Mathematics behind Internet, MAA507

Sergei Silvestrov, Christopher Engström

January 29, 2013

L2: Distance, similarity and clustering on graphst

Todays lecture:

- Distance and metrics.

L2: Distance, similarity and clustering on graphst

Todays lecture:

- Distance and metrics.
- Similarity measurements.

L2: Distance, similarity and clustering on graphst

Todays lecture:

- Distance and metrics.
- Similarity measurements.
- Centrality and rankings.

L2: Distance, similarity and clustering on graphst

Todays lecture:

- Distance and metrics.
- Similarity measurements.
- Centrality and rankings.
- Hierarchical clustering.

Distance and metrics

First we would like to define what we mean by the distance between two vertices.

- We want our distance function and the distance between two vertices in a graph to behave in the same way as the distance between two points in space using some metric.

Distance and metrics

First we would like to define what we mean by the distance between two vertices.

- We want our distance function and the distance between two vertices in a graph to behave in the same way as the distance between two points in space using some metric.
- We starrt by defining a metric. Which we will use to construct distance functions between vertices in a graph as well.

Distance and metrics

Definition

A function $f(x, y)$ is a **metric** if the following properties are satisfied:

- $f(x, y) = 0$, iff $x = y$

Distance and metrics

Definition

A function $f(x, y)$ is a **metric** if the following properties are satisfied:

- $f(x, y) = 0$, iff $x = y$
- $f(x, y) = f(y, x)$ (symmetric)

Distance and metrics

Definition

A function $f(x, y)$ is a **metric** if the following properties are satisfied:

- $f(x, y) = 0$, iff $x = y$
- $f(x, y) = f(y, x)$ (symmetric)
- $f(x, y) \geq 0$ (non-negative)

Distance and metrics

Definition

A function $f(x, y)$ is a **metric** if the following properties are satisfied:

- $f(x, y) = 0$, iff $x = y$
- $f(x, y) = f(y, x)$ (symmetric)
- $f(x, y) \geq 0$ (non-negative)
- $f(x, y) + f(y, z) \geq f(x, z)$ (triangle inequality)

Distance and metrics

A set with a metric is called a metric space, some examples are:

- $\mathbb{R}^3$ together with the euclidean distance:

$$d(a,b) = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2 + (a_3 - b_3)^2}$$

Distance and metrics

A set with a metric is called a metric space, some examples are:

- $\mathbb{R}^3$ together with the euclidean distance:

$$d(a,b) = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2 + (a_3 - b_3)^2}$$

- $\mathbb{R}^2$ together with the manhattan distance:

$$d(a,b) = |a_1 - b_1| + |a_2 - b_2|$$

Distance and metrics

A set with a metric is called a metric space, some examples are:

- $\mathbb{R}^3$ together with the euclidean distance:

$$d(a,b) = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2 + (a_3 - b_3)^2}$$

- $\mathbb{R}^2$ together with the manhattan distance:

$$d(a,b) = |a_1 - b_1| + |a_2 - b_2|$$

- Any set together with the discrete metric:

$$d(a,b) = 1, \quad a \neq b$$

Distance and metrics

However the set of vertices V in a graph can also be considered a metric space:

Distance and metrics

However the set of vertices V in a graph can also be considered a metric space:

- We obviously cannot use the euclean or manhattan distance though, since our vertices don't have coordinates.

Distance and metrics

However the set of vertices V in a graph can also be considered a metric space:

- We obviously cannot use the euclean or manhattan distance though, since our vertices don't have coordinates.
- The discrete metric obviously works, but doesn't tell us much.

Distance and metrics

Let us consider a connected undirected unweighted graph G with vertices V.

Theorem

The set of vertices V of a undirected simple graph is a metric space together with metric defines as the shortest path between vertices.

Proof.

We check if shortest path fulfills the required properties:

- The shortest path to itself is zero, hence: $d(x, x) = 0$ and for all other pairs it must be $d(x, y) > 0$ since there must be at least one edge between them.

□

Distance and metrics

- Since the graph is undirected any path from x to y is also a path from y to x by reversing the order. We have:

$$d(x, y) = d(y, x)$$

Since we could otherwise reverse the order of the shorter distance to create one of the same length in the other direction.

Distance and metrics

- Since the graph is undirected any path from x to y is also a path from y to x by reversing the order. We have:

$$d(x, y) = d(y, x)$$

  Since we could otherwise reverse the order of the shorter distance to create one of the same length in the other direction.
- The shortest path is obviosly non-negative since we can't have a negative number of edges between two vertices.

Distance and metrics

- Last to check we have the triangle inequality:

$$d(x, y) + d(y, z) \geq d(x, z)$$

Distance and metrics

- Last to check we have the triangle inequality:

$$d(x, y) + d(y, z) \geq d(x, z)$$

- If we assume that this is not true for shortest path, then there is vertices x, y, z such that:

$$d(x, y) + d(y, z) < d(x, z)$$

Distance and metrics

- Last to check we have the triangle inequality:

$$d(x, y) + d(y, z) \geq d(x, z)$$

- If we assume that this is not true for shortest path, then there is vertices x, y, z such that:

$$d(x, y) + d(y, z) < d(x, z)$$

- However this would mean that we would get a path from x to z going through y which is shorter than the shortest path from x to z. We have a contradiction and the triangle inequality must be true.

Distance and metrics

- Last to check we have the triangle inequality:

$$d(x, y) + d(y, z) \geq d(x, z)$$

- If we assume that this is not true for shortest path, then there is vertices x, y, z such that:

$$d(x, y) + d(y, z) < d(x, z)$$

- However this would mean that we would get a path from x to z going through y which is shorter than the shortest path from x to z. We have a contradiction and the triangle inequality must be true.
- All four properties hold and we see that shortest path is a metric.

Distance and metrics

- If we have a weighted graph with positive weights it can be shown that the shortest path, where the length of a path is the sum of the weights on the edges of the path is also a metric.

Distance and metrics

- If we have a weighted graph with positive weights it can be shown that the shortest path, where the length of a path is the sum of the weights on the edges of the path is also a metric.
- If we have a directed graph, then shortest path is no longer a metric since it might not be symmetric. Functions that fulfill all except the symmetry is sometimes called quasimetrics.

Distance and metrics

- Another type of metric used for tree graphs are *Zhong defined as:*

$$d(x, y) = 2m(ccp(x, y)) - m(x) - m(y)$$

$$m(c) = \frac{1}{k^{depth(c)+1}}$$

Where $m(c)$ called the milestone depend on the depth of node c in the tree and $ccp(x, y)$ is the closes common parent node. k is a positive constant.

Distance and metrics

- Another type of metric used for tree graphs are *Zhong defined as:*

$$d(x, y) = 2m(ccp(x, y)) - m(x) - m(y)$$

$$m(c) = \frac{1}{k^{depth(c)+1}}$$

Where $m(c)$ called the milestone depend on the depth of node c in the tree and $ccp(x, y)$ is the closes common parent node. k is a positive constant.

- This metric is useful when you want the depth of the nodes to influence the distance between them, for example if you want the distance of nodes higher up in the tree to be larger.

Similarity

An alternative to the distance between nodes is to instead use a similarity measurement between them.

- If two nodes are are "close" the similarity is large and if they are "far apart" the similarity is close to zero.

Similarity

An alternative to the distance between nodes is to instead use a similarity measurement between them.

- If two nodes are are "close" the similarity is large and if they are "far apart" the similarity is close to zero.
- Similarity between nodes is not as properly defined, it is however often easy to go from similarity to a metric or the other way around: If we got the distance $d(x, y)$ between nodes, we can define the similarity as $sim(x, y) = max(D) - d(x, y)$, where $max(D)$ is the maximum distance between any pair in the graph.

Similarity

An alternative to the distance between nodes is to instead use a similarity measurement between them.

- If two nodes are are "close" the similarity is large and if they are "far apart" the similarity is close to zero.
- Similarity between nodes is not as properly defined, it is however often easy to go from similarity to a metric or the other way around: If we got the distance $d(x, y)$ between nodes, we can define the similarity as $sim(x, y) = max(D) - d(x, y)$, where $max(D)$ is the maximum distance between any pair in the graph.
- Similarily we can transform a similarity measurement to a metric (possibly not fulfilling all the properties).

Similarity

Examples of similarity measurements are:

- Covariance or correlation.

Similarity

Examples of similarity measurements are:

- Covariance or correlation.
- LCH defined on tree graphs as:

$$sim(x, y) = -\log \frac{sp(x, y) + 1}{2M}$$

Where $sp(x, y)$ is the shortest path and M is the maximum depth in the graph.

Similarity

The choice of similarity or distance often comes down to the choice of method, or what the possible weights on edges represent.

- If weights on edges represent for example length or travel time, distance is probably more appropriate.

Similarity

The choice of similarity or distance often comes down to the choice of method, or what the possible weights on edges represent.

- If weights on edges represent for example length or travel time, distance is probably more appropriate.
- If weights represent correlation or number of times two objects was found or used together, similarity is probably better.

Centrality and ranking

Centrality of a node in a network is a measurement on how important the node is within the graph. This could for example be:

- How often a certain road was used in a road network?

Centrality and ranking

Centrality of a node in a network is a measurement on how important the node is within the graph. This could for example be:

- How often a certain road was used in a road network?
- How influential someone is in a social network?

Centrality and ranking

Centrality of a node in a network is a measurement on how important the node is within the graph. This could for example be:

- How often a certain road was used in a road network?
- How influential someone is in a social network?
- How important a server is for the overall network stability, which are the major potential bottlenecks?

Degree centrality

The easiest commonly used centrality measure is **degree centrality**:

Definition

Degree centrality $C_D(v)$ of a vertex v in a graph G with vertices V and edges E is defined as:

$$C_D(v) = deg(v)$$

Where the *degree* $deg(v)$ is the number of edges connected with v, loops (edge from v to v) are counted twice.

- If the graph is directed we let **indegree** be the number of edges pointing towards v. and the **outdegree** be the number of edges starting in v.

Degree centrality

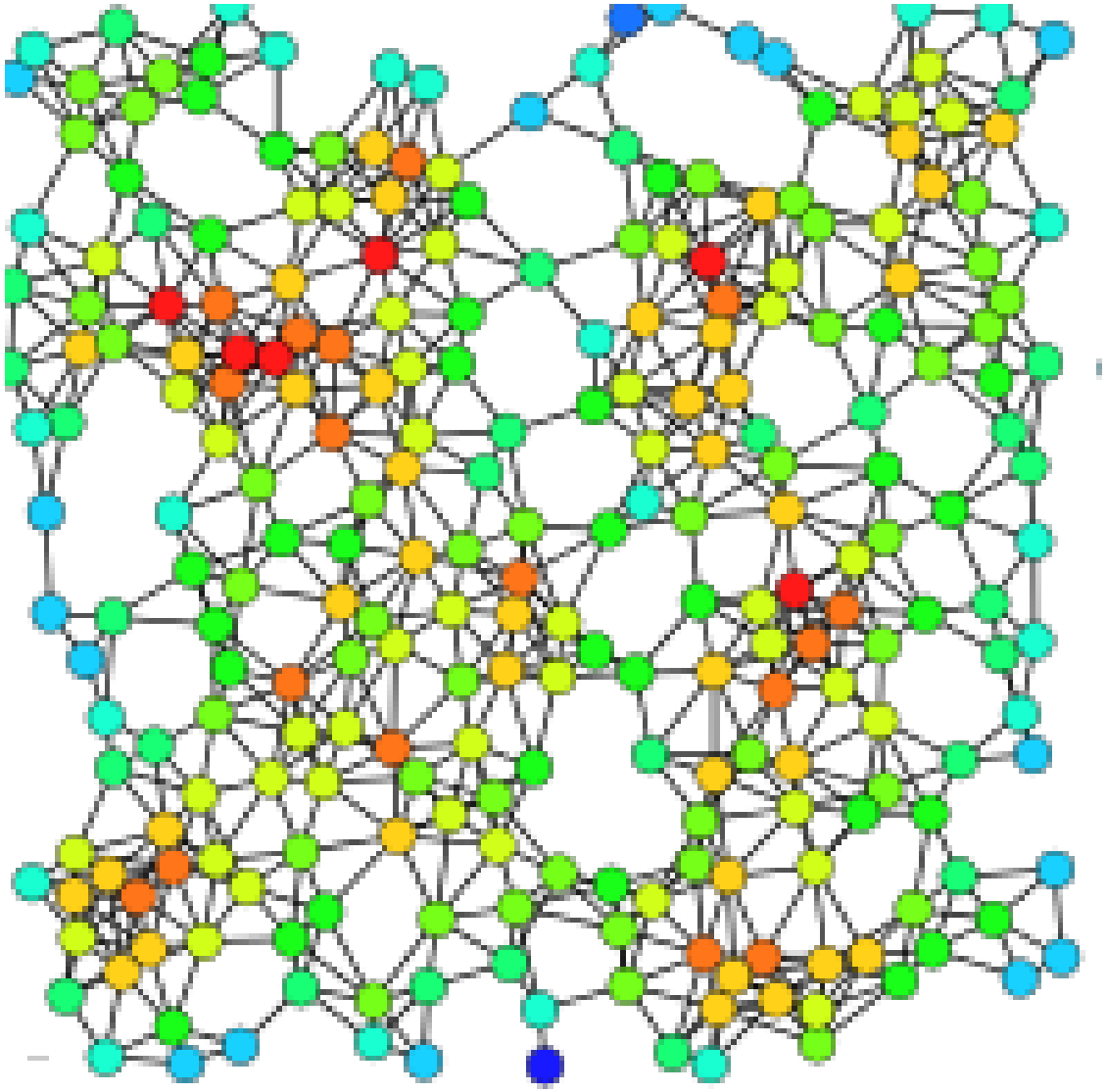

Figure: Example of degree centrality of a graph

Closeness centrality

Definition

Closeness centrality $C_C(v)$ of a vertex v in a graph G with $|V|$ vertices V and edges E is defined as:

$$C_C(v_i) = \frac{1}{\sum_{j=1}^{|V|} sp(v_i, v_j)}$$

Where $sp(v_i, v_j)$ is the length of the shortest path between v_i and v_j.

Closeness centrality

- The shorter the paths is from a vertex to all other vertices the larger is the closeness.

Closeness centrality

- The shorter the paths is from a vertex to all other vertices the larger is the closeness.
- The closeness can be seen as a measurement on how fast information can spread from the vertex to all other in for example a broadcast newtork.

Closeness centrality

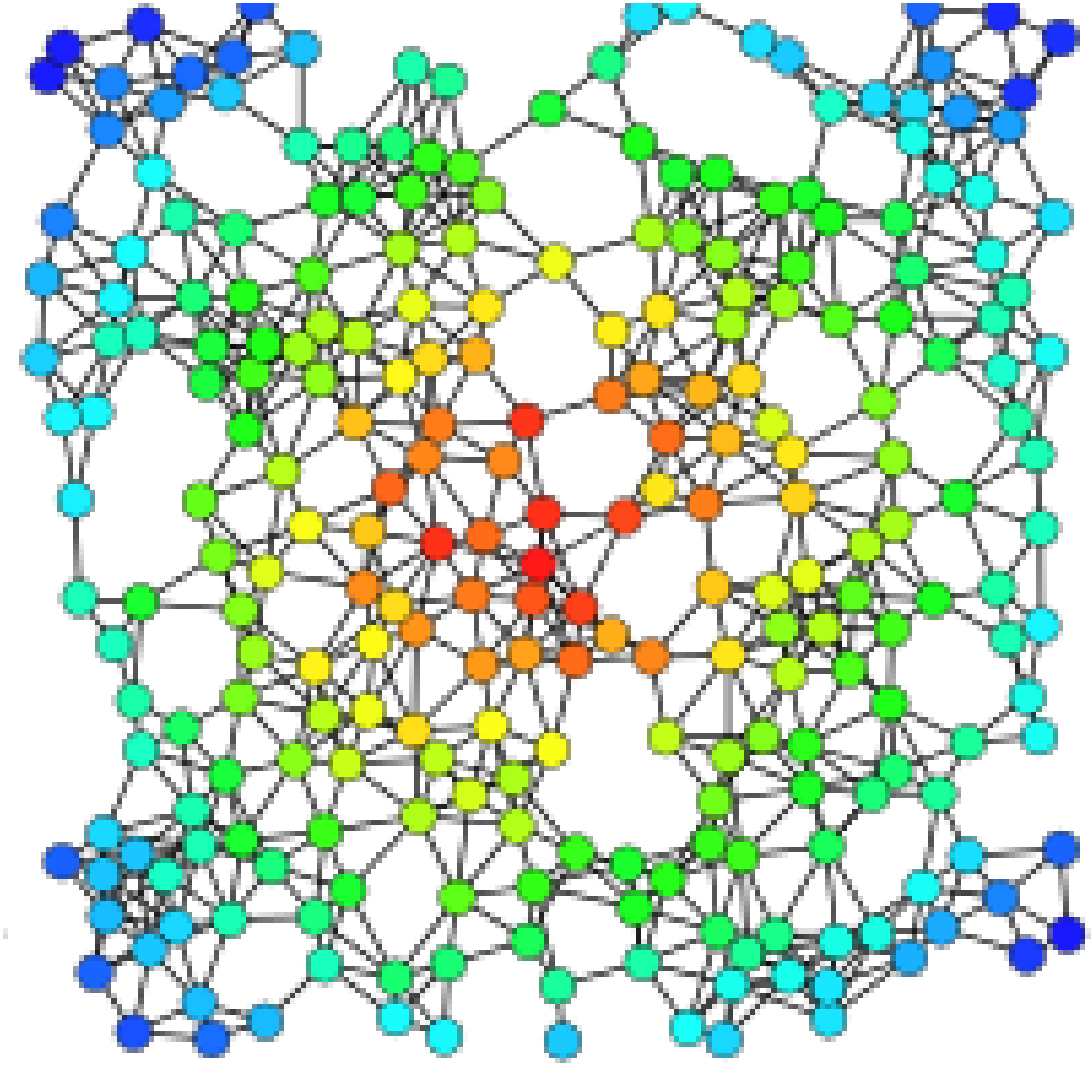

Figure: Example of closeness centrality of a graph

Betweenness centrality

Definition

Betweenness centrality $C_B(v)$ of a vertex v in a graph G with vertices V and edges E is defined as:

$$C_C(v) = \sum_{s \neq v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

Where σ_{st} is the total number of shortest paths from s to t and $\sigma_{st}(v)$ is the total number of shortest paths from s to t that pass through v.

Betweenness centrality

- Betweenness was initially used to measure the control one human have over the comminucation of others in social networks.

Betweenness centrality

- Betweenness was initially used to measure the control one human have over the comminucation of others in social networks.
- It's a good way to identify possible bottlenecks in a network.

Betweenness centrality

- Betweenness was initially used to measure the control one human have over the comminucation of others in social networks.
- It's a good way to identify possible bottlenecks in a network.
- Sometimes the betweenness is normalized by dividing by the total number of pair of vertices not including v.

Betweenness centrality

- Betweenness was initially used to measure the control one human have over the comminucation of others in social networks.
- It's a good way to identify possible bottlenecks in a network.
- Sometimes the betweenness is normalized by dividing by the total number of pair of vertices not including v.
- Calculating either the betweenness or the closeness earlier is quite expensive, since we need to find the shortest path of every pair of vertices.

Betweenness centrality

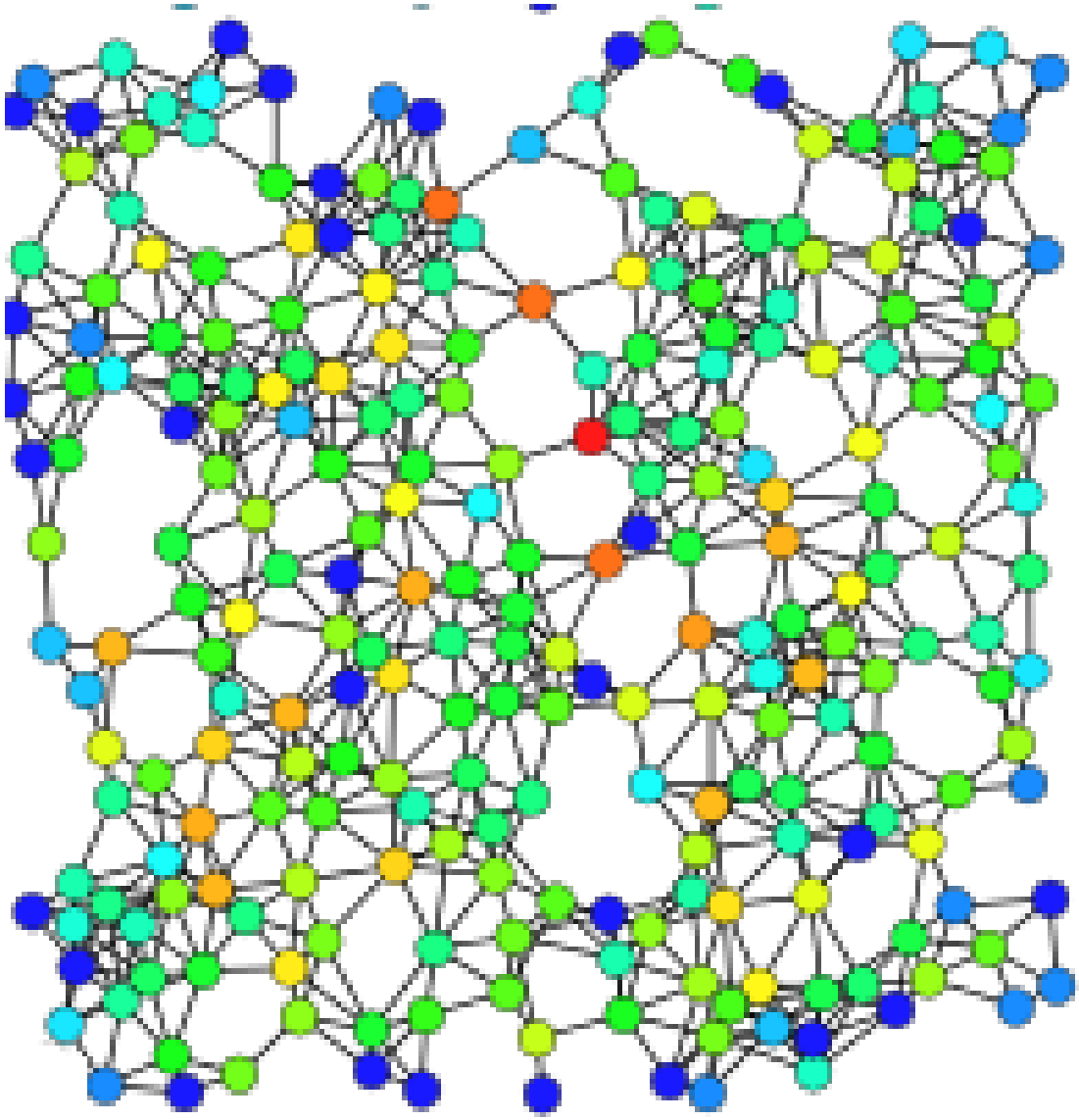

Figure: Example of betweenness centrality of a graph

Eigenvector centrality

Definition

Given a connected graph G with vertices V, edges E and adjacency matrix A. From the eigenvalue equation:

$$Ax = \lambda x$$

We get the **eigenvector centrality** of vertex v_i as the i'th element of x, where x is the eigenvector to the dominant eigenvalue of A.

Eigenvector centrality

- x is a positive vector (given by Perron-Frobenius for non-negative irreducible matrices).

Eigenvector centrality

- x is a positive vector (given by Perron-Frobenius for non-negative irreducible matrices).
- Eigenvector centrality gives a measure of the influence of nodes in a network.

Eigenvector centrality

- x is a positive vector (given by Perron-Frobenius for non-negative irreducible matrices).
- Eigenvector centrality gives a measure of the influence of nodes in a network.
- Eigenvector centrality can be computed efficiently for even extremely large sparse graphs through the use of the Power method.

Eigenvector centrality

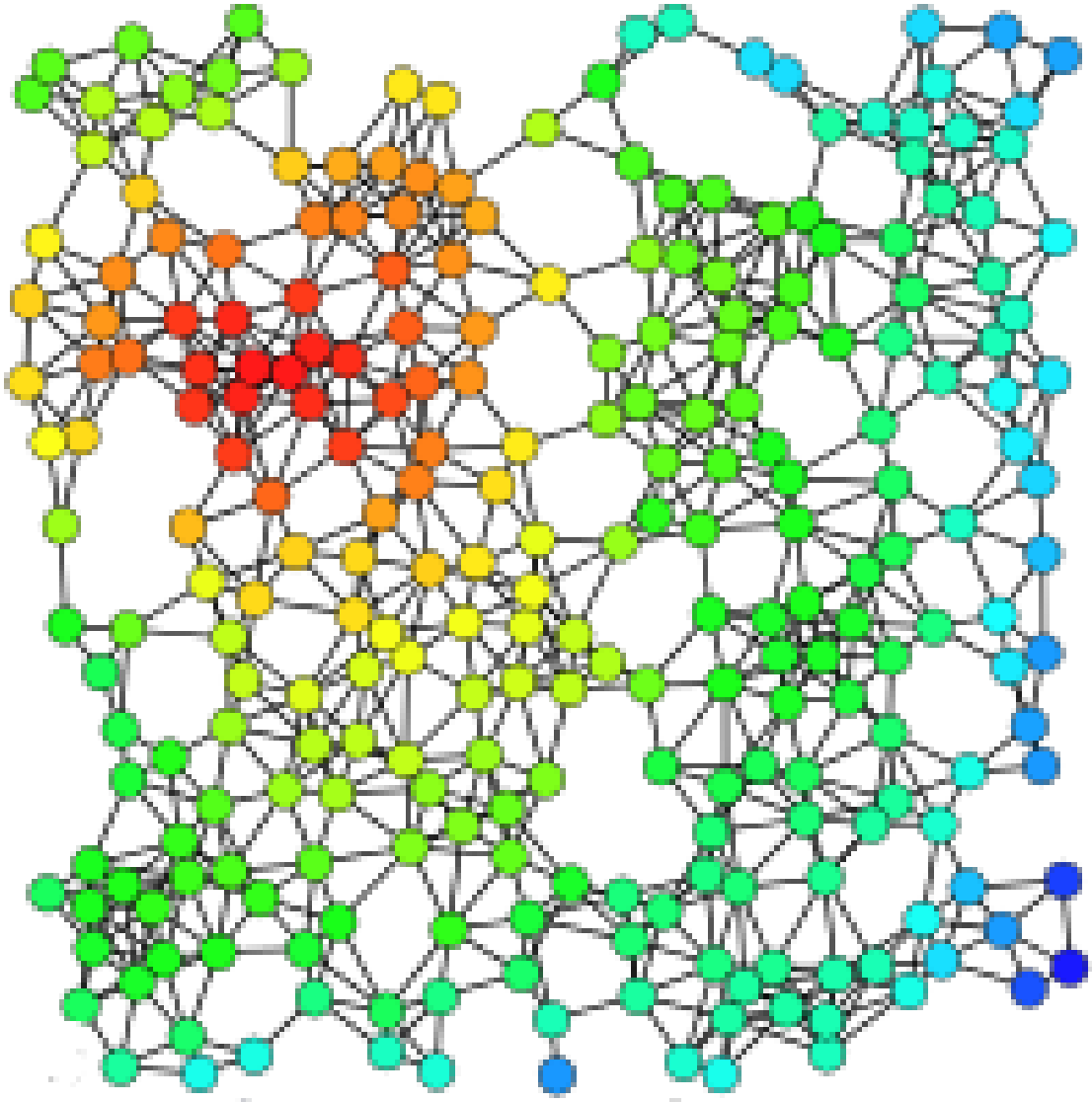

Figure: Example of eigenvector centrality of a graph

Other centrality measures

Two other centrality measures is

- **Katz centrality** which can be written as:

$$x_i = \sum_{k=1}^{\infty} \sum_{j=1}^{N} \alpha^k (A^k)_{ij}, \alpha \in (0, 1)$$

Other centrality measures

Two other centrality measures is

- **Katz centrality** which can be written as:

$$x_i = \sum_{k=1}^{\infty} \sum_{j=1}^{N} \alpha^k (A^k)_{ij}, \alpha \in (0, 1)$$

- This can be seen as a generalized version of degree centrality where vertices further away are also counted but with a discounting factor α.

Other centrality measures

Two other centrality measures is

- **Katz centrality** which can be written as:

$$x_i = \sum_{k=1}^{\infty} \sum_{j=1}^{N} \alpha^k (A^k)_{ij}, \alpha \in (0, 1)$$

- This can be seen as a generalized version of degree centrality where vertices further away are also counted but with a discounting factor α.
- **PageRank** is another centrality measure which we will look at more closely during one of the seminars.

Centralization

The **centralization** of a network is a measure of how central the most central node is in relation to how central al other are.

Definition

Centralization C_x of a graph G with vertices V and edges E is defined as:

$$C_x = \frac{\sum_{i=1}^{N} C_x(p_*) - C_x(p_i)}{\max \sum_{i=1}^{N} C_x(p_*) - C_x(p_i)}$$

Where $C_x(p_i)$ is the centrality of point p_i, $C_x(p_*)$ is the largest centrality in the network and $\max \sum_{i=1}^{N} C_x(p_*) - C_x(p_i)$ is the largest theoretical sum of differences in centrality for any graph with the same number of vertices.

Hierarchical clustering

Hierarchical clustering is a method which doesn't try to create one set of clusters, instead it aims to create a hierarchy of clusters. There are two primary types of hierarchical clustering:

Hierarchical clustering

Hierarchical clustering is a method which doesn't try to create one set of clusters, instead it aims to create a hierarchy of clusters. There are two primary types of hierarchical clustering:

- **Agglomerative**, in which each vertex starts in it's own cluster, clusters are then merged to create larger and larger clusters.

Hierarchical clustering

Hierarchical clustering is a method which doesn't try to create one set of clusters, instead it aims to create a hierarchy of clusters. There are two primary types of hierarchical clustering:

- **Agglomerative**, in which each vertex starts in it's own cluster, clusters are then merged to create larger and larger clusters.
- **Divisive**, in which all vertices starts in a single cluster, clusters are then split to create smaller and smaller clusters.

Hierarchical clustering

Hierarchical clustering is a method which doesn't try to create one set of clusters, instead it aims to create a hierarchy of clusters. There are two primary types of hierarchical clustering:

- **Agglomerative**, in which each vertex starts in it's own cluster, clusters are then merged to create larger and larger clusters.
- **Divisive**, in which all vertices starts in a single cluster, clusters are then split to create smaller and smaller clusters.
- Because of computational limits, new clusters are generally decided using a "greedy algorithm", in which the "best" new merge/split is choosen every time based on the old clusters.

Hierarchical clustering

Most forms of hierarchical clustering use some metric in order to define the distance between points in space or vertices in a graph.

- After a metric is chosen we also need some way to determine the distance between two sets of points or vertices.

Hierarchical clustering

Most forms of hierarchical clustering use some metric in order to define the distance between points in space or vertices in a graph.

- After a metric is chosen we also need some way to determine the distance between two sets of points or vertices.
- We call this a linkage criterion.

Hierarchical clustering

Some common Linkage criterions between two different sets are:

- Maximum: $max\{d(a, b) : a \in A, b \in B\}$

Hierarchical clustering

Some common Linkage criterions between two different sets are:

- Maximum: $max\{d(a,b) : a \in A, b \in B\}$
- Minimum: $min\{d(a,b) : a \in A, b \in B\}$

Hierarchical clustering

Some common Linkage criterions between two different sets are:

- Maximum: $max\{d(a,b) : a \in A, b \in B\}$
- Minimum: $min\{d(a,b) : a \in A, b \in B\}$
- Mean: $\frac{1}{|A||B|} \sum_{a \in A} \sum_{b \in B} d(a,b)$

Hierarchical clustering

Some common Linkage criterions between two different sets are:

- Maximum: $max\{d(a,b) : a \in A, b \in B\}$
- Minimum: $min\{d(a,b) : a \in A, b \in B\}$
- Mean: $\frac{1}{|A||B|} \sum_{a \in A} \sum_{b \in B} d(a,b)$
- Probability that cluster come from the same distribution.