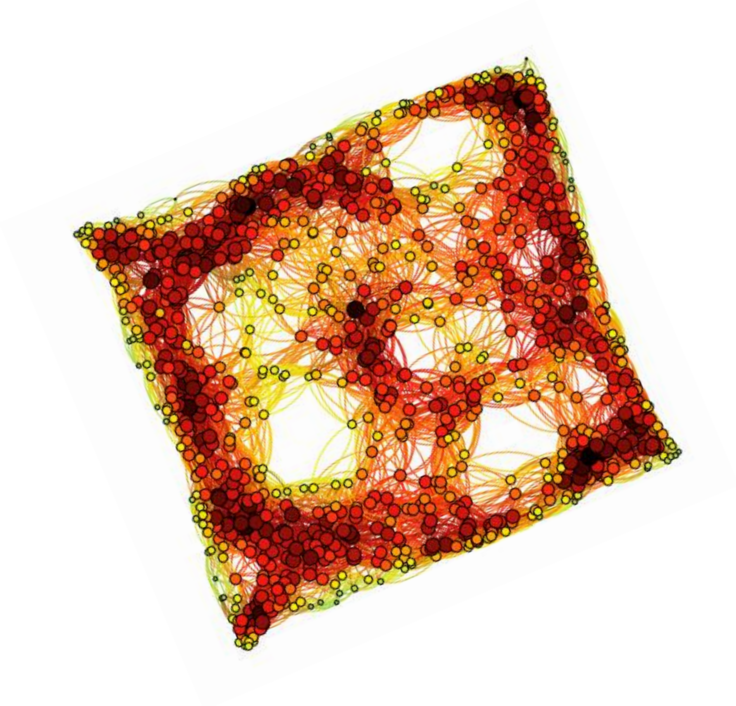
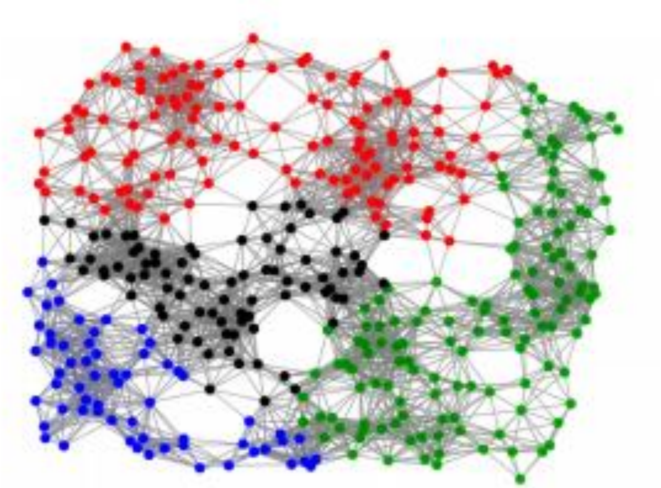
K-Means on graphs using PageRank



Mustafa Hajij

Problem

Given a graph G(V,E) (directed or undirected), we like to segment the graph into k "natural" subgraphs.

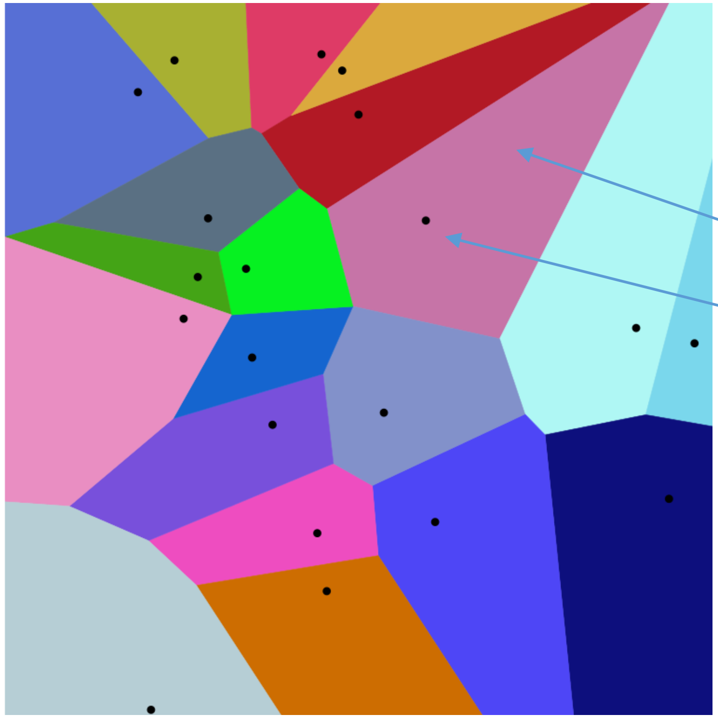


Voronoi cells

Let (X, d) be a metric space and let $C \subset X$ be subset of X, called the the subset of centroids.

The Voronoi cell at point $c \in C$, denoted by VC(c) is defined to be the set of all points $y \in X$ that are closer to c than to any other point in C.

The collection of subsets VC(c) for all c in C is called the Voronoi diagram, denoted by VD(C) of the metric space X with respect to the subset C.



The Voronoi cell of this black dot

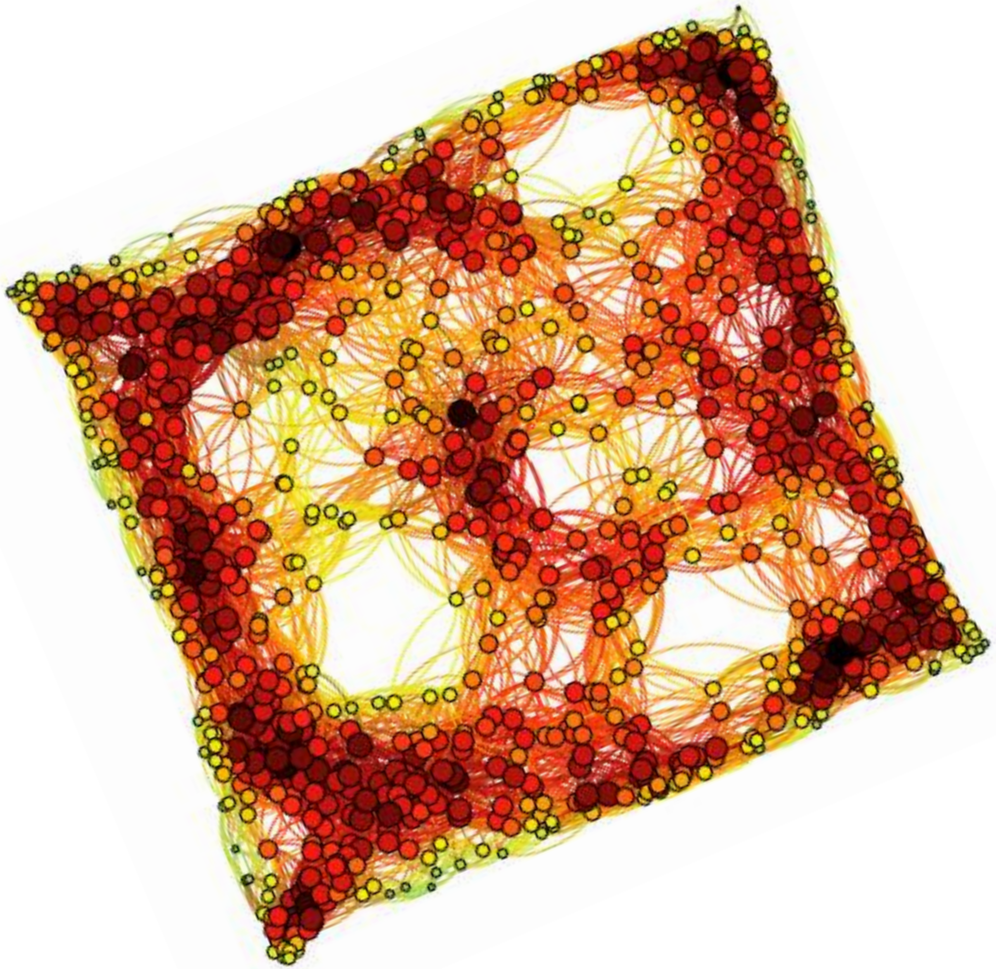
The Voronoi diagram of the black dots

Image source : https://en.wikipedia.org/wiki/File:Euclidean_Voronoi_diagram.svg

PageRank

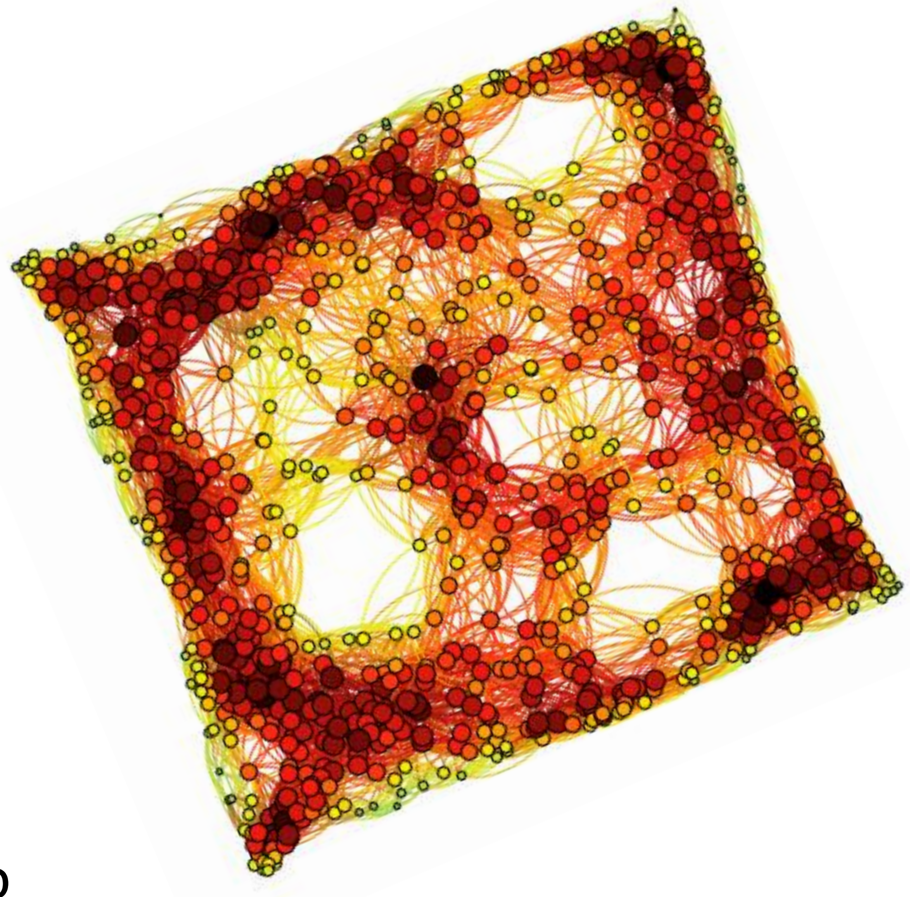
$$R(v) = \frac{(1-d)}{|V|} + d \sum_{u \in N(v)} \frac{R(u)}{|N(u)|}$$

Where N(v) is the set of neighbors of v;
0 < d < 1 is the damping factor.



PageRank

$$R(v) = \frac{(1-d)}{|V|} + d \sum_{u \in N(v)} \frac{R(u)}{|N(u)|}$$



Intuitively, a high PageRank value at a given node v usually means that v is connected to many other nodes, which also have high PageRank scores.

From this perspective, PageRank can be viewed as a measure of centrality for the nodes of the graph.

PageRank-based k-means clustering

Algorithm 1: PageRank-based k-means clustering algorithm on graphs.

Input: Graph $G(V, E)$, number of clusters k.
Output: A partition of the node set V into k subsets.
Initialize the set C by choosing k nodes from V
while *While termination criterion has not been met* **do**
  for c_i *in* C **do**
    Compute $V_i = VC(c_i)$
  end
  for V_i *in* $VD(C)$ **do**
    Compute PageRank PR_i on the subgraph (V_i, E_i)
    $c_i := argmax_{v \in V_i}(PR_i(v))$
  end
end

Examples

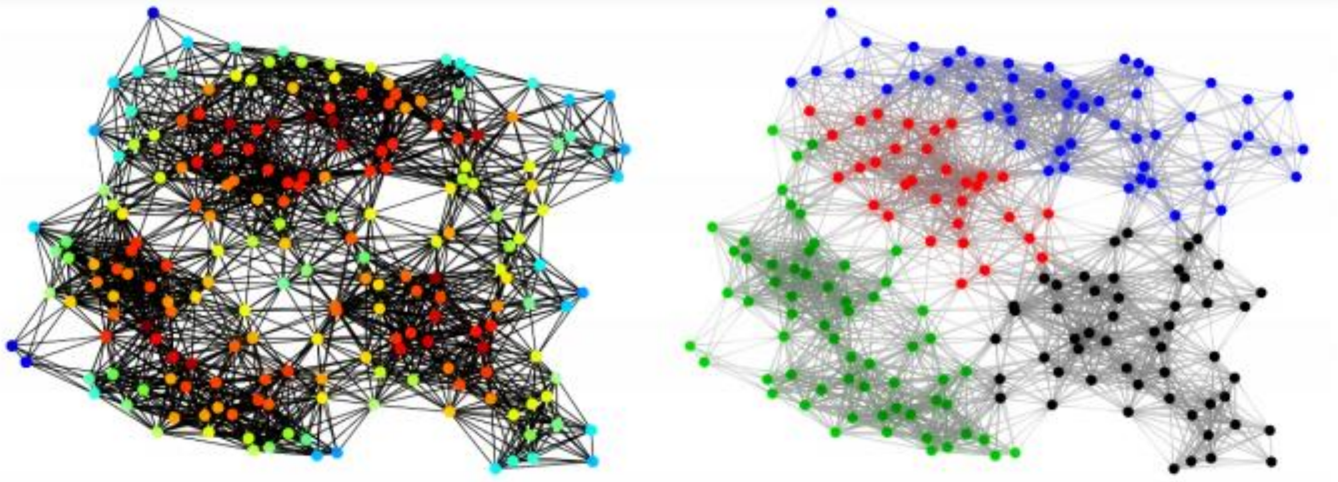
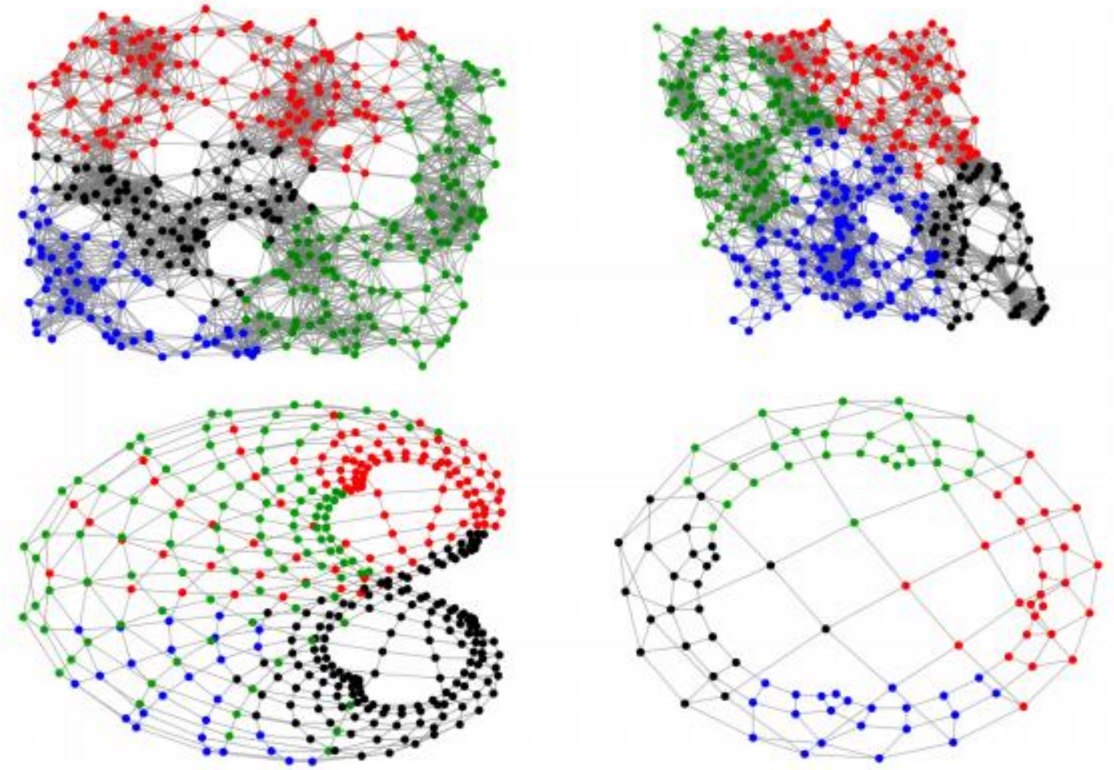
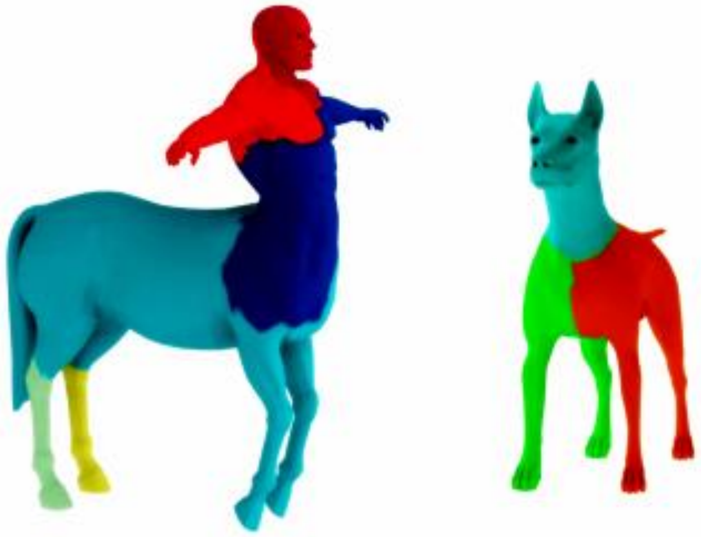


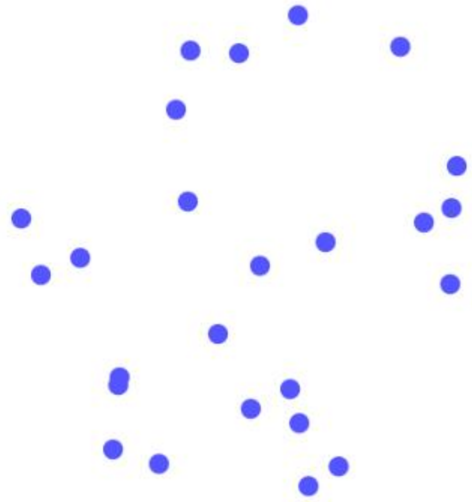
Figure 5: *The PageRank function is utilized as a centrality measure in our work. The figure shows the visualization of the PageRank function on the nodes of the graph on the left. On the right we show the application of our algorithm on the same graph with $k = 4$. The clusters are indicated by the colors of the nodes.*

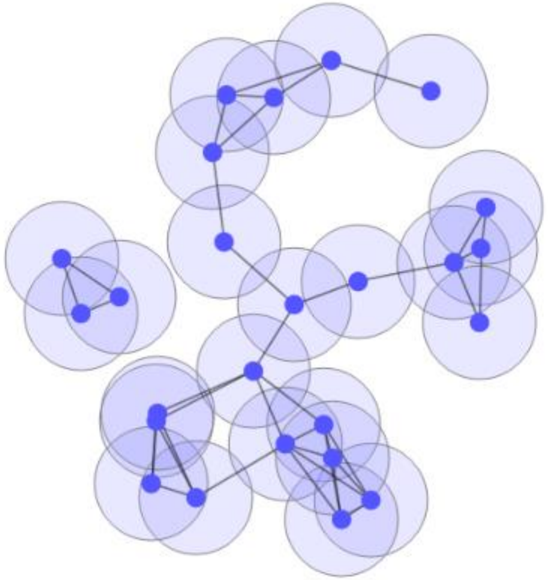
Examples

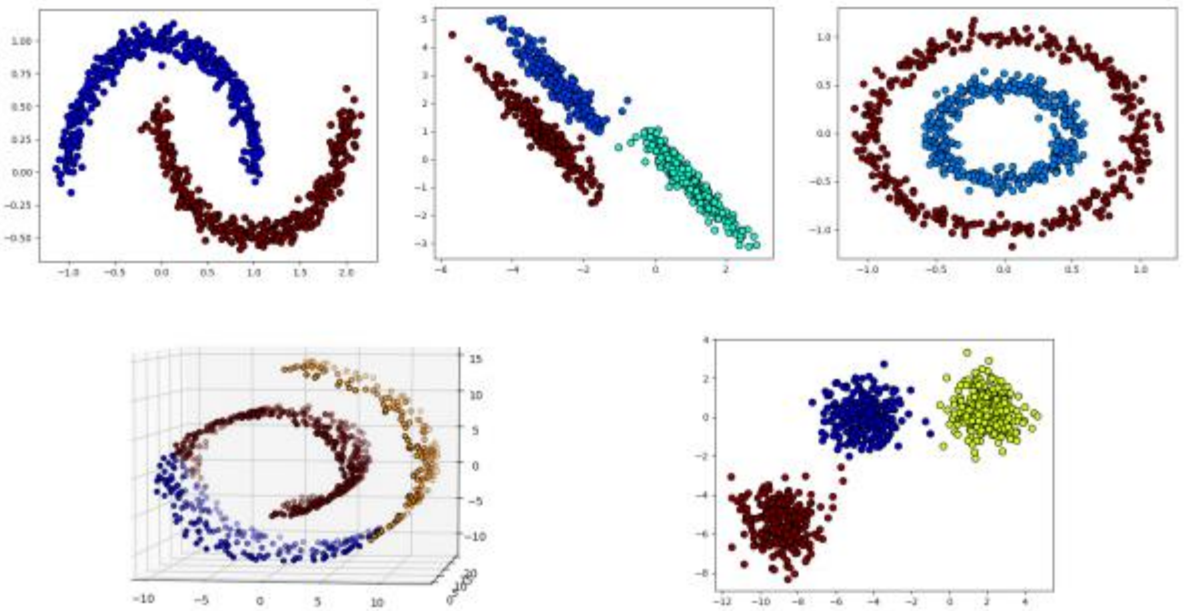




Clustering point clouds using the same method







Other centrality measures and generalization

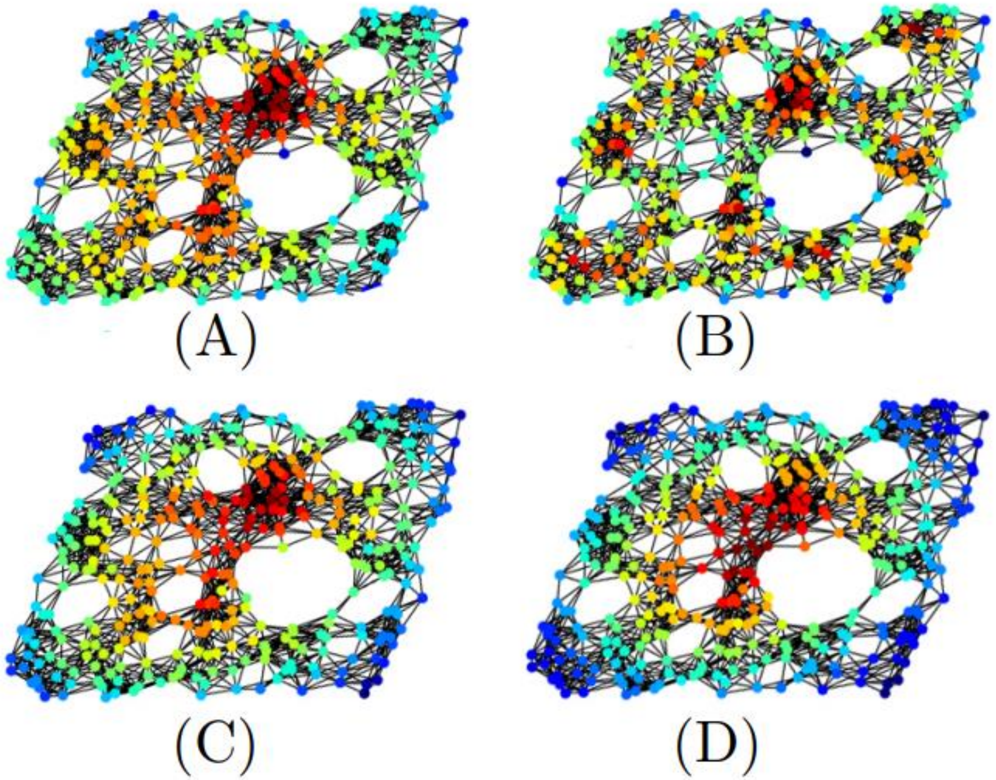


Figure 1: *Various centrality measures on graphs. (A) Information centrality, (B) PageRank. (C) Harmonic centrality. (D) Closeness centrality.*

Other centrality measures on graphs and generalization

For instance Harmonic centrality depends only on the metric:

$$H(x) = \sum_{y \neq x} \frac{1}{d(y, x)}$$

So the algorithm can be generalized to metric spaces :

while *While termination criterion has not been met* **do**
  for c_i *in* C **do**
    Compute $V_i = VC(c_i)$
  end
  for V_i *in* $VD(C)$ **do**
    Compute H_i on V_i
    $c_i := argmax_{v \in V_i}(H_i(v))$
  end
end