Graph Theory and Complex Networks: An Introduction

Maarten van Steen

VU Amsterdam, Dept. Computer Science
Room R4.20, steen@cs.vu.nl

Chapter 06: Network analysis

Version: April 28, 2014

vrije Universiteit amsterdam

Introduction

Observation

In real-world situations, graphs (or networks) may become very large, making it difficult to (visually) discover properties $\Rightarrow$ we need **network analysis** tools.

Vertex degrees: Consider the distribution of degrees: how many vertices have high degrees versus the number of vertices with low degrees.

Distance statistics: Focus on where vertices are positioned in the network: far away from each other, central in the network, etc.

Clustering: To what extent are my neighbors also adjacent to each other?

Centrality: Are there vertices that are more important than others?

Introduction

Observation

In real-world situations, graphs (or networks) may become very large, making it difficult to (visually) discover properties $\Rightarrow$ we need **network analysis** tools.

Vertex degrees: Consider the distribution of degrees: how many vertices have high degrees versus the number of vertices with low degrees.

Distance statistics: Focus on where vertices are positioned in the network: far away from each other, central in the network, etc.

Clustering: To what extent are my neighbors also adjacent to each other?

Centrality: Are there vertices that are more important than others?

Introduction

Observation

In real-world situations, graphs (or networks) may become very large, making it difficult to (visually) discover properties $\Rightarrow$ we need **network analysis** tools.

Vertex degrees: Consider the distribution of degrees: how many vertices have high degrees versus the number of vertices with low degrees.

Distance statistics: Focus on where vertices are positioned in the network: far away from each other, central in the network, etc.

Clustering: To what extent are my neighbors also adjacent to each other?

Centrality: Are there vertices that are more important than others?

Introduction

Observation

In real-world situations, graphs (or networks) may become very large, making it difficult to (visually) discover properties $\Rightarrow$ we need **network analysis** tools.

Vertex degrees: Consider the distribution of degrees: how many vertices have high degrees versus the number of vertices with low degrees.

Distance statistics: Focus on where vertices are positioned in the network: far away from each other, central in the network, etc.

Clustering: To what extent are my neighbors also adjacent to each other?

Centrality: Are there vertices that are more important than others?

Introduction

Observation

In real-world situations, graphs (or networks) may become very large, making it difficult to (visually) discover properties $\Rightarrow$ we need **network analysis** tools.

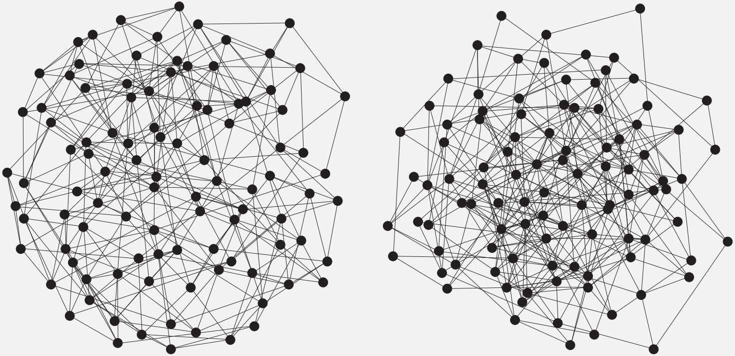
Vertex degrees: Consider the distribution of degrees: how many vertices have high degrees versus the number of vertices with low degrees.

Distance statistics: Focus on where vertices are positioned in the network: far away from each other, central in the network, etc.

Clustering: To what extent are my neighbors also adjacent to each other?

Centrality: Are there vertices that are more important than others?

Vertex degree

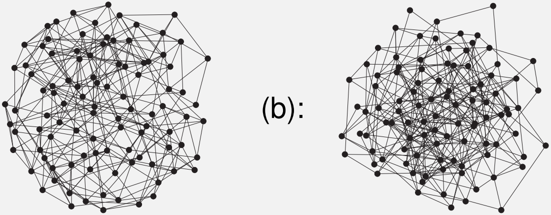


Question

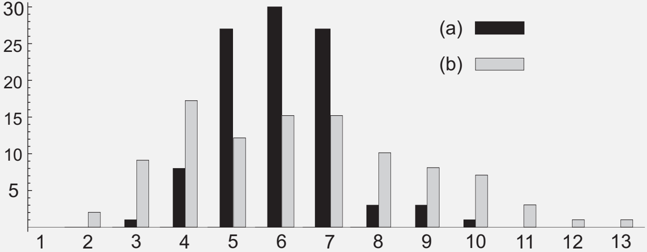
Can you visually observe real (nonisomorphic) differences?

Vertex degree: Histogram

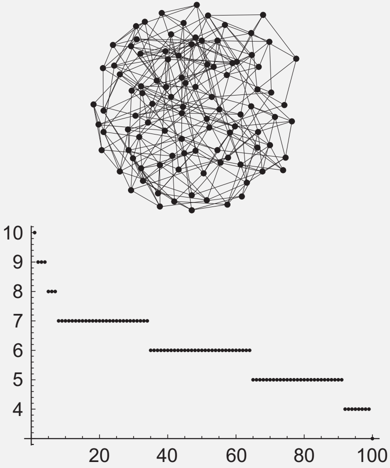
(a): (b):

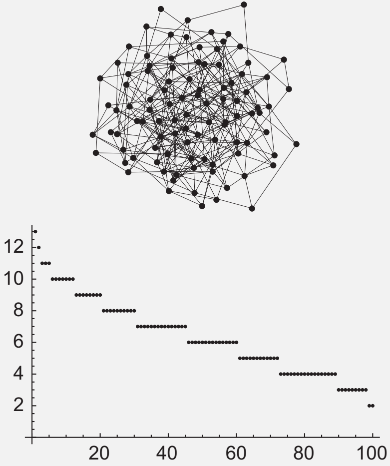


n = 100, m = 300   n = 100, m = 317



Vertex degree: Ranked histogram





Distance statistics

Definition

G is connected, $d(u, v)$ is distance between vertices u and v: the **length** of a shortest path between u and v.

Eccentricity $\varepsilon(u)$:	$\max\{d(u, v) \mid v \in V(G)\}$
Radius $rad(G)$:	$\min\{\varepsilon(u) \mid u \in V(G)\}$
Diameter $diam(G)$:	$\max\{d(u, v) \mid u, v \in V(G)\}$

Note

Note that these definitions apply to directed as well as undirected graphs.

Distance statistics

Definition

G is connected, $d(u, v)$ is distance between vertices u and v: the **length** of a shortest path between u and v.

Eccentricity $\varepsilon(u)$: $\max\{d(u, v) | v \in V(G)\}$
Radius $rad(G)$: $\min\{\varepsilon(u) | u \in V(G)\}$
Diameter $diam(G)$: $\max\{d(u, v) | u, v \in V(G)\}$

Note

Note that these definitions apply to directed as well as undirected graphs.

Path lengths

Definition

G is connected with vertex V; $\overline{d}(u)$ is average **length** of shortest paths from u to any other vertex v:

$$\overline{d}(u) \stackrel{\text{def}}{=} \frac{1}{|V|-1} \sum_{v \in V, v \neq u} d(u,v)$$

The average path length $\overline{d}(G)$:

$$\overline{d}(G) \stackrel{\text{def}}{=} \frac{1}{|V|} \sum_{u \in V} \overline{d}(u) = \frac{1}{|V|^2 - |V|} \sum_{u,v \in V, u \neq v} d(u,v)$$

Path lengths

Definition

The characteristic path length is the median over all $\overline{d}(u)$.

Note

The median over n nondecreasing values $x_1, x_2, \ldots, x_n$:

- n odd $\Rightarrow x_{(n+1)/2}$
- n even $\Rightarrow (x_{n/2} + x_{n/2+1})/2$

The median separates the higher values from the lower values into two equally-sized subsets.

Example

$\{3,4,4,6,0,6,1\} \Rightarrow [0,1,3,4,4,6,6] \Rightarrow M = x_{(7+1)/2} = x_4 = 4$

Path lengths

Definition

The characteristic path length is the median over all $\overline{d}(u)$.

Note

The median over n nondecreasing values $x_1, x_2, \ldots, x_n$:

- n odd $\Rightarrow x_{(n+1)/2}$
- n even $\Rightarrow (x_{n/2} + x_{n/2+1})/2$

The median separates the higher values from the lower values into two equally-sized subsets.

Example

$\{3,4,4,6,0,6,1\} \Rightarrow [0,1,3,4,4,6,6] \Rightarrow M = x_{(7+1)/2} = x_4 = 4$

Path lengths

Definition

The characteristic path length is the median over all $\overline{d}(u)$.

Note

The median over n nondecreasing values $x_1, x_2, \ldots, x_n$:

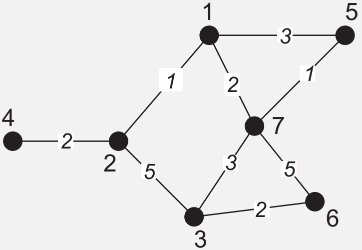
- n odd $\Rightarrow x_{(n+1)/2}$
- n even $\Rightarrow (x_{n/2} + x_{n/2+1})/2$

The median separates the higher values from the lower values into two equally-sized subsets.

Example

$\{3,4,4,6,0,6,1\} \Rightarrow [0,1,3,4,4,6,6] \Rightarrow M = x_{(7+1)/2} = x_4 = 4$

Example distance statistics



Vertex	1	2	3	4	5	6	7	$\varepsilon(u)$	$\sum_{v \neq u} d(u,v)$	$\overline{d}(u)$
1	0	1	5	3	3	7	2	7	21	3.50
2	1	0	5	2	4	7	3	7	22	3.67
3	5	5	0	7	4	2	3	7	26	4.33
4	3	2	7	0	6	9	5	9	32	5.33
5	3	4	4	6	0	6	1	6	24	4.00
6	7	7	2	9	6	0	5	9	36	6.00
7	2	3	3	5	1	5	0	5	19	3.17

Clustering coefficient

Observation

Many networks show a high degree of clustering: my neighbors are each other's neighbors.

Note

An extreme case is formed by having all my neighbors be adjacent to each other $\Rightarrow$ neighbors form a complete graph.

Question

What is the other extreme case?

Clustering coefficient

Observation

Many networks show a high degree of clustering: my neighbors are each other's neighbors.

Note

An extreme case is formed by having all my neighbors be adjacent to each other $\Rightarrow$ neighbors form a complete graph.

Question

What is the other extreme case?

Clustering coefficient

Observation

Many networks show a high degree of clustering: my neighbors are each other's neighbors.

Note

An extreme case is formed by having all my neighbors be adjacent to each other $\Rightarrow$ neighbors form a complete graph.

Question

What is the other extreme case?

Clustering coefficient

Definition

G is simple, connected, undirected. Vertex $v \in V(G)$ with neighborset $N(v)$.

- Let $n_v = |N(v)|$.
  Note: max. number of edges between neighbors is $\binom{n_v}{2}$.
- Let m_v is number of edges in subgraph induced by $N(v)$: $m_v = |E(G[N(v)])|$.

Clustering coefficient $cc(v)$:

$$cc(v) \stackrel{\text{def}}{=} \begin{cases} m_v / \binom{n_v}{2} = \frac{2 \cdot m_v}{n_v(n_v - 1)} & \text{if } \delta(v) > 1 \\ \text{undefined} & \text{otherwise} \end{cases}$$

Clustering coefficient

Definition

G is simple, connected and undirected.
Let $V^* \stackrel{\text{def}}{=} \{v \in V(G) | \delta(v) > 1\}$.

Clustering coefficient $CC(G)$ for G:

$$CC(G) \stackrel{\text{def}}{=} \frac{1}{|V^*|} \sum_{v \in V^*} cc(v)$$

Clustering coefficient: triangles

Definition

A triangle is a complete (sub)graph with exactly 3 vertices. A triple is a (sub)graph with exactly 3 vertices and 2 edges.

Definition

G is simple and connected with $n_\Delta(G)$ distinct triangles and $n_\wedge(G)$ distinct triples.
The network transitivity $\tau(G) \stackrel{\text{def}}{=} n_\Delta(G)/n_\wedge(G)$.

Notation

A triple at v: v is incident to both edges ("in the middle"). $n_\wedge(v)$: number of triples at v.

Clustering coefficient: triangles

Definition

A triangle is a complete (sub)graph with exactly 3 vertices. A triple is a (sub)graph with exactly 3 vertices and 2 edges.

Definition

G is simple and connected with $n_\Delta(G)$ distinct triangles and $n_\Lambda(G)$ distinct triples.

The network transitivity $\tau(G) \stackrel{\text{def}}{=} n_\Delta(G)/n_\Lambda(G)$.

Notation

A triple at v: v is incident to both edges ("in the middle"). $n_\Lambda(v)$: number of triples at v.

Clustering coefficient: triangles

Definition

A triangle is a complete (sub)graph with exactly 3 vertices. A triple is a (sub)graph with exactly 3 vertices and 2 edges.

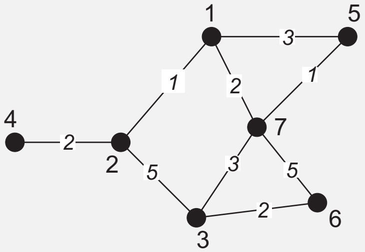
Definition

G is simple and connected with $n_{\Delta}(G)$ distinct triangles and $n_{\Lambda}(G)$ distinct triples.
The network transitivity $\tau(G) \stackrel{\text{def}}{=} n_{\Delta}(G)/n_{\Lambda}(G)$.

Notation

A triple at v: v is incident to both edges ("in the middle"). $n_{\Lambda}(v)$: number of triples at v.

Clustering coefficient: example



Vertex:	**1**	**2**	**3**	**4**	**5**	**6**	**7**
cc:	1/3	0	1/3	*undefined*	1	1	1/3
$n_\wedge$:	3	3	3	0	1	1	6

Vertex 1 $N(1) = \{2,5,7\}$; $E(G[N(1)]) = \langle 5,7 \rangle \Rightarrow cc(1) = \frac{1}{3}$

Triples at 1: $G[\{2,1,5\}], G[\{2,1,7\}], G[\{5,1,7\}]$

Clustering coefficient versus transitivity

Observation

Let $n_\Delta(v)$ be the number of triangles of which v is member $\Rightarrow$

- $cc(v) = \frac{n_\Delta(v)}{n_\Lambda(v)}$
- $n_\Lambda(v) = \binom{\delta(v)}{2}$
- $n_\Delta(G) = \frac{1}{3}\sum_{v \in V^*} n_\Delta(v)$ (Note: $V^* = \{v \in V | \delta(v) > 1\}$)

Clustering coefficient versus transitivity

Observation

Let $n_\Delta(v)$ be the number of triangles of which v is member $\Rightarrow$

- $cc(v) = \frac{n_\Delta(v)}{n_\Lambda(v)}$
- $n_\Lambda(v) = \binom{\delta(v)}{2}$
- $n_\Delta(G) = \frac{1}{3}\sum_{v \in V^*} n_\Delta(v)$ (Note: $V^* = \{v \in V | \delta(v) > 1\}$)

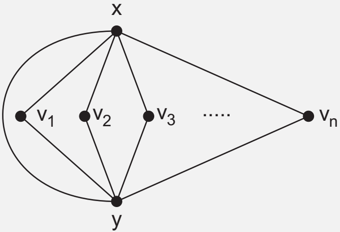
Clustering coefficient versus transitivity

Observation

Let $n_\Delta(v)$ be the number of triangles of which v is member $\Rightarrow$

- $cc(v) = \frac{n_\Delta(v)}{n_\Lambda(v)}$
- $n_\Lambda(v) = \binom{\delta(v)}{2}$
- $n_\Delta(G) = \frac{1}{3}\sum_{v \in V^*} n_\Delta(v)$ (Note: $V^* = \{v \in V | \delta(v) > 1\}$)

Clustering coefficient versus transitivity

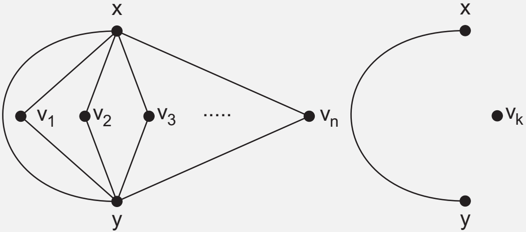


$G_k = G[\{x, y, v_1, v_2, \ldots, v_k\}] \Rightarrow$:

$$cc(u) = \begin{cases} \end{cases}$$

$$CC(G_k) = \frac{1}{k+2}(2 \cdot \frac{2}{k+1} + k \cdot 1) = \frac{k^2+k+4}{k^2+3k+2} \Rightarrow \lim_{k \to \infty} CC(G_k) = 1$$

Clustering coefficient versus transitivity

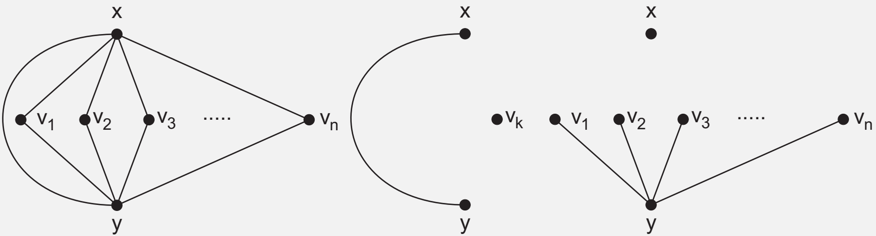


$G_k = G[\{x, y, v_1, v_2, \ldots, v_k\}] \Rightarrow$:

$$cc(u) = \begin{cases} 1 & \text{if } u = v_1, \ldots, v_k \\ & \end{cases}$$

$$CC(G_k) = \frac{1}{k+2}\left(2 \cdot \frac{2}{k+1} + k \cdot 1\right) = \frac{k^2+k+4}{k^2+3k+2} \Rightarrow \lim_{k\to\infty} CC(G_k) = 1$$

Clustering coefficient versus transitivity

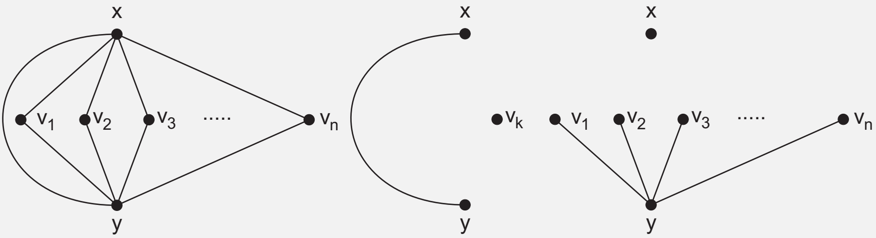


$G_k = G[\{x, y, v_1, v_2, \ldots, v_k\}] \Rightarrow$:

$$cc(u) = \begin{cases} 1 & \text{if } u = v_1, \ldots, v_k \\ \frac{k}{\binom{k+1}{2}} = \frac{k}{\frac{1}{2} \cdot k(k+1)} = \frac{2}{k+1} & \text{if } u = x \text{ or } u = y \end{cases}$$

$$CC(G_k) = \frac{1}{k+2}\left(2 \cdot \frac{2}{k+1} + k \cdot 1\right) = \frac{k^2+k+4}{k^2+3k+2} \Rightarrow \lim_{k \to \infty} CC(G_k) = 1$$

Clustering coefficient versus transitivity

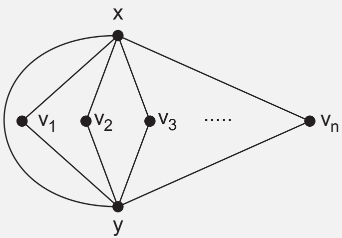


$G_k = G[\{x, y, v_1, v_2, \dots, v_k\}] \Rightarrow$:

$$cc(u) = \begin{cases} 1 & \text{if } u = v_1, \dots, v_k \\ \frac{k}{\binom{k+1}{2}} = \frac{k}{\frac{1}{2} \cdot k(k+1)} = \frac{2}{k+1} & \text{if } u = x \text{ or } u = y \end{cases}$$

$$CC(G_k) = \frac{1}{k+2}\left(2 \cdot \frac{2}{k+1} + k \cdot 1\right) = \frac{k^2+k+4}{k^2+3k+2} \Rightarrow \lim_{k\to\infty} CC(G_k) = 1$$

Clustering coefficient versus transitivity

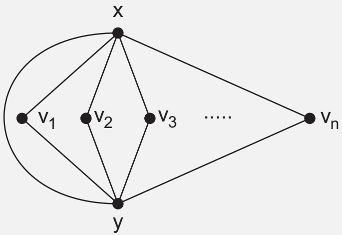


$G_k = G[\{x, y, v_1, v_2, \ldots, v_k\}] \Rightarrow$

$$n_\wedge(u) = \begin{cases} 1 & \text{if } u = v_1, \ldots, v_k \\ \binom{\delta(u)}{2} = \binom{k+1}{2} & \text{if } u = x, y \end{cases}$$

$$\tau(G_k) = \frac{n_\Delta(G_k)}{\sum n_\wedge(u)} = \frac{k}{2 \cdot \frac{1}{2} \cdot k(k+1) + k} = \frac{1}{k+2} \Rightarrow \lim_{k \to \infty} \tau(G_k) = 0$$

Clustering coefficient versus transitivity



$G_k = G[\{x, y, v_1, v_2, \ldots, v_k\}] \Rightarrow$

$$n_\Lambda(u) = \begin{cases} 1 & \text{if } u = v_1, \ldots, v_k \\ \binom{\delta(u)}{2} = \binom{k+1}{2} & \text{if } u = x, y \end{cases}$$

$$\tau(G_k) = \frac{n_\Delta(G_k)}{\sum n_\Lambda(u)} = \frac{k}{2 \cdot \frac{1}{2} \cdot k(k+1) + k} = \frac{1}{k+2} \Rightarrow \lim_{k\to\infty} \tau(G_k) = 0$$

Centrality

Issue

Are there any vertices that are more important than the others?

Definition

G is (strongly) connected. The center $C(G)$ is the set of vertices with minimal eccentricity:

$$C(G) \stackrel{\text{def}}{=} \{v \in V(G) | \varepsilon(v) = rad(G)\}$$

Intuition

At the center means at minimal distance to the farthest node.

Centrality

Issue

Are there any vertices that are more important than the others?

Definition

G is (strongly) connected. The center $C(G)$ is the set of vertices with minimal eccentricity:

$$C(G) \stackrel{\text{def}}{=} \{v \in V(G) | \varepsilon(v) = rad(G)\}$$

Intuition

At the center means at minimal distance to the farthest node.

Centrality

Issue

Are there any vertices that are more important than the others?

Definition

G is (strongly) connected. The center $C(G)$ is the set of vertices with minimal eccentricity:

$$C(G) \stackrel{\text{def}}{=} \{v \in V(G) | \varepsilon(v) = rad(G)\}$$

Intuition

At the center means at minimal distance to the farthest node.

Vertex centrality

Definition

G is (strongly) connected. The (eccentricity based) vertex centrality $c_E(u)$ of u:

$$c_E(u) \stackrel{\text{def}}{=} \frac{1}{\varepsilon(u)}$$

Intuition

The higher the centrality, the "closer" to the center of a graph.

Vertex centrality

Definition

G is (strongly) connected. The (eccentricity based) vertex centrality $c_E(u)$ of u:

$$c_E(u) \stackrel{\text{def}}{=} \frac{1}{\varepsilon(u)}$$

Intuition

The higher the centrality, the “closer” to the center of a graph.

Closeness

Definition

G is (strongly) connected. The closeness $c_C(u)$ of u:

$$c_C(u) \stackrel{\text{def}}{=} \frac{1}{\sum_{v \in V(G)} d(u,v)}$$

Intuition

How close is a vertex to all other nodes?

Closeness

Definition

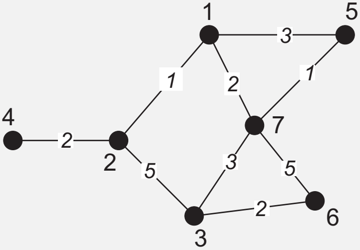
G is (strongly) connected. The closeness $c_C(u)$ of u:

$$c_C(u) \stackrel{\text{def}}{=} \frac{1}{\sum_{v \in V(G)} d(u, v)}$$

Intuition

How close is a vertex to all other nodes?

Centrality: example



Vertex:	1	2	3	4	5	6	7
$\varepsilon(u)$	7	7	7	9	6	9	5
$\sum d(u,\cdot)$	21	22	27	32	24	37	29
$c_C(u)$:	0.048	0.045	0.037	0.031	0.042	0.027	0.034

Betweenness

Intuition

Important vertices are those whose removal significantly increases the distance between other vertices. **Example**: cut vertices.

Definition

G is simple and (strongly) connected. $S(x,y)$ is set of shortest paths between x and y. $S(x,u,y) \subseteq S(x,y)$ paths that pass through u. Betweenness centrality $c_B(u)$ of u:

$$c_B(u) \stackrel{\text{def}}{=} \sum_{x \neq y \neq u} \frac{|S(x,u,y)|}{|S(x,y)|}$$

Betweenness

Intuition

Important vertices are those whose removal significantly increases the distance between other vertices. **Example**: cut vertices.

Definition

G is simple and (strongly) connected. $S(x,y)$ is set of shortest paths between x and y. $S(x,u,y) \subseteq S(x,y)$ paths that pass through u.
Betweenness centrality $c_B(u)$ of u:

$$c_B(u) \stackrel{\text{def}}{=} \sum_{x \neq y \neq u} \frac{|S(x,u,y)|}{|S(x,y)|}$$