Chapter 2
Stastical Techniques and Location Discovery

In this chapter, we survey the relevant literature of the related fields along the following line: Location Discovery in wireless ad hoc sensor networks, statistical techniques including error modeling and regression, linear, and nonlinear function optimization.

2.1 Statistical Techniques

Least linear-squares fitting/regression is one of the most popular parametric techniques for fitting a set of data to a particular compressed form. The technique is optimal under the assumption that the error distribution follows the Gaussian distribution. The standard technique for conducting least squares fitting is to use the singular value decomposition (SVD). In addition, there are various available techniques that can be applied to determine whether the least squares model is appropriate for the given set of data. Belsley et al. [1], Chatterjee and Price [2], and Good [3] provide the detailed description of these methods. In addition to parametric techniques, nonparametric techniques that do not depend on any underlying assumptions have been gaining popularity. Some standard references of nonparametric regression techniques include [4–7].

Linear programming (LP) and nonlinear programming have been the popular optimization mechanisms since the 1940s. Both of these methods involve three entities: variables, an OF, and constraints. The objective is to find a set of assignments to the variables in such a way that the OF is minimized or maximized, and at the same time have all the constraints remain satisfied. Standard references for linear programming include [8–10]. For nonlinear programming, some useful sources include [11–13]. In our research, we chose to use unconstrained nonlinear function minimization techniques. More specifically, we used the public available software package WNLIB [14]. The NLP problem instance was solved by

J. F. Sanford et al., *Localization in Wireless Networks*,
DOI: 10.1007/978-1-4614-1839-9_2,
© Springer Science+Business Media New York 2012

weighting all constraints into the OF and relying on the conjugate direction-based unconstrained nonlinear function minimization techniques provided by WNLIB.

There exist a wide variety of widely used parametric and nonparametric regression techniques [15]. Isotonic (monotone) regression is a prediction technique that for a set of real data values (x_i, y_i, w_i) is the set (x_i, z_i) that minimizes a specific norm s.t. $z_1 \leq z_2 \leq \ldots \leq z_n$. The class of unimodular regressions is closely related to isotonic regression, where the condition has form $z_1 \leq z_2 \leq \ldots \leq z_k \geq z_1 \geq z_n$. Both isotonic and unimodular regressions belong to a family of nonparametric techniques with restricted properties [16], and are often used when assumption about being monotonic is easily justifiable but there is no a priori parametric knowledge [17–19]. A number of algorithms for both polynomial and exponential complexity have been proposed for monotonic and unimodular regressions [16, 18, 20]. All of them exploit the pair adjacent violators (PAV) principle [21] in continuous domains to construct a set of piecewise horizontal monotonic or unimodular segments. PAV starts by sorting points according to their x coordinates and allocating horizontal segment to each node. It replaces each pair of adjacent segments that violate the constraint of being monotonic by a new single horizontal segment that is optimal with respect to the selected L_p norm and removes the violation.

More recently, several techniques for approximating smooth monotonic and unimodular regression have been developed [22, 23]. Pairwise consistency is the property that for a pair of variables (x_i, y_i) and (x_j, y_j), a larger x component implies a larger (or in less strict definition—not smaller) y component. Pairwise consistency among all pairs of variables strongly implies that monotonic regression is bound to be accurate. It is often used as the requirement for correct decision making with respect to selecting better y options in presence of knowledge only about x components. Pairwise consistency should not be confused with consistency, a term that is used in statistics to indicate that a considered statistical technique guarantees that in presence of asymptotically large data set, it will produce actual distribution of the data. To the best of our knowledge, no previous work has been done on constructing measurement error models based on monotonic regression that follows the pairwise consistency.

2.2 Location Discovery

Sensor networks and pervasive computing systems form one of the fastest growing computer and networking research frontiers. Once the nodes that form a network or an infrastructure are deployed, invariably there is a need that each node discovers its position. Global positioning system (GPS) greatly facilitates this task. However, because of obstacles such as trees and walls, GPS system often does not lock to satellite signals. At the same time, GPS systems are relatively expensive and consume significant amount of energy. Therefore, usually only a limited subset of nodes is equipped with GPS and other nodes deduce their locations by measuring distances between themselves.

2.2.1 Definition

The *Location Discovery (LD or localization)* problem is defined in the following way. A total of N nodes, K of which ($K \ll N$) have exact information about their positions. In addition, the measured distances, which are subject to errors, between M pairs of nodes are also available. The goal is to conclude the location (x_i, y_i) of each unknown location node i in such a way that $L(x_{ri} - x_i, y_{ri} - y_i)$ is minimized, where (x_{ri}, y_{ri}) is the real location of i. Usually, the targeted error norm L is L_1, L_2, or L_∞.

For the sake of completeness, we formally state the Location Discovery problem in the standard Garry–Johnson format:

The Location Discovery Problems

Instance: A network of N sensor nodes embedded in a k dimensional space, each has location $L_i = (x_{i1}, x_{i2}, \ldots, x_{ik})$, $i = 1, \ldots, N$, a subset of sensors P is aware of their locations $P = \{i, i \in N\}$, a subset of pairs of sensors has distance measurements $Q = \{d_{ji}, i, j \in N\}$, positive integer M.

Question: Is there a one-to-one function g: $i \rightarrow L_i' = \left(x_{i1}', x_{i2}', \ldots, x_{ik}'\right)$, where $\left(x_{i1}', x_{i2}', \ldots, x_{ik}'\right)$ is the estimated location of i, $i \notin P$, such that the overall discrepancy between the estimated locations and the real locations of nodes not in P satisfies the following condition: $\sum_{i \notin P} |L_i - L|_i' \leq M$?

Numerous problem formulations for Location Discovery that target different technologies for distance measurements, use different optimization mechanisms, and impose different sets of constraints and objectives have been proposed. These efforts formed strong foundations for addressing Location Discovery in sensor networks. They range from iterative linearization and convex programming to conjugate direction-based and multi-resolution search. In addition, a variety of distance measurement technologies have been employed, including signal strength attenuation techniques, ultra-wide-band approaches, Doppler-assisted methods, carrier-phase-based measurements, and acoustic signal-based techniques. The technologies significantly differ in terms of maximum and minimum measuring range, resilience toward obstacles, power consumption, cost of deployment, and required power budget. Nevertheless, they share a common denominator: Distance measurements are prone to both small fluctuation and occasional large errors.

2.2.2 Classification

There are many possible ways to classify the Location Discovery algorithms, such as based on (1) how the algorithms is executed—*centralized* versus *Localized,* (2) what environment is the algorithm targeting—*indoor* versus *Outdoor,* (3) network topology—*static* versus *dynamic* networks, (4) available resources—*GPS-based* versus *GPS-less,* etc. Centralized algorithms are executed at a single

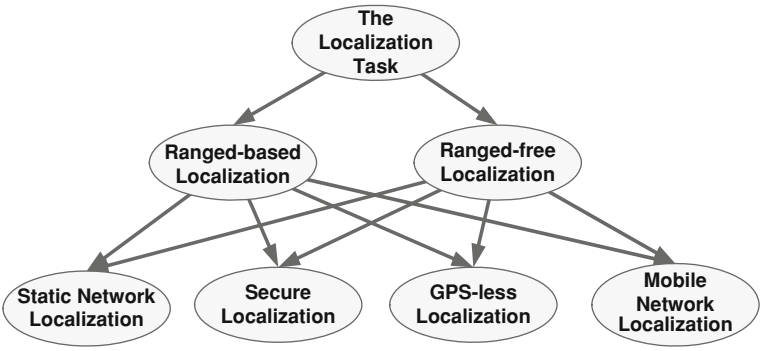


Fig. 2.1 Taxonomy of localization algorithms

place with the availability of the complete information about all measurements, and localized algorithms are executed by multiple nodes simultaneously and/or consecutively where each node has limited information provided by its neighbors. Indoor Location Discovery is in general more robust than the outdoor environment because of the presence of physical obstacles and reflections. Some networks consist of *beacons/anchors* (the nodes that have GPS devices, therefore, are aware of their absolute positions, while some networks do not). In GPS-based Location Discovery, it is possible to realize the absolute positions of the unknown nodes, whereas the positions derived are relative positions in GPS-less networks. However, it is possible to derive the absolute locations from the relative locations given the ground truth of at least three nodes. In addition, there are also algorithms specifically tailored for the mobile networks.

In this section, we classify the Location Discovery algorithms based on the inputs of the algorithm, i.e., the available resources used to derive the locations. All the algorithms are categorized to two classes: *range-based* and *range-free* Location Discovery. In the range-based Location Discovery, a set of distance measurements between a subset of pairs of nodes is available. In addition, acoustic source Location Discovery can also provide an inexpensive and easily integrated solution to the more general sensor node localization problem, by using a source-localizing infrastructure to detect and locate small inexpensive nodes that emit a characteristic calibration signal. On the other hand, no such requirement is posed in the range-free Location Discovery. Within each class, we further differentiate the algorithms according to the four possible categorizations. Figure 2.1 shows the taxonomy of the algorithms in detail.

2.2.3 Ranging Technology

For range-based techniques, the accuracy of *range estimation*, the process of measuring the Euclidean distances between nodes, strongly affects the accuracy

of the derived locations. The range estimation mainly exploits four types of technologies: (1) radio frequency (RF), (2) infrared, (3) ultrasound, and (4) acoustic. There are four main ranging techniques developed based on these four different technologies: (1) angle of arrival (AoA), (2) time of arrival (ToA), (3) received signal strength (RSS), and (4) time difference of arrival (TDoA). Whitehouse et al. [24] give a top-down approach to evaluate various models of how empirical ranging data affect the multihop Location Discovery. The authors use the Noisy Disk Model, which is well known to model ultrasound and radio signal strength, and identifies where and how it deviates from real-world characteristics. In addition, Cerpa et al. [25] constructs nonparametric statistical models that establish the relationship between location (e.g., distance) and communication properties (e.g., reception rate).

Angle of arrival (AoA) is generally measured using a multiple element antenna array, followed by computation to extract the angle information. The multiple element array requires calibration, and the phase and amplitude response of each element and cables, switches, etc. must be known and stable for measurement accuracy. Alternately, a single antenna may be moved to synthesize an array. The angle resolution is limited by the size of the array and the limitations of the processing algorithm, and many implementations of this method are very slow.

Direct measurement of AoA is possible using a rotating directional antenna. Resolution depends on the antenna beam width, which in turn depends on antenna size in terms of wavelength, and antenna travel during the measurement. Antenna front to back ratio will also limit the dynamic range of the measurement. The rotating antenna method, combined with a wide-band channel sounder, has the attraction of being direct and reasonably fast.

The differences in the arrival times of a signal at spatially separated sensors can be used to estimate the angle of arrival of the signal. Because of noise or sensor malfunction, the time difference estimates may be erroneous and thus the angle of arrival estimate may be inaccurate. If the signal to noise ratio is high enough, these differences, i.e., time delays, can be estimated by simple methods, e.g., polarity coincidence correlation.

Time of arrival (ToA) requires precise time synchronization between the sender and the receiver. The sender transmits the signal along with a time stamp to indicate the time the signal is sent. The receiver can, therefore, calculate the traveling time and estimate the distance between by multiplying the traveling time with the speed of the signal in a particular medium. Given the requirement on the desired precision, acoustic signal is usually chosen because the speed of radio or light signal is so fast that a small timing error would lead to an unacceptably large ranging error [26]. But even if the relatively slower acoustic signal is chosen, the precision requirement on ToA estimation is still very stringent. For example, one millisecond error in ToA estimation will translate to more than 30 cm error in the ranging result.

Received signal strength (RSS) measures the power of the signal at the receiver. Based on the known transmit power, the effective propagation loss can be

Table 2.1 Ranging technologies and techniques

Technology	Technique			
	AoA	ToA	RSS	TDoA
Acoustic	X	X		X (w/RF)
Infrared		X		
RF (radio)	X	X	X	X (w/ultrasound or Acoustic)
Ultrasound		X		X (w/RF)

calculated. Theoretical and empirical models are used to translate this loss into a distance estimate.

Time difference of arrival (TDoA) relies on the difference between two signal receptions. Usually, one signal is assumed to arrive at the receiver instantaneously, which serves as the reference as when the signals are sent. On receiving the second type of signal, the time difference is calculated and the estimated distance can be derived as in the case of ToA. Table 2.1 shows the four ranging technologies, the four popular ranging techniques, and which techniques are usually used on which type of technologies.

Peng et al. [26] present an acoustic-based ranging system—BeepBeep, readily applicable to many low-cost sensor platforms and to commercial-off-the-shelf mobile devices such as cell phones and PDAs. It achieves high accuracy through a combination of three techniques: two-way sensing, self-recording, and sample counting. Each device emits an acoustic signal, the signal is detected both the emitting device and the other devices. Exchanging the time duration information with its peers, each device can derive the two-way time of flight of the beeps at the granularity of sound sampling rate. The algorithm is evaluated based on ranging accuracy, confidence, and operational range, based on both indoor and outdoor environments using off-the-shelf devices. In addition, the authors also provide lower bound estimation of the sending and receiving uncertainties.

In [27], the authors present ENSBox, acoustic-embedded networked sensing box, a platform for prototyping rapidly deployable distributed acoustic sensing systems. Each ENSBox integrates a developer-friendly ARM/Linux environment with key facilities required for source localization. One of the capabilities—self-calibration—is especially important to source localization applications. The authors report a 5 cm average 2D positioning error and 1.5 degree orientation error over a partially obstructed 80×50 m^2 area, which is an order of magnitude better than prior similar acoustic work.

2.2.4 Ranged-Based Location Discovery

Given the measured Euclidean distances of pairs of nodes, we survey some of the state-of-the-algorithms. Sheng and Hu [28] present a Location Discovery approach based on the acoustic energy decay model, where the acoustic energy decays

inverse of distance square under some mild conditions. This energy-based Location Discovery problem is then solved by combining maximum likelihood (ML) estimation with expected maximization (EM) solution and projection solution. In addition, they also derived the Cramer-Rao Bound (CRB) for sensor deployment analysis.

Niculescu and Nath [29] propose a Location Discovery approach that is based on the basic idea of distance vector routing using only a fraction of beacons, with the assumption that each sensor node has some combination of ability to measure range, angle of arrival (AOF), and orientation. They propose a lower bound for positioning error for a range/angle free algorithm and examine the error characteristics of various classes of multihop ad hoc positioning systems (APS) algorithms.

The Location Discovery method proposed by Galstyan et al. [30] is distributed and on-line, which means the Location Discovery process is simultaneously conducted with an application task. Sensor nodes use their geometric constraints induced by radio connectivity and sensing to decrease the uncertainty of positions. The performance of the algorithm is compared with the centralized (convex) programming.

In addition to static networks, Hu and Evens [31] introduce the sequential Monte Carlo Location Discovery method for mobile networks, which exploits mobility to improve the accuracy and precision of positioning. The approach does not require any additional hardware and has competitive results when compared with static Location Discovery methods.

Biswas and Ye [32] propose a semi-definite programming (SDP) relaxation-based Location Discovery method where the main idea is to convert the nonconvex quadratic distance constraints into linear constraints by introducing a relaxation to remove the quadratic term. The L_1 norm of location errors is served as the optimization target.

Nasipuri and Li [33] propose a Location Discovery method that a sensor node can determine its location by noting the times when it receives the different beacon signals and evaluating its angular bearings and location with respect to the beacon nodes using triangulation.

Shang and Zhang [34] present an algorithm that uses the basic connectivity information—which nodes are in the communication ranges of which others, to derive the locations of the unknown nodes.

As location-based services become more prevalent, the Location Discovery Infrastructure will become the target of malicious attacks. These attacks are specific to Location Discovery services and are different in nature than the traditional security systems. Li et al. [35] not only suggest a list of attacks that are unique to Location Discovery algorithms but also present robust statistical methods such as least median squares estimator and adaptive least squares to make Location Discovery attack tolerant on both triangulation and RF-based fingerprinting Location Discovery methods. Liu et al. [36] argue that location determination uses certain physical features, such as the received signal strength of radio signals, makes authentication not as effective as in the traditional security applications.

They propose two methods that (i) filter out or (ii) iteratively refine the malicious beacons signals to ensure specified location accuracy.

To the best of our knowledge, no comprehensive statistical studies on measurement errors or error modeling have been conducted. Two popular assumptions regarding the measurement/ranging errors include the Gaussian distribution [29] and the L_1 norm [32]. When Location Discovery is formed in terms of an optimization instance, maximizing the likelihood of Gaussian-based errors or minimizing the L_1 norm of location errors usually serves as the optimization target. For example, Niculescu and Nath [29] derive a Cramer-Rao lower bound for positioning error of multihop distance-vector-based algorithms based on the assumption of Gaussian error measurements. Savvides et al. [37] conduct comprehensive studies on the position error behavior of multihop Location Discovery protocols based on the assumption that the measurement errors are independent Gaussian random variables with zero mean and a known variance.

Recent development of new deployable acoustic sensor platforms presents opportunities to develop automated tools for bio-acoustic field research. Ali et al. [38] report an approximate maximum likelihood (AML)-based source localization algorithm and use it to localize marmot alarm-calls. The authors assess the performance of the techniques based on results from two field experiments.

Ash and Potter [39] propose an algorithm to self-localize and orient sensors in a network based on AoA information. The proposed noniterative subspace-based method can operate given missing and noisy measurements and work regardless whether the sensor orientations are known. The computational complexity of the algorithm is $O(mn^2)$, where m is the number of measurements and n is the total number of sensors. In simulation, the authors show that their algorithm is able to achieve comparable location error with iterative maximum-likelihood estimator, but the runtime is nearly two orders of magnitude less.

Goldenberg et al. [40] present a class of simple algorithms referred to as sweeps. The authors improve the computational complexity of sweeping using consistent position combinations and shell sweeps and extend sweeping to handle angle and noisy measurements. The simulation shows that under the condition of precise distance measurements (no measurement error present), the algorithm achieves an at least 90% of all uniquely localizable nodes when the average connectivity is between 6 and 7.5 in a network of 250 nodes. The authors also consider the condition of zero-mean Gaussian noise with standard deviation of 1, 5, and 15% of the sensing range to all distance measurements with no outliers.

Taylor et al. [41] introduce a simultaneous localization and tracking algorithm. The Bayesian filter-based algorithm tracks a target in a sensor network while simultaneously localizes and calibrates the nodes of the network.

Zhong and He [42] present a multi-sequence positioning (MSP) method for large-scale stationary sensor node localization in outdoor environments. Their method utilizes two-dimensional location information for each sensor node by processing multiple one-dimensional node sequences. The authors assume that in deployments where an event source has line-of-sight (LOS) to all sensors (i.e.,

centralized). The impact of factors such as number of anchors and target node density is studied in the simulation.

Many sensor network applications require sensors' locations to function correctly. Despite the recent advances, Location Discovery for sensor networks in hostile environments has not been addressed extensively. Li et al. [35] and Liu et al. [36] address the Location Discovery problem in presence of hostile attacks. Li et al. [35] first identify a list of malicious attacks that are unique to localization algorithms and also present statistical methods to identify intentionally corrupted measurements, therefore, making localization attack-tolerant.

Stoleru et al. [43] present an outdoor localization system for wireless sensor networks—spotlight. The system does not require any additional hardware for the sensor nodes, other than what already exists. All the complexity of the system is encapsulated into a single spotlight device. More specifically, the system uses the spatiotemporal properties of well-controlled events in the network (e.g., light) to obtain the locations of sensor nodes. The main idea is to generate controlled events in the deployed field. Mica2 and XSM motes are used to evaluate the system, and the authors report an average of 20 cm localization error in an outdoor environment. In addition, with any number of nodes deployed in a 2500 m^2 area, the sensor nodes can be localized less than 10 min, using a device that costs less than $1000.

2.2.5 Range-Free Location Discovery

Because nodes in a network may not be equipped to find exact range to other nodes and coarse accuracy is sufficient for most sensor network applications, solutions in range-free Location Discovery are being pursued as a cost-effective alternative to more expensive range-based approaches. Range-free Location Discovery methods do not require accurate distance measurements, but localize the nodes based on network connectivity (proximity) information. However, because wireless connectivity is highly influenced by the environment and hardware calibration, most of the existing solutions fail to deliver encouraging empirical results, or require substantial survey and calibration on a case-by-case basis [42].

Range-free Location Discovery techniques do not require the availability of the estimated/measured Euclidean distances between pairs of communicating nodes. He et al. [44] propose a range-free Location Discovery approach, which performs the best when an irregular radio pattern and random node placement are considered. The algorithm is so called area-based, which has two phases: (i) isolating the environment into triangular regions between beacon nodes and (ii) by considering whether a node's presence inside or outside of these triangular regions allows a node to narrow down the area in which it can potentially reside. The algorithm is demonstrated in conjunction with the routing and the tracking problems.

Shang et al. [34] presents a Location Discovery method that uses the connectivity information (i.e., who is within the communication ranges of whom) to derive the positions of the unknown sensor nodes. The algorithm is based on a data

analysis technique—multidimensional scaling. The algorithm provides the flexibility of constructing relative coordinates when no beacon nodes are available.

We survey three range-free Location Discovery schemes in detail: (1) centroid, (2) DV-Hop, and (3) amorphous APIT—approximate point-in-triangulation. Before we start, here is a list of terminologies used in the following section:

- *Node Density (ND):* average number of nodes per node radio area.
- *Anchor Heard (AH):* average number of anchors heard by a node.
- *Anchor to Node Range Ratio (ANR):* average distance an anchor beacon travels divided by the average distance a regular node signal travels.
- *Anchor Percentage (AP):* The number of anchors divided by the total number of nodes.

Centroid Localization:
In this method, the idea is not to use any ranging at all, but simply deploy enough beacons. Anchors periodically broadcast their location. This method uses anchor beacons, containing location information (X_i, Y_i), to estimate node position. After receiving these beacons, a node estimates its location using the following Centroid formula:

$$(X_{\text{est}}, Y_{\text{est}}) = \left(\frac{X_1 + \ldots + X_N}{N}, \frac{Y_1 + \ldots + Y_N}{N}\right) \tag{2.1}$$

We can estimate location of a node by averaging the values of the nodes within signal range within an error of:

$$LE = \sqrt{(X_{\text{est}} - X_a)^2 + (Y_{\text{est}} - Y_a)^2} \tag{2.2}$$

The distinguished advantage of this centroid localization is its simplicity and ease of implementation. The disadvantage is that a good anchor placement is crucial.

DV-Hop Localization:
DV-Hop localization uses a mechanism that is similar to classical distance vector routing. In this work, one anchor broadcasts a beacon to be flooded throughout the network containing the anchors location with a hop-count parameter initialized to one. Each receiving node maintains the minimum counter value per anchor of all beacons it receives and ignores those beacons with higher hop-count values. Beacons are flooded outward with hop-count values incremented at every intermediate hop. Through this mechanism, all nodes in the network (including other anchors) get the shortest distance, in hops, to every anchor. The average hop-count distance is estimated by anchor i using the following formula:

$$\text{HopSize} = \frac{\sum \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}}{\sum h_j} \tag{2.3}$$

where (x_j, y_j) is the location of anchor j and h_j is the distance in hops from anchor j to anchor i. Trilateration is used to estimate location in every node. Nodes will calculate the distance estimation to more than three anchors in the plane. When distance is known, the nodes will use trilateration to estimate its location. Error will decrease with the greater number of anchors.

Amorphous Localization:
This method is very similar to DV-Hop localization. First, each node obtains the hop distance to distributed anchors through beacon propagation. It assumes that the density of the network n_{local} is known a priori, so that it can calculate HopSize offline in accordance with the Kleinrock and Silvester formula:

$$\text{HopSize} = r\left(1 + e^{-n_{\text{local}}} - \int_{-1}^{1} e^{-\frac{n_{\text{local}}}{\pi}\left(\arccos t - t\sqrt{1-t^2}\right)} dt\right) \tag{2.4}$$

Finally, after obtaining the estimated distances to three anchors, triangulation is used to estimate a node's location.

APIT Localization:
APIT (approximation point-in-triangulation) algorithm narrows down the possible area in which a target node resides. Each node chooses three anchors from all audible anchors and tests whether it is inside the triangle formed by connecting these three anchors. The steps to perform this algorithm are:

- Beacon exchange
- PIT testing
- APIT aggregation
- COG—center of gravity calculation

These steps are performed at individual nodes in a purely distributed fashion. For PIT testing, the problem is that for three given anchors, $A(ax, ay), B(bx, by)$, and $C(cx, cy)$, determine whether a point M is inside triangle ΔABC or not. The perfect solution is that M is inside triangle ΔABC if for any direction node M moves, it will be closer to A, B, or C, but not for all three nodes. The problem with this method is that nodes cannot move, and also it needs a way to determine that a node is closer to an anchor. The solution is to use neighbor information to simulate movement, assuming nodes know their relative distance to other nodes using signal strength.

Once the individual APIT tests finish, APIT aggregates the results through a grid SCAN algorithm. In this algorithm, a grid is used to represent the maximum area in which a node will likely reside. For each APIT inside decision, the values of the grid regions over which the corresponding triangle resides are incremented. For an outside decision, the grid area is decremented. Once all triangular regions are computed, the resulting information is used to find the maximum overlapping area, which is then used to calculate the center of gravity for position estimation.

Table 2.2 Comparison of four range-free localizations

	Centroid	DV-Hop	Amorp	APIT
Accuracy	Fair	Good	Good	Good
Node density	>0	>8	>8	>6
Anchor heard	>10	>8	>8	>10
ANR	>0	>0	>0	>3
DOI	Good	Good	Fair	Good
GPS error	Good	Good	Good	Good
Overhead	Smallest	Largest	Large	Small

Comparison of different algorithms:
It is natural to compare different algorithms by estimation error over different parameters:

- With varying anchor heard (AH)
- Node density
- Anchor noise ratio
- Degree of irregularity
- Tolerance to GPS error
- Communication overhead for AH and varied neighborhood density

The result of the comparison is summarized in Table 2.2. As we can observe from the table, there is no one particular technique that is superior to others in all criteria.

More recently, GPS-less positioning approaches have also emerged where the resultant locations of the unknown nodes are relative with respect to their neighboring nodes either in terms of the Euclidean distances or hops. For example, Capkun et al. [45] developed a distributed infrastructure-free (mobile) positioning algorithm that uses the measured Euclidean distances between the nodes to build a relative coordinate system in which the node positions are computed in two dimensions. The authors also demonstrated that relative coordinates are sufficient for applications such as location-aided routing and geodesic packet forwarding. In addition, there have been studies on the positioning algorithms targeting mobile sensor networks as well [31, 45].

Li and Liu [46] claim that existing range-free localization schemes often suffer from poor accuracy and low scalability. They propose a rendered path (REP) protocol for locating sensors with constant number of beacons in anisotropic sensor networks in presence of possible holes. By path rendering and virtual hole constructions in a distributed manner, REP is able to accurately estimate the node-to-node distances and calculate the node locations with only three beacons, as supposed to seven in most of the existing range-free localization algorithms. In their simulation, the authors report an average of 7.4% distance estimation error in a perturbed grid deployment of 3024 nodes with average 7.3 neighbors. In a random deployment of 3027 nodes, the distance estimation error is 11.7% given the same average 7.3 neighbors.

Rudafshani and Datta [47] present two range-free algorithms: mobile and static sensor network localization (MSL) and MSL*, that work when any number of

sensors are static or mobile. Both algorithms are variations of the Monte Carlo method. The authors model irregularity in radio range by assuming that the radio range of a sensor follows a normal distribution with variable mean and standard deviation. Convergence, accuracy, and computational time are compared across four different algorithms: MSL*, Gradient, MCL, and MSL.

StarDust, a framework for wireless sensor network localization based on passive optical components, is proposed in [48]. Each sensor node is equipped with optical retro-reflectors. An aerial device projects light towards the deployed sensor network and records an image of the reflected light. Then, an image-processing algorithm is used for obtaining the locations of the sensor nodes. The authors report localization accuracy ranges from 2 to 5 ft in an experimental setting of 26 sensor nodes in a 120×60 ft^2. The operational time ranges from 10 to 2 min.

2.2.6 Performance, Topology Control, Coverage and Connectivity

Deploying sensors deterministically is an expensive undertaking in terms of time, effort, and money. However, properties such as deployment topology, coverage, and connectivity of a network are crucial to the performance of the network.

When monitoring spatial phenomena with wireless sensor networks, selecting the best sensor placements is a fundamental task. Krause et al. [49] present a data-driven approach that addresses the three central aspects of the sensor selection problem: sensing quality communication cost; and designing an algorithm with provable quality guarantees that optimizes the NP-hard tradeoff. The authors use data from a pilot deployment to construct nonparametric probabilistic models called Gaussian processes. A polynomial-time data-driven algorithm, pSPIEL (sensor placements at informative and cost-effective locations), is developed based on the models.

Deriving the critical density (which is equivalent to deriving the critical power or radius) needed to achieve coverage and/or connectivity is a fundamental problem. It can be used to predict the behavior of a network with regard to coverage and/or connectivity. Balister et al. [50] claim that one major limitation of most existing techniques of deriving critical density is that they limit themselves to thin deployment regions such as disks and squares. They present technique to derive reliable density estimates for achieving barrier coverage in finite realistic (e.g., not squares or disks) regions, such as thin strips.

Qiu et al. [51] developed a general model to estimate the throughput between arbitrary pairs of nodes in the presence of interference from other nodes in static multihop wireless networks. The model is based on measurements from the underlying network itself and consists of three major components: N-node Markov model for capturing interactions among an arbitrary number of broadcast senders, a receiver model of packet-level loss rates, and unicast sender and receiver models.

Goldenberg et al. [40] present a class of simple algorithms referred to as sweeps. The authors improve the computational complexity of sweeping using

consistent position combinations and shell sweeps, and extend sweeping to handle angle and noisy measurements. The simulation shows that under the condition of precise distance measurements (no measurement error present), the algorithm achieves at least 90% of all uniquely localizable nodes when the average connectivity is between 6 and 7.5 in a network of 250 nodes. The authors also consider the condition of zero-mean Gaussian noise with standard deviation of 1, 5, and 15% of the sensing range to all distance measurements with no outliers.

The cricket compass system is a position and orientation indoors, for a handheld, mobile device, and informs an application running on the device of the position and orientation in a local coordinate system established by the fixed set of beacons [52]. In addition, Priyantha et al. [53] present a location-support system that allows mobile and static nodes to learn their physical location by estimating their distances to beacons thought the building. The beacons are equipped with combined RF and ultrasound signals in a decentralized uncoordinated architecture. The authors note that positioning of a beacon within a room or space plays a nontrivial role in enabling nodes to make the correct choice of their location. As the cricket system is decentralized and preserves privacy, the beacons placed to demarcate a physical or virtual boundary corresponding to a different space must be placed at a fixed distance away from the boundary demarcating the two spaces. In addition, the beacons utilize a randomized transmission schedule.

2.3 Location Discovery Computational Complexity

Recently, it has been proven that localization is a NP-complete problem by transforming the graph-embedding problem into the localization problem [54]. It was identified that mirroring and flipping in 3D space are the cause of computational intractability [55]. We provide an alternative proof of the computational intractability of the localization problem. The impetus for the development of the new proof is provided by our objective to better understand the sources of the complexity, to establish more precise conditions under which localization is NP-complete and to better understand the importance of noisy measurements. Our main conclusion is that localization is NP-complete even when no mirroring and flipping are possible as in the one-dimensional localization problem because of errors in measurements. At the same time, actual distance measurements usually have significant errors and, therefore, there is a need to address error modeling, optimization in the presence of error and to develop and employ powerful optimization mechanisms for localization.

We prove the NP-completeness of the 1D Location Discovery problem by polynomial transformation of an instance of the known NP-complete problem—optimal linear arrangement problem [56], into the 1D localization problems, in polynomial time. For the sake of completeness and readability, we state both problems using the standard Gary-Johnson format:

The Optimlinear Arrangement Problem

Instance: Graph $G = (V, E)$, positive integer $K \leq |V|$.

Question: Is there a one-to-one function $f : V \rightarrow \{1, 2, \ldots, |V|\}$ such that $\sum_{\{u,v\} \in E} |f(u) - f(v)| \leq K$?

The 1D Localization Problem

Instance: A network of N sensor nodes, measured distances between all pairs of sensors $P = \{d_{ij}, i, j \in N\}$, a subset of P also has bi-directional measurements $Q = \{d_{ji}\}, Q \in P$, positive integer M.

Question: Is there a one-to-one function g: $i \rightarrow x_i$, where x_i is the estimated location of $i, i = 1, \ldots, N$, such that the overall discrepancy between the calculated distances $|g(i) - -g(j)|$ and the measured distances d_{ij} and d_{ji} of all pairs of sensor nodes i and j satisfy the following condition:$\sum_{\{i,j\} \in P} ||g(i) - g(j)| - d_{ij}| + \sum_{\{i,j\} \in Q} ||g(i) - g(j)| - d_{ji}| \leq M$?

Proof The reduction from the linear arrangement problem to the localization problem is as follows. Let the vertices in G be the sensor nodes, i.e., $V = \{i, i = 1, \ldots, N\}$. Let the edges in G be the measurements in Q, i.e., $E = \{d_{ji}, d_{ji} \in Q\}$.

More formally, let the graph $G = (V, E)$ and the positive integer K constitute an arbitrary instance of the linear arrangement problem. The basic units of the instance of the linear arrangement problem are the vertices and the edges of G. The instance of the localization problem is completely specified by:

$$i = \{v : v \in V\}$$

$$d_{ij} = \{\{i,j\} : i,j \in V\}$$

$$d_{ji} = \{\{j,i\} : \{j,i\} \in E\}$$

$$M = K + C, \text{ where } \quad C = \frac{1}{4}(K^2 + K + 2)(K - 1)$$

It is easy to see that this instance can be constructed in linear time. Note that the measured distances d_{ij} act as an "enforcer" [56, 57], which impose additional restrictions on the ways the sensor nodes must be placed. Specifically, all the d_{ij} have value $K + h$, where h specifies the least distance between any pair of nodes. This enforcer is necessary and sufficient to prevent multiple nodes being placed on an identical location, which corresponds to the condition that each node must have a unique assignment in the linear arrangement problem, that the distance between any two nodes is at least 1 unit. Therefore, each node maps into a unique integer location between 0 and K.

Function f exists if and only if there exists a function g that satisfies the condition of

$$\sum_{\{i,j\} \in P} \left|g(i) - g(j) - d_{ij}\right| + \sum_{\{i,j\} \in Q} \left|g(i) - g(j) - d_{ji}\right| \leq M$$

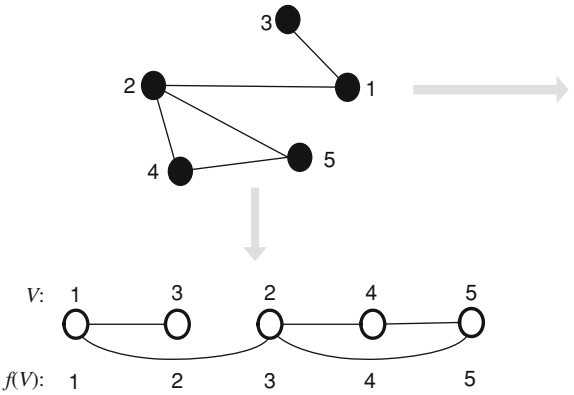


Fig. 2.2 An instance of the linear arrangement problem

Suppose g such that

$$\sum_{\{i,j\}\in P} \left|\left|g(i) - g(j)\right| - d_{ij}\right| + \sum_{\{i,j\}\in Q} \left|\left|g(i) - g(j)\right| - d_{ji}\right| \leq M$$

Consequently, there exists an f such that

$$\sum_{\{i,j\}\in P} |f(i) - f(j)| \leq K$$

Therefore, function f satisfies the condition

$$\sum_{\{u,v\}\in E} |f(u) - f(v)| \leq K \text{ since } P = \left\{d_{ji} : \{j,i\} \in E\right\}$$

Figure 2.2 shows a small instance of the linear arrangement problem with $K = 7$. G consists of five nodes and five edges. Table 2.3 demonstrates the transformation of the linear arrangement problem to the localization problem.

In the localization problem, all pairs of nodes have measured distance $d_{ij} = 8\,\text{m}$ $(K = 7\,\text{m}, h = 1\,\text{m})$. This enforcer restricts any pair of sensor nodes must be at least 1 m apart. A subset of pairs of nodes has an additional measurement d_{ji}, which corresponds to the edges in G.

It is easy to see that it requires polynomial time to verify a proposed solution. Also note that the localization problem considered in the proceeding sections is an extension of the 1D localization problem, i.e., the localization is done in 2D or 3D space.

2.4 Measurements, Error Sources and Modeling

Sensor networks frequently operate in indoor or outdoor settings where the impact of the environment on the sensing and communication capabilities of the nodes cannot be estimated in advance. Therefore, we need to ensure that network

Table 2.3 The corresponding arrangement problem localization problem

i	j	d_{ij}	d_{ji}
1	2	8	0
1	3	8	0
1	4	8	–
1	5	8	–
2	3	8	–
2	4	8	0
2	5	8	0
3	4	8	–
3	5	8	–
4	5	8	0

protocols and algorithms in WASNs perform well under a wide range of conditions before the network is actually deployed. Also, we need to know before the network deployment if it is advantageous for WASN applications to adapt their functionality to the actual conditions in the monitored area.

Although testing of a real network in various physical environments is an indispensable part of the deployment preparation, we still rely on simulations as a cheaper and a faster way to test networks under varying conditions. In this chapter, we develop error models for the input parameters of the self-localization process. For localization algorithms, one of the most important parameters affected by the environment is the ranging error because a large number of localization algorithms depend on the range estimates to derive the locations of nodes. We comment first on the sources of errors in WASNs and their relative importance for error modeling, and then we describe three different error models. Two of them are based on general assumptions about the sources of ranging error for the radio and acoustic signal, and one error model is based on the real life data.

2.4.1 Error Sources

We identify at least five main sources of error that influence optimization results in WASNs:

(1) Measurement
(2) Finite precision
(3) Objective function specific
(4) Intractable optimization tasks
(5) Localized algorithms.

The first two, measurement-related and finite precision-related errors, are inherent in all physical computing systems. Measurement errors arise because of sensing technology limitations, phenomena instability, and environment noise. Numerous well-studied techniques exist to reduce or compensate for such errors in many domains. Examples of such techniques are averaging methods that rely on several

distinct measurements and digital signal processing (DSP) techniques. Finite precision errors are caused by inaccuracies induced in the result because of the limited computation precision of digital computers. Because the WASNs hardware is typically very resource constrained, such errors can be of significant magnitude.

The third and the fourth source of errors are optimization task specific. The OF used at the heart of an optimization process may not accurately or completely capture the essence of the problem and thus can lead to erroneous conclusions. The fourth source of error is because of the inherently intractable nature of many optimization tasks tackled by WASNs. Optimal solutions to intractable problems are by definition very difficult to compute. Furthermore, exceptionally limited resources such as energy, communication bandwidth, storage, and processing power make tackling such problems and their associated errors (in results) especially challenging in WASNs.

As opposed to the sources of error described above, which exist in many computation domains, the error associated with localized algorithms is unique to distributed computation systems such as WASNs. Localized algorithms are especially suited to WASNs because of the geographic structure of the network. In localized algorithms, only nodes in spatial (or other) "proximity" collaborate and participate in a computation. Using local information to compute inherently global metrics can be a very error-prone task. Understanding such errors and their behavior is a fundamental task in designing successful localized algorithms.

Distance measurements, and in some cases GPS-generated location estimates, are the input parameters of the Location Discovery process. They are also the main source of errors in results of Location Discovery algorithms. Modeling of errors in input parameters, therefore, is of great significance for simulations of algorithms and applications in WASNs. As wireless sensor networks are intended to be used, among other environments, in remote and inhospitable areas, where the error characteristics of different measurements cannot be examined in advance, the prediction of the performances of the network depends on simulations with carefully chosen error models.

2.4.2 Theoretical Error Models

The distance estimates in WASNs are susceptible to different sources of error, including obstacles, interference, and multipath effects. The impact of these sources of error depends on a hardware technology used for measurements. As mentioned previously, there are three most frequently proposed distance measuring technologies: RSS, TDoA combining the radio and acoustic signal, and acoustic-based ranging. Each of the given technologies has different characteristics of the error. Unfortunately, there are no enough publicly available data to precisely characterize the given technologies. We opted for two error models that correspond to each of the hardware technologies. The first error model, mainly intended to capture properties of distance measurement error for RSS, is based on the path

loss models from Girod et al. [58]. The main source of error in RSS-based distance measurements is the complexity of modeling of environmental effects in the propagation model. Reflection, scattering, and diffraction, as well as the antenna gains, produce significantly different path losses for equal distances. From [58] (p. 104) the distribution of measured distances $\widehat{d}$ is:

$$10n\log\left(\frac{\widehat{d}}{d_0}\right) - 10n\log\left(\frac{d}{d_0}\right) = X_\sigma[dB] \tag{2.5}$$

where $\widehat{d}$ is the actual distance, and $X_\sigma$is a zero-mean Gaussian random variable with the standard deviation σ, both in dB (Eq. 2.5):

$$\widehat{d} = d + d(10^{\frac{X_\sigma}{10n}} - 1) \tag{2.6}$$

The second term in (2.6) represents the distance-dependent error. The second error model is based on the Gaussian distribution. The distance estimate $\widehat{d}$ is given $\widehat{d} = d + d \times G(0, \sigma)$, where $G(0, \sigma)$ is white Gaussian noise with the standard deviation σ.

Modeling of the acoustic ranging error is based on the research reported in [59]. There are three important sources of error in acoustic ranging that cannot be eliminated by averaging distance measurements over time [59]:

1. Non-line-of-sight (NLOS) error: This error is caused by obstacles between nodes. We model it as a distance-dependent uniformly distributed positive error $U_n(0, \text{NLOS ERROR}_{\max}(d))$.
2. Speed of sound error: Atmospheric changes in the environment, as well as different atmospheric conditions in various parts of the network, affect the speed of sound. We model this as distance-dependent Gaussian noise $G_1(0, \sigma(d))$.
3. Orientation error: The emitter and the sound sensor may not be aimed directly towards each other, which create an error that depends on the angle between them. We model this error as angle-dependent Gaussian noise $\alpha G_2(0, \sigma)$, where α is the angle between the emitter and the sound sensor.

Thus, the acoustic distance measurement between the nodes i and j is simulated as:

$$\widehat{d} = d + d(U_n + G_1) + \alpha G_2 \tag{2.7}$$

For each simulation, a subset of nodes that have initial estimates of their locations (beacons) is randomly selected. The initial locations of beacons are generated by superimposing an error to the real locations as follows. The real locations of sensor nodes A_i, $i = 0, \ldots, n$ are represented as points $A_i(x_i, y_i)$. Coordinates x_i and y_i are generated from two uniform distributions, one on the interval $[0, X_{max}]$ and the other on the interval $[0, Y_{max}]$. The error is generated

from the Rayleigh distribution, by generating a pair $(\Delta x_i, \Delta y_i)$, where both Δx_i and Δy_i are selected from a zero-mean Gaussian distribution $N(0, \sigma_{loc})$. The average location error is then:

$$\mu = \sqrt{\frac{\pi}{2}}\mu_{\text{loc}} \tag{2.8}$$

Now, by selecting σ_{loc}, we can generate a location error distribution with a desired mean value.

2.4.3 Real Data Error Models

The statistical model of the ranging error defined in this section is based on data sets obtained from Lewis Girod's experiments at the UCR James Reserve [60]. We demonstrate that the ranging error in the data sets is correlated with the measured distance. We then determine that a linear regression model, with the measured distance as the independent variable, fits well with the experimental data. We also propose an explanation of the physical meaning of the parameters of the model as well as possible sources of the variability of the parameters.

The data sets are gathered from a network of 10 nodes, where the nodes measure mutual distances generating one data set during intervals of one or two minutes in length, with 10 min intervals between data sets. There are 114 data sets; however, not every pair of nodes generates a measurement in a particular data set. The original data sets contain five measurements between each pair of nodes that communicated in that particular round. In [60], it was assumed that the errors are because of NLOS propagation, and therefore, the shortest of the five measurements should be selected as the one closest to the actual distance. However, the speed of sound changes with temperature, so there is a possibility that the speed of sound decreases enough that one of the excluded four measurements is actually closer to the actual distance. In the data sets, there are some negative values for error, which could confirm that using the shortest distance estimate might not always give the closest estimate of the measured distance. However, we will also show that the data sets with such negative values are most likely unreliable, so the negative error might not be a result of the incorrect calibration.

The problem of dependence on temperature is lessened in [60] by the implemented localization algorithm. The map of the network is built with the distance estimates, without any assumptions about the absolute locations of the nodes. Once the map is in the place, the absolute locations of some of the nodes are taken into account, and the map is scaled until it fits the known locations. If we assume that the environmental effects affect the distance measurements proportionally, the scaling of the map would negate the effects of the incorrect calibration. However, because temperature, wind, and obstacles affect the distance estimates by different factors, these approximations can add one more component of error.

2.4.3.1 Linear Regression Model

If d_{ij} denotes the actual distance between nodes i and j, and $\widehat{d}_{ij,k}$ denotes the corresponding distance measurement in the data set k, the ranging error $e_{ij,k}$ for the distance d_{ij} in the data set k is:

$$e_{ij,k} = \widehat{d}_{ij,k} - d_{ij} \tag{2.9}$$

Our goal is to determine a parametric model in which we can express the ranging error as

$$e_{ij,k} = F_k(d_{ij}; \theta) \tag{2.10}$$

where $F_k(d_{ij}; \theta)$ is a probability distribution, with a set of parameters θ.

The initial statistical examination of the data sets attempts to determine if there is any correlation between the ranging error and the measured distance. To detect correlation, we use the Pearson's correlation coefficient, with outliers removed from the data sets before calculating the correlation coefficient. For most measurements, the ranging error is less than 1 m, so the threshold for outliers set at 5 m for ranging error removes one or zero measurements per data set.

The results show an important difference between two groups of data sets. The first 42 data sets show a high degree of correlation between the ranging error and the measured distance. Two typical scatterplots from that set of data are given in Fig. 2.3. The first data set is Set 9, where the coefficient of correlation is 0.98, whereas the second data set, Set 29, has the coefficient of correlation of 0.94.

Figure 2.4 shows two typical data sets from the group containing the last 72 data sets. The correlation in these data sets is significantly lower. For data set 55, the coefficient of correlation is 0.39, and for the set 101, it is 0.38. It is important to note that, while the negative values of errors are shown in figures, the correlation coefficients are calculated using absolute values.

Finally, correlation coefficients for all 114 data sets are displayed in Fig. 2.5. There is a significant change in the value and the size of the 95% confidence ranges between the data sets before or around the data set 40 and the following data sets. A possible cause of this difference is malfunctioning of the nodes in the latter data sets. The batteries in some nodes ran out and a bug in the time synchronization service validated measurements after a certain point, according to L. Girod (personal communication, December 22, 2006). It is not known at what time the mentioned problems occurred. Therefore, we model only the first group because the second group most likely contains invalid measurements.

The validity of the correlation coefficient relies on the assumptions that both distances and errors are normally distributed, and that the variance of the error does not change across all measured distances. We should note here that even if the second condition is not met, its influence on the results would be to increase the confidence ranges for the correlation coefficient. Therefore, the real correlation between the distance measurements and ranging error could only be higher in that case.

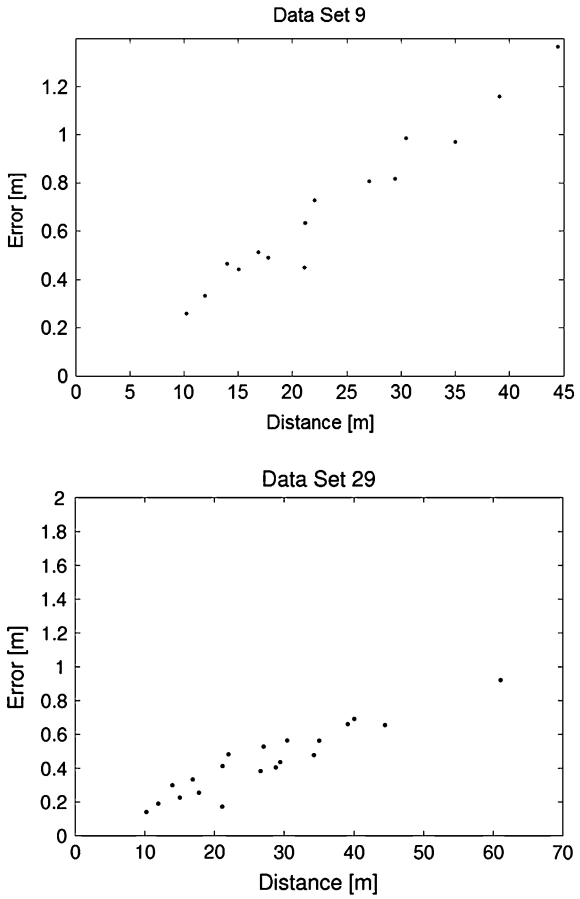


Fig. 2.3 Scatterplot of the measured distance and the ranging error for two of the first 42 data sets, where the correlation exists

For each data set in this group, we tested the measured distances and ranging errors for normality to ensure that the correlation coefficients are valid. Using both, Lilliefors and Jarque–Bera test, we can report the following results:

- Both tests do not reject the normality hypothesis for measured distances for any of the data sets at the significance level of 0.05. Although distances are not sampled from a normally distributed population, the results of the tests ensure that the high values of the correlation coefficient are not caused by a specific distribution of either variable.
- For errors, both tests rejected the normality hypothesis for only 2 out of 42 data sets.

Because the correlation coefficients confirm that there is a linear relation between measured distances and ranging errors, we determine the parameters of the linear regression model and use the model and the parameters to simulate the distance error. For a data set k, we can represent the linear regression model between the measured distance d_{ij} and the distance error $e_{ij,k}$ as:

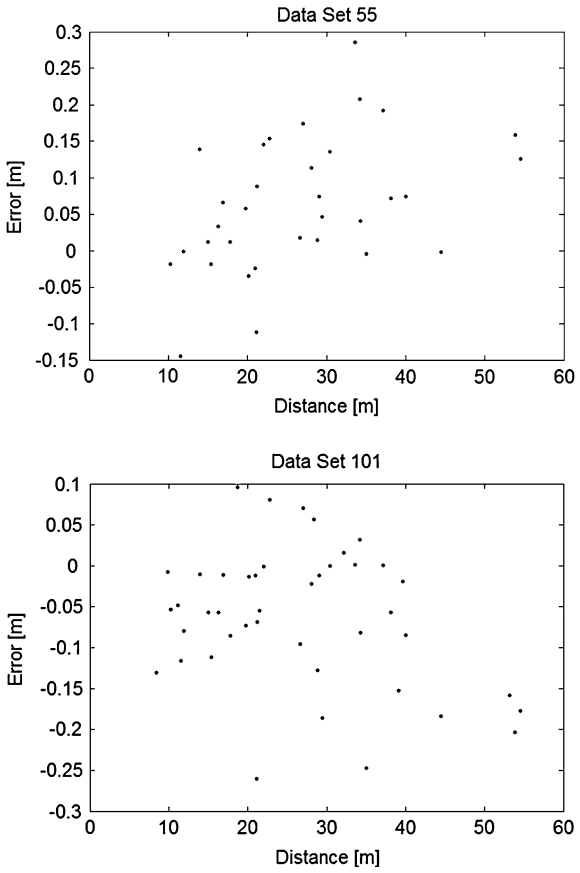


Fig. 2.4 Scatterplot of the measured distance and the ranging error for two of the last 72 data sets, where there is no correlation

$$e_{ij,k} = \alpha_k + \beta_k * d_{ij} + \epsilon_k \tag{2.11}$$

where α_k and β_k are regression parameters specific for the data set k, while ε_k is a random Gaussian component of error.

The model for the distance estimates $\widehat{d}_i$ for the data set k can be then represented as:

$$\widehat{d}_{ij,k} = d_{ij} + \alpha_k + \beta_k * d_{ij} + \epsilon_k \tag{2.12}$$

As mentioned previously, there are two more assumptions we need to verify to validate the results of the regression analysis. The assumptions state that (a) for each data set, ε_k is normally distributed, and that (b) the variance for ε_k does not change with the distance (homoscedasticity).

When testing the correlation between the measured distances d and the residuals ε, we need to confirm that the residuals are randomly dispersed when shown as a function dependent on distance. Visual observation of the 42 data sets shows that the dispersion does not show any dependence on distance. A typical case is

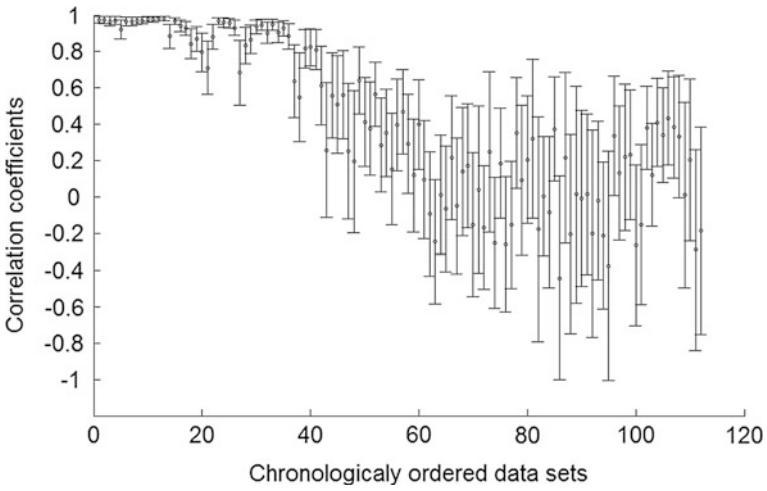


Fig. 2.5 Correlation coefficients of measured distances and ranging error with 95% confidence ranges

given in Fig. 2.6. We need to ensure not only that there is no linear dependency but also that there is no any visible pattern in data.

Additionally, we run correlation tests to formally determine that there is no correlation between distances and residuals. For 34 out of 42 data sets, the hypothesis of the correlation is rejected, as shown in Fig. 2.7. The figure shows the value of the correlation coefficient, together with the 95% intervals of confidence, which in 34 cases intersect the y-axis. For eight cases, the correlation coefficient is between 0.48 and 0.61. Combined with the visual observation of the variance of residuals in each data set separately, we accepted that residuals ε can be modeled as a normal distribution. Furthermore, the results of testing of ε for normalcy using Lilliefors and Jarque–Bera test show that in 35 data sets for Lilliefors, and in 36 data sets for Jarque–Bera, the hypothesis of the normalcy cannot be rejected at the significance level of 0.05.

As the final step, we need to establish the formulas for generating the values of the parameters α and β and the variance for ε in the model represented by (Eq. 2.12). Figure 2.8 presents the values of the parameters with the data sets chronologically ordered. The range of values for β is between 0.005 and 0.04 (0.5 and 4% of the measured distance), whereas the range for α is between -0.53 and 0.27. However, more important to notice is that there is seemingly an inverse correlation between α and β. To verify the correlation, we perform another regression analysis to find a relationship between these parameters.

From the results of that analysis, we derive the following relation between α and β:

$$\alpha = -13.4\beta + 0.25 + N(0;\, 0.14) \tag{2.13}$$

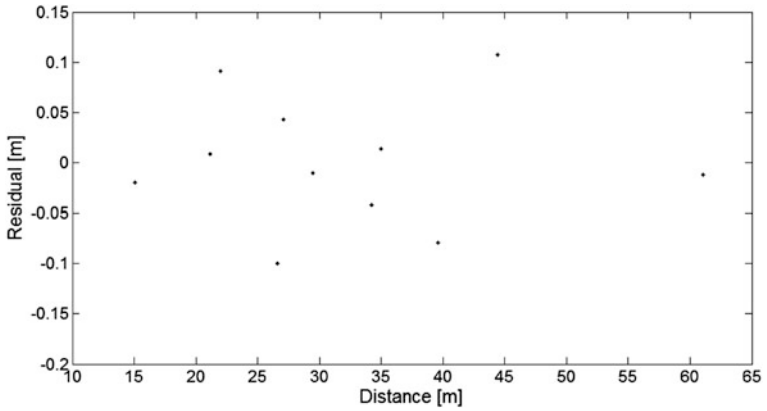


Fig. 2.6 Dispersion of distance measurement residuals for one of the datasets

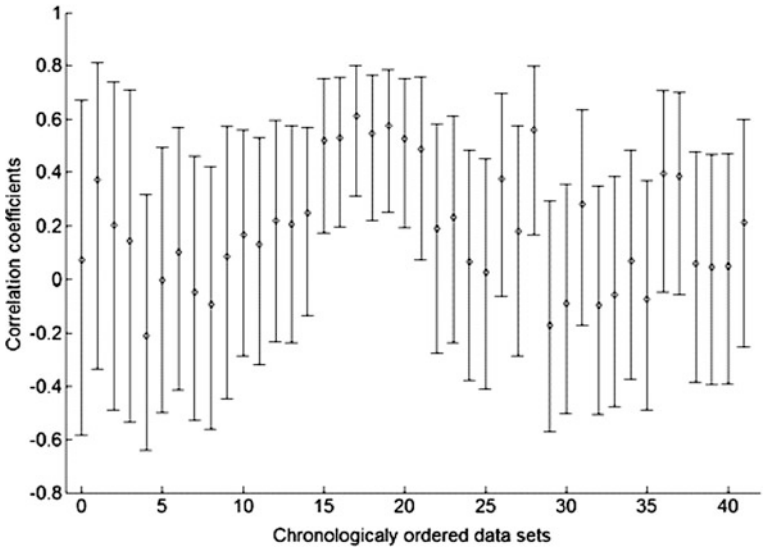


Fig. 2.7 Correlation of measured distances and regression residuals with 95% confidence ranges

where $N(0;0.14)$ is a normal distribution with $\sigma^2 = 0.14$. Therefore, when the error model is used, we randomly select β from its interval given above, and then generate the value for α from (Eq. 2.13). The same analysis is applied to variance of ε in the data sets, but there were no correlation between ε and either α or β. Therefore, the variance for ε is randomly selected from the range in the data sets, between 0.01 and 0.23.

The error model derived from Eqs. 2.12 and 2.13 does not make any assumptions about environmental conditions that could be represented by the

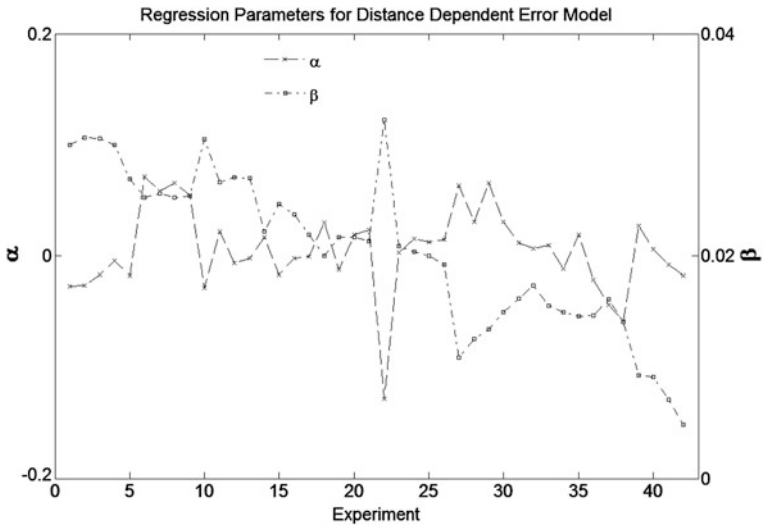


Fig. 2.8 Regression parameters for data sets where the linear dependency of the ranging error on measured distances is established

parameters of both linear models. However, we can assume that the linear relation between the measured distance and the distance error can be explained by several different factors. First, the speed of the sound depends on temperature, so the speed of the sound for which the ranging system is calibrated can differ from the actual speed. Wind and humidity can also affect the speed of sound. Second, the likelihood of foliage between nodes is higher when nodes are further apart. The parameter α represents the component of error that does not change with the distance. Such error can be caused by the phase error in clocks on the nodes. Because we do not have the measurements of the environmental conditions, further studies are needed to determine physical nature of the parameters. However, our primary goal is to define an error model that approximates well the real life data, so we can use that model to simulate networks with sizes and topologies different from the original network in which the data sets were acquired.

2.5 Distance Measurements and Experimental Platform

We construct the statistical error models and conduct Location Discovery on sets of distance measurements that are collected using the acoustic signal detection-based ranging techniques. The number of deployed sensor nodes varies from 79 to 93, with the average being 90. The sensor nodes are custom designed based on an SH4 microprocessor running at 200 MHz (Fig. 2.9a). The nodes are deployed at

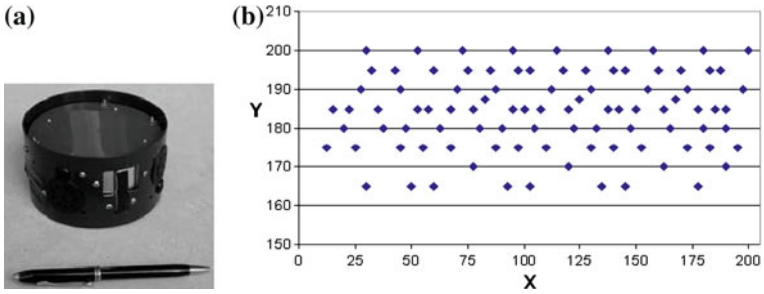


Fig. 2.9 **a** A SH4 node. **b** An example of deployment topology

the Fort Leonard Wood Self Healing Minefield Test Facility, which measures 200 × 50 m. The nodes are roughly 10.5 m apart, and the radio signal (communication) range is about 50 m. Figure 2.9b shows an example of a deployment topology. Each node is equipped with four independent speakers and microphones that are used as the ranging tool. The distance between two nodes is obtained by timing the arrival of the acoustic signals [61]. Each node in the network takes turns to transmit the acoustic signals; all the nodes that receive the signals record the ToA and convert the time of flight to distance in meters. There are 33 sets of distance measurements in total that were collected during the course of few days. Each set consists of one round of acoustic signal transmission by all the nodes. For the sake of simplicity, we demonstrate the algorithms and techniques on a randomly selected set of measurements, and we present the results for other data sets in the experimental result section of each chapter. The details on the experimental setup and the acoustic detection scheme used can be found in [62].

From the communication point of view, we distinguish two types of the communications between a pair of sensors: (1) exchange of the acoustic signals for the purpose of distance ranging and (2) transmission and reception of radio signals (in terms of bytes) for the purpose of exchanging information. More specifically, we denote L_i as a set of sensors that receive sensor i's acoustic signals, therefore can estimate the distances between themselves to sensor i. Similarly, C_i denotes a set of sensors that receive the radio signals from i. We assume that the Acoustic Signal Range (ASR) is independent from the Radio Signal Range (RSR), which means that it is possible for a sensor i to have the distance estimate to another sensor j (i has received j's acoustic signals), while i cannot exchange information with j (j is out of i's radio signal range) and vice versa. Furthermore, it is not necessary that all sensors in the network have the same ASR and RSR properties. In our study, we assume ASR and RSR are identical for all beacons and unknown location sensors. In our study, "neighbors" of sensor i refer to C_i, the set of sensors i has distance estimates, and because we assume $C_i = L_i$, therefore it is also the set of sensors i can communicate information with.

2.5.1 ILP-Based Instance Generation/Selection

In this section, we present the boolean ILP formulation, which selects a portion of the original data set to satisfy a set of neighboring requirements, while the measurement errors still follow the same distribution as the original instance. Note that the larger instance creation is very similar to the instance selection we are presenting. Because of the limitation of space, we only present the latter case. We formulated this problem as ILP because of its optimality and its ability of solving large instances. All of our experiments regarding instance selection and generation are done by the commercial ILP solver CPLEX. The input of the formulation is a set of constants that denote the existence of edges between all N nodes:

$$E_{ij} = \begin{cases} 1, & \text{if there exists edge between nodes } i \text{ and } j \\ 0, & \text{otherwise} \end{cases} \tag{2.14}$$

The outputs are the subset of nodes selected and the size of the subset. More specifically, the goal is to select a subset S of nodes from the original network in such a way that (1) each node in S has at least C_1 but not more than C_2 neighbors in S and (2) the average number of neighbors in the subset is equal to C_3. The ILP formulation has two types of variables and four types of constraints:

$$x_i = \begin{cases} 1, & \text{if node } i \text{ is chosed} \\ 0, & \text{otherwise} \end{cases} \tag{2.15}$$

$$L_{ij} = \begin{cases} 1, & \text{if edge between nodes } i \text{ and } j \text{ is chosen} \\ 0, & \text{otherwise} \end{cases} \tag{2.16}$$

Four required types of constraints are:
All variables must have value either 1 or 0.

$$x_i \geq 0; x_i \leq 1; \quad i = 1, \ldots, N$$

$$\forall E_{ij} = 1, L_{ij} \geq 0; L_{ij} \leq 1; \quad i = 1, \ldots, N; j = 1, \ldots, N \tag{2.17}$$

Each node in S must have at least C_1 and at most C_2 neighbors that also belong to S.

$$\sum_{j=1}^{N} \left(L_{ij} \wedge \left(E_{ij}(x_i \wedge y_i)\right)\right) - C_1 x_i \geq 0; \tag{2.18}$$

$$\sum_{j=1}^{N} \left(L_{ij} \wedge \left(E_{ij}(x_i \wedge y_i)\right)\right) - C_2 x_i \leq 0; \quad i = 1, \ldots, N \tag{2.19}$$

The nodes in S have an average number of neighbors close to C_3.

$$\sum_{i=1}^{N}\sum_{j=1}^{N} L_{ij} - C_3 \sum_{i=1}^{N} x_i \leq \varepsilon; \tag{2.20}$$

where ε is a small user-specified discrepancy constant.

If an edge between nodes i and j is selected, then nodes i and j must be selected too.

$$L_{ij} - (x_i \wedge y_i) \leq 0; \quad i = 1, \ldots, N; j = 1, \ldots, N \tag{2.21}$$

In the second type of constraint, the term $L_{ij} \wedge E_{ij}(x_i \wedge y_i)$ specifies whether the edge E_{ij} is chosen (value 1) or not (value 0). The summation of this term over all j is, therefore, the number of edges (or neighbors) chosen for node i. The condition of not fewer than C_1 number of chosen neighbors is enforced by constraining the difference between the summation and $(C_1 x_i)$ to be greater than or equal to zero. Similarly, constraining the difference between the summation and $(C_2 x_i)$ requires that the number of chosen neighbors does not exceed C_2. The last constraint ensures that in the situation in which an edge E_{ij} is chosen, the corresponding node i and j must belong to the subset.

The OF is to maximize the number of nodes that satisfy these conditions (i.e., the size of the selected subset):

$$\max : \sum_{i=1}^{N} x_i \tag{2.22}$$

The logical and operator ($\wedge$) operator is implemented in the following way. Consider a and b as the two operands and c is the result (i.e. $c = a \wedge b$). There are four types of constraints:

1. $\mathrm{a} \geq 0; \mathrm{a} \leq 1; \mathrm{b} \geq 0; \mathrm{b} \leq 1; \mathrm{c} \geq 0; \mathrm{c} \leq 1;$
2. $\mathrm{c} - \mathrm{a} \leq 0;$
3. $\mathrm{c} - \mathrm{b} \leq 0;$
4. $\mathrm{c} - (\mathrm{a} + \mathrm{b}) \geq -1;$

The first type of constraint forces every variable to be boolean. The second and the third type of constraint enforce c to be 0 when either a or b or both have value 0. The forth constraint enforces c to be 1 when both a and b have value 1.

References

1. D.A. Belsley, E. Kuh, R.E. Welsch, *Regression Diagnostics* (Wiley, New York, 1980)
2. S. Chatterjee, B. Price, *Regression Analysis by Example* (Wiley, New York, 1977)
3. P.I. Good, *Resampling Methods* (Springer, New York, 1999)
4. A.C. Davison, D.V. Hinkley, *Bootstrap Methods and their Application* (Cambridge University Press, Cambridge, 1997)
5. B. Efron, *The Jackknife, the Bootstrap, and Other Resampling Plans, Society for Industrial & Applied Mathematics*, 1982
6. B. Efron, R. Tibshirani, *An Introduction to the Bootstrap* (Chapman-Hall, New York, 1993)

7. D. Kahaner, C. Moler, S. Nash, *Numerical Methods and Software*. (Prentice Hall, Englewood Cliffs, New Jersey, 1989)
8. D.G. Luenberger, *Linear and Nonlinear Programming*, 2nd edn. (Addison & Wesley, Boston, 1984)
9. D. Kahaner, C. Moler, S. Nash, *Numerical Methods and Software* (Prentice Hall, Englewood, 1989)
10. W. Press, B. Flannery, S. Teukolsky, W. Vetterling, *Numerical Recipes*. (Cambridge University Press, Cambridge, 1986)
11. D.P. Bertsekas, *Nonlinear Programming*, 2nd edn. (Athena Scientific, Nashua, 1999)
12. W. Press, B. Flannery, S. Teukolsky, W. Vetterling, *Numerical Recipes* (Cambridge University Press, Cambridge, 1986)
13. R. Thisted, *Elements of Statistical Computing: Numerical Computation* (CRC Press, Boca Raton, 1988)
14. WNLIB Subroutine Library (http://www.willnaylor.com/wnlib.html)
15. T. Hastie, R. Tibshirani, J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (Springer, New York, 2001)
16. T. Roberton, F. Wright, T. Dykstra, *RL Order Restricted Statistical Inference* (John Wiley and Sons, New York, 1988)
17. R.E. Barlow, D.J. Bartholomew, J.M. Bremner, H.D. Brunk, *Statistical Inference Under Order Restrictions* (J. Wiley & Sons, New York, 1972)
18. G. Pan, Subset selection with additional order information. Biometrics **52**, 1363–1374 (1996)
19. T.R. Turner, P.C. Wollan, Locating a maximum using isotonic regression. Comput. Stat. Data Anal. **25**, 305–320 (1997)
20. Q.F. Stout, Optimal algorithms for unimodal regression. Computer Science and Statistics (University of Michigan, Michigan, 2000)
21. M. Ayer, H.D. Brunk, G.M. Ewing, W.T. Reid, E. Silverman, An Empirical Distribution Function for Sampling with Incomplete Information. Ann. Math. Stat. **26**, 641–647 (1955)
22. E. Mammen, Estimating a smooth monotone regression function. Ann. Stat. 724–740 (1991)
23. J.O. Ramsay, Estimating smooth monotone functions. J. Royal Stat. Soc. **B-60**, 365–375 (1998)
24. K. Whitehouse, A. Woo, C. Karlof, F. Jiang, D. Culler, *The Effects of Ranging Noise on Multi-Hop Localization: An Empirical Study*, IPSN, 73–80, 2005
25. A. Cerpa, J. Wong, L. Kuang, M. Potkonjak, D. Estrin, *Statistical Model of Lossy Links in Wireless Sensor Networks*, IPSN, 81–88, 2005
26. C. Peng, G. Shen, Y. Zhang, Y. Li, K. Tan, *BeepBeep: A High Accuracy Acoustic Ranging System using COTS Mobile Devices*, SenSys, 1–14, 2007
27. L. Girod, M. Lukac, V. Trafa, D. Estrin, *The Design and Implementation of a Self-Calibrating Distributed Acoustic Sensing Platform*, SenSys, 71–84, 2006
28. X. Sheng, Y.H. Hu, *Energy Based Acoustic Source Localization*, IPSN, 285–300, 2003
29. D. Niculescu, B. Nath, *Error Characteristics of Ad Hoc Positioning Systems* (*APS*), MobiHoc, 20–30, 2004
30. A. Galstyan, B. Krishnamachari, K. Lerman, S. Pattem, *Distributed Online Localization in Sensor Networks Using a Moving Target*, IPSN, 61–70, 2004
31. L. Hu, D. Evans, *Localization for Mobile Sensor Networks*, MobiCom, 45–57, 2004
32. P. Biswas, *Semidefinite Programming for Ad Hoc Wireless Sensor Network Localization*, IPSN, 46–54, 2004
33. A. Nasipuri, K. Li, *A Directionality based Location Discovery Scheme for Wireless Sensor Networks*, WSNA, 105–111, 2002
34. Y. Shang, W. Ruml, Y. Zhang. *Localization from Mere Connectivity*, MobiHoc, 201–212, 2003
35. Z. Li, W. Trappe, Y. Zhang, B. Nath, *Robust Statistical Methods for Securing Wireless Localization in Sensor Networks*, IPSN, 91–98, 2005
36. D. Liu, P. Ning, W. Du, *Attack-Resistant Location Estimation in Sensor Networks*, IPSN, 99–106, 2005

37. A. Savvides, W. Garber, S. Adlakha, R. Moses, M.B. Srivastava, *On the Error Characteristics of Multihop Node Localization in Ad-Hoc Sensor Networks*, IPSN, 317–332, 2003
38. M.A. Andreas, C. Travis, Le. Girod, K. Yao, E.Y. Chares, D.T. Blumstein, *An Empirical Study of Collaborative Acoustic Source Localization*, IPSN, 41–50, 2007
39. J. Ash, L. Potter, *Robust System Multiangulation using Subspace Methods*, IPSN, 61–68, 2007
40. D. Goldenberg, P. Bihler, M. Cao, J. Fang, B. Anderson, A.S. Morse, Y. R. Yang, *Localization in Sparse Networks using Sweeps*, MobiCom, 110–121, 2006
41. C. Taylor, A. Rahimi, J. Bachrach, H. Shrobe, A. Grue, *Simultaneous Localization, Calibration, and Tracking in an Ad Hoc Sensor Network*, IPSN, 27–33, 2006
42. Z. Zhong, T. He, *MSP*: *Multi-Sequence Positioning of Wireless Sensor Nodes*, SenSys, 15–28, 2007
43. R. Stoleru, T. He, J.A. Stankovic, D. Luebke, *High-Accuracy, Low-Cost Localization System for Wireless Sensor Network*, SenSys, 13–26, 2005
44. T. He, C. Huang, B.M. Blum, J.A. Stankovic, T. Abdelzaher, *Range-Free Localization Schemes for Large Scale Sensor Networks*, MobiCom, 81–95, 2003
45. S. Capkun, M. Hamdi, J.P. Hubaux, GPS-free Positioning in Mobile Ad Hoc Networks. Clust. Comput. J. **5**(2), 157–167 (2002)
46. M. Li, Y. Liu, *Rendered Path*: *Range-Free Localization in Anisotropic Sensor Networks with Holes*, MobiCom, 51–62, 2007
47. M. Rudafshani, S. Datta, *Localization in Wireless Sensor Networks*, IPSN, 51–60, 2007
48. R. Stoleru, P. Vicaire, T. He, J.A. Stankovic, *StarDust*: *A Flexible Architecture for Passive Localization in Wireless Sensor Networks*, SenSys, 57–70, 2006
49. A. Krause, C. Guestrin, A. Gupta, J. Kleinberg, *Sensor Selection and Placement*, IPSN, 2–10, 2006
50. P. Balister, A. Sarkar, B. Bollobas, S. Kumar, *Reliable Density Estimates for Achieving Coverage and Connectivity in Thin Strips of Finite Length*, MobiCom, 75–86, 2007
51. L. Qiu, Y. Zhang, F. Wang, M.K. Han, R. Mahajan, *A General Model of Wireless Interference*, MobiCom, 171–182, 2007
52. N.B. Priyantha, A.K.L. Miu, H. Balakrishnan, S. Teller, *The Cricket Compass for Context-Aware Mobile Applications*, MobiCom, 1–14, 2001
53. N.B. Priyantha, A. Chakraborty, H. Balakrishnan, *The Cricket Location-Support System*, MobiCom, 32–43, 2000
54. J.B. Saxe, *Embeddability of Weighted Graphs in k-Space is Strongly NP-Hard*, Conference of Communications, Control, and Computing, 1979, pp. 480–489
55. D. Moore, J. Leonard, D. Rus, S. Teller, *Robust Distributed Network Localization with Noisyrange Measurements*, SenSys, 50–61, 2004
56. M.R. Garey, D.S. Johnson, *Computers and Intractability, a Guide to the Theory of NP-Completeness* (W.H. Freeman and Company, New York, 2002)
57. T.G. Szymanski, Assembling Code for Machines with Span-Dependent Instructions. Comm. ACM **21**, 300–308 (1978)
58. T.S. Rappaport, *Wireless Communications—Principles and Practice* (Prentice-Hall, Inc., New Jersey, 1996)
59. L. Girod D. Estrin, Robust range estimation using acoustic and multimodal sensing, in IEEE/RSJ international conference on intelligent robots and systems (IROS 2001), Maui, Hawaii, Oct 2001
60. L. Girod, M. Lukac, V. Trifa, D. Estrin, The design and implementation of a self-calibrating acoustic sensing platform, in *Proceedings of the ACM Conference on Embedded Networked Sensor Systems* (SenSys 2006), pp. 71–84
61. L. Girod, *Development and Characterization of an Acoustic Rangefinder*, Tech. Rep. USC-CS-00-728, 2002
62. W. Merrill, L. Girod, J. Elson, K. Sohrabi, F. Newberg, W. Kaiser, *Autonomous Position Location in Distributed, Embedded, Wireless Systems*, IEEE CAS Workshop, 2002



http://www.springer.com/978-1-4614-1838-2

Localization in Wireless Networks
Foundations and Applications
Sanford, J.F.; Potkonjak, M.; Slijepcevic, S.
2012, XVI, 200 p., Hardcover
ISBN: 978-1-4614-1838-2