This full text paper was peer reviewed at the direction of IEEE Communications Society subject matter experts for publication in the ICC 2008 proceedings.

Robust Distributed Localization with Data Inference for Wireless Sensor Networks

Yu Zhang, Lin Zhang and Xiuming Shan
Tsinghua National Laboratory for Information Science and Technology
Department of Electronic Engineering, Tsinghua University, Beijing 100084 China
zhang-y-02@mails.tsinghua.edu.cn, {linzhang, shanxm}@tsinghua.edu.cn

***Abstract*—We consider the range-based localization with highly insufficient inter-node distance measurements in Wireless Sensor Networks (WSNs). With an error model established, a novel distributed probabilistic localization approach which exploits the network constraints through statistical inference on missing distances is proposed. The advantage of the new approach over traditional methods is proved analytically and further confirmed by extensive simulations.**

I. Introduction

Localization has become a hot topic amongst researches on Wireless Sensor Networks (WSNs), as most applications of WSNs could benefit from the awareness about position information of sensor nodes [1].

Until now, remarkable progress has been achieved in two major classes of localization algorithms.

- **Range-based** techniques directly implement the pairwise distance measurements such as radio signal [2], acoustic energy [3], hop count [4], etc. to compute node positions.
- **Range-free** techniques work with network constraints instead of direct distance measurements to infer inter-node proximities [5], [6].

As improvements of Yi Shang's work of MDS-MAP [7] in 2003, Costa's Distributed Weighted MDS (DWMDS) [8] and Patwari's Maximum Likelihood Estimation (MLE) technique [9] take various iterative approaches to map estimated Euclidean distances to available distance measurements as close as possible. Notwithstanding their different metrics of "closeness", MDS and MLE are both categorized as "case-delete" methods in this paper, because both methods follow the heuristic of discarding all missing pairwise distances whose measurements are unavailable in their calculation.

However, in realistic environment, cases of network sparsity, large obstacles, and irregular topologies would make the collection of all distance measurements infeasible; there would always be some empty entries in D_{mes}, the measurement of distance matrix D. In case of highly insufficient measurements, performances of case-delete methods degrade dramatically.

In this paper, we propose a hybrid localization method called **Inference-Assisting MLE (IA-MLE)**, with substantially enhanced performance to counteract measurement insufficiency. Under a predefined measurement model, we statistically infer the missing distances based on available distance measurements and other network constraints. Using D_{mes} and those inferences together, the localization performance is expected to be improved.

This research is supported in part by the National Science Foundation of China (Grant No.60672107), the Hi-tech Research and Development Program of China (Grant No.2006AA10Z261), and China 973 Project (Grant No.2007CB307105).

Taking the traditional **Case-delete MLE (CMLE)** in [9] as a benchmark, it can be proved, both analytically and experimentally, that our novel IA-MLE has better immunity to measurement insufficiency. Specifically, our method has comparable performance in dense networks and significant improvement in sparse networks with highly insufficient measurements.

The rest of this paper is organized as follows. In Section II, IA-MLE is proposed and theoretically analyzed, with the measurement model established in advance; a distributed scheme of IA-MLE is also considered here. Then through the extensive experimental evaluation of Section III, IA-MLE and CMLE are comprehensively compared on various uniform and irregular networks. Finally, Section IV concludes the paper and proposes some possible future work.

II. Localization with Data Inference

A. Problem Statement

In a typical sensor network, $M = p + q$ sensor nodes are deployed in an r-dimensional area with known boundary. The first p nodes $\{n_i\}_{i=1}^{p}$ are anchors with their actual positions $\{\mathbf{x}_i\}_{i=1}^{p}$ perfectly known through accurate positioning techniques such as GPS, while the rest q nodes $\{n_i\}_{i=p+1}^{p+q}$ are unknown nodes which have limited knowledge about their own positions $\{\mathbf{x}_i\}_{i=p+1}^{p+q}$.

Such a network could be represented by a simple graph $G_{r,M}(X, E, P)$, with vertice X representing true positions of sensor nodes, edges E representing the communication links between node pairs, and weights P, which are the distance measurements of the edges.

In later illustration, P is represented by D_{mes}. The distance matrix D is thus made up of D_{mes} and the set of missing distances D_{mis}. E is also translated into an $M \times M$ matrix W, in which

$$w_{ij} = \begin{cases} 1, & n_i \text{ and } n_j \text{ can communicate} \quad (1) \\ 0, & n_i \text{ and } n_j \text{ can't communicate} \quad (1') \end{cases}$$

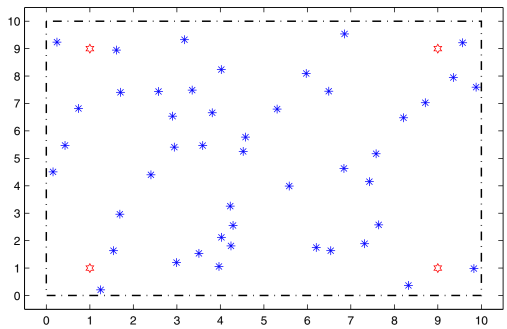

Fig. 1. 50 nodes randomly deployed in a $10m \times 10m$ square area, with 4 anchors

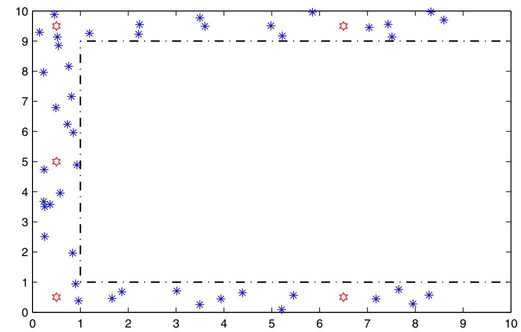

Fig. 2. Example of C-Shape Corridor

In WSNs, it is natural to assume that $w_{ij} = 1 \Leftrightarrow d_{ij} \in D_{mes}$.

With predefined metric of "closeness", the traditional case-delete methods aim for an optimal topology estimate X to minimize the difference between D_{mes} and its corresponding part in Euclidean distance matrix $D_{euc} = \{\delta_{ij}\}$, where $\delta_{ij} = \sqrt{\sum_{k=1}^{r}(x_{ik} - x_{jk})^2}$. Our effort then, as mentioned in the above section, is to infer D_{mis} first, then minimize the difference between $\widetilde{D} = D_{mes} \bigcup D_{mis}$ and D_{euc}.

B. System Model

We use RSSI (Radio Signal Strength Indicator) to measure inter-node distances, which is a cheap and widely used technique suitable for various sensor networks. To avoid obscurity, it is assumed that only line-of-sight communication is allowed, when two nodes are obstructed by radio-impenetrable obstacles thus cannot see each other, they could not directly communicate.

Based on these assumptions, the classical radio signal propagation model which Rapport proposed in [10] is used. In this model, the average signal power decreases logarithmically against transmission distance, added with shadow effect noise. After simple manipulation, it is easy to find that every pairwise distance measurement d_{ij} complies with a log-normal distribution around the true Euclidean distance δ_{ij}, with a probability distribution function (PDF) as

$$f(d_{ij} \mid \delta_{ij}(X)) = \frac{1}{\sqrt{2\pi}\sigma_d d_{ij}} e^{-\frac{(\log d_{ij} - \log \delta_{ij})^2}{2\sigma_d^2}} \tag{2}$$

$$\sigma_d = \frac{\sigma_P}{10n}, \tag{3}$$

where σ_P represents the influence of shadowing effect, and n is the path loss exponent. Both of them are assumed constant throughout the network. When σ_d^2 goes to 0, d_{ij} converges to δ_{ij} with probability 1, representing the case of no shadowing noise and thus no measurement error.

Assuming the independencies between different pairwise distance measurements, the loglikelihood generated from a specific measurement could be expressed as

$$l_{ij} = \ln f(d_{ij} \mid \delta_{ij}(X)) \tag{4}$$

C. Constraints in Sensor Network

Based on the conclusion in [11], the missing mechanism of any data could be categorized into MAR (Missing At Random) or MNAR (Missing Not At Random). In WSNs, the missing distances are not MAR, but related to their actual values. For instance, assuming P_{th} denotes the threshold power reflecting sensor nodes' radio sensitivity. For a node pair n_i and n_j, when received power P_{ij} at n_j is below P_{th}, the signal could not be correctly demodulated; hence d_{ij} is unable to be estimated from P_{ij} and thus missing from RSSI measurement. Through the missing mechanism, although d_{ij}'s accurate value could not be obtained, some side information indicating the cause of missing could still be found. For the above example, the loss of measurement imposes a convincing determinant upon d_{ij}, indicating it being larger than d_{th}, the threshold distance corresponding to P_{th}. In the following paper, such side information is denoted as *constraint* in the network.

With an MNAR pattern of missing distances D_{mis}, it is not preferable to impute random guess on them simply based on their a priori distributions $\{f(d_{ij})\}$. Given that various constraints in the network are available, they could be expected to provide significant enhancement to the localization performance, if used smartly and adaptively, especially when the "Prior Knowledge" of the network such as D_{mes} and position-known anchors are insufficient.

In this paper, the *Shortest Path* constraints are employed.

If the *Shortest Path First* algorithm was employed in the sensor network, for node pair (n_i, n_j) who could not directly communicate, their shortest path $R(n_i, n_j)$ via other nodes could be obtained. Let n_k denote the precedent of n_j on $R(i, j)$, and the lengths of $R(i, k)$ and $R(i, j)$ are represented by $d(i, k)$ and $d(i, j)$ respectively; it is justified to assume d_{ij} lying between $d(i, k)$ and $d(i, j)$ with high probability. Then the PDF of d_{ij} could be updated as

$$f^{new}(d_{ij} \mid \delta_{ij}(X)) = \frac{f^{old}(d_{ij} \mid \delta_{ij}(X))}{\int_{d(i,k)}^{d(i,j)} f^{old}(d_{ij} \mid \delta_{ij}(X))} \tag{5}$$

D. Proposed Approach

Based on the network constraints, our novel IA-MLE approach employs the idea of *Expectation-Maximization* (EM) raised by Dempster in 1978 [11] to statistically infer the missing distances, then constructs a "complete" loglikelihood before localizing the network.

Here we use a matrix C to denote the set of constraints utilized in our method, in which every entry c_{ij} denotes the corresponding constraint on d_{ij}. As mentioned in previous section, Shortest Path constraints are engaged here, then c_{ij} is simplified into a "0/1"-valued binary variable.

IA-MLE handles missing distances in an iterative fashion, consisting of four steps:

1) Utilize conventional methods such as MDS-MAP to derive an initial topology estimate $X^{(0)}$ of the network, then compute shortest paths between all node pairs. D_{mes} and C are also initialized.
2) The "Inference step". In the t^{th} iteration, the missing distances are inferred with the most up-to-date topology estimate $X^{(t-1)}$. Instead of filling a specific value for every missing d_{ij}, IA-MLE infers the update of d_{ij}'s conditional PDF, that is,
$$f^{(t)}(d_{ij} \mid D_{mes}, C, X^{(t-1)}) = \frac{r^{(t-1)}(d_{ij})}{\int_{d(i,k)}^{d(i,j)} r^{(t-1)}(d_{ij})} \quad (6)$$
$$r^{(t-1)}(d_{ij}) = \frac{1}{\sqrt{2\pi}\sigma_d d_{ij}} e^{-\frac{\left[\log d_{ij} - \log \delta_{ij}\left(X^{(t-1)}\right)\right]^2}{2\sigma_d^2}} \quad (7)$$
The updated conditinal PDF of the missing distance set D_{mis} then is
$$\begin{aligned} & f^{(t)}(D_{mis} \mid D_{mes}, C, X^{(t-1)}) \\ &= \prod_{mis} f^{(t)}(d_{ij} \mid D_{mes}, c_{ij}, X^{(t-1)}) \end{aligned} \quad (8)$$
PDF (6) and (8) are denoted as $f^{(t)}(d_{ij})$ and $f^{(t)}(D_{mis})$ respectively for short.
This step discovers the key idea of EM, that is, the inference on any missing d_{ij} to appear in the "complete" loglikelihood is not a specific value but a function of d_{ij}.
3) The "Expectation" step builds up the "complete" loglikelihood. Apparently, if missing distances D_{mis} were filled with specific values in the "Inference" step, then the "complete" loglikelihood should have the form as $L(X \mid D_{mes}, D_{mis}, C)$ where D_{mis} is fixed. Now with only D_{mis}'s PDF update available, what we obtain in this step is in fact the expected "complete" loglikelihood, with the randomness of D_{mis} integrated out, that is,
$$\begin{aligned} & Q(X \mid X^{(t-1)}) \\ &= \int L(X \mid D_{mes}, D_{mis}, C) f^{(t)}(D_{mis}) dD_{mis}, \end{aligned} \quad (9)$$
where
$$\begin{aligned} & L(X \mid D_{mes}, D_{mis}, C) \\ &= \ln f(D_{mes}, D_{mis}, C \mid X) \end{aligned} \quad (10)$$
$$\begin{aligned} & f(D_{mes}, D_{mis}, C \mid X) \\ &= \Big[\prod_{mes} f(d_{ij}, c_{ij} \mid X)\Big]\Big[\prod_{mis} f(d_{ij}, c_{ij} \mid X)\Big] \\ &= \Big[\prod_{mes} f(d_{ij} \mid X)\Big]\Big[\prod_{mis} f(d_{ij} \mid X) f(c_{ij} \mid d_{ij})\Big] \end{aligned} \quad (11)$$
$$f(c_{ij} \mid d_{ij}) = \begin{cases} 1, & \text{Condition A,} \quad (12) \\ 0, & \text{Otherwise.} \quad (12') \end{cases}$$
where Condition A represents the case when $d_{ij} \in [d(i,k), d(i,j)]$ and $c_{ij} = 1$.
4) Finally, the "Maximization" step finds the optimal topology estimate update $X^{(t)}$, which maximizes the above expected "complete" loglikelihood Q
$$X^{(t)} = \max_X Q(X \mid X^{(t-1)}) \quad (13)$$
If $\mid Q^{(t)} - Q^{(t-1)} \mid > \epsilon$, increase t by 1, go back to step 2; otherwise, iteration stops, $X^{(t)}$ is the final localization result of the whole network.

E. Analysis of IA-MLE

Here we take some close scrutiny over IA-MLE to theoretically analyze its performance.

From an information aspect of view, the procedure of localization is to reduce the positioning uncertainty based on network observations. Given that D_{mis} are unmeasurable, there're only the distance measurements D_{mes} and available network constraints C which could serve as the information source. Therefore, it is evident that if the loglikelihood $L(X \mid D_{mes}, C)$ is maximized, the most amount of information could be extracted and implemented. In contrast, traditional CMLE only employs D_{mes} with loglikelihood $L(X \mid D_{mes})$ maximized in its calculation, information in C is not utilized at all.

In IA-MLE. Based on equation (9), the expected "complete" loglikelihood Q could be decomposed as
$$Q(X \mid X^{(t-1)}) = L(X \mid D_{mes}, C) + H(X \mid X^{(t-1)}) \quad (14)$$
$$\begin{aligned} & H(X \mid X^{(t-1)}) \\ &= \int \ln f(D_{mis} \mid D_{mes}, X, C) f^{(t)}(D_{mis}) dD_{mis} \end{aligned} \quad (15)$$
An interesting relationship lies in equation (15). If we take $Q(X \mid X^{(t-1)}))$ as complete information and $L(X \mid D_{mes}, C)$ as observed information, then $H(X \mid X^{(t-1)})$ represents the missing information due to the unmeasurability of D_{mis}.

Dempster has already proved in [11] that $L(X \mid D_{mes}, C)$ increases monotonically in each iteration of maximizing $Q^{(t)}$, that is,
$$L(X^{(t)}|D_{mes}, C) \geq L(X^{(t-1)}|D_{mes}, C) \quad (16)$$
with
$$X^{(t)} = \max_X Q(X \mid X^{(t-1)}) \quad (17)$$

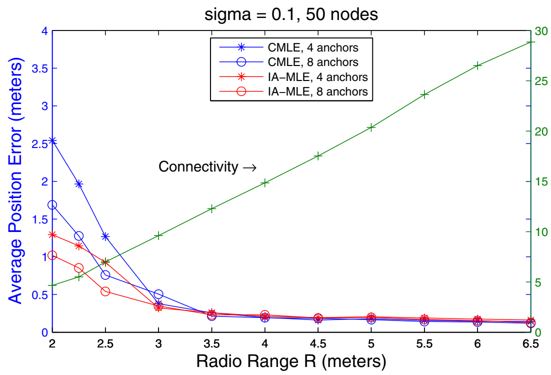

Fig. 3. Average position error and Connectivity of CMLE and IA-MLE on square topology with 50 nodes randomly deployed, $\sigma_d^2 = 0.1$

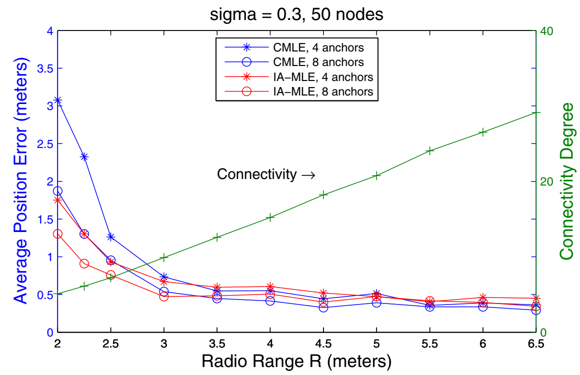

Fig. 5. Average position error and Connectivity of CMLE and IA-MLE on square topology with 50 nodes randomly deployed, $\sigma_d^2 = 0.3$

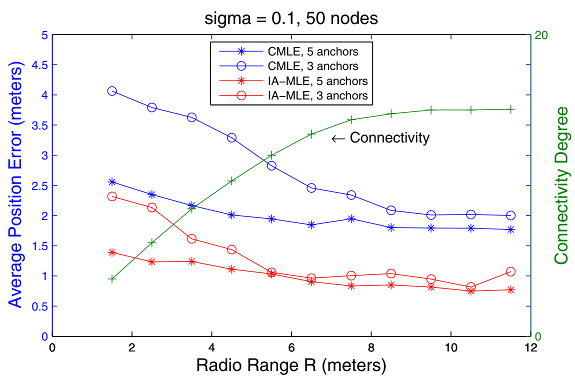

Fig. 4. Average position error and Connectivity of CMLE and IA-MLE on C-Shape Corridor with 50 nodes randomly deployed, $\sigma_d^2 = 0.1$

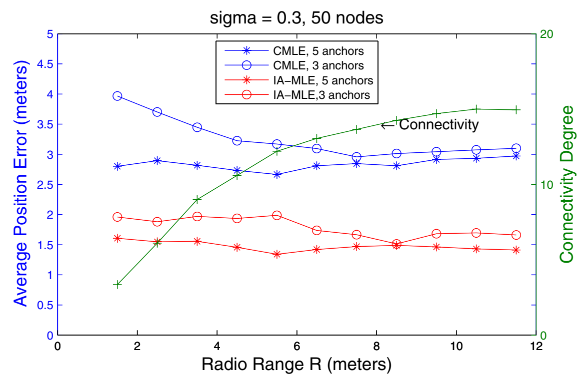

Fig. 6. Average position error and Connectivity of CMLE and IA-MLE on C-Shape Corridor with 50 nodes randomly deployed, $\sigma_d^2 = 0.3$

Therefore, although $L(X \mid D_{mes}, C)$ is not directly maximized in one step, it could be monotonically increased in IA-MLE to finally reach the global maximum.

Now the purpose of IA-MLE is getting clear. Through the missing distance inferences, the network constraints C are delicately implemented in localization. With more information included, it is straightforward that IA-MLE could achieve a lower Cramér-Rao Bound than CMLE.

III. Experimental Evaluation

We evaluated the performance of IA-MLE and CMLE in Matlab R2007 with a $3.00GHz$ Pentium(R) 4 processor. To obtain an all-sided view of IA-MLE's performance, 2 typical topologies were used: (a) A $10m \times 10m$ square as Fig.1, representing the uncluttered environment; (b) A C-Shaped corridor composed of 3 10-meter long, 1-meter wide branches as Fig.2, representing indoor networks of irregular and severely cluttered topology. The signal transmission is assumed to comply with the model established in Section II-B having a path loss exponent of 1.7.

A. Square Topology

50 nodes are randomly deployed in the square area. The network connectivity is controlled by sensor node's radio range R. Apparently, with a constant network density, the amount of available distance measurements would rise monotonically with R.

Fig.3 and Fig.5 exploit the performances of both methods with noise variance $\sigma_d^2 = 0.1$ and 0.3 respectively. The average position errors and corresponding connectivity are plotted against R. Each data point represents the average result of 30 random trials, in which two schemes using 4 and 8 random-selected anchors were considered, respectively.

Under both noise conditions, the results seem consistent. When $R > 3.5m$, D_{mes} is dense enough to provide sufficient information for localization, leading to close performances of CMLE and IA-MLE, while further increase on R would fetch low benefit.

When $R < 3m$, less inter-node connections leads to a more sparse network. As expected, performances of both methods degrade, yet the pace of CMLE is far more rapid. This phenomenon solidly demonstrates the advantage of IA-MLE in sparse network. When the sparsity is high, D_{mes} is unable to provide sufficient information for accurate localization. In this situation, IA-MLE could import information contained in the network constraints C to slow down the deterioration of localization performance. As sparsity grows, the amount of available measurements decreases, information from C turns to be more significant, leading to a monotonically broadened gap between performances of both methods.

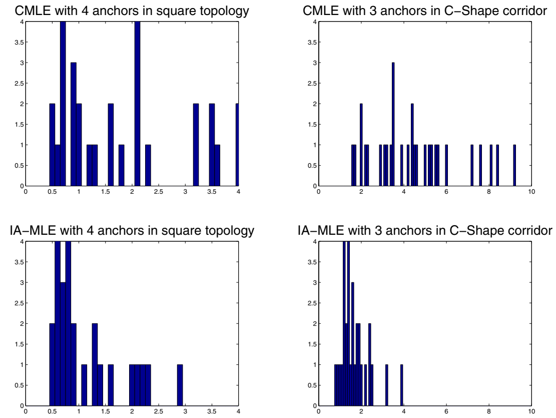

Fig. 7. Distribution of magnitude of average position error for CMLE and IA-MLE, with $\sigma_d^2 = 0.1$; in square topology, 4 anchors are selected; in C-Shape Corridor, 3 anchors are selected

B. C Topology

In C topology, the sparsity of the network is not the only reason leading to measurement insufficiency. We also randomly deploy 50 nodes in the C-shape corridor. Compared to square topology, such C-shape network would bring in 2 extra problems. The first is that small width of the corridor would constrain the distribution of sensor nodes in the same branch towards a line, therefore pairwise connections could not bring in so much rigidity for the whole topology as in square network. The second problem arises from topology irregularity. With only line-of-sight communication allowed, the irregularity renders some natural difficulty for nodes to mutually communicate, especially those lying in different branches. This sets an upper bound on the connectivity of the network, as Fig.4 and Fig.6 depict.

Fig.4 and Fig.6 show the performance of both methods in C-Shape corridor. CMLE seems don't benefit much from R's increase. This is because no matter how large R is, most available distance measurements are between nodes in the same branch, which could not render sufficient rigidity for the topology. On the contrary, because the measurement noise is proportional to the measurement value, when R increases, larger noises are then included in D_{mes} with more long distances measured, so the average position error of CMLE sometimes even rises.

In C-Shaped topology, the advantage of IA-MLE is even more remarkable. Compared to CMLE, a 50% reduction on average position error could be achieved by employing IA-MLE. Meanwhile, as R increases, the gain would not fall. This result indicates the significance of IA-MLE's inferences on crossing-branch distances. These inferences, although not accurate enough, are just crucial in the topology recovery, because they could reveal some fundamental knowledge about the spatial relationship between nodes in separate branches.

C. The Consistency of Localization Results

In sensor network localization, the consistency of localization results sometimes is even more critical than a low average position error. If the result of a localization method varies drastically from trial to trial, this method is thereby not reliable.

With 30 trials conducted at specific data point, Fig.7 shows the histogram of magnitude of average position error of both methods. In square topology, the trials are conducted at $R = 2.25m$; in C-Shape corridor, $R = 3.5m$. The result of IA-MLE has consistently lower standard error and varies less than CMLE. This, to some extent, shows that the localization result from IA-MLE would have a comparatively tighter confidence region and a lower estimation variance, to be more consistent and promising than CMLE.

IV. Conclusion

In this work, we take a first step in analyzing the localization problem where a large set of distance measurements are unavailable. Unlike a lot of preceding localization algorithms where measurement sufficiency is always guaranteed, our work brought in more realistic considerations, making itself suitable for networks under severe conditions. Empirical evaluation show that our method delivers a constantly comparable performance with sufficient measurements, and a significant improvement under measurement insufficiency. Theoretical analysis also proves that if network constraints could be appropriately used, our method is optimal, achieving a lowest Cramér-Rao Bound. Clearly, many important questions still remains open in this direction. For example, the model of the constraints could be further refined so as to extract more information from available constraints.

References

[1] Ian F. Akyildiz, Weilian Su, Yogesh Sankarasubramaniam, and Erdal Cayirci, "A Survey on Sensor Networks", *IEEE Communication Magazine*, vol. 40, no. 8, Aug.2002.

[2] Paramvir Bahl and Venkata N Padmanah han, "RADAR: An in-building RF-based user location and tracking system", in *Proc. of lnfocom*, vol. 2, Mar.2000.

[3] Chris Savarese, Jan M Rabaey, Jan Beutel, "Location in Distributed Ad-Hoc wireless sensor network", in *Proc. of IEEE International Conference on AcousticsSpeech, and Signal*, vol. 4, May.2001.

[4] Nirupama Bulusu, John Heidemann, Deborah Estrin, "GPS-less Low Cost Outdoor Localization For Very Small Devices", *IEEE Personal Communications Magazine, Special Issue on Smart Spaces and Environments*, vol.7, Oct.2000.

[5] Lance Doherty, Laurent E1 Ghaoui, Kristofer S J Pister, "Convex Position Estimation in Wireless Sensor Networks", in *Proc. of Twentieth Annual Joint Conference of the IEEE Computer and Communications Societies*, vol. 3, Apr.2001.

[6] Tian He, Chengdu Huan, Brian M Blum et a1, "Range-Free Localization Schemes in Large Scale Sensor Networks", ACM Press, 2003.

[7] Y Shang, W. Ruml, K. Zhang, and M. Fromherz.2003. "Localization from mere connectivity", in *Proc. of the 4th ACM international symposium on Mobile ad hoc networking and computing*, Jun.2003.

[8] J. A. Costa, N. Patwari, and A. O. Hero III, "Distributed multidimensional scaling with adaptive weighting for node localization in sensor networks", *IEEE/ACM Trans. Sensor Networks*, 2005.

[9] N. Patwari, A. O. H. III, M. Perkins, N. S. Correal, and R. J. ODea, "Relative location estimation in wireless sensor networks", *IEEE Transactions on Signal Processing, Special Issue on Signal Processing in Networks*, vol. 51, Aug.2003.

[10] T. S. Rappaport, "Wireless Communications: Principles and Practice", 2nd ed. Pearson Education, 2001.

[11] Dempster, A. P., Laird, N. M., and Rubin, D. B "Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion)", *Journal of the Royal Statistical Society, Series B*, vol. 39, 1978.