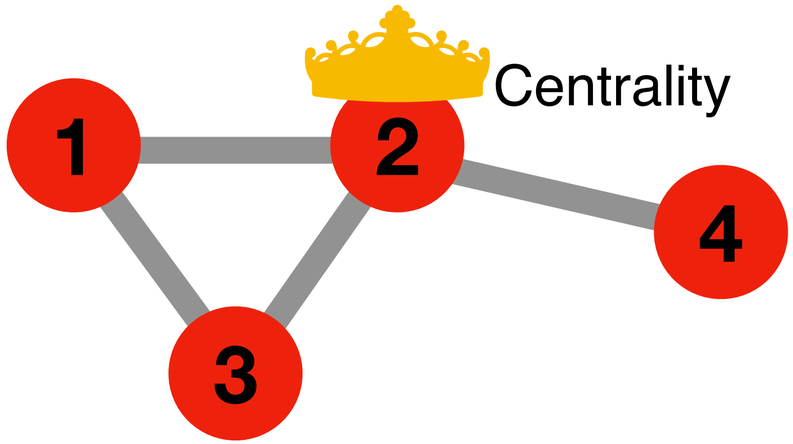
In this example:

$$k_1 = 2 \qquad k_3 = 2$$
$$k_2 = 3 \qquad k_4 = 1$$

Average degree (of the network)

$$\langle k \rangle = \frac{2+3+2+1}{4} = 2$$

Nodes N = 4
Links L = 4



Degree

Degree (of a node ***i***) is the number of incident links

k_i

In this example:

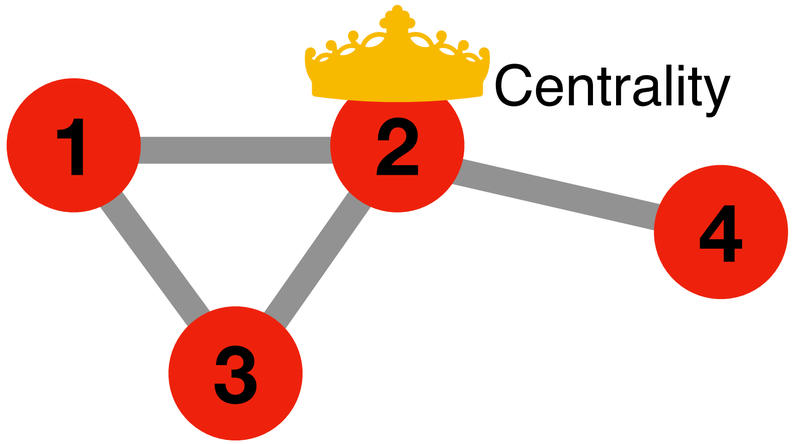
$$k_1 = 2 \qquad k_3 = 2$$
$$k_2 = 3 \qquad k_4 = 1$$

Average degree (of the network)

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

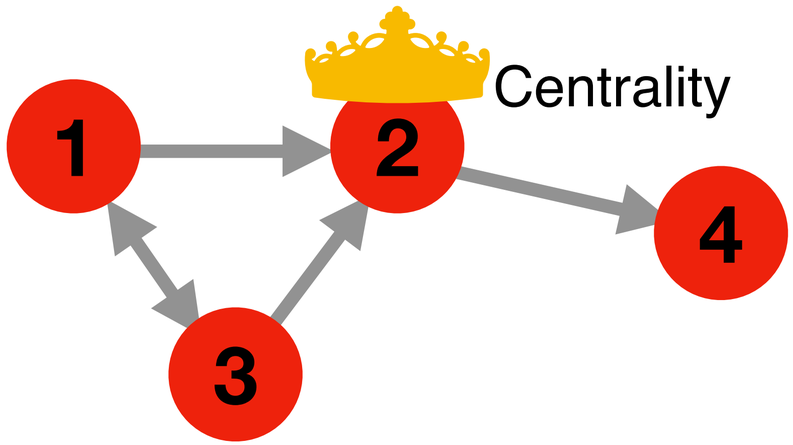
$$\langle k \rangle = \frac{2+3+2+1}{4} = 2$$

Nodes $N = 4$
Links $L = 4$



Indegree, Outdegree

What if the network is directed?

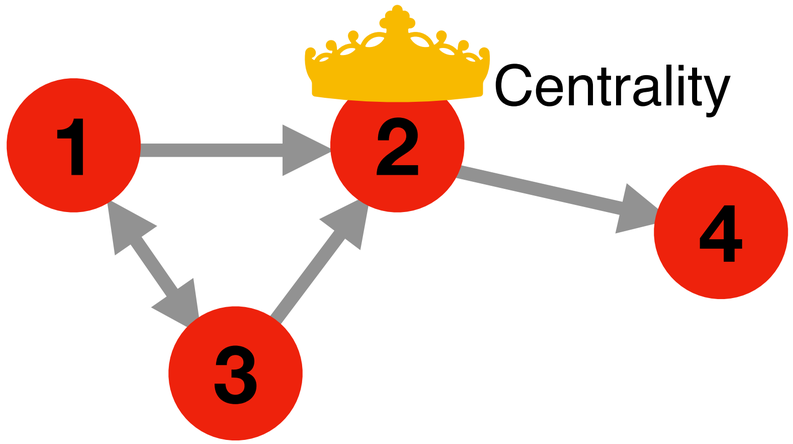


Indegree, Outdegree

Indegree (of a node ***i***) is the number of incoming links k_i^{in}

In this example:

$$k_1^{\text{in}} = 1 \qquad k_3^{\text{in}} = 1$$
$$k_2^{\text{in}} = 2 \qquad k_4^{\text{in}} = 1$$



Indegree, Outdegree

Indegree (of a node ***i***) k_i^{in}
is the number of
incoming links

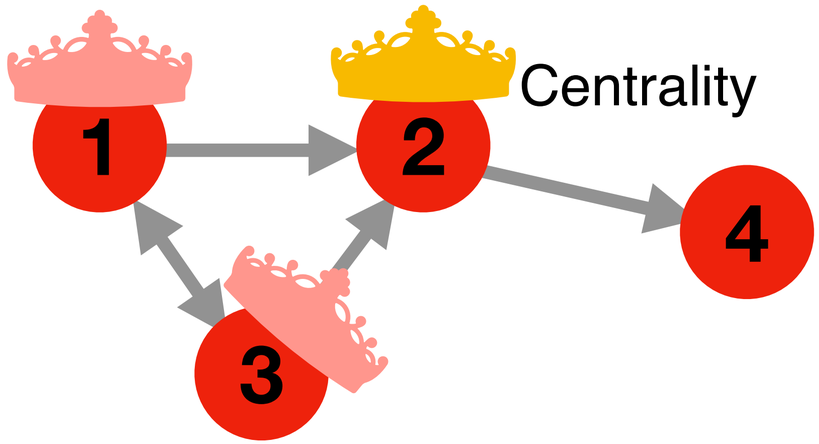
Outdegree (of a node ***i***) k_i^{out}
is the number of
outgoing links

$$k_i = k_i^{\text{in}} + k_i^{\text{out}}$$

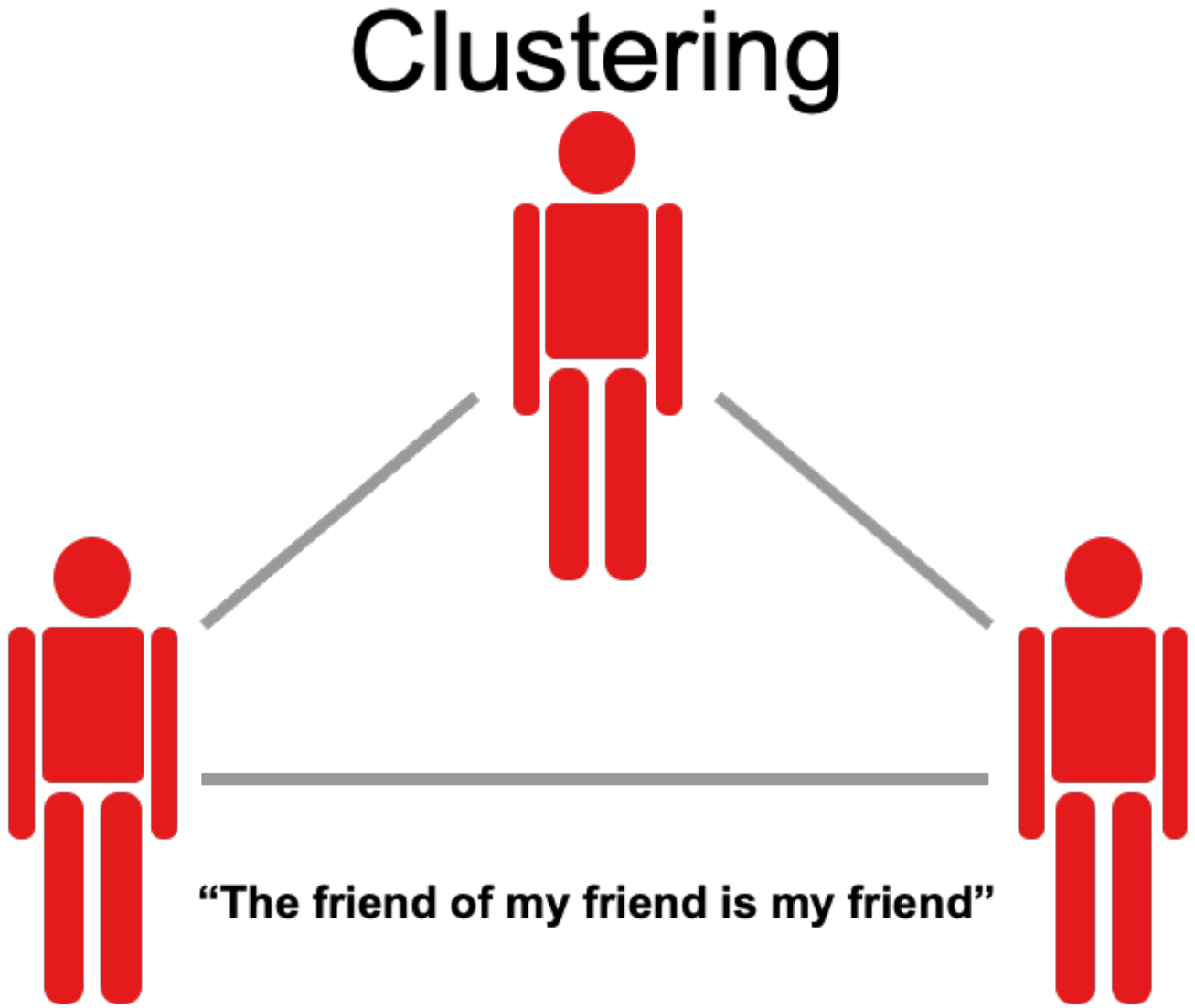
In this example:

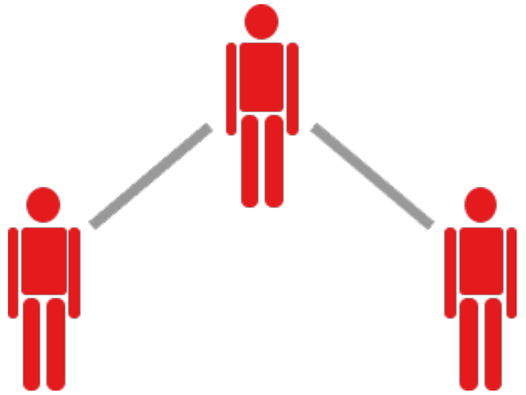
$$k_1^{\text{in}} = 1 \qquad k_3^{\text{in}} = 1$$
$$k_2^{\text{in}} = 2 \qquad k_4^{\text{in}} = 1$$

$$k_1^{\text{out}} = 2 \qquad k_3^{\text{out}} = 2$$
$$k_2^{\text{out}} = 1 \qquad k_4^{\text{out}} = 0$$



Clustering



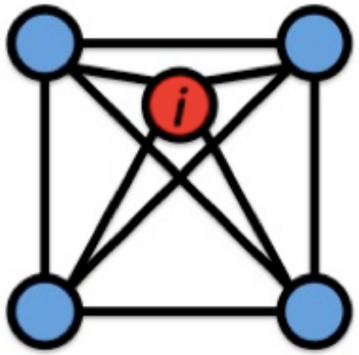


How many of these turn into triangles

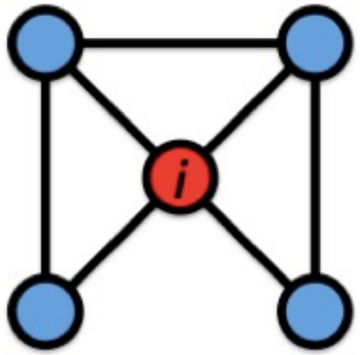
<—

Clustering

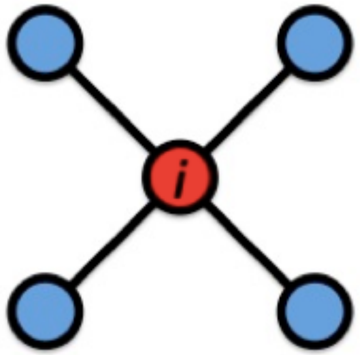
What fraction of your neighbours are connected?



$c_i = 1$



$c_i = 1/2$



$c_i = 0$

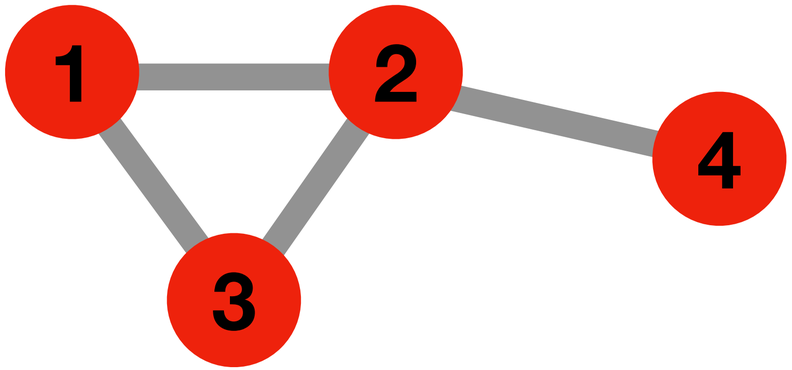
By definition, $0 \leq c_i \leq 1$

Clustering coefficient (local and global)

Clustering coefficient (local, of a node i)

If $k_i \geq 2$ $\quad c_i = \dfrac{2\mathrm{L_i}}{\mathrm{k_i}(\mathrm{k_i}-1)}$ $\quad$ else: $\quad c_i = 0$

where L_i represents the number of links between the k_i neighbours of node ***i***, *i.e.*, the subgraph induced by the neighbours of ***i***.



Clustering coefficient (local and global)

Clustering coefficient (local, of a node i)

If $k_i \geq 2$ $\quad c_i = \dfrac{2\mathrm{L_i}}{\mathrm{k_i}(\mathrm{k_i}-1)}$ $\quad$ else: $\quad c_i = 0$

where L_i represents the number of links between the k_i neighbours of node ***i***, *i.e.*, the subgraph induced by the neighbours of ***i***.

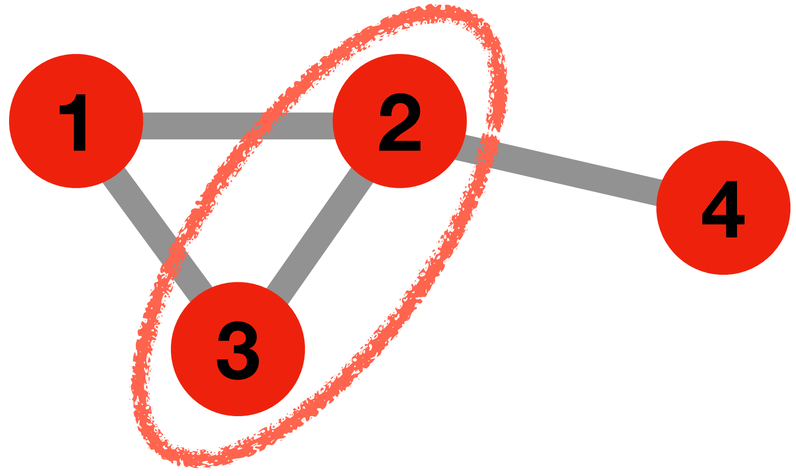
In the example:

c_1 = 2*1/(2*(2-1) = 1

…

…

c_4 = 0



Clustering coefficient (local and global)

Clustering coefficient (local, of a node i)

If $k_i \geq 2$ $\quad c_i = \dfrac{2\mathrm{L_i}}{\mathrm{k_i}(\mathrm{k_i}-1)}$ $\quad$ else: $\quad c_i = 0$

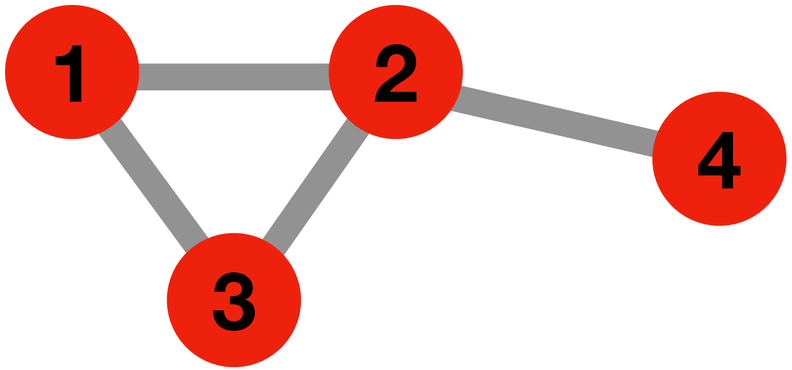
where L_i represents the number of links between the k_i neighbours of node ***i***, *i.e.*, the subgraph induced by the neighbours of ***i***.

The average clustering coefficient of the network is the average of local coefficients:

$$< C > = \frac{1}{N} \sum_{i=1}^{N} c_i$$

In the example:

c_1 = 2*1/(2*(2-1) = 1
c_2 = 1/3 = 0.33
c_3 = 1/1 = 1
c_4 = 0



Break

Centrality

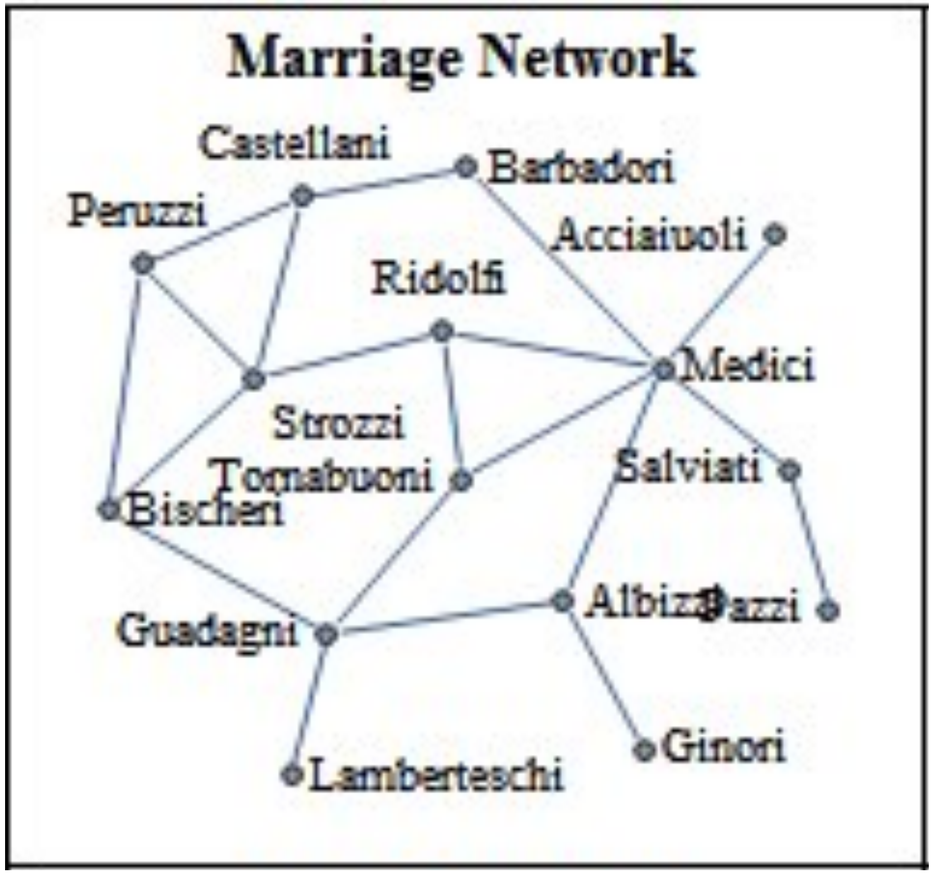
Centrality

Identify important nodes in a network
using structural information

- Nodes whose failure could disrupt the functioning of the all network
- Important nodes for the spreading of information
- "Strategic" nodes in the network

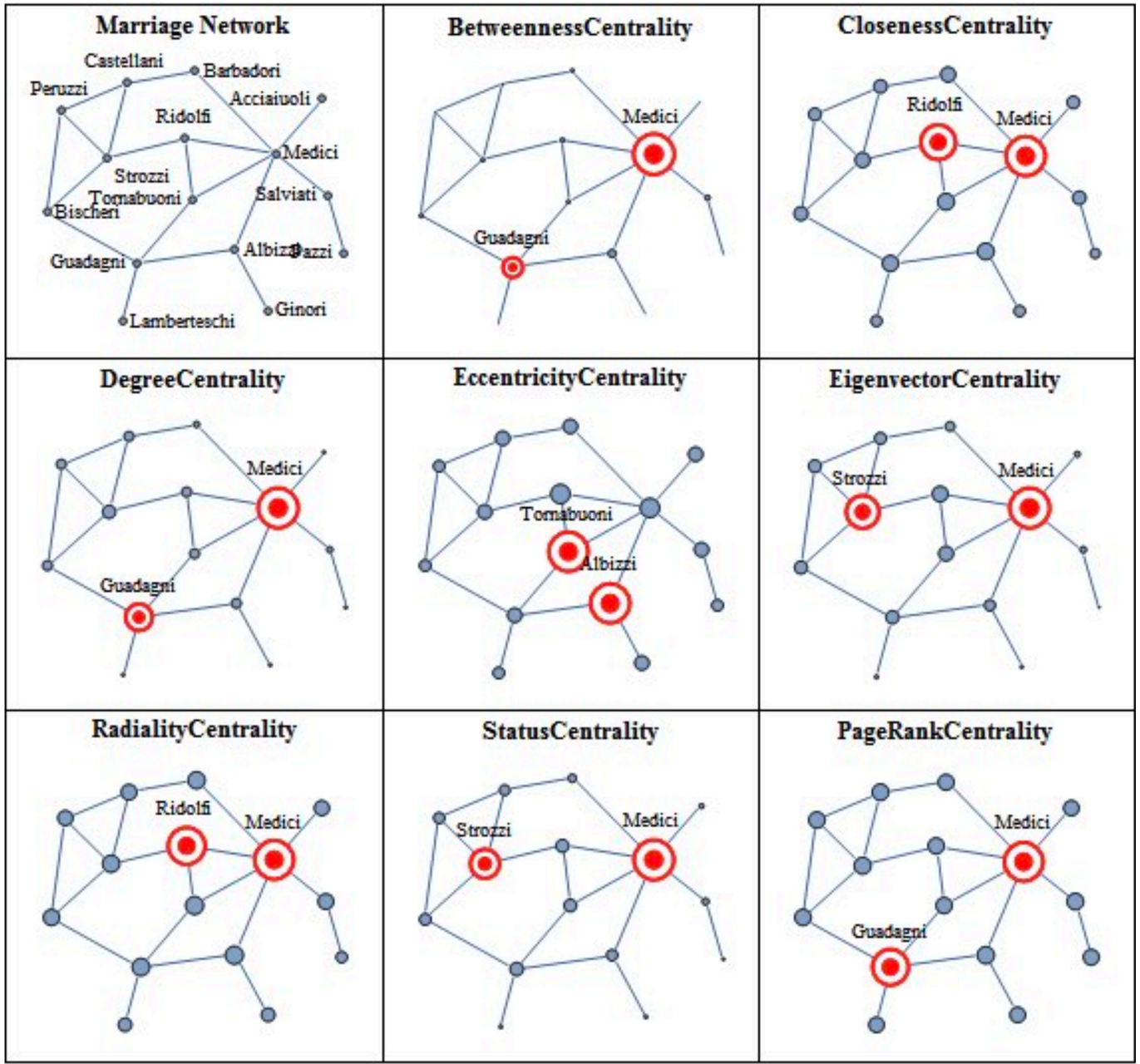
Centrality

Centrality



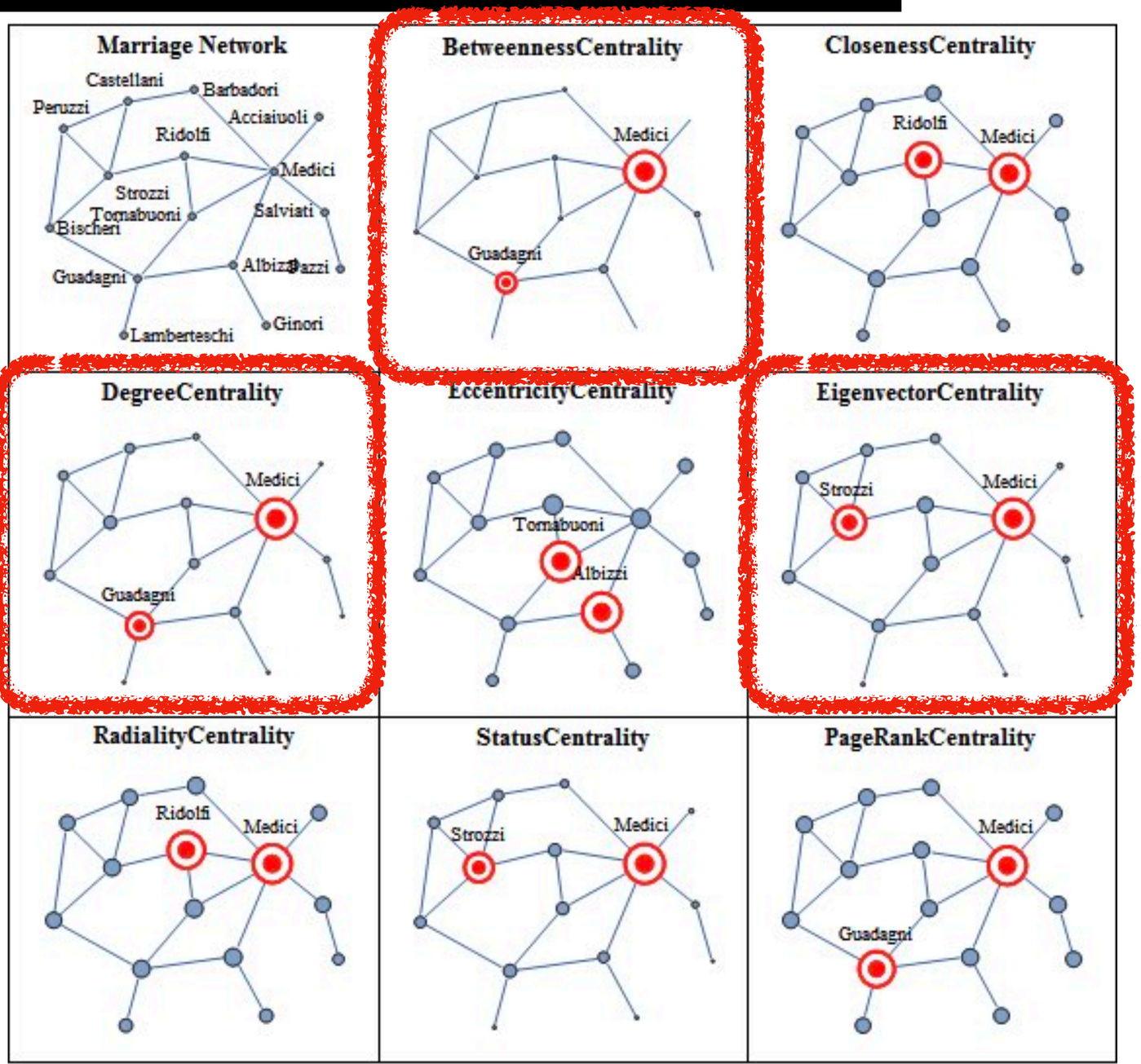
Padgett and Ansell, American Journal of Sociology, (1993).

Centrality



Padgett and Ansell, American Journal of Sociology, (1993).

Centrality



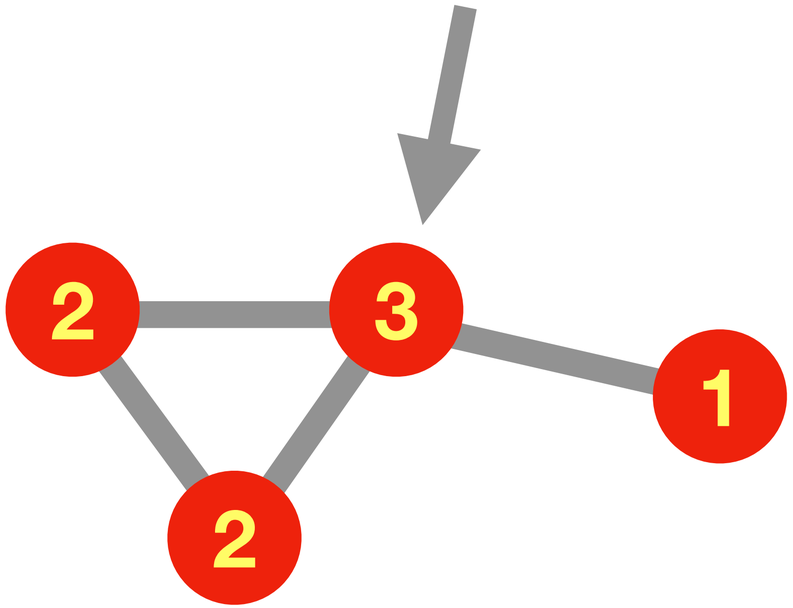
Padgett and Ansell, American Journal of Sociology, (1993).

Degree centrality

The most important node is the one with the most connections.

k_i = number of neighbors of node i

High-degree nodes are called **hubs**

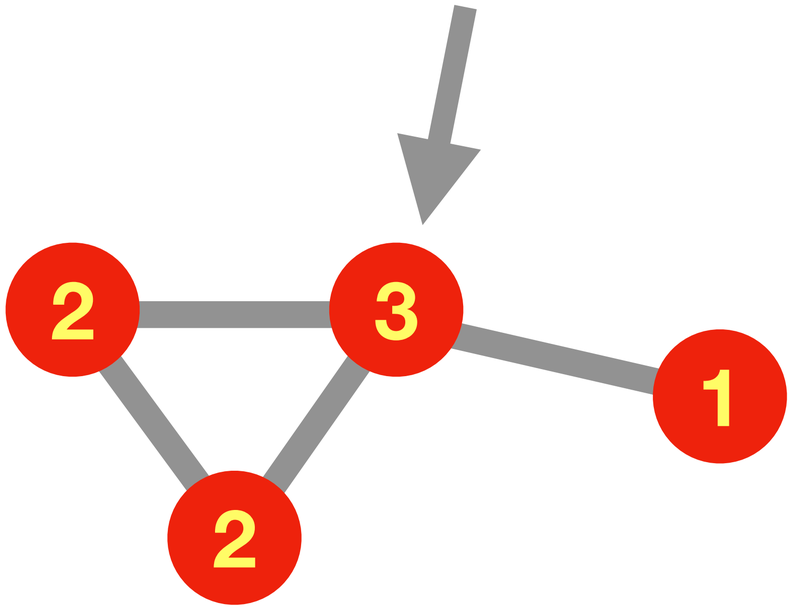


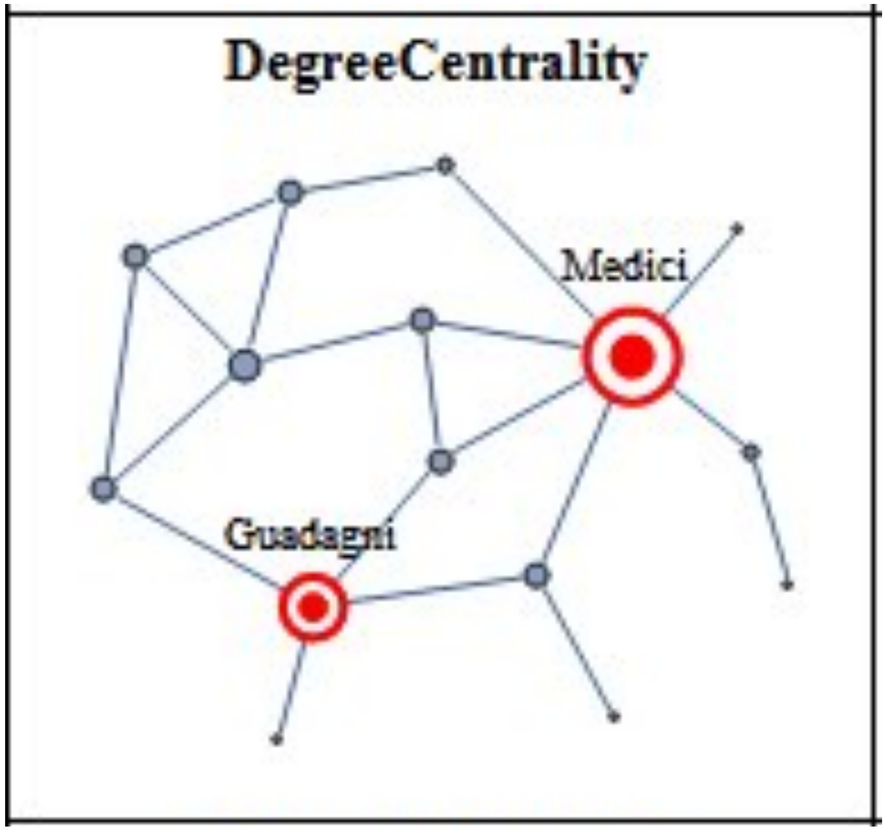
Degree centrality

The most important node is the one with the most connections.

k_i = number of neighbors of node i

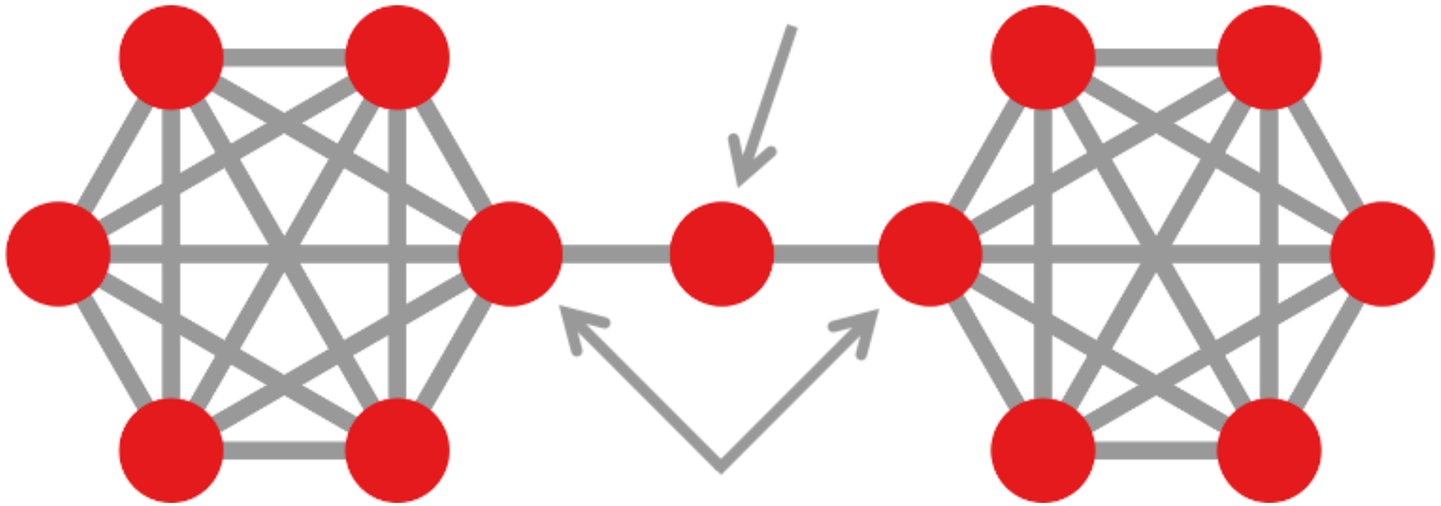
High-degree nodes are called **hubs**



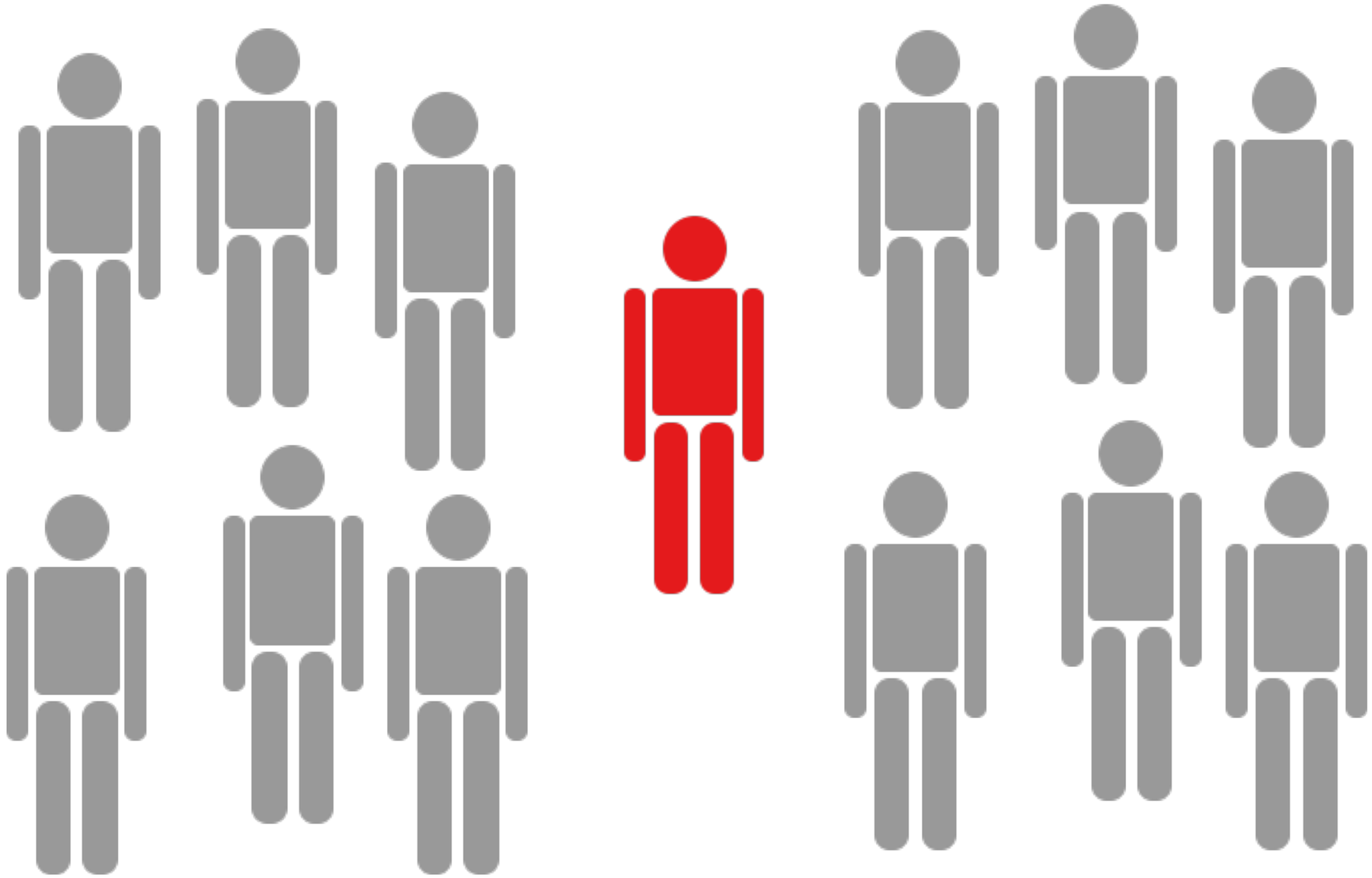


Degree centrality

Who is more important now?

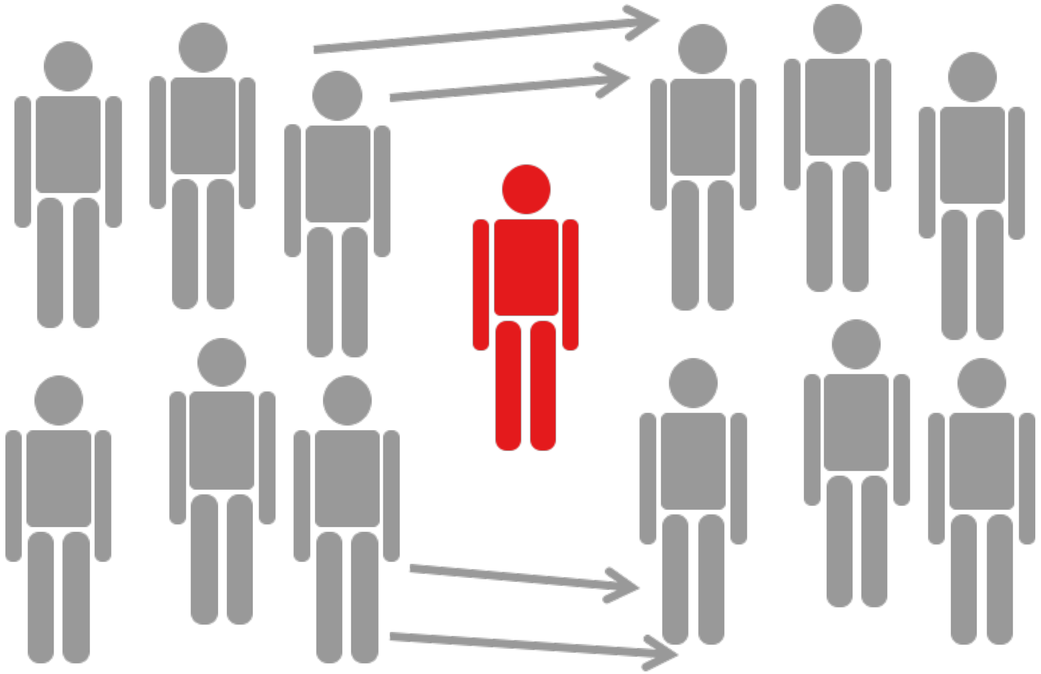


Betweenness centrality



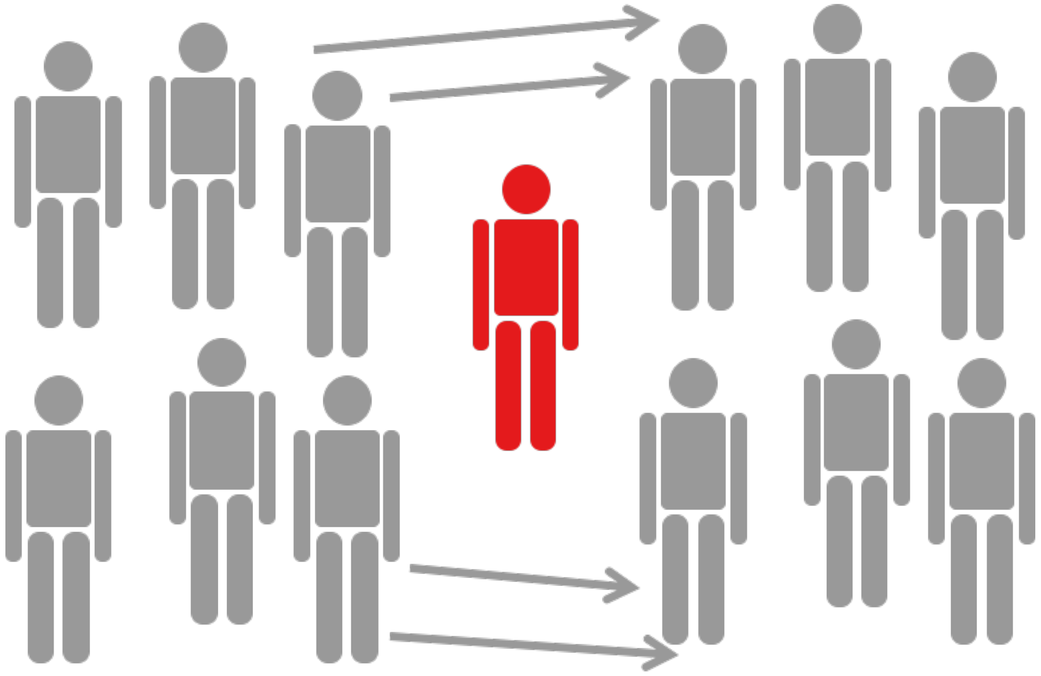
Betweenness centrality

No shortest path passes through you: **not important**, i.e., nobody is affected if you leave.

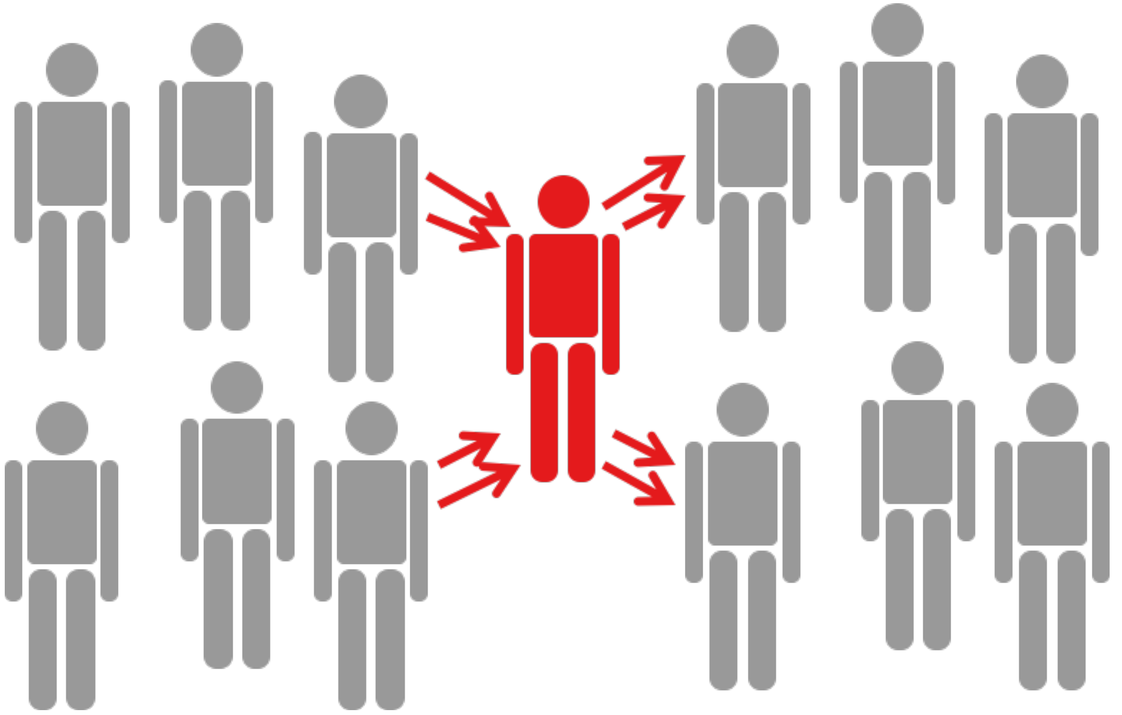


Betweenness centrality

No shortest path passes through you: **not important**, i.e., nobody is affected if you leave.



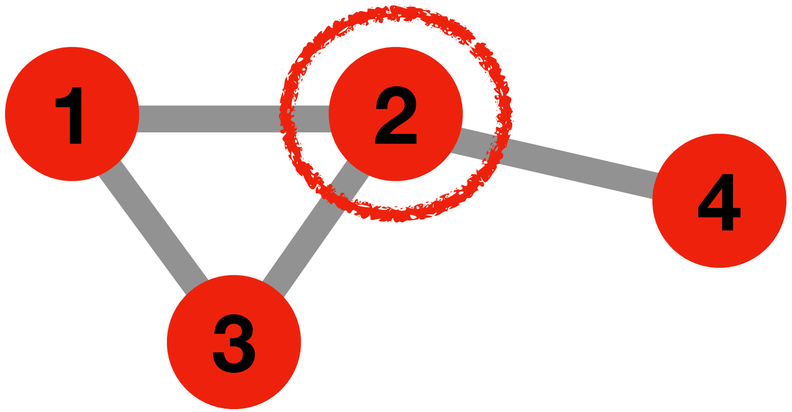
Shortest path passes through you: **important**, i.e., your leaving can disconnect people.



Betweenness centrality

Betweenness centrality counts the number of geodesics **(shortest paths)** that pass through a particular vertex i …

… divided by the number of possible geodesics from j —> k:

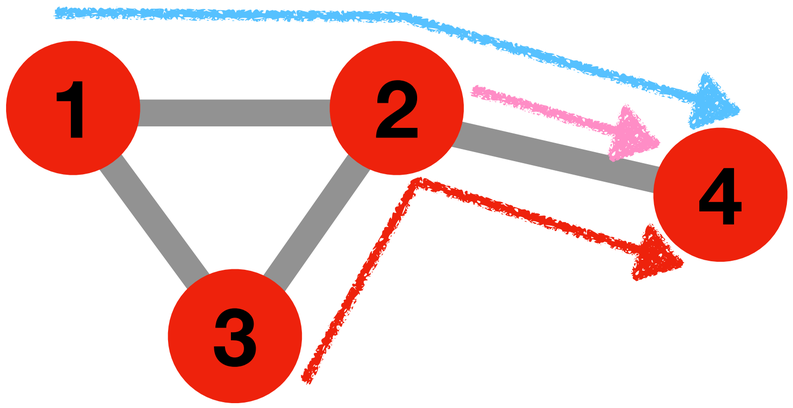


Betweenness centrality

Betweenness centrality counts the number of geodesics **(shortest paths)** that pass through a particular vertex i …

… divided by the number of possible geodesics from j —> k:

Paths (2x):
[1, 2, 4]
[3, 2, 4]
[1, 3]



Counts twice: j —> k and k —> j

Betweenness centrality

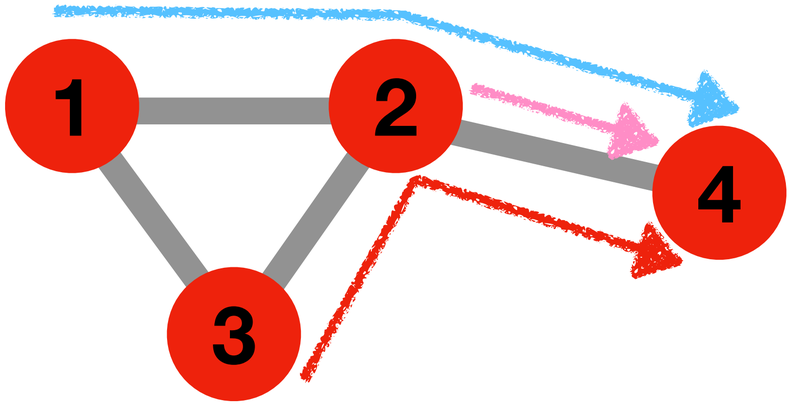
Betweenness centrality counts the number of geodesics **(shortest paths)** that pass through a particular vertex i …

… divided by the number of possible geodesics from j —> k:

In the example:

$b_1 = 0$
$b_2 = 4/6 = 0.66$
$b_3 = 0$
$b_4 = 0$

Paths (2x):
[1, **2**, 4]
[3, **2**, 4]
[1, 3]



Counts twice: j —> k and k —> j

Betweenness centrality

Betweenness centrality counts the number of geodesics **(shortest paths)** that pass through a particular vertex ***i*** …

… divided by the number of possible geodesics from ***j*** —> ***k***:

$$b_i = \sum_{jk} \frac{\#\{\text{geodesic paths } j \to \cdots \to i \to \cdots \to k\}}{\#\{\text{geodesic paths } j \to \cdots \to k\}}$$

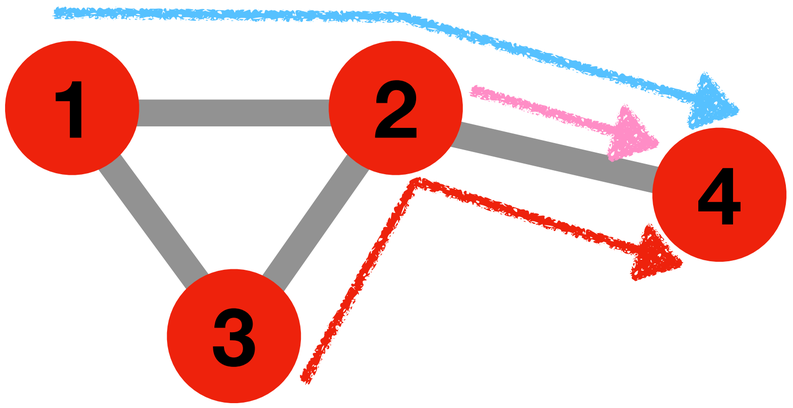
$$= \sum_{jk} \frac{\sigma_{jk}(i)}{\sigma_{jk}} \quad ,$$

Counts twice: ***j*** —> ***k*** and ***k*** —> ***j***

In the example:

$b_1 = 0$
$b_2 = 4/6 = 0.66$
$b_3 = 0$
$b_4 = 0$

Paths (2x):
[1, **2**, 4]
[3, **2**, 4]
[1, 3]



Betweenness centrality

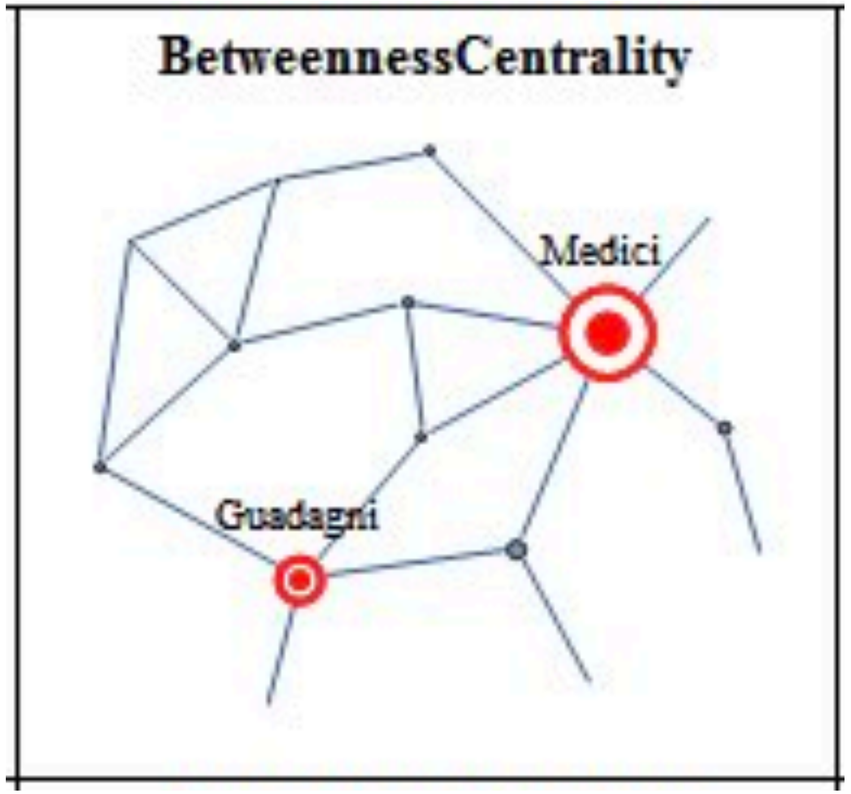
Betweenness centrality counts the number of geodesics **(shortest paths)** that pass through a particular vertex ***i*** …

… divided by the number of possible geodesics from ***j*** —> ***k***:

$$b_i = \sum_{jk} \frac{\#\{\text{geodesic paths } j \to \cdots \to i \to \cdots \to k\}}{\#\{\text{geodesic paths } j \to \cdots \to k\}}$$

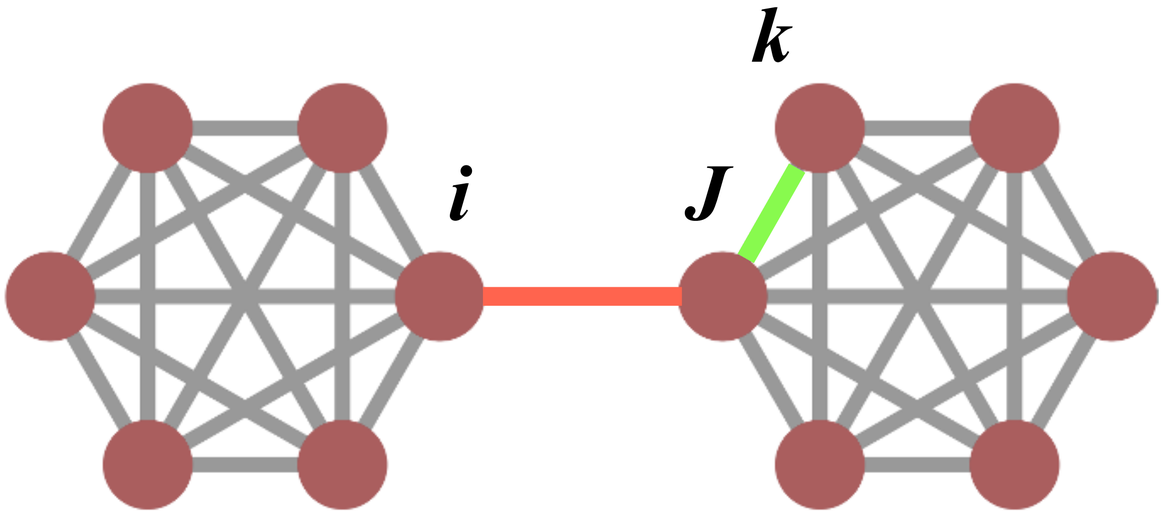
$$= \sum_{jk} \frac{\sigma_{jk}(i)}{\sigma_{jk}} \quad ,$$

Same as degree centrality!



Betweenness centrality - Edge

Freeman (1977): Betweenness centrality measures how many times a link is part of all shortest paths in the network.



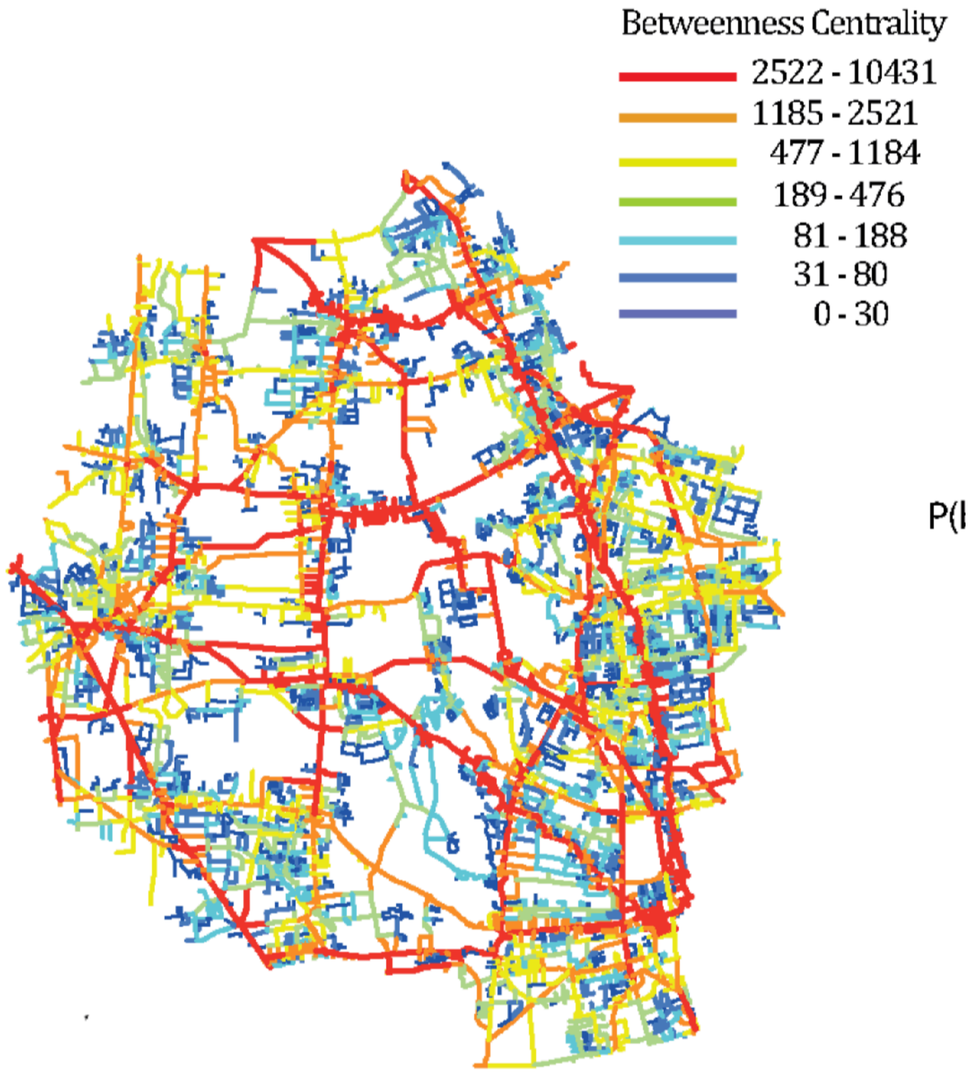
ij: large centrality

jk: small centrality

$$g(ij) = \sum_{s,t} \frac{\sigma_{st}(ij)}{\sigma_{st}}$$

Betweenness centrality - Edge

Freeman (1977): Betweenness centrality measures how many times a link is part of all shortest paths in the network.



P(l

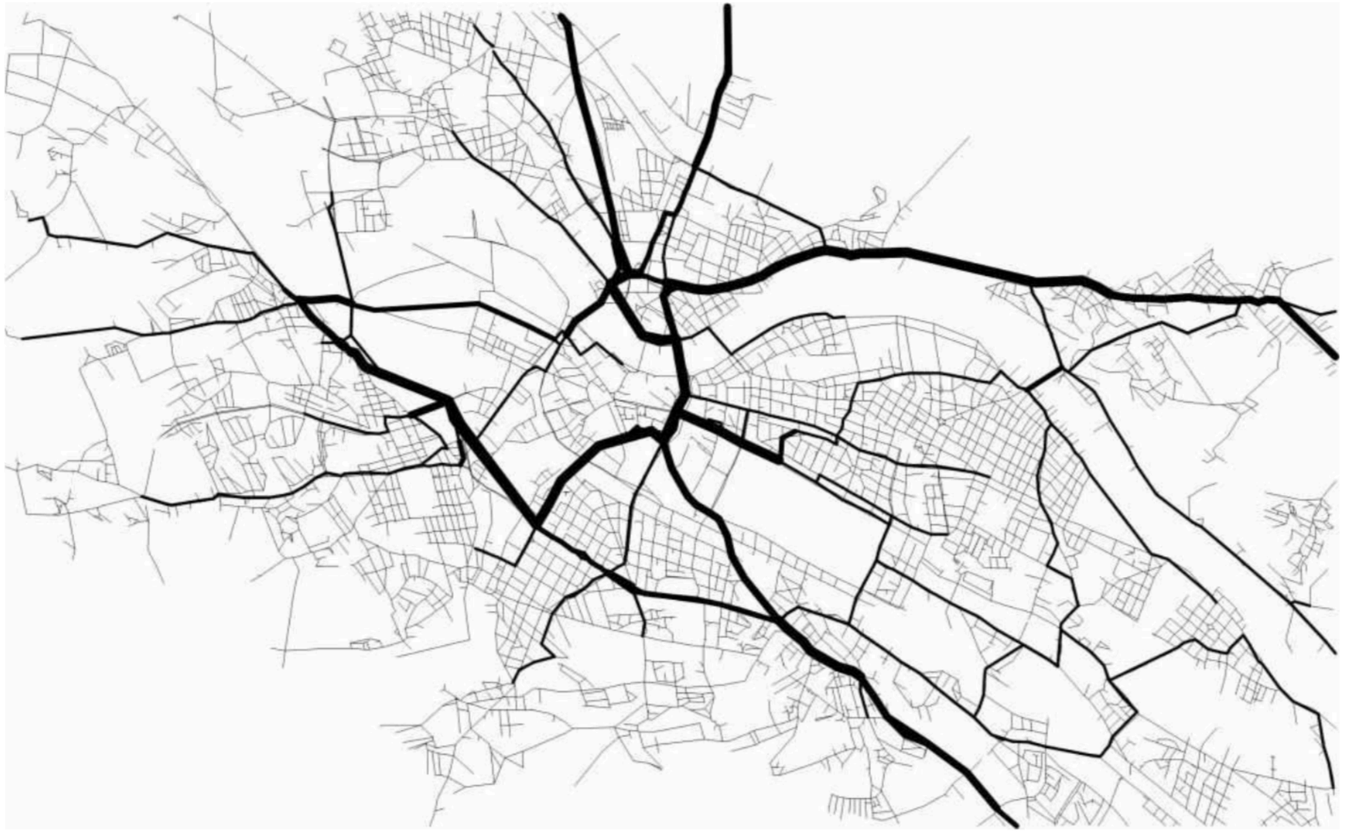
Measures the importance of a segment in the shortest paths flow

Gives the backbone of stable central roads

Betweenness centrality - Edge

Freeman (1977): Betweenness centrality measures how many times a link is part of all shortest paths in the network.

Can point to problems with **congestion**:

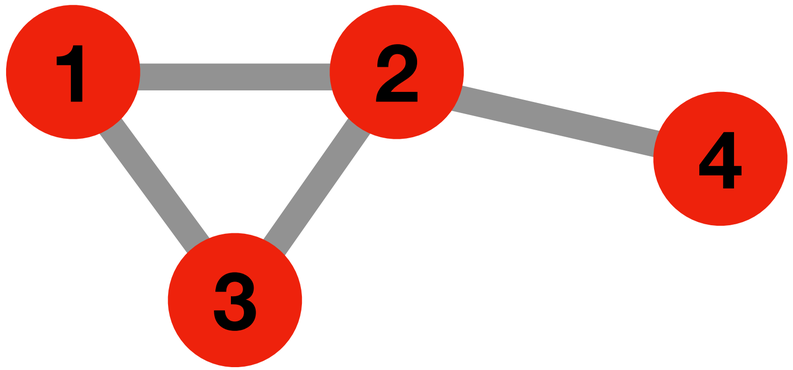


Lämmer, Stefan, Björn Gehlsen, and Dirk Helbing. *Physica A* 363.1 (2006): 89-95.

Eigenvector centrality

Assigns a vertex an importance score that is proportional to the importance scores of its neighbours:

Self-referential

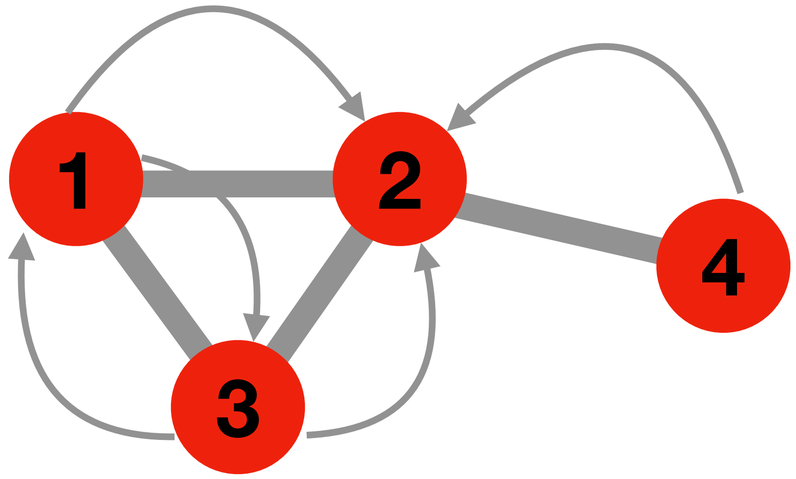


Eigenvector centrality

Assigns a vertex an importance score that is proportional to the importance scores of its neighbours:

Self-referential

1. Start with $x_i^{(t=0)} = 1$ for all i.
2. At each time t, nodes “vote” on the importance of neighbours.
3. Ends when values stop changing (converge).



Eigenvector centrality

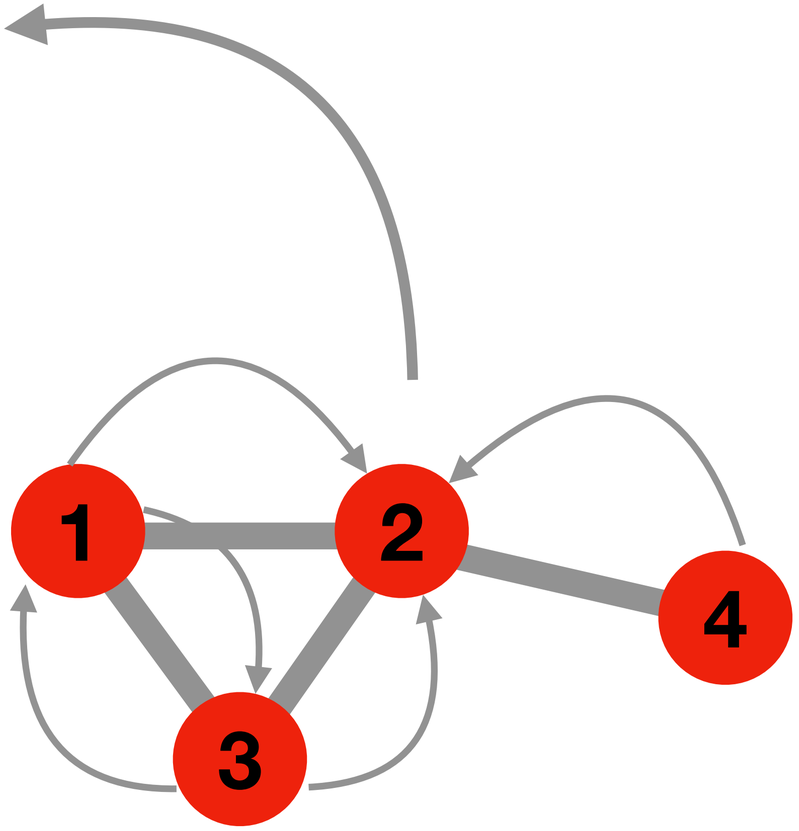
Assigns a vertex an importance score that is proportional to the importance scores of its neighbours:

Self-referential

$$x_i^{(t+1)} = \sum_{j=1}^{n} A_{ij} x_j^{(t)}$$

where A_{ij} is an element of the adj. matrix.

1. Start with $x_i^{(t=0)} = 1$ for all i.
2. At each time t, nodes "vote" on the importance of neighbours.
3. Ends when values stop changing (converge).



Eigenvector centrality

Assigns a vertex an importance score that is proportional to the importance scores of its neighbours:

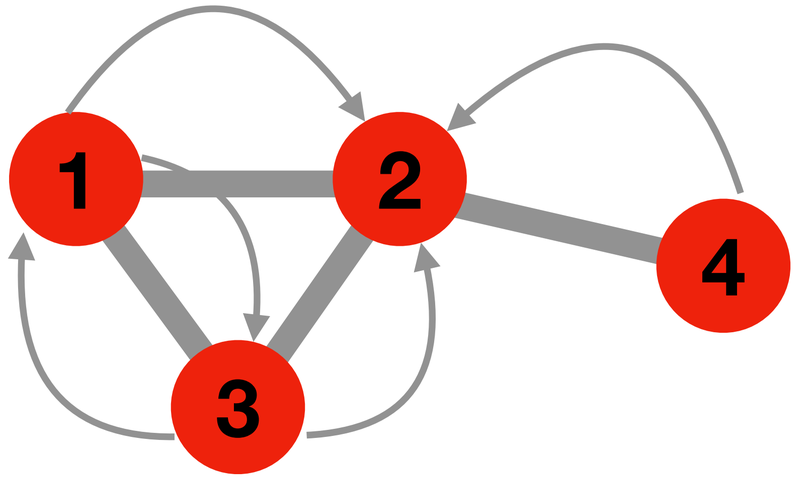
Self-referential

$$x_i^{(t+1)} = \sum_{j=1}^{n} A_{ij} x_j^{(t)}$$

j contributes to ***i*** importance only if they are connected

where A_{ij} is an element of the adj. matrix.

1. Start with $x_i^{(t=0)} = 1$ for all ***i***.
2. At each time ***t***, nodes “vote” on the importance of neighbours.
3. Ends when values stop changing (converge).



Eigenvector centrality

Assigns a vertex an importance score that is proportional to the importance scores of its neighbours:

Self-referential

$$x_i^{(t+1)} = \sum_{j=1}^{n} A_{ij} x_j^{(t)}$$

j contributes to ***i*** importance only if they are connected

where A_{ij} is an element of the adj. matrix.

Perron-Frobenius theorem: if the graph is undirected, convergence is guaranteed to the principal eigenvector of ***A*.**

Vector of eigenvector centralities

$$\mathbf{A}\mathbf{x} = \lambda_1 \mathbf{x}$$

Largest eigenvalue

Eigenvector centrality

Assigns a vertex an importance score that is proportional to the importance scores of its neighbours:

Self-referential

$$x_i^{(t+1)} = \sum_{j=1}^{n} A_{ij} x_j^{(t)}$$

j contributes to ***i*** importance only if they are connected

This is one of the reasons why I tortured you with the Eigen-stuff in the last class:

- All matrix multiplications
- Converges quite quickly
- Much more efficient than searching shortest paths

Vector of eigenvector centralities

$$\mathbf{A}\mathbf{x} = \lambda_1 \mathbf{x}$$

Largest eigenvalue

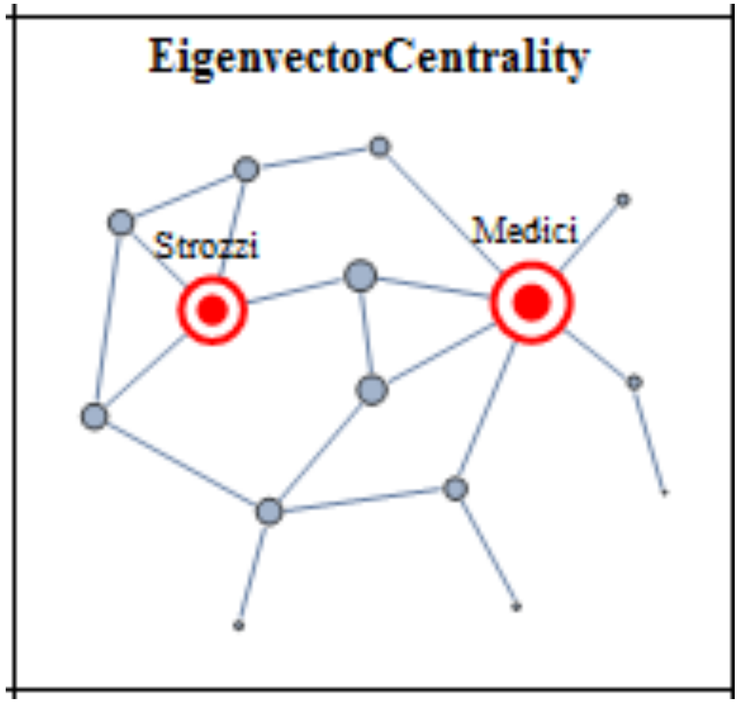
Eigenvector centrality

Assigns a vertex an importance score that is proportional to the importance scores of its neighbours:

Self-referential

$$x_i^{(t+1)} = \sum_{j=1}^{n} A_{ij} x_j^{(t)}$$

j contributes to ***i*** importance only if they are connected



Vector of eigenvector centralities

$$\mathbf{A}\mathbf{x} = \lambda_1 \mathbf{x}$$

Largest eigenvalue

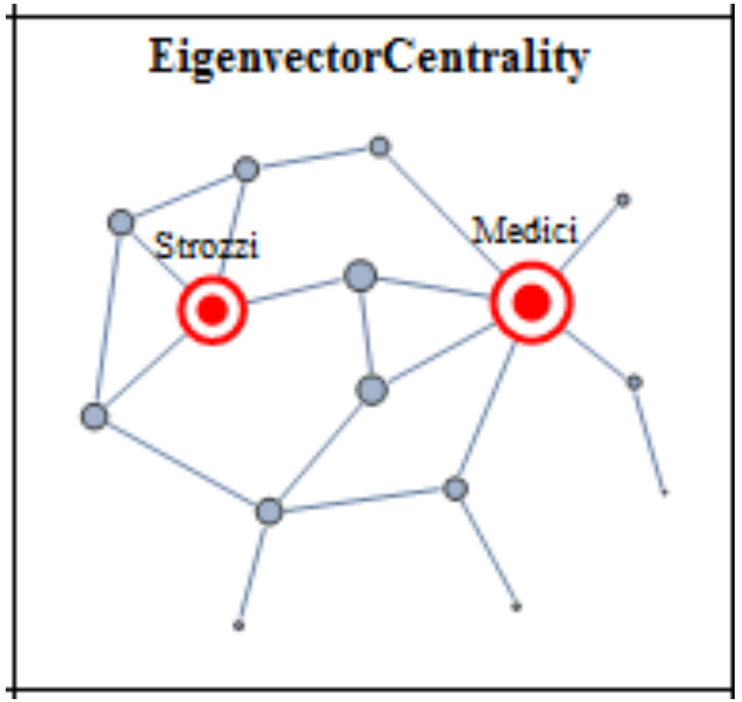
Eigenvector centrality

Assigns a vertex an importance score that is proportional to the importance scores of its neighbours:

Self-referential

$$x_i^{(t+1)} = \sum_{j=1}^{n} A_{ij} x_j^{(t)}$$

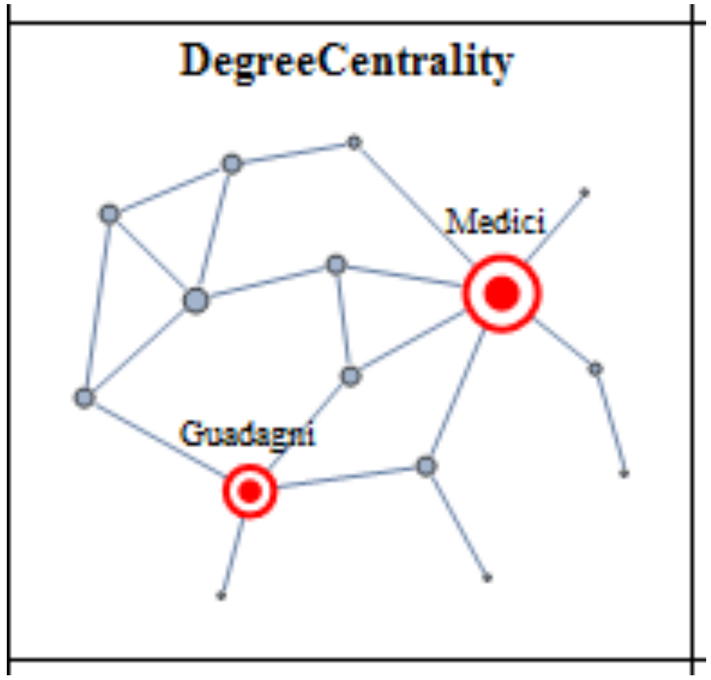
j contributes to ***i*** importance only if they are connected

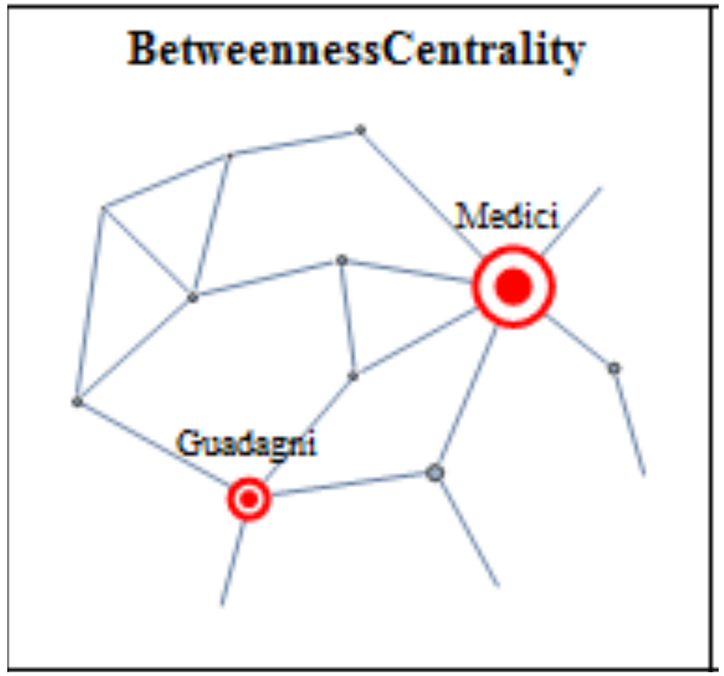


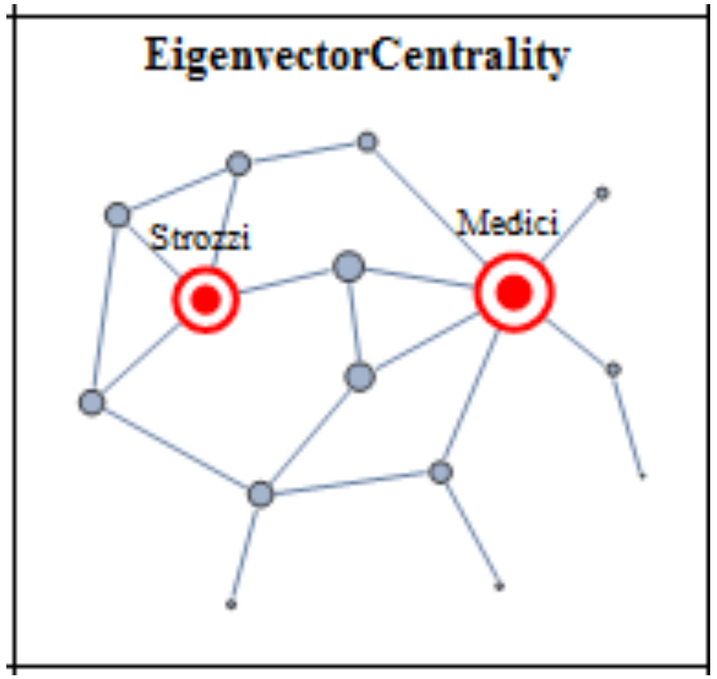
Got something else this time!

Centralities

2. What are the most important differences about the centralities: degree, betweenness, eigenvector?



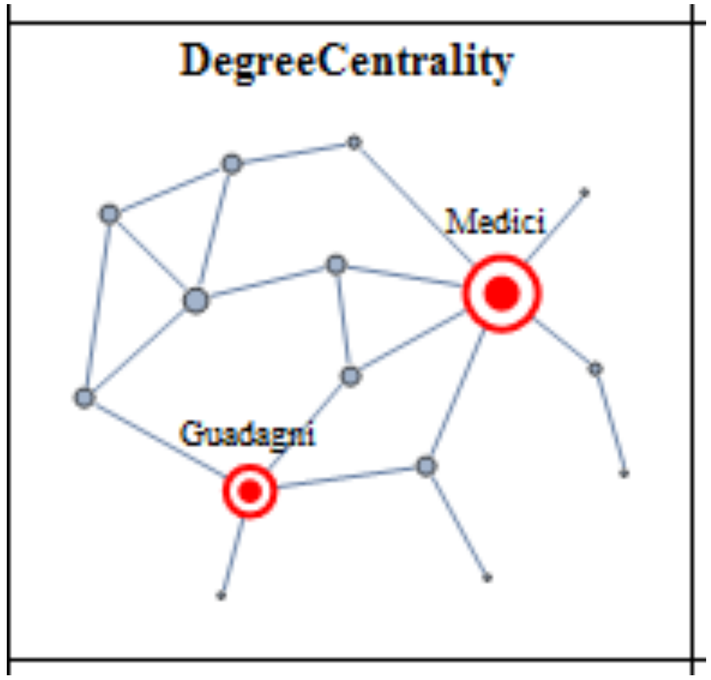




Centralities

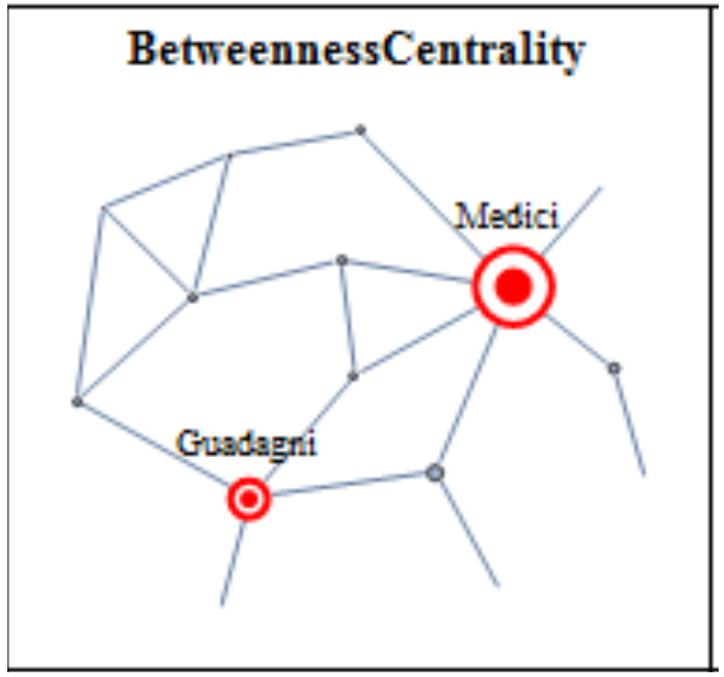
2. What are the most important differences about the centralities: degree, betweenness, eigenvector?

Quick & Easy



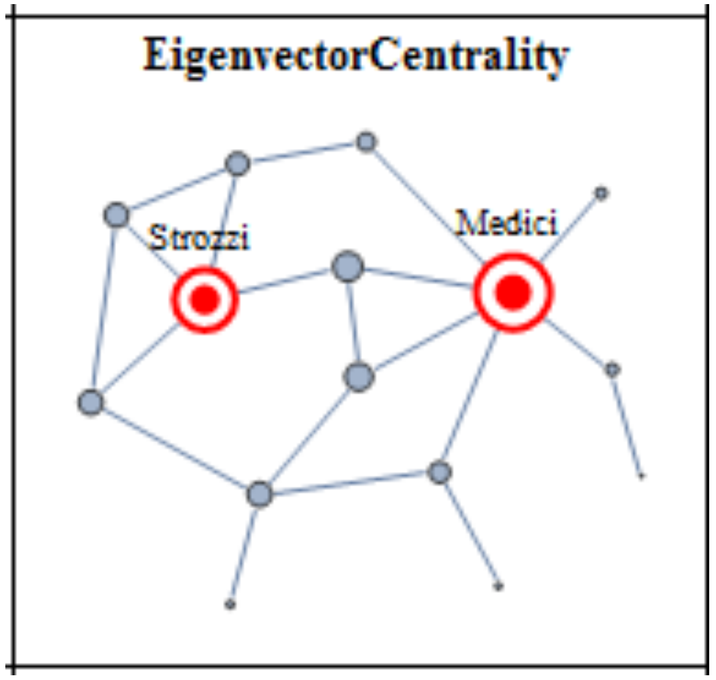
Local info

Computationally expensive



Shortest paths

Efficient



Neighbours importance

Lab time!