Cluster Identification in Nearest-Neighbor Graphs

Markus Maier, Matthias Hein, and Ulrike von Luxburg

Max Planck Institute for Biological Cybernetics, Tübingen, Germany
first.last@tuebingen.mpg.de

Abstract. Assume we are given a sample of points from some underlying distribution which contains several distinct clusters. Our goal is to construct a neighborhood graph on the sample points such that clusters are "identified": that is, the subgraph induced by points from the same cluster is connected, while subgraphs corresponding to different clusters are not connected to each other. We derive bounds on the probability that cluster identification is successful, and use them to predict "optimal" values of k for the mutual and symmetric k-nearest-neighbor graphs. We point out different properties of the mutual and symmetric nearest-neighbor graphs related to the cluster identification problem.

1 Introduction

In many areas of machine learning, neighborhood graphs are used to model local relationships between data points. Applications include spectral clustering, dimensionality reduction, semi-supervised learning, data denoising, and many others. However, the most basic question about such graph based learning algorithms is still largely unsolved: which neighborhood graph to use for which application and how to choose its parameters. In this article, we want to make a first step towards such results in a simple setting we call "cluster identification". Consider a probability distribution whose support consists of several high density regions (clusters) which are separated by a positive distance from each other. Given a finite sample, our goal is to construct a neighborhood graph on the sample such that each cluster is "identified", that is each high density region is represented by a unique connected component in the graph. In this paper we mainly study and compare mutual and symmetric k-nearest-neighbor graphs. For different choices of k we prove bounds on the probability that clusters can be identified. In toy experiments, the behavior of the bounds as a function of k corresponds roughly to the empirical frequencies. Moreover, we compare the different properties of the mutual and the symmetric nearest-neighbor graphs. Both graphs have advantages in different situations: if one is only interested in identifying the "most significant" cluster (while some clusters might still not be correctly identified), then the mutual kNN graph should be chosen. However, if one wants to identify many clusters simultaneously the bounds show no difference between the two graphs. Empirical evaluations show that in this case the symmetric kNN graph is to be preferred due to its better connectivity properties.

There is a huge amount of literature with very interesting results on connectivity properties of random graphs, both for Erdős-Rényi random graphs (Bollobas, 2001) and for geometric random graphs (Penrose, 2003). Applications include percolation theory (Bollobas and Riordan, 2006), modeling ad-hoc networks (e.g., Santi and Blough, 2003, Bettstetter, 2002, Kunniyur and Venkatesh, 2006), and clustering (e.g., Brito et al., 1997 and Biau et al., 2007). In all those cases the literature mainly deals with different kinds of asymptotic results in the limit for $n \to \infty$. However, what we would need in machine learning are finite sample results on geometric random graphs which can take into account the properties of the underlying data distribution, and which ideally show the right behavior even for small sample sizes and high dimensions. In this paper we merely scratch the surface of this long-term goal.
Let us briefly introduce some basic definitions and notation for the rest of the paper. We always assume that we are given n data points $X_1, ..., X_n$ which have been drawn i.i.d. from some underlying density on $\mathbb{R}^d$. Those data points are used as vertices in a graph. By $\mathrm{kNN}(X_j)$ we denote the set of the k nearest neighbors of X_j. The different neighborhood graphs, which are examples of geometric random graphs, are defined as

- the ε-neighborhood graph $G_{\mathrm{eps}}(n, \varepsilon)$: X_i and X_j connected if $\|X_i - X_j\| \leq \varepsilon$,
- the symmetric k-nearest-neighbor graph $G_{\mathrm{sym}}(n, k)$:
  X_i and X_j connected if $X_i \in \mathrm{kNN}(X_j)$ or $X_j \in \mathrm{kNN}(X_i)$,
- the mutual k-nearest-neighbor graph $G_{\mathrm{mut}}(n, k)$:
  X_i and X_j connected if $X_i \in \mathrm{kNN}(X_j)$ and $X_j \in \mathrm{kNN}(X_i)$.

2 Between- and within-cluster connectivity of mutual and symmetric kNN-graphs

This section deals with the connectivity properties of kNN graphs. The proof ideas are basically the same as in Brito et al. (1997). However, since we are more interested in the finite sample case we have tried to make the bounds as tight as possible. We also make all constants explicit, which sometimes results in long expressions, but allows to study the influence of all parameters. In Brito et al. (1997) the main emphasis was put on a rate of k which is sufficient for connectedness of the mutual kNN graphs, resulting in a choice of k that is proportional to $\log(n)$. However, if one is interested in identifying the clusters as the connected components of the mutual kNN graph one should optimize the trade-off between having high probability of being connected within clusters and high probability of having no edges between the different clusters. Most importantly, integrating the properties of the mutual and symmetric kNN graph we derive bounds which work for each cluster individually. This allows us later on to compare both graphs for different scenarios: identification of all clusters vs. the "most significant" one.
We assume that our clusters $C^{(1)}, \ldots, C^{(m)}$ are m disjoint, compact and connected subsets of $\mathbb{R}^d$. The distance of $C^{(i)}$ to its closest neighboring cluster

$C^{(j)}$ is denoted by $u^{(i)}$, where the distance between sets S_1, S_2 is measured by $d(S_1, S_2) = \inf\{\|x - y\| \,|\, x \in S_1,\, y \in S_2\}$. Let $p^{(i)}$ be a probability density with respect to the Lebesgue measure in $\mathbb{R}^d$ whose support is $C^{(i)}$. The sample points $\{X_i\}_{i=1}^n$ are drawn i.i.d. from the probability density $p(x) = \sum_{j=1}^m \beta_{(j)}\, p^{(j)}(x)$, where $\beta_{(j)} > 0$ for all j and $\sum_{j=1}^m \beta_{(j)} = 1$. We denote by $n^{(i)}$ the number of points in cluster $C^{(i)}$ $(i = 1, \ldots, m)$. The kNN radius of a point X_i is the maximum distance to a point in kNN(X_i). $R_{\min}^{(i)}$ and $R_{\max}^{(i)}$ denote the minimal and the maximal kNN radius of the sample points in cluster $C^{(i)}$. Bin(n, p) denotes the binomial distribution with parameters n and p. Since we often need tail bounds for the binomial distribution, we set $D\,(k; n, p) = \mathrm{P}\,(U \le k)$ and $E\,(k; n, p) = \mathrm{P}\,(U \ge k)$ for a Bin(n, p)-distributed random variable U. Finally, we denote the ball of radius r around x by $B(x, r)$, and the volume of the d-dimensional unit ball by η_d.
In the following we will need upper and lower bounds for the probability mass of balls around points in clusters. These are given by continuous and increasing functions $g_{\min}^{(i)}, \tilde{g}_{\min}^{(i)}, g_{\max}^{(i)} : [0, \infty) \to \mathbb{R}$ with $g_{\min}^{(i)}(t) \le \inf_{x \in C^{(i)}} \mathrm{P}\,(B(x, t))$, $\tilde{g}_{\min}^{(i)}(t) \le \inf_{B(x,t) \subseteq C^{(i)}} \mathrm{P}\,(B(x, t))$ and $g_{\max}^{(i)}(t) \ge \sup_{x \in C^{(i)}} \mathrm{P}\,(B(x, t))$.

2.1 Within-cluster connectivity of mutual kNN graphs

The analysis of connectedness is based on the following observation: If for an arbitrary $z > 0$ the minimal kNN radius is larger than z, then all points in a distance of z or less are connected in the kNN graph. If we can now find a covering of a cluster by balls of radius $z/4$ and every ball contains a sample point, then the distance between sample points in neighboring balls is less than z. Thus the kNN graph is connected. The following proposition uses this observation to derive a bound for the probability that a cluster is disconnected under some technical conditions on the boundary of the cluster. These technical conditions ensure that we do not have to cover the whole cluster but we can ignore a boundary strip (the collar set in the proposition). This helps in finding a better bound for the probability mass of balls of the covering.

Proposition 1 (Within-cluster connectedness of $G_{\mathbf{mut}}(n, k)$). *Assume that the boundary $\partial C^{(i)}$ of cluster $C^{(i)}$ is a smooth $(d-1)$-dimensional submanifold in $\mathbb{R}^d$ with maximal curvature radius $\kappa^{(i)} > 0$. For $\varepsilon \le \kappa^{(i)}$, we define the collar set $C^{(i)}(\varepsilon) = \{x \in C^{(i)} \,\big|\, d(x, \partial C^{(i)}) \le \varepsilon\}$ and the maximal covering radius $\varepsilon_{\max}^{(i)} = \max_{\varepsilon \le \kappa^{(i)}} \{\varepsilon \mid C^{(i)} \backslash C^{(i)}(\varepsilon)$ connected $\}$. Let $z \in \big(0, 4\,\varepsilon_{\max}^{(i)}\big)$. Given a covering of $C^{(i)} \backslash C^{(i)}(\frac{z}{4})$ with balls of radius $z/4$, let $\mathcal{F}_z^{(i)}$ denote the event that there exists a ball in the covering that does not contain a sample point. Then*

$$\mathrm{P}\big(\text{Cluster } C^{(i)} \text{ disconnected in } G_{mut}(n, k)\big) \le \mathrm{P}\big(R_{\min}^{(i)} \le z\big) + \mathrm{P}\big(\mathcal{F}_z^{(i)}\big). \quad (1)$$

Proof. The proof is based on the fact that the event $\{R_{\min}^{(i)} > z\} \cap \mathcal{F}_z^{(i)}$ implies connectedness of $C^{(i)}$. Namely, sample points lying in neighboring sets of the

covering of $C^{(i)} \backslash C^{(i)}(\frac{z}{4})$ have distance less than z. Therefore they are connected by an edge in the mutual kNN graph. Moreover, all sample points lying in the collar set $C^{(i)}(\frac{z}{4})$ are connected to some sample point in $C^{(i)} \backslash C^{(i)}(\frac{z}{4})$. □

The proof concept of Propositions 1 and 3 does not require smoothness of the boundary of the cluster. However, the more general case requires a different construction of the covering which leads to even more technical assumptions and worse constants.

Proposition 2 (Minimal kNN radius). *For all $z > 0$*

$$\mathrm{P}\big(R_{\min}^{(i)} \leq z\big) \;\leq\; n\,\beta_{(i)}\,E\big(k; n-1, g_{\max}^{(i)}(z)\big).$$

Proof. Assume without loss of generality that $X_1 \in C^{(i)}$ (after a suitable permutation). Define $M_s = |\{j \neq s \,|\, X_j \in B\,(X_s, z)\}|$ for $1 \leq s \leq n$. Then

$$\mathrm{P}\big(R_{\min}^{(i)} \leq z \,|\, n^{(i)} = l\big) \;\leq\; l\,\mathrm{P}\big(M_1 \geq k\big).$$

Since $n^{(i)} \sim \mathrm{Bin}\,\big(n, \beta_{(i)}\big)$, we have

$$\mathrm{P}\big(R_{\min}^{(i)} \leq z\big) \leq \sum_{l=0}^{n} l\,\mathrm{P}\big(M_1 \geq k\big)\mathrm{P}(n^{(i)} = l) = n\beta_{(i)}\mathrm{P}\big(M_1 \geq k\big).$$

Since $M_1 \sim \mathrm{Bin}\,(n-1, \mathrm{P}\,(B(X_1, z)))$, with $\mathrm{P}\,(B(X_1, z)) \leq g_{\max}^{(i)}\,(z)$ we obtain $\mathrm{P}\,(M_1 \geq k) \leq E\left(k; n-1, g_{\max}^{(i)}\,(z)\right)$. □

Proposition 3 (Covering with balls). *Under the conditions of Proposition 1 there exists a covering of $C^{(i)} \backslash C^{(i)}(\frac{z}{4})$ with N balls of radius $z/4$, such that $N \leq \big(8^d \,\mathrm{vol}\,\big(C^{(i)}\big)\big) \,/\, \big(z^d \eta_d\big)$ and*

$$\mathrm{P}\left(\mathcal{F}_z^{(i)}\right) \;\leq\; N\,\left(1 - \tilde{g}_{\min}^{(i)}\left(\frac{z}{4}\right)\right)^n.$$

Proof. A standard construction using a $z/4$-packing provides us with the covering. Due to the conditions of Proposition 1 we know that balls of radius $z/8$ around the packing centers are disjoint and subsets of $C^{(i)}$. Thus the sum of the volumes of these balls is bounded by the volume of the cluster and we obtain $N\,(z/8)^d\,\eta_d \;\leq\; \mathrm{vol}\,\big(C^{(i)}\big)$. Using a union bound over the covering with a probability of $\big(1 - \tilde{g}_{\min}^{(i)}\,\big(\frac{z}{4}\big)\,\big)^n$ for one ball being empty we obtain the bound. □

The following proposition gives an easy extension of the result of Proposition 1 to the symmetric k-nearest-neighbor graph:

Proposition 4 (Within-cluster connectedness of $G_{\mathbf{sym}}(n,k)$). *We have*

$$\mathrm{P}\left(\text{Cluster } C^{(i)} \text{ conn. in } G_{sym}(n,k)\right) \geq \mathrm{P}\left(\text{Cluster } C^{(i)} \text{ conn. in } G_{mut}(n,k)\right)$$

Proof. The edge set of $G_{\text{mut}}(n,k)$ is a subset of the edges of $G_{\text{sym}}(n,k)$. Hence connectedness of $G_{\text{mut}}(n,k)$ implies connectedness of $G_{\text{sym}}(n,k)$. □

Note that this bound does not take into account the better connectivity properties of the symmetric kNN graph. Therefore one can expect that this bound is quite loose. We think that proving tight bounds for the within-cluster connectivity of the symmetric kNN graph requires a completely new proof concept. See Section 3 for more discussion of this point.

2.2 Between-cluster connectivity of kNN graphs

In this section we state bounds for the probability of edges *between* different clusters. The existence of edges between clusters is closely related to the event that the maximal k-nearest-neighbor radius is greater than the distance to the next cluster. Therefore we first give a bound for the probability of this event in Proposition 5. Then we apply this result to the mutual k-nearest-neighbor graph (in Proposition 6) and to the symmetric k-nearest-neighbor graph (in Proposition 7). It will be evident that the main difference between mutual kNN graphs and symmetric kNN graphs lies in the between-cluster connectivity.

Proposition 5 (Maximal nearest-neighbor radius). *We have*

$$\mathrm{P}\big(R_{\max}^{(i)} \geq u^{(i)}\big) \leq n\beta_{(i)} D\Big(k-1; n-1, g_{\min}^{(i)}\big(u^{(i)}\big)\Big).$$

The proof is omitted here because it is very similar to the proof of Proposition 2. It can be found in Maier et al. (2007). The previous proposition can be used to compare $G_{\text{mut}}(n,k)$ and $G_{\text{sym}}(n,k)$ with respect to cluster isolation. We say that a cluster $C^{(i)}$ *is isolated in the graph* if there are no edges between sample points lying in $C^{(i)}$ and any other cluster. In $G_{\text{mut}}(n,k)$ isolation of a cluster only depends on the properties of the cluster itself:

Proposition 6 (Cluster isolation in $G_{\mathbf{mut}}(n,k)$). *We have*

$$\begin{aligned}\mathrm{P}\Big(\text{Cluster } C^{(i)} \text{ isolated in } G_{mut}(n,k)\Big) &\geq 1 - \mathrm{P}\big(\, R_{\max}^{(i)} \geq u^{(i)}\,\big) \\ &\geq 1 - \, n\,\beta_{(i)} D\Big(k-1; n-1, g_{\min}^{(i)}\big(u^{(i)}\big)\Big).\end{aligned}$$

Proof. Since the neighborhood has to be mutual, we have no connections between $C^{(i)}$ and another cluster if the maximal kNN radius fulfills $R_{\max}^{(i)} < u^{(i)}$. □

The next theorem shows that the probability for connections between clusters is significantly higher in the symmetric kNN graph.

Proposition 7 (Cluster isolation in $G_{\mathbf{sym}}(n,k)$). *We have*

$$\begin{aligned}\mathrm{P}\Big(C^{(i)} \text{ isolated in } G_{sym}(n,k)\Big) &\geq 1 - \sum_{j=1}^{m} \mathrm{P}\big(\, R_{\max}^{(j)} \geq u^{(j)}\,\big) \\ &\geq 1 - n \sum_{j=1}^{m} \beta_{(j)} D\Big(k-1; n-1, g_{\min}^{(j)}\big(u^{(j)}\big)\Big).\end{aligned}$$

Proof. Let u^{ij} be the distance of $C^{(i)}$ and $C^{(j)}$. The event that $C^{(i)}$ is connected to any other cluster in $G_{\mathrm{sym}}(n,k)$ is contained in the union $\{R^{(i)}_{\max} \geq u^{(i)}\} \cup \{\cup_{j\neq i}\{R^{(j)}_{\max} \geq u^{ij}\}\}$. Using a union bound we have

$$\mathrm{P}\Big(C^{(i)} \text{ not isolated in } G_{\mathrm{sym}}(n,k)\Big) \leq \mathrm{P}\big(R^{(i)}_{\max} \geq u^{(i)}\big) + \sum_{j\neq i} \mathrm{P}\big(R^{(j)}_{\max} \geq u^{ij}\big).$$

Using first $u^{(j)} \leq u^{ij}$ and then Proposition 6 we obtain the two inequalities. □

Note that the upper bound on the probability that $C^{(i)}$ is isolated is the same for all clusters in the symmetric kNN graph. The upper bound is loose in the sense that it does not respect specific geometric configurations of the clusters where the bound could be smaller. However, it is tight in the sense that the probability that cluster $C^{(i)}$ is isolated in $G_{\mathrm{sym}}(n,k)$ always depends on the *worst* cluster. This is the main difference to the mutual kNN graph, where the properties of cluster $C^{(i)}$ are independent of the other clusters.

3 The isolated point heuristic

In the last sections we proved bounds for the probabilities that individual clusters in the neighborhood graph are connected, and different clusters in the neighborhood graph are disconnected. The bound on the disconnectedness of different clusters is rather tight, while the bound for the within-cluster connectedness of a cluster is tight if n is large, but has room for improvement if n is small. The reason is that the techniques we used to prove the connectedness bound are not well-adapted to a small sample size: we cover each cluster by small balls and require that each ball contains at least one sample point (event $\mathcal{F}^{(i)}_z$ in Section 2). Connectedness of G_{mut} then follows by construction. However, for small n this is suboptimal, because the neighborhood graph can be connected even though it does not yet "cover" the whole data space. Here it would be of advantage to look at connectivity properties more directly. However, this is not a simple task. The heuristic we propose makes use of the following fact from the theory of random graph models: in both Erdős-Rényi random graphs and random geometric graphs, for large n the parameter for which the graph stops having isolated vertices coincides with the parameter for which the graph is connected (e.g., Bollobas, 2001, p. 161 and Theorem 7.3; Penrose, 2003, p.281 and Theorem 13.17). The isolated point heuristic now consists in replacing the loose bound on the within-cluster connectivity from Section 2 by a bound on the probability of the existence of isolated vertices in the graph, that is we use the heuristic

$$\mathrm{P}(C^{(i)} \text{ connected}) \approx \mathrm{P}(\text{no isolated points in } C^{(i)}).$$

This procedure is consistent for $n \to \infty$ as proved by the theorems cited above, but, of course, it is only a heuristic for small n.

Proposition 8 (Probability of isolated points in $G_{\mathbf{eps}}$). *We have*

$$\mathrm{P}\big(\text{ex. isol. points from } C^{(i)} \text{ in } G_{eps}(n,\varepsilon)\big) \;\leq\; \beta_{(i)}\, n\, (1 - g^{(i)}_{\min}(\varepsilon))^{n-1}.$$

Proof. Suppose there are l points in cluster $C^{(i)}$. Then a point X_i from $C^{(i)}$ is isolated if $\min_{1\le j\le n, j\neq i} \|X_i - X_j\| > \varepsilon$. This event has probability less than $(1 - g^{(i)}_{\min}(\varepsilon))^{n-1}$. Thus a union bound yields

$$\mathrm{P}\big(\text{ex. isol. points from } C^{(i)} \text{ in } G_{\mathrm{eps}}(n,\varepsilon) \,|\, n^{(i)} = l\big) \le l\,(1 - g^{(i)}_{\min}(\varepsilon))^{n-1},$$

and we sum over the $\mathrm{Bin}(n, \beta_{(i)})$-distributed random variable $n^{(i)}$. □

For the mutual nearest-neighbor graph, bounding the probability of the existence of isolated points is more demanding than for the ε-graph, as the existence of an edge between two points depends not only on the distance of the points, but also on the location of all other points. We circumvent this problem by transferring the question of the existence of isolated points in $G_{\mathrm{mut}}(n,k)$ to the problem of the existence of isolated vertices of a particular ε-graph. Namely

$$\{\text{ex. isolated points in } G_{\mathrm{mut}}(n,k)\} \implies \{\text{ex. isolated points in } G_{\mathrm{eps}}(n, R_{\min})\}.$$

Proposition 9 (Probability of isolated points in $G_{\mathbf{mut}}$). *Let* $v = \sup\big\{d(x,y)\big|x,y \in \cup_{i=1}^m C^{(i)}\big\}$ *and* $b : [0,v] \to \mathbb{R}$ *be a continuous function such that* $\mathrm{P}\big(R^{(i)}_{\min} \le t\big) \le b(t)$. *Then,*

$$\mathrm{P}\big(\text{ex. isol. points from } C^{(i)} \text{ in } G_{mut}(n,k)\big) \quad \le \beta_{(i)}\, n \int_0^v \Big(1 - g^{(i)}_{\min}(t)\Big)^{n-1} \,\mathrm{d}b\,(t).$$

Proof. Let $A_{\mathrm{mut}} = \big\{\text{ex. isolated points from } C^{(i)} \text{ in } G_{\mathrm{mut}}(n,k)\big\}$ and $A_{\mathrm{eps}}(t) = \big\{\text{ex. isolated points from } C^{(i)} \text{ in } G_{\mathrm{eps}}(n,t)\big)\big\}$. Proposition 8 implies $\mathrm{P}\big(A_{\mathrm{mut}} \,|\, R^{(i)}_{\min} = t\big) \le \mathrm{P}\big(A_{\mathrm{eps}}(t)\big) \le \beta_{(i)}\, n \big(1 - g^{(i)}_{\min}(t)\big)^{n-1}$. $c(t) = \beta_{(i)}\, n \big(1 - g^{(i)}_{\min}(t)\big)^{n-1}$ is a decreasing function that bounds $\mathrm{P}\big(A_{\mathrm{mut}} \,|\, R^{(i)}_{\min} = t\big)$. Straightforward calculations and standard facts about the Riemann-Stieltjes integral conclude the proof. For details see Maier et al. (2007). □

Note that in the symmetric nearest-neighbor graph isolated points do not exist by definition. Hence, the isolated points heuristic cannot be applied in that case.

4 Asymptotic Analysis

In this section we study the asymptotic behavior of our bounds under some additional assumptions on the probability densities and geometry of the clusters. Throughout this section we assume that the assumptions of Proposition 1 hold and that the densities $p^{(i)}$ satisfy $0 < p^{(i)}_{\min} \le p^{(i)}(x) \le p^{(i)}_{\max}$ for all $x \in C^{(i)}$. We define the *overlap function* $O^{(i)}(r)$ by $O^{(i)}(r) = \inf_{x \in C^{(i)}} \big(\mathrm{vol}\left(B(x,r) \cap C^{(i)}\right) / \mathrm{vol}\left(B(x,r)\right)\big)$. With these assumptions we can establish a relation between the volume of a ball and its probability mass,

$$g^{(i)}_{\min}(t) = \beta_{(i)} O^{(i)}(t) p^{(i)}_{\min} \eta_d t^d \quad \text{and} \quad \tilde{g}^{(i)}_{\min}(t) = \beta_{(i)} p^{(i)}_{\min} t^d \eta_d,$$

$$g_{\max}^{(i)}(t) = \begin{cases} t^d \eta_d \beta_{(i)} p_{\max}^{(i)} & \text{if } t \leq u^{(i)} \\ \left(u^{(i)}\right)^d \eta_d \left(\beta_{(i)} p_{\max}^{(i)} - p_{\max}\right) + t^d \eta_d p_{\max} & \text{if } t > u^{(i)}, \end{cases}$$

where $p_{\max} = \max_{1 \leq i \leq m} \beta_{(i)} p_{\max}^{(i)}$.
In Proposition 1 we have given a bound on the probability of disconnectedness of a cluster which has two free parameters, k and the radius z. Clearly the optimal value of z depends on k. In the following proposition we plug in the expressions for $\tilde{g}_{\min}^{(i)}(t)$ and $g_{\max}^{(i)}(t)$ above and choose a reasonable value of z for every k, in order to find a range of k for which the probability of disconnectedness of the cluster asymptotically approaches zero exponentially fast.

Proposition 10 (Choice of k for asymptotic connectivity). *Define* $1/D^{(i)} = 1 + 4^d(e^2 - 1)p_{\max}^{(i)}/p_{\min}^{(i)}$ *and*

$$k' = \frac{1}{D^{(i)}}(n-1)\beta_{(i)} p_{\min}^{(i)} \eta_d \min\left\{ \left(\varepsilon_{\max}^{(i)}\right)^d, \left(\frac{u^{(i)}}{4}\right)^d \right\}.$$

Then if $n \geq e/\left(2^d \beta_{(i)} \operatorname{vol}(C^{(i)}) p_{\min}^{(i)}\right)$ *there exists* $0 < \gamma < 1$ *such that*

$$\mathrm{P}\Big(C^{(i)} \text{ conn. in } G_{sym}(n,k)\Big) \geq \mathrm{P}\Big(C^{(i)} \text{ conn. in } G_{mut}(n,k)\Big) \geq 1 - 2\, e^{-\gamma\, D^{(i)}\, k},$$

for all $k \in \{1, \ldots, n-1\}$ *with*

$$k' \geq k \geq \frac{1}{D^{(i)}} \frac{\log(2^d \operatorname{vol}(C^{(i)}) p_{\min}^{(i)} \beta_{(i)}\, n(1-\gamma))}{(1-\gamma)}. \tag{2}$$

Proof. We give an outline of the proof for the mutual k-nearest-neighbor graph. For details see Maier et al. (2007). The statement for the symmetric k-nearest-neighbor graph then follows with Proposition 4. In the following we set $z^d = 8^d \operatorname{vol}\left(C^{(i)}\right) \alpha\theta / \left(\beta_{(i)} \eta_d\right)$ for a $\theta \in (0, \theta_{\max})$ with $\theta_{\max} = \left(8^d \operatorname{vol}\left(C^{(i)}\right) p_{\max}^{(i)}\right)^{-1}$ and $\alpha = k/(n-1)$. For θ in this interval we can apply a tail bound for the binomial distribution from Hoeffding (1963). Let z denote the radius that corresponds to θ and k. With the tail bound for the binomial and standard inequalities for the logarithm we can show that $\log\left(\mathrm{P}\left(R_{\min}^{(i)} \leq z\right)\right) \leq g(\theta)$, where

$$g(\theta) = \log\left(\frac{\beta_{(i)}}{\theta\alpha}\right) + n\alpha\left(2 + \log 8^d \operatorname{vol}\left(C^{(i)}\right) p_{\max}^{(i)} \theta - 8^d \operatorname{vol}\left(C^{(i)}\right) p_{\max}^{(i)} \theta\right)$$

and $\log\left(\mathrm{P}\left(\mathcal{F}\right)\right) \leq h(\theta)$, where

$$h(\theta) = \log\left(\frac{\beta_{(i)}}{\theta\alpha}\right) - 2^d n \alpha p_{\min}^{(i)} \operatorname{vol}(C^{(i)}) \theta.$$

With straightforward calculations and standard inequalities for the exponential function one can show that for $\theta^* = D^{(i)}/\left(2^d \operatorname{vol}(C^{(i)}) p_{\min}^{(i)}\right)$

we have $g(\theta^*) \leq h(\theta^*)$. Straightforward calculations show that under the conditions $\gamma \in (0,1)$, $n \geq e/\left(2^d \operatorname{vol}(C^{(i)})(1-\gamma)\beta_{(i)}p_{\min}^{(i)}\right)$ and that k is bounded from below as in Equation (2), we have $h(\theta^*) \leq -\gamma k D^{(i)}$. For all $n \geq e/\left(2^d \beta_{(i)} \operatorname{vol}(C^{(i)})p_{\min}^{(i)}\right)$ we can find $\gamma \in (0,1)$ such that $n \geq e/\left(2^d \operatorname{vol}(C^{(i)})(1-\gamma)\beta_{(i)}p_{\min}^{(i)}\right)$. Using $g(\theta^*) \leq h(\theta^*) \leq -\gamma k D^{(i)}$ we have shown that $\mathrm{P}\left(R_{\min}^{(i)} \leq z\right) \leq \exp\left(-\gamma k D^{(i)}\right)$ and $\mathrm{P}(\mathcal{F}) \leq \exp\left(-\gamma k D^{(i)}\right)$. Reformulating the conditions $z/4 \leq \varepsilon_{\max}^{(i)}$ and $z \leq u^{(i)}$ in terms of θ^* gives the condition $k \leq k'$. □

The result of the proposition is basically that if we choose $k \geq c_1 + c_2 \log(n)$ with two constants c_1, c_2 that depend on the geometry of the cluster and the respective density, then the probability that the cluster is disconnected approaches zero exponentially in k.
Note that, due to the constraints on the covering radius, we have to introduce an upper bound k' on k, which depends linearly on n. However, the probability of connectedness is monotonically increasing in k , since the k-nearest-neighbor graph contains all the edges of the $(k-1)$-nearest-neighbor graph. Thus the value of the within-connectedness bound for $k = k'$ is a lower bound for all $k > k'$ as well. Since the lower bound on k grows with $\log(n)$ and the upper bound grows with n, there exists a feasible region for k if n is large enough.

Proposition 11 (Maximal kNN **radius asymptotically).** *Let* $p_2^{(i)} = \beta_{(i)} O^{(i)}\left(u^{(i)}\right)\, p_{\min}^{(i)}\, \eta_d\, (u^{(i)})^d$ *and* $k \leq (n-1)p_2^{(i)} + 1$. *Then*

$$\mathrm{P}\left(R_{\max}^{(i)} \geq u^{(i)}\right) \leq n\beta_{(i)} e^{-(n-1)\left(\left(p_2^{(i)}\right)^2 e^{-p_2^{(i)}} + p_2^{(i)} - \frac{k-1}{n-1}\right)}.$$

Proof. Using a standard tail bound for the binomial distribution (see Hoeffding, 1963) we obtain from Proposition 5 for $(k-1) \leq (n-1)p_2^{(i)}$

$$\mathrm{P}\left(R_{\max}^{(i)} \geq u^{(i)}\right) \leq n\beta_{(i)} e^{-(n-1)\left(\frac{k-1}{n-1}\log\frac{(k-1)}{(n-1)p_2^{(i)}} + (1-\frac{k-1}{n-1})\log\frac{1-(k-1)/(n-1)}{1-p_2^{(i)}}\right)}.$$

Using $\log(1+x) \geq x/(1+x)$ and that $-w^2 e^{-w}$ is the minimum of $x\log(x/w)$ (attained at $x = we^{-w}$) we obtain the result by straightforward calculations. □

4.1 Identification of clusters as the connected component of a mutual and symmetric kNN graph

We say that a cluster $C^{(i)}$ is identified if it is an isolated connected component in the kNN graph. This requires the cluster $C^{(i)}$ to be connected, which intuitively happens for large k. Within-cluster connectedness was considered in

Proposition 10. The second condition for the identification of a cluster $C^{(i)}$ is that there are no edges between $C^{(i)}$ and other clusters. This event was considered in Proposition 6 and is true if k is small enough. The following theorems consider the tradeoff for the choice of k for the identification of one and of all clusters in both kNN graph types and derive the optimal choice for k. We say that k is *tradeoff-optimal* if our bounds for within-cluster connectedness and between-cluster disconnectedness are equal.

Theorem 12 (Choice of k for the identification of one cluster in $G_{\mathbf{mut}}(n,k)$). *Define $p_2^{(i)}$ as in Proposition 11 and let n and γ be as in Proposition 10. The tradeoff-optimal choice of k in $G_{mut}(n,k)$ is given by*

$$k-1=(n-1)\,p_2^{(i)}\frac{1-p_2^{(i)}e^{-p_2^{(i)}}}{1+\gamma D^{(i)}}-\frac{\log\left(\frac{1}{2}n\beta_{(i)}\right)}{1+\gamma D^{(i)}},$$

if this choice of k fulfills the conditions in Proposition 10 and $k<(n-1)p_2^{(i)}+1$. For this choice of k we have

$$\mathrm{P}\Big(C^{(i)}\text{ ident. in }G_{mut}(n,k)\Big)\geq 1-4e^{-(n-1)\frac{\gamma D^{(i)}}{1+\gamma D^{(i)}}\left[p_2^{(i)}(1-p_2^{(i)}e^{-p_2^{(i)}})-\frac{\log(\frac{1}{2}n\beta_{(i)})}{(n-1)}\right]}$$

Proof. We equate the bounds for within-cluster connectedness of Proposition 10 and the bound for between-cluster edges of Proposition 6 and solve for k. □

The result of the previous theorem is that the tradeoff-optimal choice of k has the form $k=c_3n-c_4\log(n)+c_5$ with constants $c_3,c_4\geq 0$ and $c_5\in\mathbb{R}$, which depend on the geometry of the cluster and the respective density. Evidently, if n becomes large enough, then k chosen according to this rule fulfills all the requirements in Proposition 10 and Theorem 12.
Theorem 12 allows us to define the "most significant" cluster. Intuitively a cluster is more significant the higher its density and the larger its distance to other clusters. Formally the "most significant" cluster is the one with the best rate for identification, that is the maximal rate of the bound:

$$\max_{1\leq i\leq m}\frac{\gamma D^{(i)}}{1+\gamma D^{(i)}}\left[p_2^{(i)}(1-p_2^{(i)}e^{-p_2^{(i)}})-\frac{\log(\frac{1}{2}n\beta_{(i)})}{n}\right]$$

The term in front of the bracket is increasing in $D^{(i)}$ and thus is larger, the closer $p_{\max}^{(i)}$ and $p_{\min}^{(i)}$ are, that is for a fairly homogeneous density. The second term in the brackets approaches zero rather quickly in n. It is straightforward to show that the first term in the bracket is increasing in $p_2^{(i)}$. Thus a cluster becomes more significant, the higher the probability mass in balls of radius $u^{(i)}$, that is, the higher $\beta_{(i)}$, $p_{\min}^{(i)}$, $u^{(i)}$ and the higher the value of the overlap function $O^{(i)}(u^{(i)})$.
We would like to emphasize that it is a unique feature of the mutual kNN graph that one can minimize the bound independently of the other clusters. This is

not the case for the symmetric kNN graph. In particular, in the case of many clusters, a few of which have high probability, the differences in the rates can be huge. If the goal is to identify not all clusters but only the most important ones, that means the ones which can be detected most easily, then the mutual kNN graph has much better convergence properties than the symmetric kNN graph. We illustrate this with the following theorem for the symmetric kNN graph.

Theorem 13 (Choice of k for the identification of one cluster in $G_{\mathbf{sym}}(n,k)$). *Define $\rho_2 = \min_{1\le i\le m} p_2^{(i)}$ and let n and γ be as in Proposition 10. The tradeoff-optimal choice of k in $G_{sym}(n,k)$ is given by*

$$k - 1 = (n-1)\,\rho_2 \frac{1-\rho_2 e^{-\rho_2}}{1+\gamma D^{(i)}} - \frac{\log\left(\frac{n}{2}\right)}{1+\gamma D^{(i)}},$$

if this choice of k fulfills the conditions in Proposition 10 and $k < (n-1)\rho_2 + 1$. For this choice of k

$$\mathrm{P}\Big(C^{(i)} \text{ identified in } G_{sym}(n,k)\Big) \ge 1 - 4e^{-(n-1)\frac{\gamma D^{(i)}}{1+\gamma D^{(i)}}\left[\rho_2(1-\rho_2 e^{-\rho_2}) - \frac{\log(\frac{n}{2})}{(n-1)}\right]}.$$

Proof. Combining Proposition 7 and Proposition 11 we obtain

$$\mathrm{P}\Big(\text{Cluster } C^{(i)} \text{ not isolated in } G_{\mathrm{sym}}(n,k)\Big) \le n\sum_{i=1}^{m} \beta_{(i)} e^{-(n-1)\left(\rho_2^2 e^{-\rho_2} + \rho_2 - \frac{k-1}{n-1}\right)}.$$

Equating this bound with the within-connectedness bound in Proposition 10 we obtain the result. □

A comparison with the rate of Theorem 12 for the mutual kNN graph shows that the rate of the symmetric kNN graph depends on the "worst" cluster. This property would still hold if one found a tighter bound for the connectivity of the symmetric kNN graph.

Corollary 14 (Choice of k for the identification of all clusters in $G_{\mathbf{mut}}(n,k)$). *Define $p_{\mathrm{ratio}} = \max_{1\le i\le m}\left(p_{\max}^{(i)}/p_{\min}^{(i)}\right)$ and $\rho_2 = \min_{1\le i\le m} p_2^{(i)}$. Let $1/D = 1 + 4^d(e^2-1)p_{\mathrm{ratio}}$ and n, γ be as in Proposition 10. The tradeoff-optimal k for the identification of all clusters in $G_{mut}(n,k)$ is given by*

$$k - 1 = (n-1)\rho_2 \frac{1-\rho_2 e^{-\rho_2}}{1+\gamma D} - \frac{\log\left(\frac{n}{2m}\right)}{1+\gamma D},$$

if this choice of k fulfills the conditions in Proposition 10 for all clusters $C^{(i)}$, $i = 1,\ldots,m$ and $k < (n-1)\rho_2 + 1$. For this choice of k we have

$$\mathrm{P}\Big(\text{All clusters ident. in } G_{mut}(n,k)\Big) \ge 1 - 4\,m\,e^{-(n-1)\frac{\gamma D}{1+\gamma D}\left[\rho_2(1-\rho_2 e^{-\rho_2}) - \frac{\log(\frac{n}{2m})}{(n-1)}\right]}.$$

Proof. Using a union bound, we obtain from Proposition 10 and Proposition 11

$$\mathrm{P}\Big(\bigcup_{i=1}^{m} C^{(i)} \text{ not isolated} \Big) \leq n \sum_{i=1}^{m} \beta_{(i)} e^{-(n-1)\left(\rho_2^2 e^{-\rho_2} + \rho_2 - \frac{k-1}{n-1}\right)}$$

$$\mathrm{P}\Big(\bigcup_{i=1}^{m} \text{Cluster } C^{(i)} \text{ disconnected in } G_{\mathrm{mut}}(n,k) \Big) \leq 2m\, e^{-\gamma k\, D}$$

We obtain the result by equating these two bounds. □

The result for the identification of *all* clusters in the mutual kNN graph is not much different from the result for the symmetric kNN graph. Therefore the difference in the behavior of the two graph types is greatest if one is interested in identifying the most important clusters only.

5 Simulations

The long-term goal of our work is to find rules which can be used to choose the parameters k or ε for neighborhood graphs. In this section we want to test whether the bounds we derived above can be used for this purpose, at least in principle. We consider a simple setting with a density of the form $f(x) = \beta\tilde{f}(x) + (1-\beta)\tilde{f}(x-(u+2)e_1)$, where $\beta \in (0,1)$ is the weight of the individual clusters, $\tilde{f}$ is the uniform density on the unit ball in $\mathbb{R}^d$, $e_1 = (1,0,\ldots,0)'$, and u is the distance between the clusters.
First we compare the qualitative behavior of the different bounds to the corresponding empirical frequencies. For the empirical setting, we randomly draw n points from the mixture density above, with different choices of the parameters. For all values of k we then evaluate the empirical frequencies P_{emp} for within-cluster connectedness, between-cluster disconnectedness, and the existence of isolated points by repeating the experiment 100 times. As theoretical counterpart we use the bounds obtained above, which are denoted by P_{bound}. To evaluate those bounds, we use the true parameters n, d, β, u, $p_{\min}$, $p_{\max}$. Figure 1 shows the results for $n = 5000$ points from two unit balls in $\mathbb{R}^2$ with a distance of $u = 0.5$ and $\beta = 0.5$. We can see that the bound for within-cluster disconnectedness is loose, but still gets into a non-trivial regime (that is, smaller than 1) for a reasonable k. On the other hand the bound for the existence of isolated points indeed upper bounds the within-cluster disconnectedness and is quite close to the true probability. Hence the isolated point heuristic works well in this example. Moreover, there is a range of values of k where both the empirical frequencies and the bounds for the probabilities become close to zero. This is the region of k we are interested in for choosing optimal values of k in order to identify the clusters correctly. To evaluate whether our bounds can be used for this purpose we sample points from the density above and build the kNN graph for these points. For each graph we determine the range of $k_{\min} \leq k \leq k_{\max}$ for which both within-cluster connectedness and between-cluster disconnectedness are satisfied, and compute $\hat{k}_{\min}$ and $\hat{k}_{\max}$ as the mean values over 100 repetitions.

To determine "optimal" values for k we use two rules:

$$k_{bound} := \operatorname*{argmin}_k \left(\mathrm{P}_{bound} \left(\text{connected within} \right) + \mathrm{P}_{bound} \left(\text{disconnected between} \right) \right)$$
$$k_{iso} := \operatorname*{argmin}_k \left(\mathrm{P}_{bound} \left(\text{no isolated points} \right) + \mathrm{P}_{bound} \left(\text{disconnected between} \right) \right).$$

The following table shows the results for $G_{\mathrm{mut}}(n, k)$.

n	k_{iso}	k_{bound}	$\hat{k}_{\mathrm{min}}$	$\hat{k}_{\mathrm{max}}$
500	17	25	7.2 ± 1.2	41.0 ± 6.5
1000	29	46	7.3 ± 1.2	71.7 ± 9.4
5000	97	213	8.5 ± 1.2	309.3 ± 16.9
10000	101	425	8.8 ± 1.1	596.6 ± 21.1

We can see that for all values of n in the experiment, both k_{iso} and k_{bound} lie well within the interval of the empirical values $\hat{k}_{\mathrm{min}}$ and $\hat{k}_{\mathrm{max}}$. So in both cases, choosing k by the bound or the heuristic leads to a correct value of k in the sense that for this choice, the clusters are perfectly identified in the corresponding mutual kNN graph.

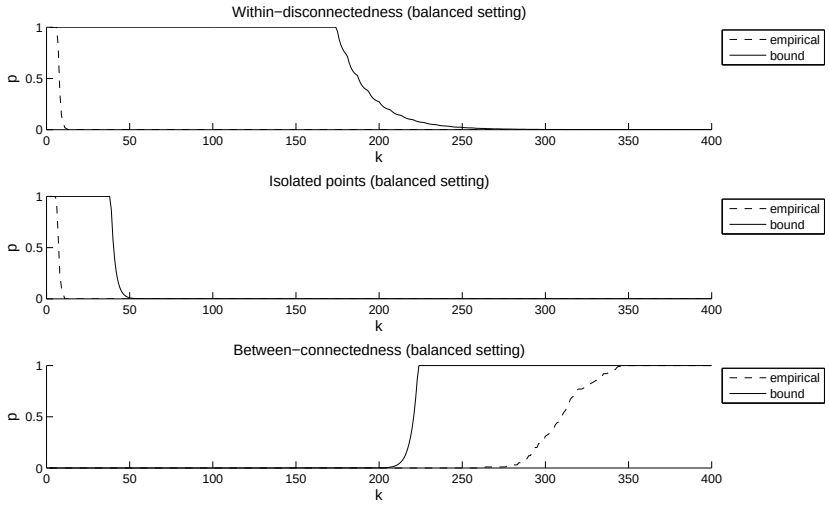

Fig. 1. Bounds and empirical frequencies for $G_{\mathrm{mut}}(n, k)$ for two clusters with $\beta = 0.5$ and $u = 0.5$ (for plotting, we set the bound to 1 if it is larger than 1).

Finally we would like to investigate the difference between G_{mut} and G_{sym}. While the within-cluster connectivity properties are comparable in both graphs, the main difference lies in the between-cluster connectivity, in particular, if we only want to identify the densest cluster in an unbalanced setting where clusters have very different weights. We thus choose the mixture density with weight parameter $\beta = 0.9$, that is we have one very dense and one very sparse cluster. We now investigate the identification problem for the densest cluster. The results are shown in Figure 2. We can see that $G_{\mathrm{sym}}(n, k)$ introduces between-cluster

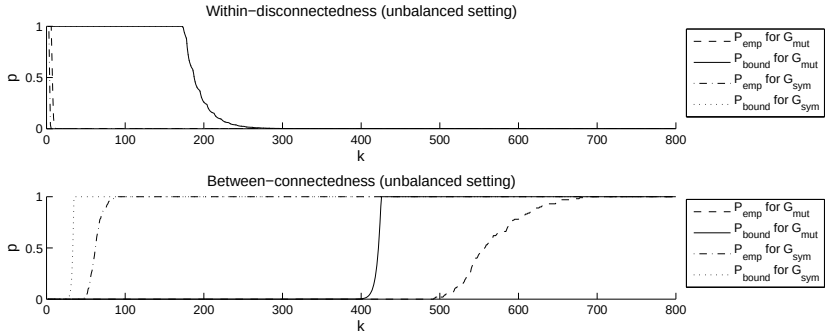

Fig. 2. Within- and between-cluster connectivity for $G_{\mathrm{mut}}(n,k)$ and $G_{\mathrm{sym}}(n,k)$ for two unbalanced clusters with $\beta = 0.9$ and $u = 0.5$. Note that the curves of P_{bound} for $G_{\mathrm{mut}}(n,k)$ and $G_{\mathrm{sym}}(n,k)$ lie on top of each other in the top plot. The scale of the horizontal axis is different from Figure 1.

edges for a much lower k than it is the case for $G_{\mathrm{mut}}(n,k)$, which is a large disadvantage of G_{sym} in the identification problem. As a consequence, there is only a very small range of values of k for which the big cluster can be identified. For G_{mut}, on the other hand, one can see immediately that there is a huge range of k for which the cluster is identified with very high probability. This behavior is predicted correctly by the bounds given above.

6 Conclusions and further work

We studied both G_{sym} and G_{mut} in terms of within-cluster and between-cluster connectivity. While the within-cluster connectivity properties are quite similar in the two graphs, the behavior of the between-cluster connectivity is very different. In the mutual kNN graph the event that a cluster is isolated is independent of all the other clusters. This is not so important if one aims to identify *all* clusters, as then also in the mutual graph the worst case applies and one gets results similar to the symmetric graph. However, if the goal is to identify the most significant clusters only, then this can be achieved much easier with the mutual graph, in particular if the clusters have very different densities and different weights.
It is well known that the lowest rate to asymptotically achieve within-cluster connectivity is to choose $k \sim \log(n)$ (e.g., Brito et al., 1997). However, we have seen that the optimal growth rate of k to achieve cluster identification is not linear in $\log(n)$ but rather of the form $k = c_3 n - c_4 \log(n) + c_5$ with constants $c_3, c_4 \geq 0$ and $c_5 \in \mathbb{R}$. This difference comes from the fact that we are not interested in the lowest possible rate for asymptotic connectivity, but in the rate for which the probability for cluster identification is maximized. To this end we can "afford" to choose k higher than absolutely necessary and thus improve the "probability" of within-connectedness. However, as we still have to keep in mind the between-cluster disconnectedness we cannot choose k "as high as we want". The rate now tells us that we can choose k "quite high", almost linear in n.

There are several aspects about this work which are suboptimal and could be improved further:
Firstly, the result on the tradeoff-optimal choice of k relies on the assumption that the density is zero between the clusters We cannot make this assumption if the points are disturbed by noise and therefore the optimal choice of k might be different in that case.
Secondly, the main quantities that enter our bounds are the probability mass in balls of different radii around points in the cluster and the distance between clusters. However, it turns out that these quantities are not sufficient to describe the geometry of the problem: The bounds for the mutual graph do not distinguish between a disc with a neighboring disc in distance u and a disc that is surrounded by a ring in distance u. Obviously, the optimal values of k will differ. It would be possible to include further geometric quantities in the existing proofs, but we have not succeeded in finding simple descriptive expressions. Furthermore, it is unclear if one should make too many assumptions on the geometry in a clustering setting.
Finally, we have indicated in Section 5 how our bounds can be used to choose the optimal value for k from a sample. However, in our experiments we simply took most of the parameters like u or β as given. For a real world application, those parameters would have to be estimated from the sample. Another idea is to turn the tables: instead of estimating, say, the distance u between the clusters for the sample and then predicting an optimal value of k one could decide to go for the most significant clusters and only look for clusters having cluster distance u and cluster weight β bounded from below. Then we can use the bounds not for parameter selection, but to construct a test whether the clusters identified for some value of k are "significant".

Bibliography

C. Bettstetter. On the minimum node degree and connectivity of a wireless multihop network. In *Proceedings of MobiHoc*, pages 80–91, 2002.

G. Biau, B. Cadre, and B. Pelletier. A graph-based estimator of the number of clusters. *ESIAM: Prob. and Stat.*, 11:272–280, 2007.

B. Bollobas. *Random Graphs*. Cambridge University Press, Cambridge, 2001.

B. Bollobas and O. Riordan. *Percolation*. Cambridge Universiy Press, 2006.

M. Brito, E. Chavez, A. Quiroz, and J. Yukich. Connectivity of the mutual k-nearest-neighbor graph in clustering and outlier detection. *Stat. Probabil. Lett.*, 35:33–42, 1997.

W. Hoeffding. Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.*, 58:13–30, 1963.

S. S. Kunniyur and S. S. Venkatesh. Threshold functions, node isolation, and emergent lacunae in sensor networks. *IEEE Trans. Inf. Th.*, 52(12):5352–5372, 2006.

M. Maier, M. Hein, and U. von Luxburg. Cluster identification in nearest-neighbor graphs. Technical Report 163, MPI for Biological Cybernetics, Tübingen, 2007.

M. Penrose. *Random Geometric Graphs*. Oxford University Press, Oxford, 2003.

P. Santi and D. Blough. The critical transmitting range for connectivity in sparse wireless ad hoc networks. *IEEE Trans. Mobile Computing*, 02(1):25–39, 2003.