Chapter 2

Preliminaries

In Chapter 2.1, a system model is presented. Then, we formulate the similarity of sensor behaviors in Chapter 2.2.

2.1 System Model

In this paper, assume that each sensor is well positioned and is aware of its own location. Sensors are deployed to sense environmental measurements with sampling frequency given [21]. Generally speaking, sensors has its valid sensing readings ranging in $[x_l, x_u]$, where x_l is the lower bound and x_u is the upper bound. The values for the lower and upper bounds either are determined by users or depend on the capability of sensors. Since a sensor has its valid range of sensor reading, sensing readings not fallen into the range are clearly viewed as faulty readings. Therefore, faulty readings dealt with in this paper are arbitrary readings in the valid sensing range $[x_l, x_u]$. These arbitrary readings could be faulty readings reported by faulty sensors or noise readings sensed by normal sensors. Furthermore, a faulty sensor always reports arbitrary readings, while a normal sensor randomly generates noise readings. In this paper, both faulty readings and noise readings are considered to be filtered out so as to improve the data accuracy of query results. In addition, once sensors are identified as faulty

Description	Symbol
The reading behavior of sensor s_i	$b_i(t)$
The length of the reading behavior	Δt
The similarity between two reading behaviors $b_i(t)$ and $b_j(t)$	$sim(b_i(t), b_j(t))$
The trust relation between sensors s_i and s_j	$tr_{i,j}$
SensorRank of sensor s_i	$rank_i$
The vote from sensor s_i to s_j	$vote_i(j)$
The decision made by sensor s_i	dec_i

Table 2.1: The description of notations

sensors that frequently report arbitrary readings, these sensors will not be inquired any more.

2.2 Similarity of Sensor Behaviors

As mentioned before, readings of nearby sensors are correlated [17]. Prior works only take the distance between sensors into consideration in formulating the correlation of sensor behaviors. However, such a similarity function could not truly reflect the similarity of sensor behaviors. For example, two sensors are deployed closely near the window of a building. Assume that one sensor exposes to sun and the other is in shadow of the window. Though the distance of these two sensors is very close, their readings are extremely different. Thus, to truly reflect the similarity of sensor behaviors, the similarity of sensors should be consider the reading behaviors of sensors.

Definition 1. Reading behavior: The reading behavior of sensor s_i consists of a series of sensing readings within a sliding window Δt. The reading behavior of s_i is expressed as $b_i(t) = \{x_i(t - \Delta t + 1), x_i(t - \Delta t + 2), \ldots, x_i(t)\}$, where $x_i(t)$ is the reading sensed by s_i at the time t.

Note that the width of the sliding window is set to Δt for all sensors and assume that the sensing frequency of all sensors are the same. Therefore, the behavior of a sensor is represented as a vector. Since the trends of readings of two sensors are more important, we employ the Extended Jaccard similarity function [22]. In addition, this similarity function has the property that longer vectors turn out to be more tolerant in terms of similarity than smaller vectors. If a reading behavior contains few and small noises, these noises may affect similarity of two reading behaviors. However, if Δt is large, the Extended Jaccard similarity function is more robust against these noise data. The Extended Jaccard similarity function for calculating the similarity of the reading behaviors of two sensors s_i and s_j is expressed by $Jacc\left(b_i\left(t\right), b_j\left(t\right)\right)$ and defined as follows.

$$\begin{aligned} Jacc\left(b_i\left(t\right), b_j\left(t\right)\right) &= \frac{b_i\left(t\right) \cdot b_j\left(t\right)}{\left\|b_i\left(t\right)\right\|_2^2 + \left\|b_j\left(t\right)\right\|_2^2 - b_i\left(t\right) \cdot b_j\left(t\right)}, \\ \text{where } \left\|b_i\left(t\right)\right\|_2^2 &= \left|x_i\left(t - \Delta t + 1\right)\right|^2 + \cdots + \left|x_i\left(t\right)\right|^2. \end{aligned} \tag{2.1}$$

It is expected that the similarity of two sensors should satisfy two criteria. On the one hand, when the readings of two sensors are extremely random, the value of $Jacc\left(b_i\left(t\right), b_j\left(t\right)\right)$ should be close to 0. On the other hand, the value should be set to 1 while reading behaviors of two sensors are exactly the same. Thus, the similarity function is normalized as follows.

$$sim\left(b_i\left(t\right), b_j\left(t\right)\right) = \max\left(0, \frac{Jacc\left(b_i\left(t\right), b_j\left(t\right)\right) - c_{rand}}{1 - c_{rand}}\right), \tag{2.2}$$

where c_{rand} is a constant.

After collecting a sufficient amount of $b_i\left(t\right)$ and $b_j\left(t\right)$, the constant c_{rand} is determined by deriving the average value of $Jacc\left(b_i\left(t\right), b_j\left(t\right)\right)$. For example, assume that $\Delta t = 3$ and the reading range is $[0, 100]$, the average of the similarity of two random reading behaviors is about 0.5. If two reading behaviors are $\{23, 28, 34\}$ and $\{28, 35, 41\}$, the similarity will be $\max\left(0, \frac{0.96-0.5}{1-0.5}\right) = 0.92$.

Note that sensors are required to periodically sense data. Consequently, the reading behaviors of sensors may vary with time. The sensing readings collected recently should be more important than those occurred long time ago. Thus, by exploiting the dynamic feature of reading behaviors, the temporal similarity is defined.

Definition 2. Temporal similarity: The temporal similarity of environments sensed by two sensors s_i and s_j is generally the similarity of $b_i(t)$ and $b_j(t)$. The initial temporal similarity is defined as $s_{temp}^{(0)}(i,j) = sim(b_i(\Delta t - 1), b_j(\Delta t - 1))$, where $b_i(\Delta t - 1) = \{x_i(0), x_i(1), \ldots, x_i(\Delta t - 1)\}$ is the initial reading behavior of s_i. In essence, the temporal similarity will be calculated with a fixed period. Assume that each time we calculate the temporal similarity is referred to as one round. The temporal similarity of $k+1$ round is derived as follows:

$$s_{temp}^{(k+1)}(i,j) = \rho \cdot s_{temp}^{(k)}(i,j) + (1-\rho) \cdot sim(b_i(t), b_j(t)), \tag{2.3}$$

where $\rho \in [0,1]$ is an adjustable variable.

To facilitate the presentation of temporal similarity between s_i and s_j, the most updated temporal similarity is defined the trust relation, denoted as $tr_{i,j}$. In light of trust relations between sensors, the correlation network of sensors is defined as follows:

Definition 3. Correlation network: The correlation network is an overlay network which can be modeled as a graph $G = (V, E)$, where V represents the sensor nodes in the deployment region and $E = \{(s_i, s_j) | s_i, s_j \in V \text{ and } tr_{i,j} > 0\}$. The neighbors of a sensor s_i in the correlation network are defined as $nei(i) = \{s_j | (s_i, s_j) \in E\}$.

Once the correlation network of sensors are constructed, we shall utilize the correlation network to identify the faulty readings when queries are issued. In the correlation network, one can easily justify similarities among sensors. As mentioned before, when an unusual reading is reported by a sensor, this sensor could ask its neighbors to justify whether this reading is faulty or not. However, only using the trust relation could not correctly identify faulty readings in that the most similar sensor nearby may be a faulty sensor, thereby dominating the voting processing. Thus, based on the correlation network devised, we should further develop algorithm SensorRank to rank sensors in terms of the similarity with its neighbors

The value of SensorRank could be a reputation for a sensor when this sensor participates voting processes.