On the Selection of Testbeds for the Evaluation of Sensor Network Protocols and Applications

Pablo E. Guerrero*, Iliya Gurov*, Silvia Santini§ and Alejandro Buchmann*
* Databases and Distributed Systems
Email: {guerrero, gurov, buchmann}@dvs.tu-darmstadt.de
§ Wireless Sensor Networks Lab
Email: {santinis}@wsn.tu-darmstadt.de
Technische Universität Darmstadt, Darmstadt, Germany

***Abstract*—Wireless sensor network protocols and applications typically need to be evaluated and tested not only using simulators but also on testbeds. While simulations allow studying the performance of protocols and applications in a controlled environment, they usually do not provide a sufficient level of realism. On the contrary, testbeds allow exposing protocols and applications to the real vagaries of wireless communication as well as to other "non-idealities" that typically occur in real wireless sensor network scenarios. Experimental results gathered in one or more testbeds are thus typically reported in most recent research papers on wireless sensor networks.**

Usually, the higher the number of different testbeds has been used, the more significant the obtained results are considered to be. However, it is often unclear whether the used testbeds actually expose protocols and applications to significantly different experimental conditions. Experiments involving a high number of testbeds are very time-consuming and cumbersome to run, but cannot guarantee that a protocol or application has been evaluated or tested in significantly different scenarios. In this paper, we argue that a systematic methodology that allows describing how significant the differences between testbeds actually are is needed. As a first step towards the definition of this methodology, we define several quantitative properties that can be used to describe and compare testbeds in a coherent manner. We show how a representative subset of these properties can be computed in real testbeds and present the results obtained by running this computation on two different testbeds. Furthermore, we describe how we plan to use our methodology to allow researchers to automatically select – out of a set of testbeds – those that are most adequate to run a specific experiment.

I. Introduction

Protocols and applications for wireless sensor networks (WSNs) usually require a careful and thorough evaluation before they can be used in real deployments. This evaluation is typically conducted both through simulation studies and testbed-based experiments. In particular, the use of testbeds allows to expose protocol and applications to realistic experimental conditions.

When using testbeds, researchers typically upload their code on sensor nodes using a centralized (and often Web-based) interface. Nodes then run the code as in a real scenario, using their wireless communication module to interact with each other. Usually, a cable connection allows to power the nodes – so as to avoid frequent battery replacements – and a data backchannel is used to collect experimental data. Over the last years, several different testbeds, including MoteLab[1], TWIST [2], and TUDμNet [3] have been built. The availability of these facilities allows researchers to perform testbed-based evaluations of their protocols and applications. Recent studies have however shown that experimental results obtained on different testbeds are often inconsistent [4], [5]. In other terms, the results of an evaluation conducted on a testbed A might bring to different conclusions with respect to the same evaluation conducted on a testbed B.

WSN researchers thus typically evaluate their protocols and applications not only on one but at least on a few testbeds (e.g., three as in [6]). In some cases, even several different testbeds are considered (e.g., twelve as in [7]). In general, a *more is better* policy is implicitly used, i.e., an evaluation run on many testbeds is considered more significant than one run on just one or few. However, the use of several testbeds is not only impractical – since running experiments on testbeds is cumbersome and time-consuming – but also does not guarantee that the protocol or application under examination has been evaluated against a set of significantly different experimental conditions. Indeed, the rationale according to which these testbeds are chosen is often arbitrary. In particular, ease of access to – or the proficiency of a researcher with – a particular testbed typically guides the choice of one site over the other. An accurate choice of the testbeds is however crucial to ensure an evaluation to be performed under experimental conditions that differ significantly from each other. Basic *physical* properties of a testbed such as its spatial extension or the number of available nodes are important parameters, but not sufficient to completely characterize an experimental facility.

To address this issue, we introduce a methodology to describe the characteristics of a testbed in a comprehensive and coherent manner. Using this description, we aim at enabling researchers to select the best suited subset of testbeds to use for a specific evaluation. We thus address both the issue of testbed description and testbed selection. The remainder of the paper is organized as follows: Section II summarizes related work; Section III introduces our approach to testbed selection and description; Section IV describes preliminary results obtained for our description methodology based on experiments run on two different testbeds. Section V concludes the paper and discusses possible directions for future research.

978-1-4673-5577-3/13/$31.00 ©2013 IEEE

II. Related work

To date, a plethora of testbeds – such as MoteLab[1], TWIST [2], and Indriya [8] – have been built and made available to the public. Furthermore, testbed federations, like KanseiGenie [**?**], and more recently TUDμNet [3], offer a uniform interface to work with a number of individual testbeds. These sites offer limited possibilities to emulate various topologies (e.g., by changing transmission power levels or disabling particular nodes). There is no explicit support for specifying test topologies, leaving it to the researchers to find out useful configurations.

A number of studies have surveyed experimentation sites qualitatively. Gluhak et al. [9] are the first to provide a testbed taxonomy; they employ 7 axes: *heterogeneity*, *federation*, *user involvement*, *repeatibility*, *mobility*, *scale* and *concurrency*. These properties provide high-level insights about the architecture and organization of a testbed, but fall short in considering dynamic aspects of the topologies which are vital to a methodical selection of experimentation sites.

To *understand* network properties, Cerpa et al. [10] presented SCALE, a wireless network measuring and visualization tool that characterizes packet reception ratio (PRR) between nodes. They deployed the system in three different environments and quantified link asymmetry, non-isotropic connectivity and non-monotonic distance decay. Werner-Allen et al. [1] implement a similar tool called Connectivity Daemon and integrate it as a background process into the MoteLab testbed. While these tools enable understanding node connectivity, they have not been conceived to enable a direct comparison between sites nor to allow identifying the potential test sites for a protocol.

The END metric, proposed by Puccinelli et al. in [11], aims at capturing the effects of network properties (node and sink placement as well as and link dynamics) on data collection protocols. The metric combines PRRs from the links in the shortest path tree into a single number that can be used to better understand whether changes in the observed protocol performance (e.g., goodput) are due to protocol mechanisms or rather to topology-related effects. This work partially shares our goal towards methodically comparing site properties but it focuses on collection protocols only.

All the approaches mentioned above assume the existence of a testbed backchannel, i.e., of an infrastructure that both powers the nodes and allows to log data from them. This limits their applicability to experimentation sites that can rely on such an infrastructure (i.e., excludes stand-alone WSNs). We argue that a software tool that captures network properties should not necessarily assume this infrastructure to be available. In the next section we sketch our approach, which, although is evaluated on testbeds, does not rely on a backchannel.

III. Our Approach

Our approach, illustrated in Fig. 1, is composed of *two steps*. In step `a)` we combine the (rather static) *site properties* with a number of (dynamic) *site measurements*, which allows

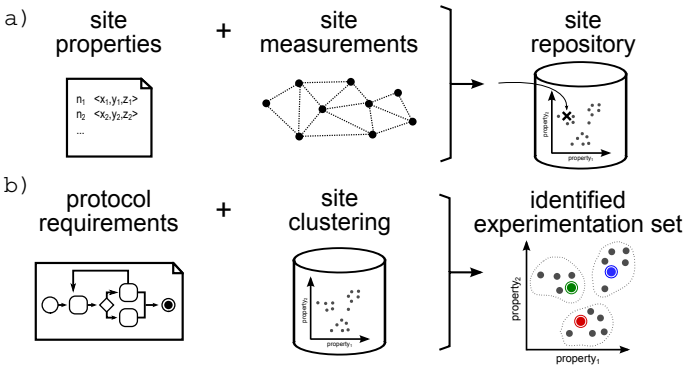


Fig. 1: High level approach

positioning the experimentation site in an n-dimensional space (the new cross in the *site repository*). When, later in step `b)`, researchers want to perform an evaluation, the site properties of relevance to the protocol (i.e., the protocol's requirements) are fed to a *clustering* algorithm, which identifies a set of sites that span a range of properties. It is up to the researchers (and their resources) to decide which sites will effectively be used.

This approach relies on two pillars: first, a catalog of site properties that allow detecting differences and similarities between experimentation sites needs to be identified. These properties must be *quantitative* (to allow a systematic comparison between sites) and relevant not to a single protocol, but to families of protocols. Furthermore, they have to be chosen after carefully studying protocols' main mechanisms. Second, the site repository needs to be populated with a large number of entries of up-to-date information from (publicly accessible) testbeds, to effectively be used by the research community. In the following, we first present our initial version of the properties' catalog of the site repository. Then, we exemplarily show the importance of a subset of these properties through measurements performed in two different testbeds.

A. Site Repository's Property Catalog

Our catalog is organized into a number of *properties*. An important aspect of these is that they are all *quantitative* and *finite*. This allows a direct, objective comparison between sites. While, eventually, all of these properties change over time, we divide them into static and dynamic, depending to the timeframe in which this occurs.

Static properties change only over long time periods (e.g., months or years). Attributes such as the number of nodes available in a site or their actual positions change as nodes break and get serviced/repaired, nodes get moved (e.g., due to building construction/renovations) or when a site grows or shrinks in size. These properties are generally recorded via inspection (in some cases advanced positioning tools have been employed, which allow a precision of $\pm$ 1cm). *Dynamic* properties change much more frequently, and are a result of complex phenomena such as internal or external interference or multi-path fading. We list these properties in Table I.

B. Site Catalog

The values of the static and dynamic properties of a site are represented in a vector $< v_1, v_{2.a}, ..., v_{6.b.ii} >$ (where v_1

TABLE I: Proposed site properties

static	1) number of nodes 2) node positions a) inter-node distance (min., avg., max.) (m) b) density (nodes/m^3)
dynamic	3) packet reception ratio (PRR) a) node degree (min, avg, max) b) % of links with non-zero PRR c) % transitional links d) % significantly asymmetrical links 4) received signal strength (RSSI) a) RSSI (min, avg, max) b) correlation between RSSI and distance 5) link quality indicator (LQI) (min, avg, max, variance) 6) networking a) broadcast/convergecast i) network delivery ii) network diameter b) point-to-point i) network delivery ii) network diameter

is the number of nodes of a site, $v_{2.a}$ is the minimum inter-node distance, $v_{6.b.ii}$ is the point-to-point network diameter, etc.). Since an experimentation site can be operated using different configurations, we employ one entry for each of these configurations in our site catalog. A site's configuration is composed by the chosen transmission power level and radio channel. An entry in our catalog, thus, will look as follows:

$$< site, date, tx\_power, radio\_channel, v_1, v_2, ..., v_n >$$

The data for this catalog is provided from different sources, e.g., running a site characterization process in a testbed. The frequency and the duration of the necessary site measurements present an important trade off in our approach: *frequent* updates to the catalog help obtain more representative results, but are costly (e.g., block a testbed for a certain time running the site characterization software); *occasional* updates might fail to capture details (in particular of dynamic properties), leading to stale information and thus to invalid results. This is an open challenge that we will address in future work.

IV. Empirical Evidence

In this section we exemplify how the selected properties are useful in distinguishing between experimental sites. Next, we describe two testbeds used for this initial study, and describe the methodology employed to collect site data.

A. Methodology

We have employed two indoor testbeds, *Piloty* and *Arena*, which are integrated into the TUDμNet federation [3], and are publicly accessible to the research community[1]. Both sites are instrumented with TelosB nodes [12], equipped with a CC2420 wireless radio.

The *Piloty* testbed spans two floors of the Computer Science department of the Technische Universität Darmstadt. Nodes are placed on windows or over fluorescent tubes inside the department's offices. (Fig. 2 depicts the first floor, the second floor is very similar.) The *Arena* testbed, depicted in Fig. 3, covers an area of approximately $220m^2$, including a disaster arena built following the Arena Assembly Guide from RoboCup Rescue's competition. The site comprises 60 nodes mounted on the ceiling in a 5x12 grid.

[1] www.tudunet.tu-darmstadt.de

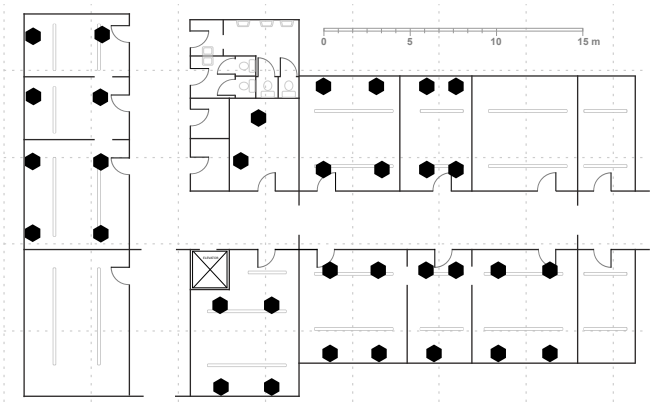


Fig. 2: The *Piloty* testbed's 1st. floor

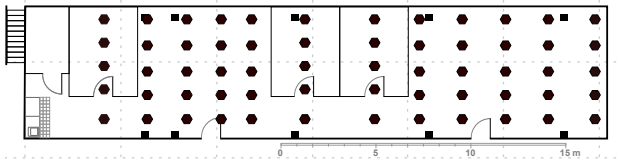


Fig. 3: The *Arena* testbed

In order to evaluate the properties of the links, we developed a Contiki application that exchanges probes (radio packets) among nodes to measure link quality. A *master* node coordinates the activities, and repetitively (in a round-robin fashion):

1) designates one node as *sender* and all other as *receivers*,
2) instructs the sender node to begin sending the probes –while all other nodes remain idle–, and
3) requests the collection of statistics centrally.

Statistics collected from the nodes include the PRR, as well as RSSI and LQI of each successfully received probe. Sender nodes transmit 200 probes, a test run executes for several hours. We repeat the experiments with 4 different transmission power levels (0, -5, -15 and -25 dBm). The tool further allows adjusting probe size (in our tests we used 64-byte packets) and the inter-probe interval. While, functionally, this application is similar to the one presented in [10], it is more flexible since it does not require a wired connection to the nodes for instructing measuring actions and collecting data.

B. Static properties

The physical properties of both sites are summarized in Table 4a. While the average inter-node distance of these two sites is similar (13.9m vs 10.4m), the Arena testbed has a high-density WSN. This is also shown in the histogram presented in Fig. 4b. The Piloty testbed has a more normal distribution (higher number of intermediate length links), whereas the Arena testbed shows a heavy tailed distribution (higher number of short links, fewer long links).

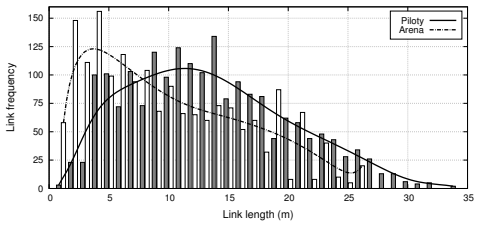
C. Dynamic properties

Although both sites are located in indoor, office premises, they exhibit similarities but also considerably different be-

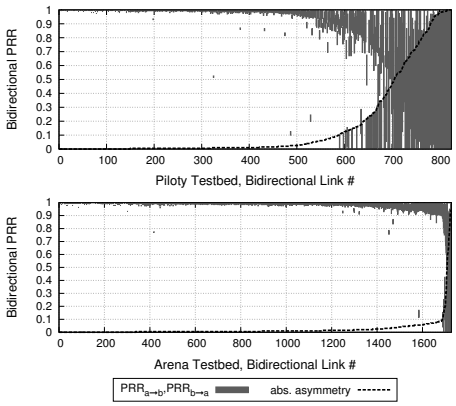
978-1-4673-5577-3/13/$31.00 ©2013 IEEE

(a) Two TUDμNet Indoor Sites

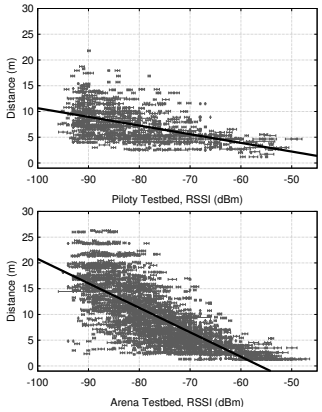
Site	Nodes	Physical Size (m^3)	Distance (m) min	avg	max	Density (n/m^3)
Piloty	63	30×20×8	1.2	13.9	34.4	0.01
Arena	60	31×7×3	1.2	10.4	26.5	0.09



(b) Distribution of link lengths



(a) Bidirectional links (TX power 0dBm)



(b) RSSI vs. distance (TX power -5 dBm)

Fig. 5: Sites' static and dynamic properties

haviour. Table II presents the obtained dynamic characteristics of both sites for different configurations; we elaborate on 4 of them in the next subsections.

1) Connectivity Regions: Adhering to the definition of link quality in terms of PRR as in [13] (*poor*, [0%-10%); *intermediate*, [10%-90%); and *good*, [90%-100%)), three *regions* are distinguished: disconnected, transitional and connected, respectively. Table II lists, for different TX power levels, the percentage of directed links (out of all possible links in a site) for which at least one probe was received (PRR>0). This number clearly depends on spatial properties of the site and the position and antenna orientation of the nodes. The spatial distribution of nodes (sparse in Piloty, dense in Arena) leads to differences in the proportion of links (i.e., 41% vs. 89%, respectively). Generally, increasing the transmission power yields a higher percentage of links with PRR>0. This has, however, other effects described next.

In our catalog we capture also the percentage of links in the transitional region (out of those links with PRR>0). The actual proportion of the transitional region is relevant because protocols that do not filter out this type of links typically yield a poor performance. Studies with an earlier sensor platform (Mica1) identified considerably large transitional regions (even greater than 50% in some cases [14]). More recent studies using nodes with IEEE 802.15.4 compliant-radios have shown it to be much smaller, except in uncommon environments like road tunnels, where it again was high. In our studied sites, the transitional region is much narrower, and is observed increasingly at lower TX power levels. This symbolizes the importance of a packet-based evaluation of link quality to compare sites.

2) Link Asymmetry: We consider link asymmetry as another important dynamic site property. It refers to the situation where probability of successful transmission from one node to another is different in each direction. Link asymmetry, for instance, leads to the existence of multiple minimum hop-count paths with poor throughput (as shown by De Couto et al. in [15]). As a result, routes with significantly less capacity are often preferred in minimum-hop-count routing protocols instead of choosing the best paths in the network.

We consider asymmetry to be significant when $|PRR_{a\to b} - PRR_{b\to a}| > 10\%$. Figure 5a presents the bidirectional links (those with PRR>0 in both directions) for the highest TX power level; each vertical bar represents the span in PRR difference for each link, while the dotted line is the absolute PRR difference. As we can see from the measured results for the highest TX power level, 29% of the bidirectional links exhibit significant asymmetry in the Piloty testbed, whereas in Arena it is only 2%. Thus, if a routing protocol does not consider link asymmetry and is only evaluated in the Arena testbed, it will not run into situations where nodes select an unfavorable next-hop neighbor (since there are almost no asymmetric links), and it will be concluded that the protocol performs very well in a WSN. However, when running such an algorithm in a site like the Piloty testbed, where much more bidirectional links are significantly asymmetric, it will perform poorly.

3) Hardware-based Link Quality Estimators: We now turn to a property relevant to localization algorithms: the correlation between RSSI and distance, depicted in Fig. 5b. While the correlation between distance and RSSI in the Arena site is relatively good (higher RSSI values map to shorter distances and vice versa), in the Piloty testbed it has a more random nature. Each data set tends to cluster around a straight non-horizontal line, which we approximate using least-squares regression. In order to quantify the goodness of fit of this line to the given data set, we calculated the linear correlation coefficient, commonly denoted as r. Indeed, a purely RSSI-based localization protocol under test will behave correctly in the Arena testbed (r=-0.79), whereas, when testbed in the Piloty site (r=-0.57), it will not entirely work as expected.

As a result, if a researcher performed an evaluation of an RSSI-based localization protocol, our clustering algorithm would position the two sites in different sets and recommend to test the protocol against both of them.

4) Networking: Lastly, we consider relevant properties in the network protocols family. We divide them into two main categories: broadcast/convergecast and point-to-point networking, where the first one is relevant to 1-to-n routing protocols, whereas the second one applies to n-to-n routing protocols.

978-1-4673-5577-3/13/$31.00 ©2013 IEEE

TABLE II: The obtained dynamic properties of the two sites entered as entries in our site catalog

Site	TX power (dBm)	Radio Ch.	Node Degree			% of links with PRR>0	% of links in transitional region	% of links w. significant asymmetry	RSSI				LQI			Broadcast/ Convergecast		Point-to-Point	
			min	avg	max				min	avg	max	r	min	avg	max	Ø	Net. Del.	Ø	Net. Del.
Piloty	0	26	12.00	25.30	39.00	41	17	29	-95.08	-78.88	-37.24	-0.57	56.00	98.78	107.92	5	0.964	9	0.990
	-5	26	5.00	15.71	26.00	25	16	33	-95.50	-80.61	-49.03	-0.52	60.00	98.73	107.95	7	0.959	16	0.980
	-15	26	3.00	8.03	17.00	13	22	33	-95.04	-82.84	-42.57	-0.40	48.85	96.57	107.91	10	0.917	17	0.764
	-25	26	0.00	0.31	3.00	1	40	60	-94.00	-90.24	-85.78	0.42	57.00	84.84	104.39	-	-	-	-
Arena	0	26	51.00	57.61	59.00	89	1	2	-105.00	-69.67	-39.02	-0.79	58.00	105.48	107.99	4	0.976	5	0.991
	-5	26	38.00	51.91	59.00	80	11	13	-95.00	-75.76	-47.36	-0.76	52.00	102.66	107.90	4	0.960	9	0.990
	-15	26	18.00	35.81	52.00	55	34	25	-95.00	-81.31	-54.34	-0.65	52.00	98.71	107.84	5	0.919	18	0.921
	-25	26	0.00	1.06	3.00	2	36	58	-93.13	-88.41	-82.05	-0.18	54.00	85.09	104.84	-	-	-	-

In order to calculate the broadcast/convergecast network diameter, we first calculate all possible Dijkstra trees taking all nodes as a source, and for each of those trees, we calculate the maximum path length. Then, we define the network diameter as the longest path out of all maximum path lengths. The point-to-point network diameter is defined as the path with maximum length out of all possible paths between nodes in the minimum spanning tree (MST). Network delivery is the expected network delivery (END) calculated as proposed by Puccinelli et al. in [11].

V. Conclusions and Outlook

In this paper we discussed the first steps towards the definition of a systematic methodology to both describe and select WSN testbeds. We introduced the first sketch of an approach that allows selecting testbeds based on a set of relevant properties. Also, we describe a first catalog of properties that can be used to systematically describe a testbed and discuss how these properties can be computed in real scenarios. We show how this catalog can be used to characterize different experimental sites by using data collected in two different testbeds.

Future work includes: the extension of the set of properties discussed in this paper; the extension and refinement of the testbed selection algorithm; the use of experimental data from further testbeds to validate the effectiveness of our description and selection methodology and the development of a web front-end for researchers that facilitates the specification of experiments requirements and which uses the site repository to identify the experimentation set.

Acknowledgments

This work has been partially supported by: the LOEWE Priority Program Cocoon (www.cocoon.tu-darmstadt.de) funded by the State of Hesse, Germany; the Research Training Group "Cooperative, Adaptive and Responsive Monitoring in Mixed Mode Environments" (GRK 1362, www.gkmm.de) funded by the German Research Foundation (DFG); and the Collaborative Research Center MAKI (Sonderforschungsbereich 1053, www.maki.tu-darmstadt.de), also funded by the DFG.

References

[1] G. Werner-Allen, P. Swieskowski, and M. Welsh, "MoteLab: a Wireless Sensor Network Testbed," in *Proceedings of the 4th International Symposium on Information Processing in Sensor Networks (IPSN '05)*, April 2005.

[2] V. Handziski, A. Köpke, A. Willig, and A. Wolisz, "TWIST: A Scalable and Reconfigurable Testbed for Wireless Indoor Experiments with Sensor Networks," in *Proceedings of the 2nd International Workshop on Multi-hop Ad Hoc Networks: From Theory to Reality (REALMAN '06)*, May 2006.

[3] P. E. Guerrero, A. P. Buchmann, A. Khelil, and K. Van Laerhoven, "TUDμNet, a Metropolitan-Scale Federation of Wireless Sensor Network Testbeds," in *Adjunct Proceedings of the 9th European Conference on Wireless Sensor Networks (EWSN '12)*, February 2012.

[4] K. Langendoen, "Apples, Oranges, and Testbeds," in *Proceedings of the 3rd International Conference on Mobile Adhoc and Sensor Systems (MASS '06)*, October 2006.

[5] L. Mottola, G. P. Picco, M. Ceriotti, S. Gună, and A. L. Murphy, "Not All Wireless Sensor Networks are Created Equal: A Comparative Study on Tunnels," *ACM Transactions on Sensor Networks*, vol. 7, no. 2, pp. 15:1–15:33, September 2010.

[6] F. Ferrari, M. Zimmerling, L. Thiele, and O. Saukh, "Efficient Network Flooding and Time Synchronization with Glossy," in *Proceedings of the 10th International Conference on Information Processing in Sensor Networks (IPSN '11)*, April 2011.

[7] O. Gnawali, R. Fonseca, K. Jamieson, D. Moss, and P. Levis, "Collection Tree Protocol," in *Proceedings of the 7th ACM International Conference on Embedded Networked Sensor Systems (SenSys '09)*, November 2009.

[8] M. Doddavenkatappa, M. C. Chan, and A. L. Ananda, "Indriya: A Low-Cost, 3D Wireless Sensor Network Testbed," in *Proceedings of the 7th ICST International Conference on Testbeds and Research Infrastructure for the Development of Networks and Communities (TridentCom '11)*, April 2011.

[9] A. Gluhak, S. Krco, M. Nati, D. Pfisterer, N. Mitton, and T. Razafindralambo, "A Survey on Facilities for Experimental Internet of Things Research," *IEEE Communications Magazine*, vol. 49, no. 11, pp. 58–67, November 2011.

[10] A. Cerpa, N. Busek, and D. Estrin, "SCALE: A Tool for Simple Connectivity Assessment in Lossy Environments," University of California, Los Angeles, Tech. Rep. TR-0021, September 2003. [Online]. Available: http://andes.ucmerced.edu/papers/Cerpa03a.pdf

[11] D. Puccinelli, O. Gnawali, S. Yoon, S. Santini, U. Colesanti, S. Giordano, and L. Guibas, "The Impact of Network Topology on Collection Performance," in *Proceedings of the 8th European Conference on Wireless Sensor Networks (EWSN '11)*, February 2011.

[12] J. Polastre, R. Szewczyk, and D. Culler, "Telos: Enabling Ultra-low Power Wireless Research," in *Proceedings of the 4th International Symposium on Information Processing in Sensor Networks (IPSN '05)*, April 2005.

[13] K. Srinivasan, P. Dutta, A. Tavakoli, and P. Levis, "An Empirical Study of Low-Power Wireless," *ACM Transactions on Sensor Networks*, vol. 6, no. 2, pp. 16:1–16:49, March 2010.

[14] J. Zhao and R. Govindan, "Understanding Packet Delivery Performance in Dense Wireless Sensor Networks," in *Proceedings of the 1st ACM International Conference on Embedded Networked Sensor Systems (Sensys '03)*, November 2003.

[15] De Couto, Douglas S. J. and Aguayo, Daniel and Chambers, Benjamin A. and Morris, Robert, "Performance of Multihop Wireless Networks: Shortest Path is Not Enough," *SIGCOMM Computer Communication Review*, vol. 33, no. 1, pp. 83–88, January 2003.