A Statistical Geometry Approach to Distance Estimation in Wireless Sensor Networks

Valerio Freschi, Emanuele Lattanzi, Alessandro Bogliolo
Department of Basic Sciences and Foundations
University of Urbino
Urbino, IT 61029
Email: {valerio.freschi, emanuele.lattanzi, alessandro.bogliolo}@uniurb.it

***Abstract*—Algorithmic approaches to the estimation of pairwise distances between the nodes of a wireless sensor network are highly attractive to provide information for routing and localization without requiring specific hardware to be added to cost/resource-constrained nodes. This paper exploits statistical geometry to derive robust estimators of the pairwise Euclidean distances from topological information typically available in any network. Extensive Monte Carlo experiments conducted on synthetic benchmarks demonstrate the improved quality of the proposed estimators with respect to the state of the art.**

I. Introduction

Distance estimation is a key computational primitive in many algorithms for wireless sensor networks (WSN). Localization algorithms are in fact crucial for many applications, ranging from routing to data delivery and management [1]. Moreover, localization algorithms heavily depend on the availability of (possibly approximated) euclidean distances between sensor nodes [1], [2], [3], [4]. These distances can be obtained by means of special-purpose hardware which exploits, for instance, the Received Signal Strenght Indication (*RSSI*) of radio signals [5] or integrates radio and ultrasound signals [6], [7]. While these systems often provide accurate distance and positioning information, they have the inherent drawback of being dependent from more sophisticated and costly sensor equipment. This has motivated the growth of a line of research aimed at designing algorithmic approaches for distance estimation between nodes of sensor networks, only assuming minimal node hardware requirements. A common feature to many of these works is the use of the number of shared neighboring nodes between two given nodes to derive an estimate of the true distance between them [8].

An empirical method for deriving the distance from the ratio between the number of common neighbors between two nodes and the total number of nodes in their neighborhoods has been proposed by Villafuerte *et al.* [9], while an analytical evaluation for this mapping has been obtained by means of a first order Taylor series expansion [10]. Finally, Merkel *et al.* proposed an alternative derivation of the distance from the above mentioned ratio using regression [11]. In general, these methods restrict distance estimation to couples of neighboring nodes. Distances among nodes that do not directly communicate with each other can be obtained using the connectivity graph. For instance, Merkel *et al.* proposed a distributed algorithm for extending the computation to non-neighboring (i.e. out of radio-range) nodes [11]. The algorithm essentially works by firstly computing estimates among communicating nodes and then by propagating this information along shortest paths.

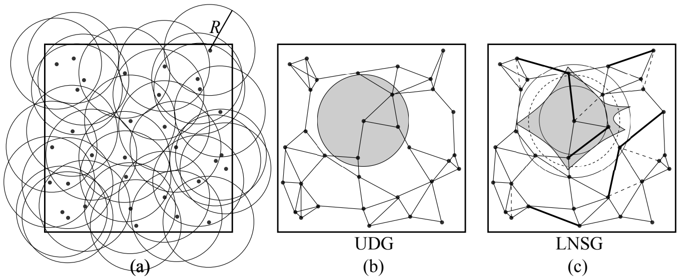

Fig. 1. Example of a random geometric graph: (a) circles of radius R represent the communication range of the nodes, (b) the UDG has undirected edges between any pair of nodes with distance lower than R, (c) the log-normal shadowing model introduces a further variation allowing pairs of nodes with distance greater than R to be connected (dark edges) and pairs of nodes with shorter distance to be not (dashed edges).

In this work we present two algorithmic methods to estimate the Euclidean distance between any pair of nodes in a graph based only on the connectivity matrix and on the transmission range of the nodes. Simulation experiments show that the proposed algorithms improve the quality of the most recent approach [11].

II. Proposed approach

A. Network model

We consider a static WSN composed of nodes randomly and uniformly distributed on a planar surface according to a homogeneous Poisson process with average density λ, so that the probability of finding n nodes in an area A of the deployment region can be expressed as:

$$Pr(n, A) = \frac{(\lambda A)^n}{n!} e^{-\lambda A} \tag{1}$$

In other words, the number of nodes in a region of area A is a Poisson random variable with mean λA.

Hereafter we consider a square deployment region with edges of unit length. This is in contrast with the infinite plane assumption which is typically adopted to avoid boundary effects, but it adds to the realism of the model.

Two nodes are connected by an undirected edge if and only if their Euclidean distance is below a fixed threshold R, so

978-1-4673-5753-1/13/$31.00 ©2013 IEEE

that the resulting graph is a *random geometric graph* (RGG), or *unit disk graph* (UDG), of parameters λ and R drawn on a square region with unitary edges [12]. Such undirected edges provide a suitable representation of the wireless links that could be established in a WSN composed of nodes with the same transmit power, under the assumption of isotropic propagation without shadowing effects. An example of RGG graph is provided in Figures 1.a and 1.b.

Although the UDG model will be used throughout this section for the sake of explanation, the effect of the additional uncertainty introduced by a more realistic *log-normal shadowing* model (LNSG) will be discussed in Section III [13]. An example is provided in Figure 1.c.

We use G to denote the RGG under analsys, N to denote the total number of nodes, and E to denote the number of edges among them. The topology of the graph is fully represented by its connectivity matrix C, which is a $N \times N$ symmetric matrix with entries $C(i, j)$ taking value 1 iff there is an edge between node i and node j in G.

In this paper we are interested in estimating the Euclidean distance between any pair of nodes of G starting only from topological information. Hence, we assume matrix C to be known, together with the length of the edge of the deployment region and with the (average value of the) communication range of the nodes. The problem of building matrix C for a randomly deployed WSN is out of the scope of this work.

The geodetic distance between two nodes i and j of G, also called *hop distance* (HD), is defined as the minimum number of edges to be traversed to go from i to j or vice versa. The *hop distance matrix* (D_H) for G can be easily determined from connection matrix C in $O(N^3)$ by solving an instance of the *all-pairs shortest path problem* [14]. Hence, for our purposes we can also assume that a $N \times N$ symmetric matrix D_H is available with entries $D_H(i, j)$ representing the hop distance between i and j.

B. Correlation between hop distance and Euclidean distance

The strong correlation between HD and Euclidean distance (ED) in RGG is a well known empirical result which has been widely exploited to develop distance estimators and localization algorithms. In order to make it possible to use such a correlation in our setting, we need to determine the ED/HD ratio starting from the only measures available, which are the edge of the deployment region and the communication range of the nodes.

We discuss in this section the suitability of four different estimators of the ED/HD ratio:

- the communication range R, which is an inherent upper bound of the actual distance covered at each hop
$$c1 = \frac{R}{1}$$
- the average radial distance of points within the communication range R of a given node, which provides an estimate of the average length of each link in the graph
$$c2 = \frac{2R/3}{1}$$

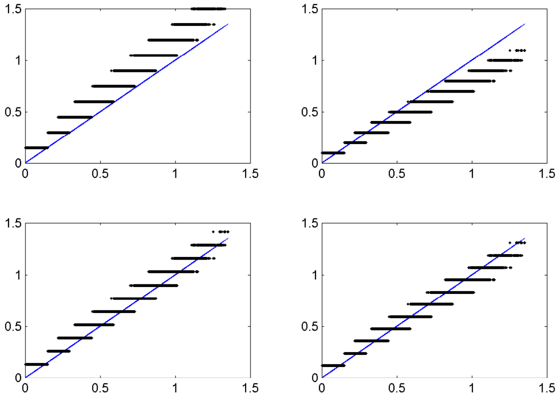

Fig. 2. Performance of coefficients estimators of ED/HD ratio. Upper side, from left to right: c1, c2. Lower side, from left to right: c3, c4.

- the diagonal of the unit square, which is the upper bound of the Euclidean distance between nodes, divided by the maximum value in D_H
$$c3 = \frac{\sqrt{2}}{\max(D_H)}$$
- the average distance between points in a unit square, which provides an estimate of the average Euclidean distance between any pair of nodes in the graph, divided by the mean of D_H entries
$$c4 = \frac{(2 + \sqrt{2} + 5\log(1 + \sqrt{2}))/15}{\text{mean}(D_H)}$$

The expressions of the average radial distance of points in range R (used to compute $c2$) and of the average distance between nodes randomly distributed in a unit square (used to compute $c4$) come from known results in statistical geometry. All the coefficients are expressed as a ratio between the estimator of the Euclidean distance and that of the correspoding HD. Notice however that $c1$ and $c2$ are apparent fractions since the Euclidean distances at the numerator refer to a single hop, so that the corresponding hop distance at the denominator is 1.

Figure 2 compares the performance of the four coefficients by means of scatter plots the points of which are associated with pairs of nodes of a given RGG (with $N = 200$ and $R = 0.15$) and represent the relation between their ED (x coordinate) and the estimates (y coordinate) provided by the coefficient under test multiplied by the corresponding HD. The bisector lines, representing ideal estimators, are reported for comparison. As expected, coefficient $c1$ tends to overestimate the actual ED, since it assumes that each hop covers the maximum range. Coefficient $c2$, on the contrary, underestimates the value of ED because of the net effect of two systematic errors: first, links along the path with the minimum number of hops are usually longer than average, so that the mean distance between two neighboring nodes underestimates the mean distance between two nodes along the shortest path; second, the shortest path between two nodes in a RGG is a broken line, the overall length of which overestimates the Euclidean distance between the end points. In spite of their

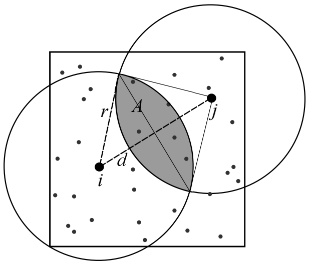

Fig. 3. Intersection between two circles of radius r centered in two of the inner nodes (namely, i and j) of a RGG, with distance $d < 2r$.

opposite effects, the two systematic errors are not guaranteed to compensate each other, since they depend on density and range. For the case of Figure 2 underestimation dominates. Coefficient $c3$ leads to an overestimate of EDs since it makes use of the diagonal of the square region, which is an upper bound for the maximum distance between node pairs. Finally, coefficient $c4$ provides an unbiased estimator built by taking into account all the node pairs. The higher number of data used to compute $c4$ adds both to the accuracy and to the statistical significance of the estimator.

Table I reports the mean value and the standard deviation of the estimation errors made by the ED/HD coefficients (namely, $c1, c2, c3$, and $c4$) computed on a sample of 100 RGGs of 400 nodes with communication range $R = 0.15$. Coefficient $c0$ is the slope of the best fitting line that can be found by having a complete knowledge of all ED, HD pairs. Hence, it represent a lower bound for the errors made by any linear estimator. Data on the first 4 rows are stratified on the basis of the value of HD compared with the maximum value of HD for that particular RGG: "very short" means less than 25%, "short" means between 25% and 50%, "long" means between 50% and 75%, "very long" means greater than 75%. The last two rows show the correlation of the average error with the parameters of the RGG, namely, λ and R. Results are based on Monte Carlo simulations performed in a $\pm 10\%$ range around the $(\lambda = 400, R = 0.15)$ point in the design space.

Simulation results reported in Table I show that the errors made by $c4$ for any class of distances are very close to the minimum errors achievable by any linear estimator using only HD as independent variable, which is around 5%. One of the main limitations of such an estimator comes from the discrete nature of HD, which is in contrast with the continous nature of ED. This explains the strong positive correlation between the communication range R and the estimation error for all the coefficient but $c1$, which benefits from the reduction of the average number of hops caused by the increased hop length.

C. Disk intersection

Figure 3 shows two overlapping circles with the same radius r, centered in i and j, which are inner nodes of an RGG. The area A of the intersection between the two circles can be expressed as a function of two independent variables: the radius r and the distance d between the centers.

$$A(d, r) = r^2(q - \sin q) \quad \text{with} \quad q = 2\arccos\left(\frac{d}{2r}\right) \tag{2}$$

Equation 2 is directly derived by observing the area of the sum of the areas of the two segments of circle which share the same chord with central angle q. For a given value of d, Equation 2 holds for any value of $r \geq d/2$. When $r = d/2$ the intersection is empty, while for $r >> d$ the area of the intersection approaches πr^2 since the two circles tend to overlap completely.

Since A grows monotonically with r and decreses monotonically with d, function $A(d, r)$ can be inverted to obtain d from A and r. In the context of RGGs, this provides a way for estimating $d_{i,j}$, i.e., the ED between nodes i and j, from the estimates of A and r.

According to the model discussed in Section II-A, the number of nodes found in a region of area A is a Poisson random variable with mean λA. Hence the ratio between the number of nodes in A (hereafter dented by N_A) and the density of the underlying homogeneous Poisson process λ can be used as an area estimator. Moreover, as long as the nodes are deployed over a unitary square region, node density is equal to the overall number of nodes in the network (N). In symbols:

$$\tilde{A} = \frac{N_A}{\lambda} = \frac{N_A}{N} \tag{3}$$

In order to exploit such an estimator in our setting, we need to find a way of counting the number of nodes in the intersection of the two circles centered in i and j using only the pairwise hop distances provided by matrix D_H. Given a RGG G, we call *geodetic circle* of radius K centered in i the subset of the nodes of G which have geodetic distance from i less or equal an K:

$$\mathcal{C}_i^{(K)} = \{j \in G | D_H(i, j) \leq K\} \tag{4}$$

The number of nodes in the intersection between two geodetic circles of radius K centered in i and j is then defined as:

$$N_{i,j}^{(K)} = |\mathcal{C}_i^{(K)} \cap \mathcal{C}_j^{(K)}| \tag{5}$$

where $|X|$ denotes the cardinality of X.

Even if we don't know the actual position of the nodes of G, from Equations 3 and 5 we can estimate the area of the region which contains the nodes in the intersection between the two geodetic circles, as:

$$\tilde{A}_{i,j}^{(K)} = \frac{N_{i,j}^{(K)}}{N} \tag{6}$$

Referring to Figure 3, if we take $\tilde{A}_{i,j}^{(K)}$ as an area estimator, we need to know the corresponding value of $\tilde{r}^{(K)}$, defined as the radius of the circle in the Euclidean plane which contains the points of a geodetic circle of radius K.

While for $K = 1$ such a value is directly provided, by construction, by the communication range of the nodes ($r^1 = R$), for $k > 1$ it needs to be estimated in its turn. There are two methods that can be adopted to this purpose. The first estimator, derived from the area of a circle, can be expressed

	$c0$		$c1$		$c2$		$c3$		$c4$	
	mean	std	mean	std	mean	std	mean	std	mean	std
very short	0.0414	0.0004	0.0822	0.0004	0.0397	0.0004	0.0549	0.067	0.0416	0.0017
short	0.0450	0.0009	0.1401	0.0054	0.0811	0.0051	0.0712	0.0141	0.0458	0.0015
long	0.0510	0.0015	0.1993	0.0096	0.1428	0.0082	0.0822	0.0240	0.0522	0.0048
very long	0.0561	0.0045	0.2669	0.0153	0.1948	0.0121	0.1005	0.0357	0.0579	0.0093
all distances	0.0466	0.0008	0.1566	0.0053	0.1028	0.0022	0.0740	0.0198	0.0476	0.0020
corr with λ	-0.3513		-0.6326		0.6931		-0.0931		-0.3156	
corr with R	0.8829		-0.5146		0.7722		0.2023		0.6405	

TABLE I. MEAN VALUE AND STANDARD DEVIATION OF THE ESTIMATION ERRORS MADE BY THE ED/HD COEFFICIENTS ON A SMAPLE OF 100 RGGS MADE OF 400 NODES WITH COMMUNICATION RANGE $R = 0.15$.

as square root of the ratio between the average number of nodes in $\mathcal{C}^{(K)}$ and πN.

$$\tilde{r}^{(K)} = \sqrt{\frac{|\mathcal{C}^{(K)}|}{\pi N}} \tag{7}$$

The second estimator can be obtained by using the most accurate scaling factor introduced in Section II-B to convert from HD to ED the value of K which eccedes the first hop:

$$\tilde{r}^{(K)} = R + c_4 \cdot (K - 1) \tag{8}$$

At this point, the ED between i and j can be numerically computed as the value of d which satisfies equation 2 for $r = \tilde{r}^{(K)}$ and $A = \tilde{A}_{i,j}^{(K)}$. Such an estimator is hereafter denoted by $\tilde{d}_{i,j}^{(K)}$ to retain the information about the size of the geodetic circles used in the computation.

Even if, in principle, the estimate should work properly for any value of $K \geq d_H(i,j)/2$, the value of K impacts the accuracy of the estimator because of the combination of two effects: the statistical significance (i.e., the confidence) of the estimator increases for larger values of K thanks to the larger number of nodes which fall into the intersection, the risk of errors caused by boundary effects increases with the value of K because of the higher probability of including in te intersection regions which are outside the deployment area.

In order to reduce the incidence of boundary effects, we take the minimum value of K which provides a non-empty intersection between the circles centered in i and i: $K = \lceil d_H(i,j)/2 \rceil$.

III. EXPERIMENTAL RESULTS

This section provides comparative results obtained on a representative set of RGGs by three different estimators of ED: *Algorithm 1*) a linear estimator directly obtained as the product between the GD and coefficient c_4 introduced in Section II-B, $d_H \cdot c_4$, *Algorithm 2*) the estimator derived in Equation (8) from the intersection of geodetic circles with radius K, $\tilde{d}^{(K)}$, and *Algorithm 3*) the estimator proposed by Merkel *et al.* [11] based on the shortest-path propagation of the distances computed from the intersections between geodetic circles of radius 1.

The three methods were implemented in Matlab and tested by means of Monte Carlo simulations. Each experimental trial entailed: the generation of a RGG with given parameters, the application of the three methods to estimate all pairwise distances, and the computation of the errors made by each method with respect to the actual values of the ED between each pair of nodes. Estimation accuracy was evaluated both in terms of *mean absolute error* (MAE) and in terms of *mean absolute percentage error* (MAPE), as defined by *Merkel et al.* [11].

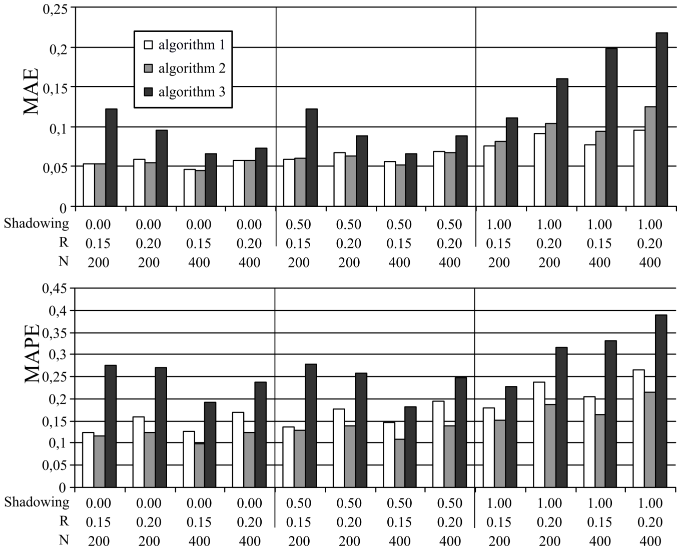

Fig. 4. Comparative results showing the mean absolute error (MAE) and the mean absolute percentage error (MAPE) of the proposed estimators for different values of the parameters used to generate the graphs.

Three parameters were used to generate the benchmarks: the overall number of nodes (N) which corresponds to the density of the Poisson process λ, the communication range (R), and the parameter ξ used to control the shadowing effect according to Equation (2) in [13]. Hereafter we denote by $\xi = 0$ the ideal case of a UDG. For all other cases ($\xi > 0$) the comunication range R has to be regarded as the distance at which the probability of having an edge is 50%, while ξ is proportional to the standard deviation of shadowing.

Figure 4 reports the results obtained by the three estimation algorithms for different configurations of the parameters used to generate the graphs. For each configuration (annotated on the x axis) three bars are used to denote the performance of the estimators. The first set of results, composed of 4 configurations, refer to the ideal case of UDG (shadowing = 0) with different number of nodes (N) and communications ranges (R). The second and third sets refer to incremental values of shadowing corresponding to $\xi = 0.5$ and $\xi = 1.0$.

Both the proposed algorithms (namely, algorithms 1 and 2) outperform the previous approach (algorithm 3) for all

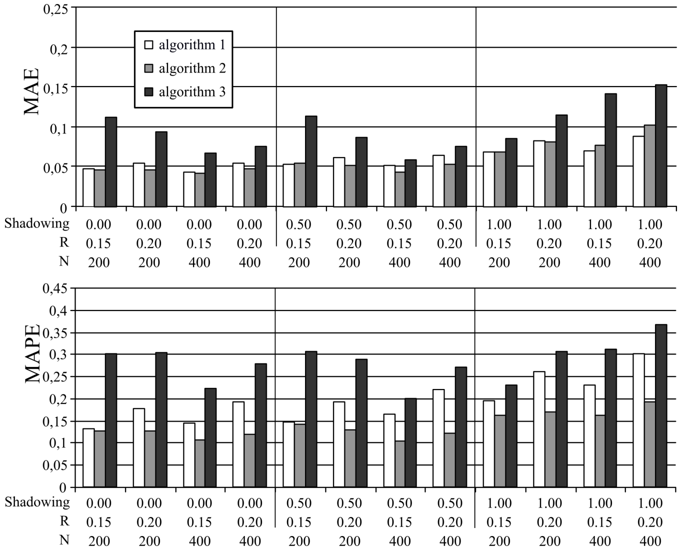

Fig. 5. Comparative results obtained by taking into account only the central nodes of the graphs used as benchmarks.

the configurations and according to both metrics (namely, the MAE reported in the upper bar graph and the MAPE reported in the lower one). While algorithms 1 and 2 are almost equivalent in terms of MAE, Algorithm 2 is consistently more accurate in terms of relative errors (MAPE). This is mainly due to the discrete nature of Algorithm 1, which makes it inherently unable to cope with the continuous distribution of Euclidean distance. As expected, al the algorithms are less accurate when applied to graphs affected by higher levels of shadowing.

In order to evaluate the impact of boundary effects, the same metrics were computed considering, for each graph, only the pairwise distances between nodes falling within a central square of edge 0.5, while all the nodes (including those falling in the outer frame) were considered to compute the intersection areas in Algorithms 2 and 3. The results, reported in Figure 5, confirm the advantage of the proposed approaches. By comparing Figures 5 with 4 we can observe that the restriction to the central nodes is always beneficial in terms of absolute errors, while sometimes the MAPE increases because of the average reduction of the distances under estimation, which appear at the denominator in the computation of percentage errors.

IV. Conclusion

Statistical geometry provides a suitable framework for algorithmic distance estimation in that it enables the exploitation of all available information to minimize the effects of noise and measurement errors. In this paper we applied statistical geometry to derive new estimators of the pairwise distance between the nodes of a graph from the topological information contained in the connectivity matrix.

The experimental results achieved on a representative set of synthetic benchmarks, including boundary effects and shadowing, have shown that the proposed algorithms are consistently more accurate than existing ones. The superior quality comes from two statistical arguments. First, we make use of as much data as possible to estimate each distance value, thus compensating the inaccuracy of the original data (consisting of hop distances in our setting). Second, we avoid error propagation, by adopting a direct method to compute distances between out-of-range nodes.

References

[1] J. Gao and L. Guibas, "Geometric algorithms for sensor networks," *Phil. Trans. R. Soc.*, vol. 370, no. 1958, pp. 27–51, 2012.

[2] L. Zhang, L. Ligang, C. Gotsman, and S. J. Gortler, "An as-rigid-as-possible approach to sensor network localization," *ACM Trans. Sen. Netw.*, vol. 6, no. 4, pp. 35:1–35:21, 2010.

[3] J. Bachrach and C. Talylor, "Localization in sensor networks," in *Handbook of Sensor Networks: Algorithms and Architectures*. Hoboken, NJ, USA: John Wiley & Sons, 2005, ch. 9, pp. 277–310.

[4] Y. Shang, W. Ruml, Y. Zhang, and M. Fromherz, "Localization from connectivity in sensor networks," *IEEE Trans. on Parall. and Distr. Systems*, vol. 15, no. 10, pp. 961–974, 2004.

[5] P. Bahl and V. N. Padmanabhan, "Radar: An in-building rf-based user location and tracking system," in *Proceedings IEEE Congerence on Computer Communications (INFOCOM 200)*, 2000, pp. 775–784.

[6] G. Oberholzer, P. Sommer, and R. Wattenhofer, "Spiderbat: Augmenting wireless sensor networks with distance and angle information," in *Proceedings of the 10th International Conference on Information Processing in Sensor Networks, (IPSN 2011)*, 2011, pp. 211–222.

[7] N. B. Priyantha, A. Chakraborty, and H. Balakrishnan, "The cricket location-support system," in *Proceedings of the 6th annual international conference on Mobile computing and networking, (MobiCom '00)*, 2000, pp. 32–43.

[8] C. Buschmann, H. Hellbrück, S. Fischer, A. Kröller, and S. P. Fekete, "Radio propagation-aware distance estimation based on neighborhood comparison," in *Wireless Sensor Networks, 4th European Conference, (EWSN '07)*, 2007, pp. 325–340.

[9] F. L. Villafuerte, K. Terfloth, and J. H. Schiller, "Using network density as a new parameter to estimate distance," in *Seventh IEEE International Conference on Networking (ICN 2008)*, 2008, pp. 30–35.

[10] B. Huang, C. Yu, B. D. O. Anderson, and G. Mao, "Connectivity-based distance estimation in wireless sensor networks," in *Proceedings of the Global Communications Conference, 2010. GLOBECOM (GLOBECOM '10)*, Miami, Florida (USA), 2010, pp. 1–5.

[11] S. Merkel, S. Mostaghim, and H. Schmeck, "Distributed geometric distance estimation in ad hoc networks," in *Proceedings of the 11th international conference on Ad-hoc, Mobile, and Wireless Networks, (ADHOC-NOW'12)*, Belgrade, Serbia, 2012, pp. 28–41.

[12] S. Nath, V. N. Ekambaram, A. Kumar, and P. V. Kumar, "Theory and algorithms for hop-count-based localization with random geometric graph models of dense sensor networks," *ACM Trans. Sen. Netw.*, vol. 8, no. 4, pp. 35:1–35:38, 2012.

[13] R. Hekmat and P. Van Mieghem, "Interference power statistics in ad-hoc and sensor networks," *Wirel. Netw.*, vol. 14, no. 5, pp. 591–599, Oct. 2008.

[14] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to Algorithms*, 3rd ed. The MIT Press, 2009.