Random geometric graphs

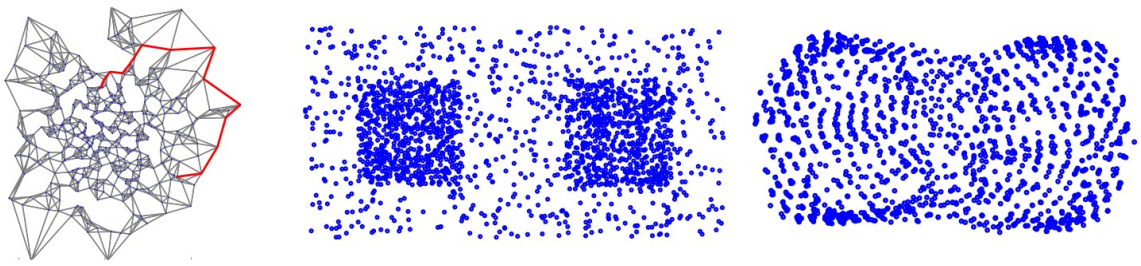

Figure 1.1: Left: Shortest path (in red) in an unweighted k-nearest neighbor graph on a random set of points. The path is far from the straight line between the two end points. Middle and right: a toy data set (middle) and its Isomap reconstruction based on the shortest path distance in the unweighted kNN graph. The reconstrution does not preserve the density information of the original data set.

Random geometric graphs are built by first sampling a set of points from some underlying distribution, and then connecting each point to its k nearest neighbors. In this project we investigated the behavior of distance functions on random geometric graphs when the sample size n goes to infinity (and the connectivity parameter k scales appropriately).

It is well known that in graphs where the edges are suitably weighted according to their Euclidean lengths, the **shortest path distance** converges to the underlying Euclidean distance. However, it turned out this is not the case for unweighted kNN graphs [461]. In this case, the shortest path distance converges to a distance function that is weighted by the underlying density and takes wide detours to avoid high density regions (see left figure above). In machine learning applications, this behavior of the shortest path distance can be highly misleading. As an example, consider the Isomap algorithm and the data set shown in the middle figure. If we build an unweighted kNN graph based on this data and apply Isomap to recover the point configuration, we get the figure on the right. Obviously, it is grossly distorted and cannot serve as a faithful representation of the original data.

The **commute distance** (aka **resistance distance**) between vertex u and v is defined as the expected time it takes the natural random walk starting in vertex u to travel to vertex v and back. It is widely used in machine learning because it supposedly satisfies the following, highly desirable property: Vertices in the same cluster of the graph have a small commute distance, whereas vertices in different clusters of the graph have a large commute distance to each other. We studied the behavior of the commute distance as the number of vertices in the graph tends to infinity [599], proving that the commute distance between two points converges to a trivial quantity that only takes into account the degree of the two vertices. Hence, all information about cluster structure gets lost when the graph is large enough.

To alleviate this shortcoming, we proposed the family of **p-resistances** [497]. For $p = 1$ it reduces to the shortest path distance, for $p = 2$ it coincides with the resistance distance, and for $p \to \infty$ it is related to the minimal s-t-cut in the graph. The family shows an interesting phase transition: there exist two critical thresholds p^* and p^{**} such that if $p < p^*$, then the p-resistance depends on meaningful global properties of the graph, whereas if $p > p^{**}$, it only depends on trivial local quantities and does not convey any useful information. In particular, the p-resistance for $p = p^*$ nicely reveals the cluster structure.

More information: https://ei.is.tuebingen.mpg.de/project/distance-functions-on-random-geometric-graphs