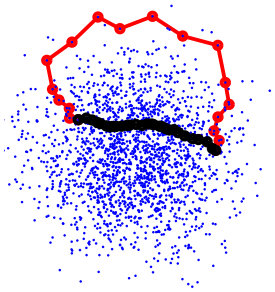
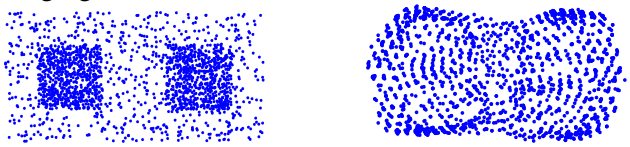
Distance functions on random geometric graphs

Finding appropriate distance functions on data is an important issue in data preprocessing. In the context of graphs, distance functions between vertices can be used to reveal global geometric properties of a graph, such as manifold structure or cluster structure. In a statistical context it is important to find out how such distance functions behave as the amount of data "grows": does more data really reveal more information about the underlying geometry of the graph? We studied this question for a number of graph distance functions. Surprisingly, it turns out that in the limit of infinitely many vertices, many distance functions become trivial and do not encode the kind of global information we would like.

The best known distance function on graphs is the **shortest path distance**. It is well known that in ε-graphs or k nearest neighbor graphs with Gaussian weights this distance converges to the underlying Euclidean distance function. However, it turned out this is not the case for unweighted kNN graphs [2]. In this case, the shortest path distance converges to a distance function that is weighted by the underlying density. The "shortest paths" are the ones that take wide detours to avoid high density regions. This is illustrated in the following figure, which shows the shortest path according to Euclidean distance (black) and the shortest path in the kNN graph (red).



In machine learning applications such as manifold learning, this behavior of the shortest path distance can be highly misleading. As an example consider the following figure.



The left side shows a two-dimensional data set with non-uniform density. 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 data on the left.

The **commute 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 equivalent (up to a constant) to the resistance distance, which interprets the graph as an electrical network and defines the distance between vertices u and v as the effective resistance between these vertices. 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 [3], 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 **p-resistance** [1]. While the standard commute distance (resistance distance) can be expressed as a 2-norm optimization problem, we consider a p-norm optimization problem instead. It turns out that the corresponding family of distances has nice properties. First, the family includes several well-known distances as special cases: 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 converges to the inverse of the minimal s-t-cut in the graph. Secondly, the family shows an interesting phase transition when we study its properties for $n \to \infty$: We prove that 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 parameter p^* nicely reveals the cluster structure.

Publications

Articles in Conference Proceedings

[1] M Alamgir and U von Luxburg. Phase transition in the family of p-resistances. In J Shawe-Taylor, RS Zemel, P Bartlett, F Pereira, and KQ Weinberger, editors, *25th Annual Conference on Neural Information Processing Systems (NIPS)*, pages 379–387, Granada, Spain, 2011. 1

[2] M Alamgir and U von Luxburg. Shortest path distance in random k-nearest neighbor graphs. In *Proceedings of the 29th International Conference on Machine Learning*, Edinburgh, Scotland, 2012. International Machine Learning Society. 1

[3] U von Luxburg, A Radl, and M Hein. Getting lost in space: Large sample analysis of the resistance distance. In J Lafferty, C KI Williams, J Shawe-Taylor, R S Zemel, and A Culotta, editors, *24th Annual Conference on Neural Information Processing Systems (NIPS)*, pages 2622–2630, Vancouver, BC, Canada, 2010. Curran. 1