Statistical Model of Lossy Links in Wireless Sensor Networks

Alberto Cerpa, Jennifer L. Wong, Louane Kuang, Miodrag Potkonjak, Deborah Estrin
Computer Science Department, University of California, Los Angeles, CA 90095

Abstract—Recently, several landmark wireless sensor network deployment studies clearly demonstrated a large discrepancy between experimentally observed communication properties and properties produced by widely used simulation models. Our first goal is to provide sound foundations for conclusions drawn from these studies by extracting the relationship between pairs of location (e.g distance) and communication properties (e.g. reception rate) using non-parametric statistical techniques and by calculating intervals of confidence for all claims. The objective is to determine not only the most likely value of one feature for an alternate given feature value, but also to establish a complete characterization of the relationship by providing a probability density function (PDF). The PDF provides the likelihood that any particular value of one feature is associated with a given value of another feature. Furthermore, we study not only individual link properties, but also their correlation with respect to common transmitters and receivers and their geometrical location.

The second objective is to develop a series of wireless network simulation environments that produce networks of an arbitrary size and under arbitrary deployment rules with realistic communication properties. For this task we use an iterative improvement-based optimization procedure to generate instances of the network that are statistically similar to empirically observed networks. We evaluate the accuracy of the conclusions drawn using the proposed model and therefore comprehensiveness of the considered properties on a set of standard communication tasks, such as connectivity maintenance and routing.

Index terms: sensor networks, wireless channel modeling, simulations, network measurements, experimentation with real networks/testbeds, statistics.

I. INTRODUCTION

It is well known that the performance of many protocols and localized algorithms for wireless multi-hop sensor networks greatly depend on the underlying communication channel. Hence, to evaluate performance in simulation, we must have an accurate communication model. Until recently, only two approaches have been in widespread use in the sensor network community: unit disk modelling and empirical data traces.

The unit disk model states that the reception rate between two nodes is solely a function of the distance between them and that communication is conducted without any loss of packets if the nodes are closer than a specified communication range. The unit disk model has a number of noble features. It is ideally suited for theoretical analysis, for the derivation of a lower bound, and the development of optimization algorithms. Moreover, it is straight-forward to develop a wireless network simulator that generates network instances that follow this communication model. However, the unit disk model also has a large number of serious drawbacks. For example, it implies complete correlation between the properties of geometric space and the topology of the network, a property refuted by numerous experiments in real-life deployments [1], [6], [14].

At the other end of the spectrum are networks and communication patterns that are empirical traces of deployed systems. These networks are, of course, completely accurate samples of real life wireless communications. However, it is difficult and expensive to create a large number of large networks that are properly characterized. Therefore, neither probabilistic nor statistical analysis of large networks is feasible. In addition, since a given trace is the result of communication over a specific topology, such a trace does not permit a simulator to reposition nodes. Finally, without validated communication analysis for these cases, theoretical analysis is not possible.

In an effort to address a middle ground, recently there have been a number of efforts to empirically capture communication patterns in wireless sensor networks. In particular, there have been several studies that use different low power, narrow band radio transceivers chipsets [3], [11] to deduce properties of communication links in wireless networks in several environments, such as open space and laboratories. These hybrid models introduce empirically observed factors that modify the communication patterns based on the unit disk communication model.

While these models are a significant step forward with respect to the unit disk model, they are only an initial step in the exploration of the space. These initial models do

not capture many important features of communication links in empirically observed networks. For example, they do not address the correlation in communication reception rates between nodes that originated at the same transmitter or differences in the quality of transmitters. Moreover, these hybrid models were not subjected to statistical validation tests against empirical data. Intervals of confidence on how well the hybrid models actually model the empirical data is also lacking.

Our goal is to develop accurate simulations of sensor network communication environments that are statistically compatible with respect to several features that impact network protocols and algorithms in real networks. To generate these simulated environments, we construct a set of models that map communication properties such as absolute physical location, relative physical proximity and radio transmission power into probability density functions describing packet reception likelihood. For all of these models, we calculate an interval of confidence. These models not only serve to generate simulated environments, they themselves have lent support to many hypotheses relating to variation in communication link quality [1], [6]. We note that the analysis of the MAC impact (e.g. contention, throughput) and temporal properties of the links is part of future work.

II. RELATED WORK

There is a large body of literature on mobile radio propagation models that have influenced this work. The emphasis has been on large scale path loss models that predict the average received signal strength at a given distance from the transmitter and the variability of the signal strength in proximity to a particular location [10]. Furthermore, the models are used to predict the coverage area of a transmitter. In addition, small scale fading models are used for modeling localized time durations (a few microseconds) and space locations (usually one meter) changes. All these models are based on the Fries free space equation and indicate that reception quality decays with the inverse of distance raised to a small power [10].

Differences between the classical models and our approach are numerous and include different modeling objectives (reception rate of packets vs. signal strength), our radios have different features (e.g. communication range in meters instead of km), we capture phenomena that is not addressed by the classical channel models (asymmetry, different quality of receivers and transmitters, correlations between reception rate of links), we use different modeling techniques (free of assumptions, non-parametric vs. parametric), and we use unique evaluation techniques (resubstitution and evaluation of multi-

hop routing). We believe existing and new techniques have complementary objectives, tools, and applications.

More recently there have been many studies of significant-scale deployments in several environments. Majority of these studies used the TR1000 [11] and CC1100 [3] low power RF transceivers (used by the Mica 1 [7] and Mica 2 [4] motes respectively). For example, Ganesan et al. [6] performed large scale experiments with 150 Mica 1 motes in order to capture features of link, MAC, and application layers in wireless communication. Their data firmly established empirical evidence that radio connectivity is not isotropic and there is a significant percentage of asymmetric links. In a series of studies conducted at Berkeley, Woo et al. [13] used experimental measurements to derive a packet loss model based on aggregate statistical measures such as mean and standard deviation of reception rates. Using these models, they analyzed its consequences on a number of communication protocols and neighborhood management policies. More recently, Zhao et al. [14] has demonstrated heavy variability in packet reception rate for a wide range of distances between a transmitter and receiver. Furthermore, Cerpa et al. [1] used heterogeneous hardware platforms consisting of Mica 1 and Mica 2 motes in three different environments to collect comprehensive data about the dependency of reception rates with respect to a variety of parameters and conditions of operation. They also provided substantial empirical evidence that the main cause of asymmetric links is due to differences in hardware calibration as it was previously hypothesized. Another notable work in this area was performed earlier this year by Zhou et al. [15] that studied radio symmetry in wireless communication networks. They postulated that variance in receive signal strength is mainly random, and used the empirical data to develop a new radio model (RAM). Using RAM, asymmetry is generated due to the variance in RF sending power and the differences in path losses for different directions of propagation. Zhou at al. used the parametric goodness-of fit statistical testing to determine that the variance of strength of the received signal that follows a Weibull distribution has the maximal likelihood of matching their experimental data.

There are three major differences between the models developed in this paper and the previously published models. The first is that we study the impact of a significantly large number of factors that impact reception rate and attempt to model not only isolated pairs of transmitters and receivers, but also the correlation between different pairs and different subsets of links. The second major difference is that our goal is not only to establish a model, but also to establish statistically sound measures

```
1. Conduct exploratory data analysis;
2. while (interval of confidence > criteria) {
3.    Collect new data or define new windows;
4.    Sort all points according to distance;
5.    for(from smallest to largest distance) {
6.       Define sliding window for distance;
7.       Apply weight function to distances inside of sliding window;
8.       Sort all points according reception rate;
9.       for(from smallest to largest reception rate ) {
10.         Define sliding window for reception rate;
11.         Apply weight function to reception rates sliding window; } }
12.   Build mapping function;
13.   Build normalized mapping function;
14.   Establish intervals of confidence; }
```

Fig. 1.   Pseudo-code of the PDF model generation for two features.

(a) Data set

(b) Zoom data box (a)

Fig. 2.   Scatter plot of distance vs. reception rate

as to what extent the model corresponds to experimentally captured data. Our statistical techniques are generic in the sense that they can be used in other studies with minimal changes. Also, we have developed a procedure for creating instances of an arbitrary size for simulation and mechanisms to ensure that they are accurate models with respect to the collected data.
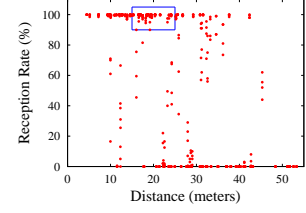
All of our techniques are based on non-parametric procedures. In particular, we directly leverage on smoothing and density kernel estimators [8], [12]. In addition, we use resubstitution [5] and bootstrap techniques.
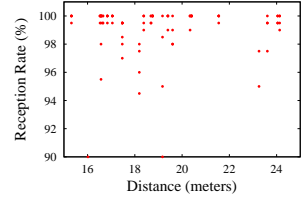
III. INDIVIDUAL LINK MODELS

In this Section, we present a new non-parametric statistical approach for building communication link models in wireless multi-hop networks. The goal is to find a statistically sound mapping between two user-specified features that characterize communication links. The core of the section is a generic non-parametric statistical procedure for establishing a mapping between two properties.

A. Design Guidelines

Our starting task is to analyze the dependency between two properties of wireless networks. We note that exactly the same procedure described below can be used to find the dependency between *any* two wireless communication features, but for the sake of brevity and clarity, we focus on two specific features: distance and reception rate. The objective is to find the PDF of reception rate for any distance with intervals of confidence. For example, we could

use our model to find that the probability of reception rate being 95% at a distance of 25 ft is 0.05 ± 0.000012.

We are guided by two principles, smoothness and prediction ability. The smoothness property states that if two pairs of receivers have very similar distance, their reception rates also often have rather similar probabilistic distribution. In other words, instances of reception rates may be different from one distance to the other, but the underlying reception rate probabilistic distribution is similar. There are two fundamental justifications for this standard assumption. The first is that at an intuitive level one expects that small changes in one variable (in our case, distance) should have limited impact on the probability distribution of the other parameter (reception rate). In addition, it is important to recognize that both distance and reception rate are subject to errors in measurement which additionally smoothes the mapping function.

Secondly, the only sound criteria for any prediction model is its ability to predict. Therefore, we can alter various parameters in our procedures so long as the finally adopted parameter values produce a model that withstands standard evaluations of accuracy. Specifically, we use the resubstitution rule which builds additional models using a variety of randomly selected subsets of the data set. If the resulting models from all subsets are similar we can conclude that the used parameters are properly selected.

From a technical point of view, when building a model of individual links we have two major difficulties: (i) we do not have enough measurements for each distance of interest, and (ii) for a given distance and given reception rate, we also do not have enough collected data samples. As we already stated our main tool to resolve these diffi-
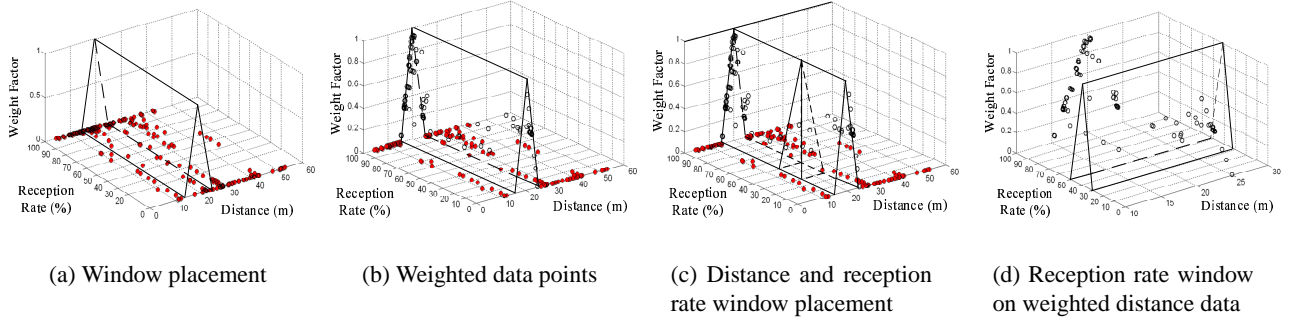
(a) Window placement   (b) Weighted data points   (c) Distance and reception rate window placement   (d) Reception rate window on weighted distance data

Fig. 3. Application of sliding window on distance and reception rate axis.

culties is to use the smoothness principle.

B. Methodology

The global flow of our approach is shown in Figure 1 using pseudo-code format. The starting point for the procedure is exploratory data analysis. As the first step of this phase, we examine a scatter plot of all available data points. Specifically, we position each communication link in a two dimensional space, placing on the x-axis distance and on the y-axis the reception rate. The goal is to identify if there are any specific trends in the data and to determine whether it is advantageous to split the data into two or more subsets that have specific features.

Figure 2(a) illustrates a scatter plot of distance versus reception rate at medium power for the outdoor case. Figure 2(b) shows a zoomed version of a subset of data which was examined during exploratory data analysis.

Phase two consists of three steps shown in Figure 1 in lines 4-8. In the step shown in line 4, we sort all available data according to distance to identify data points that are similar with respect to this parameter. Next, we use a sliding window for all points which are within a similarity range of a given point (distance). Each point within this range is weighted according to its quantified similarity to a given point. Note that both the scope of the window as well as the weighting function can be defined in an arbitrary fashion, so long as the monotonicity property imposed by the smoothing principle is maintained.

Figures 3(a) and 3(b) show how the data is weighted after the application of a triangular window to the set of points around distance 20 meters with window size ± 5 meters. The intuition behind the weighting function is that data points that are closer to the center of the window are better approximations for the point at the center of the window.

For each distance of interest we also build another system of sliding windows this time along the y-axis corresponding to the reception rate. The points within the win-
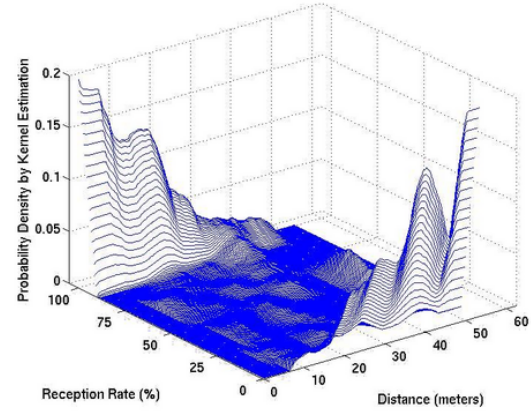

Fig. 4. PDF for distance versus reception rate.

dow are weighted as the product of the weight factor of both the distance window and the reception rate window. Exactly the same techniques that are used to determine the parameters of the distance window are applicable and used for the shaping and sizing the reception rate window. Figures 3(c) and 3(d) show a window of data for the point centered at distance 20m and reception rate 33%.

Once the first eleven lines of the pseudo-code are executed we have enough information to begin building a PDF that indicates how likely a particular reception rate is for a given distance. For this task we used quadratic least linear squares fitting for a particular pair of distance and reception rate.

Once the model is built, the next step is PDF normalization that ensures that for a given distance the integral of the function below the PDF mapping function is equal to one. Figure 4 illustrates how the normalized reception rate PDF changes with respect to distance.

C. Evaluation

The final step of our procedure is the evaluation of the quality of the developed statistical model for the PDF. The
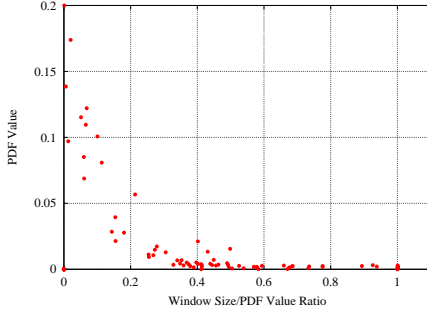
Fig. 5. PDF values of the different random points as a function of the confidence interval/PDF value ratio. Outdoor Urban, 90% Conf. Level.

Environment	Conf. Level	H.L.PDF Value	Conf. Intervals
Indoor	90	0.997627	±0.325969
Outdoor	90	1.064365	±0.381719
Indoor	95	1.023886	±0.723887
Outdoor	95	1.022372	±0.691752

fidence levels. In general, the global highest likelihood PDF values are centered around one, which is a good sign of the statistical soundness of the model.

IV. EXPERIMENTAL DATA COLLECTION

In this section we discuss the methodology used for our experimental data collection. We used an existing data set and performed additional experiments using the SCALE wireless measuring tool [1]. The basic data collection experiments work as follows. Each node transmits a certain number of packet probes in a round robin fashion (one transmitter at a time). Each probe packet contains the sender's node id and a sequence number. The rest of the nodes record the packets received from each neighbor and keep updated connectivity statistics, using the sequence numbers to detect packet losses. We refer to [1] for further details about the tool.

The topology used for our experiments consisted of 16 nodes distributed in an ad-hoc manner in different environments. We also used up to 55 nodes for our indoor experiments distributed in the ceiling of our lab. For outdoor experiments, nodes were placed in a variety of different positions, such as near the ground or elevated off the ground, with or without line of sight (LOS) between them, and with different levels of obstructions (furniture, walls, trees, etc.). The placement of the nodes also took into account the distance between them, in order to create a rich set of links at distances varying from 2 to 50 meters and in multiple different directions from any particular sender.[1] In most of our experiments, each node sends up to 200 packets per round, transmitting 2 packets per second.[2].

Using this setup, we varied five factors in our experiments: the choice of environments, the radio type (and frequency), the output transmit power settings, the packet size settings, and the antenna height. The first factor we

evaluation procedure itself has three components: Monte Carlo sampling, resubstitution, and establishment of interval of confidence. Monte Carlo sampling selects k (in our experimentation we use 200) randomly selected pairs of distance and reception rate points.

Resubstitution is the process where a statistical model is built using the exact same procedure (same kernel window scope and weight function) on randomly selected subsets of data. Specifically, in our simulations, we select 70% of the available data to build a model on each resubstitution run. For each resubstitution run we record the value of the PDF function at each of the k selected points. After conducting m resubstitution runs (in our experimentation m was 100), we are ready to establish an interval of confidence for our statistical PDF model. This is performed in two stages. We first establish an interval of confidence for each point individually, and then by combining information from all local interval of confidence we establish a global interval of confidence. Figure 5 shows the relationship between the different confidence intervals for each random sample tuple (reception rate and distance) and the highest likelihood PDF value for different confidence levels. Each point in the graphs show the highest likelihood PDF value with its confidence interval. For example, the top left point in the graph of Figure 5 corresponds to sample point of distance 52 meters, reception rate 0% with highest likelihood PDF value of 0.2 ± 0.0001953 with confidence level of 90%. The final step of resubstitution is to build a global measure of the model's accuracy. To build a global interval of confidence we use the following procedure. First, for each separate point in k, we use the highest likelihood PDF value and normalize all other values against this value. After that, we combine all data from all sampling points into one set of the size k x m. Finally, we calculate the confidence intervals of the normalized global array. Table I shows the overall interval of confidence for indoors and outdoors with different con-

---

[1]There are algorithms to find the optimal placement of nodes to get a uniform range of distances in the area of interest. We did not perform that optimization in our experiments.

[2]We have left for future work the evaluation of how accurate is an average reception rate

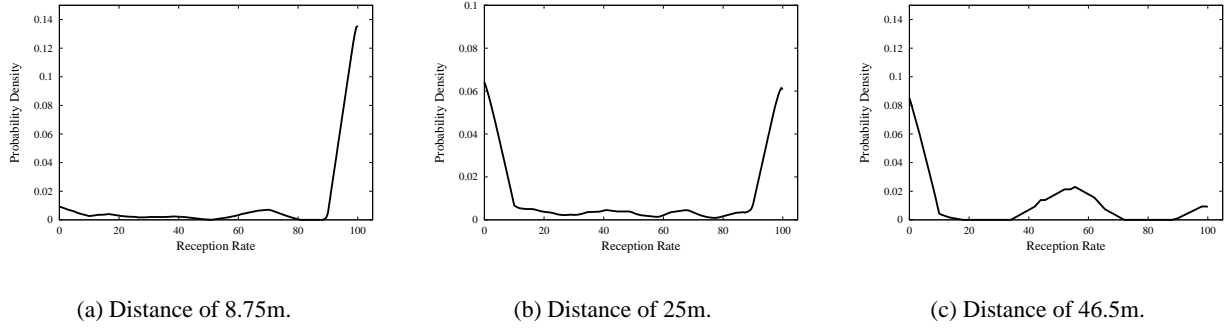|(a) Distance of 8.75m.|(b) Distance of 25m.|(c) Distance of 46.5m.|

Fig. 6. PDFs for reception rate for various values of distance.

varied was the environment type. We selected three environments for our experimentation. We chose our lab to perform some *indoor* experiments. The choice of this environment is motivated by sensing applications in indoor environments [9]. We also picked our Engineering building courtyard as an *outdoor urban* environment. This environment is an intermediate measuring point between indoor places and outdoor natural habitats, due to more interference, reflections, etc. Finally, we used a 200m by 150m section of the Will Rogers State Park, Pacific Palisades, California. There have been several efforts to monitor habitats in sensor networks [2], which motivate the choice of this environment. The second factor we varied was the node platform and corresponding radio type. The Mica 1 mote [7] uses the RFM TR1000 radio chip [11], transmits in the 916MHz band, and uses an amplitude shift keying (ASK) modulation scheme. The Mica 2 mote [4] uses the Chipcon CC1100 radio chip [3], transmits in the 433MHz band, and uses a frequency shift keying (FSK) modulation scheme. The third factor we varied was the output transmission power. The dynamic range of the output power selection with Mica 1 ranges from -10dBm to 0dBm and with Mica 2 ranges from -20dBm to 10dBm. The fourth factor we varied was the packet size. The set of different packet sizes used was 25, 50, 100, 150 and 200 bytes. Finally, the last factor we varied was the antennae heights. We set the nodes' antennae on the ground, 0.25ft, 1ft, 3ft, and 5ft above the ground. Note that the height is measured with respect to the ground level at the node's location which may be different from the absolute z coordinate.

In summary, the data set used in this paper consisted of packet delivery data from more than 450,000 packet probes in experiments performed in 3 different environments, with 2 different type of radios, with 6 different power settings, 5 different packet sizes, and 4 different antennae heights. We used up to 16 nodes in our outdoor experiments and up to 55 nodes in our indoor experiments.

In each experiment, we measured the packet delivery performance of 240 links for the outdoor experiments and 2970 links for the indoors experiments.

V. PROPERTIES OF INDIVIDUAL LINKS

In this Section, we outline the analyzed features of individual communication links in wireless sensor networks.

At the highest level of consideration the features can be classified into two groups: physical properties of the network, and communication features of the network. Physical properties include distance, direction as a function of angle with respect to reference direction, and areas. Communication properties include reception rate between receiver A and transmitter B, asymmetry in communication which refers to the difference in reception rates between a pair of nodes (transmitter A and receiver B and transmitter B and receiver A), and temporal variation of reception rate between receiver A and transmitter B.

We have analyzed two types of mapping functions between properties. The first is the established one-way mapping relationship between a given structural property and a targeted communication feature. The second analyzes the one-way dependency between two communication features. It is important to once again emphasize that our goal is to not only establish the most likely value of one property for a given value of another property, but also to obtain probability distribution functions for all expected values of the second property for a given value of the first feature. We have studied the following pairs of properties.

- **Dependency of reception rate as a function of distance.** This property is selected mainly because there is a wide consensus that distance significantly impacts reception rate.
- **Dependency of asymmetric reception rate as a function of distance.** Note that in the previous case we assumed that reception rate between transmitter A to receiver B is the same as from transmitter B

(a) Asymmetric links vs distance   (b) Asymmetric links vs reception rate   (c) Temporal variation vs reception rate
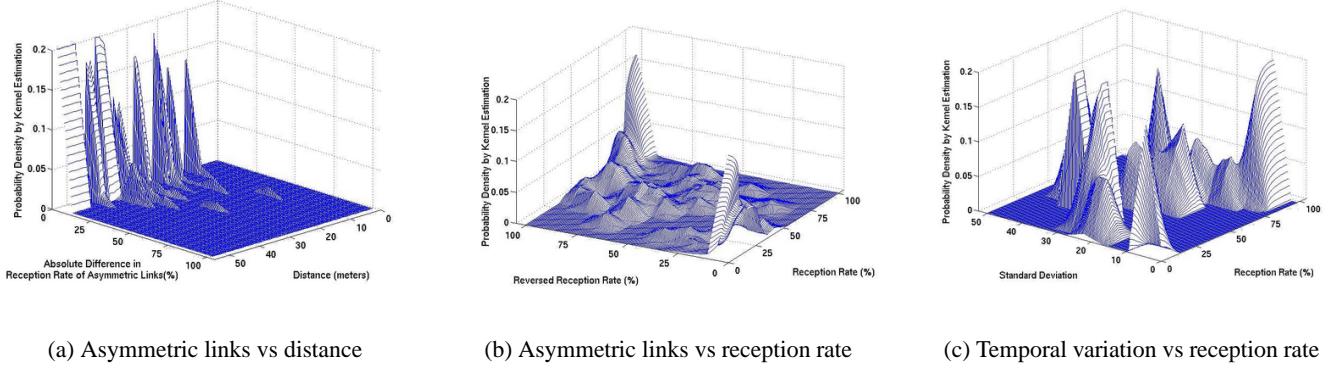
Fig. 7. PDFs for Asymmetric Link and Temporal Variability features.

to receiver A, but recently several empirical studies demonstrated this is not the case [1], [6]. Our goal is to quantitatively capture how frequently there is asymmetry in reception rates as a function of distance.

- **Dependency of asymmetric reception rate as a function of reception rate.** This property is an example where we study functional dependencies between two communication properties. Our goal is to identify if it is more likely that high asymmetry happens when links have high, low, or medium reception rates. For example, we are interested if it is more likely to have a pair of nodes with reception rates of 95% and 75%, or with 30% and 10% reception rates.
- **Dependency of reception rate standard deviation as a function of the average reception rate.** The final property is an example where we study temporal dependencies between two communication properties. An empirical study [1] has shown that such correlation exists. Our goal is to quantitatively capture this relationship and provide some initial results on how this property affects the link estimation algorithms used for online quality estimation.

In addition to the listed properties, we also studied link quality dependency on angle, but were not able to identify any statistically interesting patterns with significantly strong intervals of confidence.

In section III, Figure 4 we have illustrated how the reception rate changes as a function of distance. The figures in 6 show normalized PDFs for three typical distances for 8.75, 25, and 46.5 meters. These results confirm the findings of several studies in the literature that show that there is a significant percentage of the radio range where links are highly variable, with similar probabilities of having very high or low reception rates. In addition, we show that even for very short distances, the probability of hav-

ing very low reception rate links is not zero, and it starts growing fast as distance increases. More importantly, it is clear from the graph in Figure 6(b) that the average and standard deviation values of reception rate are insufficient parameters to model reception rate as a function of distance. While the average reception rate is around 50% in this case, most of the links have either very high or low reception rates. A communication model built using only average and standard deviation values of reception rate may not represent the underlying communication properties found in real environments.

Figures 7(a) and 7(b) show the PDF of how asymmetric reception rate depends on distance and average reception rate. Figure 7(a) shows that there is no clear correlation between link asymmetries and distance. Figure 7(b) shows an interesting pattern; links with very high or very low reception rates tend to be highly symmetrical, as it can be observed by the two peaks in the PDFs. Links with medium reception rate tend to be much less symmetrical.

Figure 7(c) shows the temporal variability of the links as a function of the reception rate. We clearly see that links with very low or very high reception rates tend to be more stable over time (smaller standard deviation), while the links with intermediate values of reception rate tend to be more unstable (higher standard deviation). One interesting question we wanted to answer is how long a node needs to measure the communication channel in order to get an accurate estimate of reception rate with a certain confidence interval. This has a profound impact in the design of algorithms for topology control that need to measure the channel as little as possible in order to save energy by periodically turning the radio off. To evaluate this, we took long time series of reception rate data, and picked k window sizes. For each window size, we took p (set to 100) initial random points of measurements from the time series, generating a reception rate estimate for each
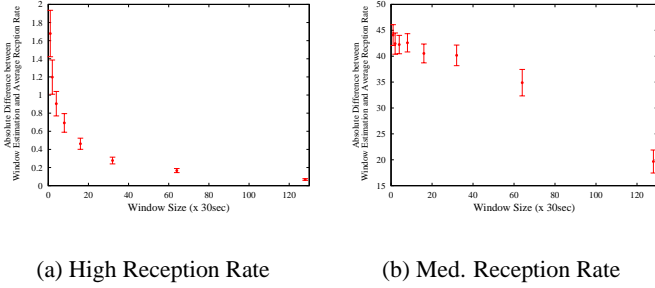
(a) High Reception Rate   (b) Med. Reception Rate

Fig. 8. Time series for on-line link quality estimation.

p using only a window of size k (ranging from 30 seconds to 64 minutes) of data from the starting point. Then we compare the absolute difference between each of the $p \times k$ estimates with the absolute reception rate calculated using the entire time series of data. Figure 8 shows the results of the previous analysis on two qualitatively different type of links. Figure 8(a) shows that links with very high reception rate need very short window sizes to get an accurate estimate of the reception rate, and they converge quite fast to an accurate estimate (low reception rate links show similar behavior). Figure 8(b) shows that links with intermediate reception rates take much larger window sizes to converge to accurate estimate values. We have left for future work the issue of optimal on-line link characterization using statistical methods.

From the spatial, asymmetrical, and temporal properties presented in Figures 4, 7(b) and 7(c) we can see an interesting pattern that has emerged. For a large range of distances there is a low but non-zero probability of links with medium reception rates. These links are also the ones that present the most highly asymmetrical and temporal variability properties. We believe these links may introduce serious stability and convergence problems for several routing algorithms, and it might be useful to design mechanisms to detect these types of links and filter them out. Another interesting point is that reliable, highly symmetrical and stable links exist even at very long distances. Online detection and use of these type of links could affect algorithm design.

VI. GROUP LINK PROPERTIES

Group link properties are joint properties of related subsets of links. These links may be links that originate from the same transmitter or received by the same receiver, processed by the same radio, or communicated by nodes that are geometrically close. These properties are of crucial importance for any analysis and answer the frequently asked fundamental questions about reasons for particular behavior of communication patterns. These questions include whether the performance of a particular node as a transmitter mainly depends on the quality of its radio or its geometric position. Another frequently asked question is whether asymmetry is a consequence of different radio properties between two nodes or their location. However, with the exception of the property which examines pairs of links between two nodes, group link properties have been rarely studied due to their perceived complexity.

The first question we answer is to what extent the quality of transmitters and receivers on different nodes is uniform. We normalized the quality of each link at each node versus the average link quality at the corresponding distance in terms of reception rate. After that, we calculated the geometric mean of all links that originate or end at a particular node.

Figures 9(a) and 9(b) show the PDFs for normalized transmitter and receiver quality of nodes in indoor and outdoor environments. We see that a large percentage of nodes are either significantly better transmitters or receivers than average, in particularly for the outdoor environment. The result is important because if nodes were completely uniform, deployment strategy would be based solely on topology. However, our results show that further communication optimization can be made by considering transmitting/receiving quality.

The second question we answer is whether there is correlation between the quality of the transmitter and receiver on the same node. Figures 9(c) and 9(d) show a summary of our results. In this case, stronger statistical proof was established when considering the arithmetic mean between the transmitter and receiver correlations. All studies indicate that there is a significant positive correlation of transmitting and receiving capability of the nodes. Therefore, a designer could optimize the placement of few very good transmitters and/or receivers in strategic locations (in particular for broadcast and multicast routing) at the cost of fairness (those few nodes may be over utilized).

Once we conclude that some nodes are much better transmitters or receivers than other nodes, the natural question is to what extent they are uniformly better transmitters or receivers with respect to all their links. In order to answer this question we calculated the correlation between all transmitting (receiving) links related to the same node. Table II shows the correlation value r, the t-test value, the degree of freedom (DofF) which is equal to the number of samples reduced by 2, and the probability that this correlation is accidental. For both indoor and outdoor environments we see essentially very small or no correlation with very high probability (the probability of this result being accidental is lower than 0.1% for the indoor case). This essentially means that no node has perfectly
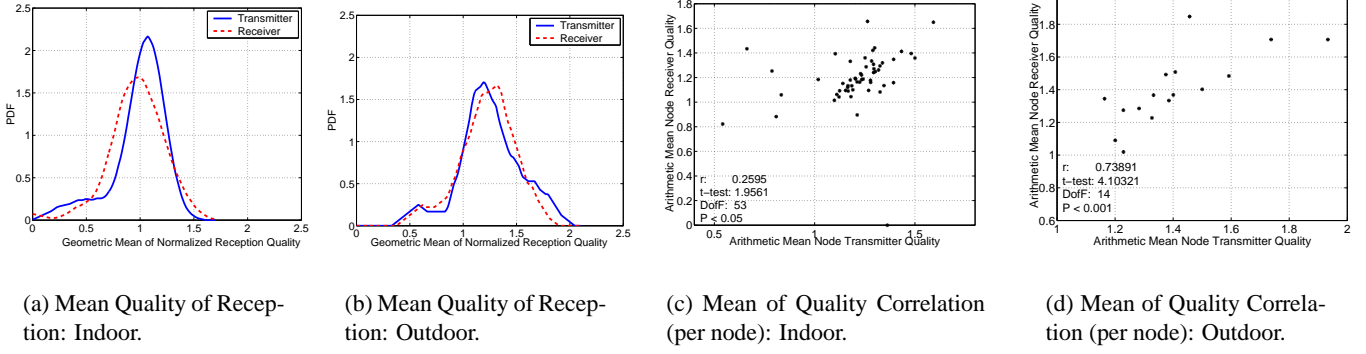
(a) Mean Quality of Reception: Indoor.

(b) Mean Quality of Reception: Outdoor.

(c) Mean of Quality Correlation (per node): Indoor.

(d) Mean of Quality Correlation (per node): Outdoor.

Fig. 9.   PDF for Normalized Transmitter and Receiver Quality and Correlation

TABLE II

CORRELATION OF ALL PAIRS FOR INDOOR AND OUTDOOR.

	Outdoor			Indoor		
	r	t-test	DofF	r	t-test	DofF
Trans.	-0.004	0.121	887	0.0592	11.824	39770
Rcvr	0.012	0.370	885	0.0590	11.859	40183

TABLE III

CORRELATION BY DISTANCE FOR OUTDOOR EXPERIMENTS.

	Transmitter Correlation			Receiver Correlation		
Dist.	r	t-test	DofF	r	t-test	DofF
4-6	-0.009	0.174	357	-0.003	0.050	336
6-10	0.003	0.117	1185	0.039	1.305	1094
10-20	0.000	0.029	4826	-0.009	0.637	4938
20-30	-0.005	0.312	3291	-0.015	0.884	3423
30+	0.016	0.585	1331	0.014	0.474	1120

TABLE IV

CORRELATION BY DISTANCE FOR INDOOR EXPERIMENTS.

	Transmitter Correlation			Receiver Correlation		
Dist.	r	t-test	DofF	r	t-test	DofF
0-2	0.008	3.851	261756	0.010	4.963	262225
2-4	0.012	9.673	629573	-0.001	1.104	627600
4-6	0.007	4.118	374363	-0.011	6.421	368896
6+	0.004	2.522	348187	-0.007	4.089	354141

TABLE V

COVARIANCE MATRIX FOR INDOOR EXPERIMENTS.

	B	C	D	E	F	G	H
A	0.780	0.778	0.578	0.652	0.164	0.114	-0.338
B		0.791	0.545	0.721	0.095	0.144	-0.322
C			0.435	0.640	0.329	0.133	-0.293
D				0.878	0.206	0.202	-0.246
E					0.357	0.684	0.774
F						0.764	0.327
G							0.634

good links to all other nodes in some distance range, and even the best nodes will have average or very poor links. In addition, almost all nodes will have good links to some neighbor in the same distance range.

The situation is even more interesting when we try to examine the correlation between links that belong to nodes which are geographically close. We see with very high confidence that the correlation even for very small distances is rather low.

The correlation data of geographically close receivers and senders is shown in Tables III and IV. In all cases the probability of accidental conclusion is less than 0.01%. The results show that even if two receivers are very close to each other, the fact that one of them has high reception rate from a transmitter does not guarantee that the second receiver would also have good reception rate with respect to the same transmitter. This could also be explained by constructive and destructive interference of multi-path affecting these low power narrow band radios; small variations in location can drastically affect reception. Algo-

rithms based on geographic correlation assumptions may not work well in practice for this same reason.

The last question we would like to answer is whether there are subsets of nodes that communicate well with each other while communicating at significantly lower levels with other nodes in the network. Table V shows the covariance matrix for 9 nodes in the indoor environment. We clearly see that nodes A, B, C and D form one group, nodes D, E, and H another, and F and G, the third group. All nodes in these groups are highly correlated in terms of normalized communication with respect to other nodes. The data was obtained in the following way. For each node we sorted in decreasing order the quality of its links to other nodes. After that, for each pair of nodes, we found a subset of corresponding receivers that hear both nodes, and eventually found rank correlation for these two lists. As part of the table indicates, very often the correlation between two nodes is rather high, close to positive 1

or very low close to -1. The Spearman t-test indicates that all covariance values have probability of accidentally happen well below 0.1%. In other words, group of nodes in a particular distance range can communicate to each other *significantly* better than other group of nodes in the *same* distance range. Identification of these groups of nodes could be important for tree-based routing algorithms; it would be convenient that at least one node in each of these groups join the tree since it could communicate better to the other nodes in the group than any other node.

VII. WIRELESS NETWORKS GENERATORS

Using the knowledge gained from analysis of single and multiple link properties, we have built a series of wireless multi-hop network instance generators to be used in simulation environments. In this Section, we present three models, increasing in complexity, which create communication links for an arbitrary network that are statistically similar to observed networks. The basic model assigns communication links based solely on the relationship between reception rate and distance. To build the more complex models, we introduce an iterative improvement-based procedure for creating communication links which abide by multiple link properties. The starting point for all models is the generation of a specified number of nodes in the given area. We allow the user to either specify specific locations for each node or to specify which distribution the placement of the nodes in the given area must follow.

A. Probabilistic Disk

The basic model, probabilistic disk, considers only the dependency between distance and reception rate. It is created by generating for each calculated distance between two nodes a randomly selected reception rate according to the PDF (Figure 4) for the respected distance. We first translate the PDF into the corresponding cumulative distribution function (CDF) and use a uniform random generator between 0 and 1 to generate a value of CDF. The resulting value is then mapped into the corresponding reception rate.

B. Bi-Directional Correlated Probabilistic Disk Model

In this model, we consider two functional dependencies. In addition to the dependency between reception rate and distance (Figure 4), we also consider the dependency between asymmetric reception rate and reception rate (Figure 7(b)). Our goal is to generate an instance of the network where all communication links follow the PDF for a corresponding distance *and* for any given pair of nodes, we have a reception rate between transmitter A

and receiver B and transmitter B and receiver A that follows the PDF for asymmetric reception rate. An instance of the network which follows this model is generated in the following way. We first generate for each pair of nodes the reception rate between the transmitter of node A and receiver of node B, where the notation of nodes A and B for a given pair of nodes is randomly conducted. Next, we generate the reception rate of the transmitter of node B and the receiver of node A following the PDF for reception rate into probabilistically selected asymmetric rate using one of the previously mentioned methods. One can prove using Bayesian rules that the network generated using this procedure does follow both PDF functions.

C. Non-parametric Statistical Model

While the the Bayesian rule is powerful enough to generate instances of the network that follow one and in some cases two PDFs, it is easy to see that when a large number of statistical measures must be followed, it does not provide an adequate solution. For example, it is not clear how to simultaneously generate a network which follows PDFs for reception rate versus distance, asymmetric reception versus distance, and non-uniform quality of transmitters and receivers. In order to overcome this difficulty we have developed an iterative improvement-based algorithm that generates an instance of the network that approximately, or arbitrarily closely, follows an arbitrary number of interacting PDFs defined on arbitrary pairs of network and communication properties. The key idea is to first separately generate an instance of the network that follows each of the considered PDFs and to randomly select one of them as a starting point for the iterative improvement procedure. At each step of the iterative improvement procedure, we attempt all possible changes at all possible pairs of nodes A and B and select one which makes the overall discrepancy between the parameters of that network more similar to a combination of the originally generated networks that separately considers the PDF of only a single property. The similar PDF function is defined using standard L_2 measure. The procedure is repeated until no further improvement can be found. In order to improve the quality of the generated network, one can perform restarts or employ probabilistic mechanisms for escaping local minima (e.g. simulated annealing). If a restart is performed, there is the option to probabilistically select one of the final solutions for the restart according to their maximum likelihood expectation. These expectations are generated from the space that contains all networks that follow all the specified PDFs.

TABLE VI

Comparison of four statistical models using Floyd-Warshall all pair shortest path algorithm.

	Unit	Unit Real	Prob.	Prob-Real	Assym	Assym Real	Statistical	Statistical-Real
MIN	2	2	2.0079	2.0079	2.00188	2.00188	2.00002	2.00002
MAX	26	20.0569	41.881	43.354	45.9964	44.1535	42.99	42.9285
AVE	6.87574	5.78918	14.687	15.002	14.8176	14.6217	14.6991	14.6928

(a) Empirical Network Data.

(b) Probabilistic Model Generated Network.

(c) Large Network with Empirical Data

Fig. 10. Illustration of three phases for large network generation.

Fig. 11. Similarity between path weights in large networks.

D. Generation of Large Network Instances

Scalability is one of the key issues in wireless sensor networks both during deployment as well as during protocol and algorithm development. In particular for the demonstration of localized algorithms it is important to have instances of the wireless network with a large number of nodes. Unfortunately, it is both expensive and time consuming to deploy large networks solely for the purpose of building a model or developing a localized protocol. Therefore, there is a need to develop a methodology and approach that creates and validates networks of an arbitrary size. While their creation is often rather straightforward, since all that is needed is to place the communication between a pair of nodes according to our statistical model, we can not compare their statistical similarity to existing networks directly. At first glance it seems that empirical evaluation of the accuracy of a large generated network is an infeasible task.

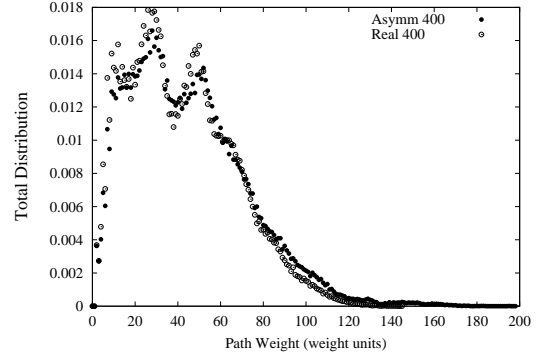However, we have developed a perturbation-based anal-

ysis that facilitates sound statistical validation of networks of an arbitrary size with respect to experimentally available and characterized networks. The key idea is to begin by creating an instance of the network using a specific communication model, and run the algorithm or protocol of interest on the network instance. Next, we replace randomly selected subparts of the instance with instances of data from actually deployed networks. Figure 10 illustrates how the replacements is conducted. Figure 10(a) shows small actually deployed networks. For the sake of clarity we show only the area in which nodes are deployed and represent each node by a dot. Figure 10(b) shows an instance generated by one of the models using the same format. Finally, Figure 10(c) presents the perturbation-based compound instance ready for evaluation. The last step of the procedure is to generate, using the selected statistical model, the connectivity between nodes in the patches of the real networks and the neighboring nodes from the generator. After the procedure is completed, we compare the initial and perturbed networks with respect to results they produce on a task of interests. The extent to which the results are similar, the large instance is representative of the real-life networks. Figure 11 illustrates the similarity in terms of all-pairs shortest path between two large network examples of 400 nodes built using the asymmetric link model.

We compared using the perturbation-based method four models: unit disk model, probabilistic disk model, asymmetric probabilistic disk model, and a non-parametric statistical model. For this purpose we compare the length of all-pairs shortest paths for an instance with 400 nodes. Ta-

ble VI provides a summary for the length of the minimal, maximal, and average path. Notice, that all three newly developed models are much more statistically sound. In particular the non-parametric statistical model provides a very close match before and after perturbations.

VIII. DESIGN CONSIDERATIONS AND CONCLUSION

From the conceptual point of view, the first important observation is that the distribution of lossy links can greatly affect routing algorithms based on geometric concepts. For example, all local avoidance approaches that reduce the routing problem to traversal on Gabriel or local neighborhood graphs may no longer be applicable in practice. Another, possibly more impacting ramification is that no deterministic method can be used to guarantee packet delivery in stateless routing protocols. This is justified by the small but non-zero probability of having links with *very small* or close to *zero* reception rate even at very small distances (Figure 4). The third major conceptual change is that there is a strong benefit of observing at least some percentage of links on-line. This is because some of the most effective links in terms of metrics of travel distance versus required number of messages are links that have a reception rate between 40-60%. In addition, we can observe from Figures 4, 7(b) and 7(c) that it is perfectly possible to find high reception rate links that are stable and highly symmetrical that cover medium to long distances.

The complex and correlated nature of links implies that newly developed routing protocols should be simulated for much longer periods of time in order to ensure that overall they perform well. The existence of superior nodes in terms of both transmitters and receivers capabilities implies that fairness will become one of the major issues for any routing, multicast, and broadcasting approach, because all of these protocols have a tendency to disproportionately use a subset of nodes.

The statistically demonstrated space correlation will also greatly impact the development of routing protocols, as well as power management techniques. For example, since nodes are naturally clustered in subsets that efficiently communicate with each other and poorly with the rest of the network, it will be important that power management strategies, simultaneously turn down or up the majority of the nodes in one of such subsets. Furthermore, clustering techniques might be even more efficient than in networks modeled with the unit disk communication model.

In summary, we have developed a set of non-parametric statistical models for characterizing links in wireless sensor networks. In addition to modeling properties of a single link our statistical models also capture features of groups of links associated with a particular receiver, a particular transmitter, a particular radio, and links associated with a group of radios that are geographically close. The models are the basis for new generators of wireless networks to be used in simulations that are statistically similar to deployed networks. The models and generators are analyzed in terms of their accuracy. The insight gained while building these models has helped identifying future directions for developers of protocols and localized algorithms for wireless sensor networks.

REFERENCES

[1] A. Cerpa, N. Busek, and D. Estrin. SCALE: A tool for simple connectivity assessment in lossy environments. Technical Report CENS Technical Report 0021, Center for Embedded Networked Sensing, UCLA, September 5 2003.

[2] A. Cerpa, J. Elson, D. Estrin, L. Girod, M. Hamilton, and J. Zhao. Habitat monitoring: application driver for wireless communications technology. *SIGCOMM Comput. Commun. Rev.*, 31(2 supplement):20–41, 2001.

[3] Chipcon. Cc1000 low power radio tranciever, http://www.chipcon.com.

[4] Crossbow. Mica2 wireless measurement system, http://www.xbow.com.

[5] B. Efron and R. Tibshirani. *An introduction to the bootstrap*. Chapman & Hall, Ltd., 1993.

[6] D. Ganesan, B. Krishnamachari, A. Woo, D. Culler, D. Estrin, and S. Wicker. Complex behavior at scale: An experimental study of low-power wireless sensor networks. Technical Report UCLA CSD-TR 02-0013, Center for Embedded Networked Sensing, UCLA and Intel Research Lab, UCB, February 2002.

[7] J. Hill and D. Culler. Mica: A wireless platform for deeply embedded networks. *IEEE Micro*, 22(6):12–24, Nov/Dec 2002.

[8] A. B. Owen. *Empirical Likelihood*. Chapman and Hall, 2001.

[9] J. Rabaey, E. Arens, C. Federspiel, A. Gadgil, D. Messerschmitt, W. Nazaroff, K. Pister, S. Oren, and P. Varaiya. Smart energy distribution and consumption: Information technology as an enabling force. Technical report, Center for Information Technology Research in the Interest of Society (CITRIS), 2001.

[10] T. S. Rappaport. *Wireless Communication: Principles and Practice*. Prentice Hall, 2000.

[11] RFM. Tr1000 low power radio system, http://www.rfm.com.

[12] R. A. Thisted. *Elements of statistical computing*. Chapman & Hall, Ltd., 1986.

[13] A. Woo, T. Tong, and D. Culler. Taming the underlying challenges of reliable multihop routing in sensor networks. In *Proceedings of ACM Sensys 2003*, pages 14–27, Los Angeles, CA, USA, November 5–7 2003. ACM.

[14] Y. Zhao and R. Govindan. Understanding packet delivery performance in dense wireless sensor networks. In *Proceedings of ACM Sensys 2003*, pages 1–13, Los Angeles, CA, USA, November 5–7 2003. ACM.

[15] G. Zhou, T. He, S. Krishnamurthy, and J. A. Stankovic. Impact of radio irregularity on wireless sensor networks. In *International Conference on Mobile Systems, Applications and Services*, pages 125–138, 2004.