Network Connectivity and Data Quality in Crowd Assisted Networks

Izzet Fatih Senturk[1] and Metin Bilgin[2]

[1]Dept. Computer Engineering, Bursa Technical University, Bursa, TURKEY
[2]Dept. Mechatronic Engineering, Bursa Technical University, Bursa, TURKEY

Abstract

Proliferation of smartphones equipped with several on-board sensors leveraged people-centric sensing. Human mobility, combined with the advanced computing and wireless communication capabilities available on mobile devices, enables mobile crowdsensing and paves the way for a wide range of novel applications. Through public involvement, crowdsensing facilitates large scale sensing tasks through collective intelligence. Inherent crowd mobility, on the other hand, offers ample opportunities such as improving connectivity and coverage in a network. Mobile crowd may even be employed to provide connectivity to an otherwise disconnected network of sensor nodes and enable a crowd assisted network (CrAN). However, dynamic conditions of the crowd also pose a challenge to the quality of the collected data in terms of accuracy, latency, and integrity. Moreover, sampled data can be manipulated by malicious people. In such a case, the application will likely receive conflicting data from different participants. Thus, a truth discovery model is required to resolve data conflicts and determinine the sampled data in order to improve data quality. In this chapter, we focus on connectivity and data quality issues in CrANs and present two novel approaches to maintain data quality while ensuring network connectivity. We also define two define two different metrics to assess the data quality, namely accuracy and integrity. Accuracy evaluates the disparity between the obtained data and the expected data. Depending on the application, minor deviation from the expected data may be acceptable especially if the sample values are incremental (e.g. brightness obtained from a light sensor). On the other hand, certain applications require precise measurements (e.g. directions obtained from a gesture sensor) and we employ integrity metric for this case.

1 Introduction

Ubiquity of various sensors on pervasive computing devices has enabled observing the physical world, real-time, with a plethora of sensors. Thanks to human involvement, mobility which is inherent to human-accompanied devices makes Mobile Crowd Sensing (MCS) possible. Smartphones, wearable devices, and vehicular systems are some of the human-accompanied mobile devices with a variety of on-board sensors as listed in Table 1. Wearables market is already diversified with a dazzling array of products for various applications including entertainment, fitness, medical, etc. More than 400 different wearable devices are already available [1] to change the way we work, exercise, and interact. According to Gartner, the market trend indicates a 90 percent increase by 2021 in the worldwide wearable device sales [2].

Table 1: Mobile devices equipped with sensors.

Sensor	iPhone 8 [3]	Samsung S8 [4]	Garmin Vívoactive 3 [5]	Tesla Model X [6]
GPS	✓	✓	✓	✓
Accelerometer	✓	✓	✓	
Barometer	✓	✓	✓	
Compass		✓	✓	
Gyro	✓	✓		
Proximity	✓	✓		✓
Ambient light	✓	✓		✓
Fingerprint	✓	✓		
Thermometer			✓	
Heart rate		✓	✓	
Iris		✓		
Hall		✓		
Camera	✓	✓		✓
Microphone	✓	✓		✓
Radar				✓

In MCS, generated data is consumer-centric in the sense that a certain degree of user participation is essential at different stages of the application. MCS process can be defined as a series of steps: task allocation, sampling, and data collection.

- *Task allocation:* In the first step, sensing tasks are defined and assigned to participants. Depending on the application, the number of participants may be crucial to provide a certain level of service quality. For instance, while a single participant is sufficient to monitor the movement pattern (i.e. transportation mode) of an individual for a personal health application, the phenomena at a larger scale (e.g. traffic congestion monitoring)

require collective sensing of many individuals. Considering the overhead to be incurred to perform the assigned tasks and the privacy concerns due to revealing sensitive information such as location, people can be reluctant to participate to the system. Therefore, incentive mechanisms should be applied to attract more participants.

- *Sampling:* During the sampling phase, environmental conditions are observed through employed sensors. Sensors to be employed are subject to the sensing context defined by the sensing task. Apart from the sensing context, other requirements such as location, time, and the sampling rate are defined by the sensing task. However, fulfillment of the task request is contingent on the participant's approval. A major concern in this phase is the accuracy of the indicated value at the output of the employed sensor(s). Due to the heterogeneity of devices, accuracy may vary between sensors of different manufacturers. Besides, malicious participants may send manipulated data deliberately without performing the actual task.

  Based on the degree of the user involvement in the sampling phase, crowdsensing can be classified into two categories: participatory crowdsensing and opportunistic crowdsensing. In the participatory sensing, user involvement (participation) is explicit (i.e. human work is required to satisfy the request). On the other hand, sensing tasks are automated and the data is collected without active user involvement in the opportunistic sensing. MCS can also be classified based on the data generation modes. Besides sensors, MCS can also leverage user-contributed data from social networks. However, this chapter focuses on the mobile sensing data.

- *Data collection:* In the last step, sensor readings are collected by the remote server (i.e. data center). Different communication models can be applied based on the wireless communication methods provided to the MCS application by the participant. Note that, despite its availability, participants may not opt to use cellular data (e.g. LTE) considering the communication cost which needs to be covered by individuals and the overhead on the battery. If long range wireless communication means are provisioned, data can be forwarded to the remote server immediately. Though, communication can be limited to WiFi or Bluetooth as well. In such a case, despite progress in sampling, data collection will be postponed until the mobile device is connected to a network. Such limitations introduce delay in data collection which may not be acceptable for some MCS applications. Delay also occurs when the data collection frequency is set lower than the sampling frequency in order to minimize the communication overhead.

User involvement in MCS not only pose challenges but also offers unprecedented opportunities. Unlike traditional sensor networks which require deployment of custom hardware, crowd-sensing applications leverage devices in sheer numbers that are already deployed in the field. This not only avoids the deployment cost of specialized sensing infrastructure but also minimizes the time

required to launch the application. On the other hand, the control on the data quality is rather limited. Two major concerns are the sparsity of participants at a certain location on a given time, and the quality of the generated data in terms of accuracy, integrity, and latency. Considering the fact that rational users control mobile devices, selfish users may be reluctant to participate in crowd sensing applications in order to conserve energy, storage and computing resources. Furthermore, malicious users may generate fabricated data on purpose. Therefore, new methods must be developed to assure a certain degree of data quality while increasing the number of participants by applying incentive mechanisms in conjunction.

MCS employs heterogeneous devices with diverse sensing capabilities, various wireless communication standards, and different mobility models. Besides mobile devices, MCS may also comprise stationary sensor nodes. In such a case, mobile devices can be used to improve some of the network performance metrics such as connectivity and coverage. Network connectivity is a fundamental issue that needs to be tackled in wireless networks and a major concern of this chapter. If sensor nodes form a partitioned network, mobile devices in MCS can be exploited to provide intermittent connectivity between sensor nodes and the remote server. This is similar to MSNs which are intermittently connected through Mobile Data Collectors (MDCs) [7, 8]. The resulting network model is referred to as crowd assisted network (CrAN) and the mobile devices in the network signifies corresponding participants. A sample CrAN can be found in Fig. 1.

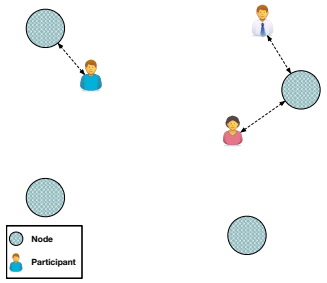


Figure 1: A sensor network with 4 disconnected nodes. Sensor readings cannot be delivered to the remote server due to the limited transmission range. Participants offer intermittent connectivity for otherwise disconnected nodes and enable a crowd assisted network.

This chapter discusses how public crowd can be exploited to intermittently connect nodes in a partitioned network and enable a CrAN. First, we assume a sensor network with disjoint nodes such that none of the nodes are reachable from the rest of the network. Then we introduce humans to the network. Humans offer inherent mobility and the accompanying devices provide a means of wireless communication. Humans, involved in the network, are regarded as participants. Availability of wireless communication combined with inherent

mobility enables employing participants as mobile data collectors. Participants collect data from sensor nodes within their proximity and relay the collected data to the remote server. This scheme provides intermittent connection to the nodes so that the data sampled in the network can be forwarded to the remote server. The resulting intermittently connected network is regarded as crowd-assisted network. Considering the fact that the sampled data can be manipulated or even fabricated by malicious people, we introduce malicious participants into the network. Then we investigate data quality and define two different metrics, namely accuracy and integrity to evaluate the quality of the collected data. In order to resolve data conflicts and determine the actual data that is most likely sampled from individual sensors, we present two novel approaches based on arithmetic average and frequency. Both approaches are evaluated in terms of integrity and accuracy.

The rest of the chapter is organized as follows. We highlight the key differences between wireless/mobile sensor networks and mobile crowd sensing in Section 2. Connectivity and coverage issues in CrANs are discussed in Section 3. Two different approaches to assess reliability of the collected data are presented in Section 4. Proposed approaches are evaluated in Section 5. The chapter is concluded and open issues are discussed in Section 6.

2 From Sensor Networks to MCS

Availability of low cost sensor nodes with wireless communication capabilities has enabled wireless sensor networks (WSNs) comprising sensors in large quantities to monitor an area of interest and track certain events or phenomena. Mobile sensor networks (MSNs) have emerged with the deployment of mobile sensors to take advantage of mobility and sensing at the same time. Unless sensor nodes have inherent mobility capabilities, node mobility can be enabled by attaching nodes to mobile robots [9] as well. Considering the limited onboard batteries of the mobile nodes and excessive energy overhead of mobility compared to other network activities such as messaging [10], mobility should be limited and employed in a controlled manner upon needed to extend the lifetime of the mobile. There are several solutions where mobility is employed as a means for optimizing the network performance in terms of connectivity [11], coverage [12], and lifetime [13, 14]. Mobility is also exploited to tolerate node failures [15].

Compared to sensor networks where energy is constrained and usually non-rechargeable, mobile crowd sensing leverages human-accompanied devices such as smartphones, wearables, and intelligent vehicles which are less restricted in terms of power supply, computational power, communication capacity, memory, and storage. Furthermore, such devices are maintained by users through charging as needed and they usually provide direct access to the internet. Since the cost of ownership is addressed by its users, MCS is more cost-effective compared to MSNs. Also, possibility of user involvement in large numbers offers a scalable and flexible solution that can be easily extended to cover across large areas.

Financial cost of providing the same coverage with traditional sensor networks in large areas renders the application infeasible. For instance, CO_2 monitoring application within the 5th ring in Beijing (about 900 km^2), would require the deployment of at least 90,000 sensor nodes and around 1,000,000 relay nodes to maintain full area coverage and communication connectivity [16] which is undesirable. On the other hand, it is possible to provide 90 percent coverage for the same region by employing 6300 taxis [16].

In the MCS, mobility is inherently exploited through human mobility. This avoids the high mobility cost that exists in MSNs which is much more compared to the messaging cost [10] and therefore should be controlled carefully. In MCS, on the other hand, human-companioned devices with sensing, localization, and wireless communication capabilities collect samples whenever they are within the pre-determined area to be monitored and report their readings to a remote server to enable large-scale sensing tasks. Human mobility not only improves coverage but also offers intermittent connectivity for otherwise disconnected nodes as focused in this chapter.

Several advantages of MCSs exist over MSNs. The primary advantage is the involvement of humans to cover the cost of devices, handle mobility, take care of the communication cost, and maintain devices (e.g. recharging) to sustain their operations. Millions of smart devices and intelligent vehicles already exist and they are ready to be employed around the world. The second advantage is the abundance of resources. Due to the limited form factor of the sensor nodes, WSNs are limited in terms of computation, communication, and energy. Most of the WSNs employ low-rate short range wireless technologies such as IEEE 802.15.4 [17] to communicate within the network and thus network wide collaboration is critical to sustain connectivity with the Base Station (BS) which acts as a gateway between the network and the remote user. However, due to the depletion of limited on-board batteries and the exposure of the nodes to harsh environmental conditions, network can be subject to random node failures. While some of the failures can be compensated with redundancy, failure of the cut-vertex nodes renders the network partitioned. Such problems do not exist, most of the times, for MCS applications since individuals have direct access to the internet through one of the long range wireless communication methods such as LTE [18] or WiFi.

MCS also poses several challenges that need to be addressed. Unlike traditional sensor networks where the number of nodes and their locations are known in advance, controlling data quality is more challenging in MCS. Stability is a major concern typically. The number of participants is expected to fluctuate due to random user mobility. Eagerness to participate may also change based on the actual condition of the device such as battery life and user preferences. Heterogeneity of devices is another challenge. In the ideal case, sensors are expected to produce the same output for the same input. However, sensor readings for the same environmental conditions can be different even for sensors from the same manufacturer depending on the sensor calibration. Ambient conditions surrounding the device and physical alignment with the phenomena to be monitored may also impact the sensor accuracy adversely. Consider an application

where ambient noise is to be monitored. Sensor readings will be subject to the location of the device (e.g. hand, bag, pocket, etc.). Besides, some users may deliberately send false data to earn money without performing the assigned task. Spatial redundancy is another issue which both poses a challenge and offers opportunities. On the one hand, duplicate data from multiple participants must be eliminated. On the other hand, redundancy can be exploited to assess reliability of the collected data. The idea is evaluating disparity in the data collected from the same region by different participants. For a detailed discussion please refer to Section 4. In any case, drastic measures must be taken accordingly to address the mentioned challenges and ensure data quality in terms of accuracy, latency, and integrity.

Another concern from the participants' perspective is privacy. Collected data may contain sensitive information. In general, sensor readings are tagged with location and time. Moreover, collected data can be analyzed to reveal patterns such as trajectories and extract sensitive information such as participants' home and office addresses [19]. Anonymization is an option to provide preservation. Providing anonymity, on the other hand, may encourage users to send incorrect data due to the complexity of taking action on anonymous users. Privacy in MCSs is an open issue and new methods should be developed to ensure user privacy at a certain level.

Besides the risk of privacy issues, users also consume their own resources for data collection. The total cost of ownership includes the initial purchase price of the device, maintenance (e.g. charging the device as needed), mobility cost, and communication cost. To compensate the associated costs and improve participation, incentive mechanisms should be developed. Otherwise users will be reluctant to participate. Several incentive strategies exist which can be classified into entertainment, service, and money [20]. Besides such incentives, social recognition can be another motivating factor for participation.

Some of the key features of MSNs and MCS are summarized in Table 2.

3 Connectivity and Coverage in CrANs, Challenges and Opportunities

Mobile crowd can be employed to provide intermittent connectivity to an otherwise disconnected network of stationary sensor nodes and enable a CrAN. This model offers unprecedented opportunities in network connectivity and coverage. In this section, we briefly describe the challenges regarding connectivity and coverage in MSNs first and then discuss how these issues are addressed by CrANs.

In MSNs, limited energy supplies on the nodes enforce limited transmission range to minimise the communication overhead. Limited transmission range, on the other hand, requires nodes to collaborate with each other in order to send their data to the BS. Therefore, connectivity of the sensor nodes with the BS must be maintained at all times in order to sustain network operations. How-

Table 2: Some of the key differences between MSNs and MCS.

	Mobile sensor networks	Mobile crowd sensing
Operators	Institutions.	Individuals.
Network	Homogeneous network with a fixed network size.	Dynamic network with diverse devices.
Control	Autonomous control with a possible intervention.	User-controlled with limited or no access to the hardware.
Maintenance	Self-organizing. Energy is limited and usually not rechargeable.	User-maintained.
Mobility	Limited autonomous movement.	Inherent mobility with no control.
Context	Limited to the deployed region and the employed hardware.	Location and available sensors may change.
Sampling and Collection	Full control.	User-dependent.
Scalability	Good.	Best.

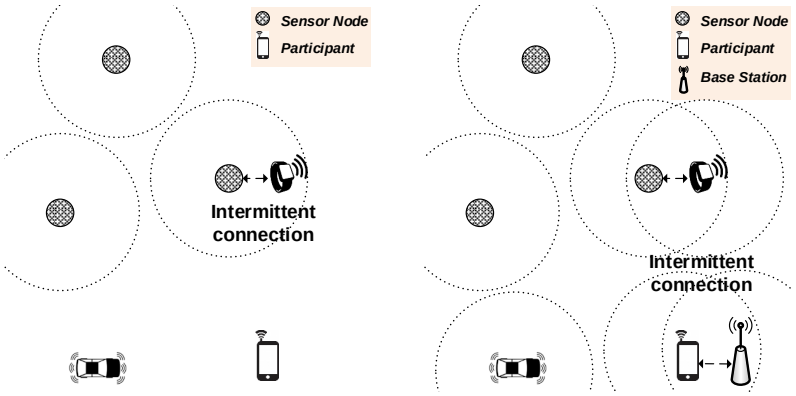
ever, nodes may fail arbitrarily due to various reasons such as battery depletion, hardware malfunction, or an external damage. Such failures may partition the network into disjoint subsets if the failed nodes are cut-vertices. When a partition is isolated from the rest of the network, collected data within the partition cannot be delivered to the BS and the sensing coverage drops drastically.

Several solutions exist in the literature which deals with the connectivity restoration problem through employing mobility. To restore the connectivity of a partition with the rest of the network, network topology should be adjusted accordingly. Three of the most common approaches are as follows:

- *Restructuring network topology through relocation of the existing mobile nodes*: Since mobility imposes significant energy cost on the limited batteries of the nodes, movement distance should be minimized. In addition, if the scope of the damage is too wide, determining the nodes to be relocated and their final locations is another challenge.

- *Deploying additional nodes between the partitions*: In the second approach, determining the minimum number of nodes to be introduced to ensure recovery is crucial. Multiple batches of deployment will be inevitable if the number of deployed nodes is not sufficient to guarantee connectivity. Besides, a self-configuring scheme is required to determine movement destinations of the nodes.

- *Employing mobile data collectors*: MDCs must be assigned to partitions uniformly in such a way that the tour lengths of MDCs are minimized and the load among MDCs are balanced..



(a) Participants enable LTE to forward data obtained from sensor nodes. Participants do not have to collaborate and visit the BS since they have direct internet access to forward the collected data.

(b) Participants employ one of the short range communication methods. Only the BS has direct internet access. Participants should collaborate and visit the BS in order to forward the collected data.

Figure 2

It can be noticed that all solutions pose certain challenges and the primary challenge in MSNs is limited energy. While the first two approaches provide a stable connection, the last one offers intermittent connectivity. Compared to MSNs, we assume sensor nodes to be stationary in CrANs. Therefore, the first approach is not applicable to CrANs. We also assume that additional nodes will not be deployed between the existing nodes to connect them. Thus, the second approach is not an option for CrANs. On the other hand, the last solution is very similar to a CrAN. However, there are a few differences. First of all, mobility is inherent to participants with no additional cost in a CrAN. We also assume no control on the participants' mobility pattern. Therefore, unlike MSNs, we cannot optimize the movement path of a participant and balance the overhead between participants. The lack of control on the mobility simplifies the problem in CrANs. On the other hand, success of the data transmission cannot be guaranteed due to the random mobility. Apparently, the number of participants and the size of the application area for mobility is crucial on the data delivery success. To demonstrate the correlation, we have evaluated network connectivity in CrANs with varying number of participants. The details and further discussion can be found in Section 5.

Despite limited energy supplies of MSNs, CrANs are less restricted in terms of energy. First of all, most of the human-accompanied mobile devices have higher energy capacity compared to tiny sensor nodes. Furthermore, they can be easily recharged as needed. On the other hand, sensor networks often operate unattended in environments where human intervention is limited. Thus, recharging is not an option for MSNs typically. Also, the lack of control on

the hardware and random human mobility render any mobility optimization algorithm inapplicable. Collaboration in message passing, another challenge inherent to MSNs, is not a case most of the times in CrANs. Such a collaboration is not imposed in CrANs since devices are equipped with long range communication technologies such as LTE which provides direct access to the rest of the world. Considering possible limitations on the employed communication method, we demonstrate a scenario where participants need to collaborate in order to forward their data as illustrated in Fig. 2b. Despite its availability, participants may not opt to leverage long range communication means and impose limitations on available communication methods to be employed considering the communication overhead in terms of cost and energy. In the scenario given in Fig. 2b, only low-rate short range wireless communication technologies are assumed to be available for the CrAN. Therefore, participants are employed as mobile data collectors to relay data from nodes to the BS. BS has direct internet access and it is employed as a gateway between the network and the remote server. This scheme requires participants to visit the BS periodically in order to forward the collected data. Assuming limited or no control on the mobility patterns of the participants, data collection rate will decline drastically compared to the direct communication.

Reliability of the collected data is a major challenge in a CrAN. Malicious participants may alter samples they obtain and send falsified data on purpose. The situation can easily deteriorate further if participant collaboration is assumed as in Fig. 2b. Note that, in collaboration networks, one malicious participant has potential to manipulate other participants' data while relaying. In case of malicious participants, the data center will likely receive conflicting data from different participants. If direct internet access is assumed as in Fig. 2a, it can be possible to validate the data sampled from the corresponding sensor node. We present two different approaches in Section 4 for data validation. To demonstrate the correlation between the ratio of malicious participants and the data quality, we have evaluated the proposed data validation approaches in Section 5.

4 Approach

Due to sensor calibration or malicious participants, data center may receive conflicting data from different participants. However, a certain level of data reliability is essential in order to meet application-level objectives. Therefore, we need a truth discovery model so that conflicting data can be resolved. In this chapter, we consider a model where data is sampled by sensor nodes and forwarded by participants. Sensor nodes are assumed to maintain high level of accuracy such that individual nodes sustain producing the same output for the same input. This provides repeatable measurements for the same environmental conditions. In the proposed model, we assume a sample space to represent the set of possible values that can be sampled from sensors. Each sensor node is assumed to report ambient conditions denoted with a value randomly selected

from the sample space. Different sensors may report different values, however, we assume the sampled data to be invariant for individual sensors. As part of the threat model, we assume malicious participants manipulating the actual data obtained from sensors. Since participants are mobile, single malicious participant may alter one or more sensors data depending on the trajectory. Data center may receive true and false values for each sensor node. By analyzing the set of values corresponding to different nodes, we aim to determine the true value for each node by employing data validation methods.

In this section, we introduce two different data validation approaches, namely Arithmetic Average based Data Validation (AADV) and Frequency based Data Validation (FDV) to resolve conflicting data. By analyzing the collected data, we can determinine the value that most likely sampled and improve data quality. We consider the connectivity issue as well in Section 5. But, apart from that, we introduce two different metrics, namely accuracy and integrity to evaluate data quality. To clarify both metrics, let us consider two different scenarios where light sensors and gesture sensors are employed to obtain ambient light information and to detect gestures (e.g. left to right, up to down, etc.) respectively. While precise measurements are required for the gesture sensor, some deviation from the actual value may still be acceptable for the light sensor. Thus, we classify sensors into two categories based on the reported data: incremental and particular. Considering this classification we define accuracy and integrity metrics to evaluate sensors generating incremental and particular samples respectively. The details can be found in Section 5.

Considering privacy concerns, we assume anonymization of the participants. Thus, data center does not collect information regarding participants. Data center only collects sensor readings and the corresponding node id where the data is sampled. Note that, samples of a sensor node may not be received by the data center due to random mobility of the participants. On the other hand, some nodes may be visited multiple times by the same participant or various participants. The more the data is collected from a node, the more the chance is available to determine the actual data. The ratio of malicious participants is also crucial on the success of the data validation.

In the presented model, without loss of generality, five different values are assumed in the sample space: A, B, C, D, and E as given in Table 3. Sample space can represent both incremental and particular data such as *very low*, *low*, *medium*, *high*, and *very high* for ambient light levels or *left to right*, *right to left*, *up to down*, *down to up*, and *wave* for gesture directions. We assume participants to have direct internet access to forward data and therefore no collaboration is required between participants. AADV and FDV approaches are detailed next.

4.1 Arithmetic Average based Data Validation (AADV)

A set of conflicting data may be reported to the data center for the same node due to various reasons such as malicious participants or the lack of sensor accuracy. To maintain data quality, we aim to resolve data conflict and determine the actual data sampled by the corresponding nodes. Thus, characteristics of

Table 3: Notations used in the algorithms.

n_i	Sensor node $i \in N$
R_i	The set of data collected from n_i
v_j	Sample $j \in R_i$
P	The set of possible sensor data values: A, B, C, D, E.
S	The set of actual data reported from corresponding nodes.
A	The set of data determined by AADV for corresponding nodes.
F	The set of data determined by FDV for corresponding nodes.

the obtained data set must be described. One approach is to pursue quantitative methods and apply frequency analysis. The idea is evaluating the number of occurrences of each data and reveal central tendency of the overall data set. This approach enables representing the data set through a single value with the most accuracy. One of the common measures of central tendency is the mean value. While mean has various definitions depending on the context, we consider arithmetic average in this approach. AADV, as the name suggests, applies arithmetic average of the data reported for corresponding nodes. Since numerical data is required for mean, each data in the sample space is mapped to a numerical value starting from one and incremented by one. The algorithm can be found in Algorithm 1.

Algorithm 1 AADV(N, R)

```
1: for i = {1, 2, ..., |N|} do
2:     R_i = getSensorReadings(n_i)
3:     if |R_i| == 0 then                          ▷ No data from n_i
4:         continue
5:     end if
6:     sum = 0
7:     counter = 0
8:     for ∀v_j ∈ R_i do
9:         sum += v_j
10:        counter++
11:    end for
12:    A_i = sum/counter
13: end for
```

4.2 Frequency based Data Validation (FDV)

Mod is another popular measure of central tendency that can be applied to identify a single value to represent the whole data set with the most accuracy. The idea is employing probability density function to determine the data that is most likely to be sampled from corresponding nodes. Thus, we evaluate how frequently each data is reported. If the data appears more, its relative likelihood

to be the actual data is assumed to be increased. FDV, considers mod of the data reported for corresponding nodes and set the data that appears the most as the validated data. The algorithm can be found in Algorithm 2.

Algorithm 2 FDV(N, R)

```
1: for i = {1, 2, ..., |N|} do
2:     R_i = getSensorReadings(n_i)
3:     if |R_i| == 0 then                          ▷ No data from n_i
4:         continue
5:     end if
6:     counts[] = 0
7:     for ∀v_i ∈ R_i do
8:         counts[ordinal(v_i)]++
9:     end for
10:    max = 0
11:    maxIndex = 0
12:    for j = {1, 2, ..., |P|} do
13:        if counts[j] > max then
14:            max = counts[j]
15:            maxIndex = j
16:        end if
17:    end for
18:    F_i = maxIndex
19: end for
```

5 Experimental Evaluation

This section explains the experiment setup, performance metrics and the obtained results.

5.1 Experiment Setup

Efficiency and validity of the presented approaches are tested through simulations. We have considered stationary sensor nodes to monitor the surrounding physical phenomena and a remote server (i.e. data center) to collect and process the data. The nodes are deployed randomly in such a way that none of the nodes are reachable from the rest of the network. To provide intermittent connection between the nodes and the data center, mobile devices are introduced into the network. Mobile devices represent participants in the crowd and the resulting intermittently connected network is regarded as a crowd-assisted network (CrAN). Random way point mobility model is applied to mobile devices.

We have varied the number of nodes (i.e. 4-10), the number of mobiles (i.e. 1-4), and the size of the application area (i.e. 200 meters $\times$ 200 meters - 500

meters × 500 meters) during experiments. For each setup, experiments were carried out 30 times and the average is reported for significance.

5.2 Performance Metrics

We have considered four different performance metrics for assessment.

- *Connectivity*: This metrics reveals the number of nodes that were able to establish intermittent connection to the data center at least once. Higher number of connected nodes denotes improved network coverage.
- *Message Count*: This metric indicates the total number of messages collected from the network and successfully delivered to the data center. Unlike the first metric, message count implies the duration of the network connection.
- *Accuracy*: This metric evaluates the deviation from the expected value. This metric is especially useful to assess data quality when sensors with incremental sample values (e.g. brightness level obtained from ambient light sensor) are employed.
- *Integrity*: This metric signifies whether the collected data is consistent with the generated data. This metric does not tolerate deviation and expects the same value with the actual data. This metric is essentially useful when precise measurements are needed (e.g. directions obtained from a gesture sensor).

Algorithm 3 elaborates how accuracy and integrity metrics are computed.

Algorithm 3 Evaluate_Quality(S, A, F)

```
1: a = 0, v = 0
2: for i = {1, 2, ..., |S|} do
3:     if S_i == A_i then
4:         i++
5:     else
6:         v = |S_i − A_i|
7:     end if
8: end for
9: Accuracy = a/|N|, Integrity = v/|N|
```

5.3 Performance Results

Figs. 3 and 4 present the number of connected nodes with varying number of participants and network size respectively. To observe the relation between the network density and the network connectivity, the size of the application area is varied between 200 meters × 200 meters and 500 meters × 500 meters. The

number of sensor nodes is set to 10 in Fig. 3. As Fig. 3 suggests, network connectivity improves when the number of participants is increased. This is expected since participants follow a random movement pattern and introducing additional participants into the network increases the chance of an encounter with the sensor nodes.

Though, the size of the application area impacts the network connectivity adversely. It can be observed from Fig. 3 that the network connectivity suffers in networks with low node density especially with a single participant. The number of connected nodes declines almost 55% when the size of the application area is increased from 200 meters $\times$ 200 meters to 500 meters $\times$ 500 meters in a network with one participant. On the other hand, for the same scenario, connectivity drops 11% when 4 participants exist in the network. It can be concluded that redundancy in mobility alleviates the adverse effects of the sparse networks.

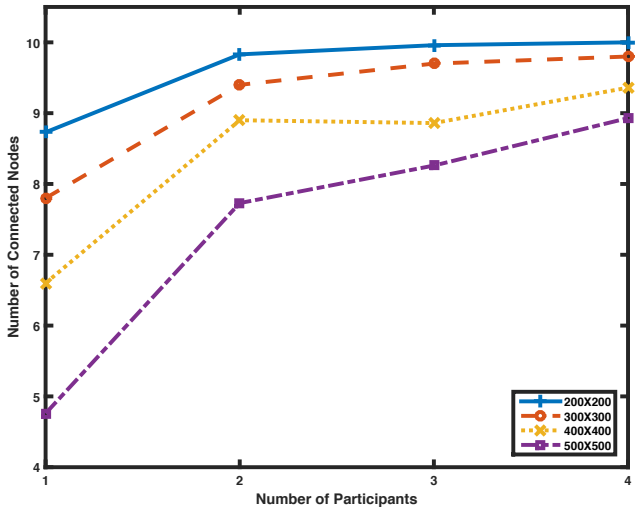


Figure 3: Network connectivity with respect to the number of participants and the size of the monitored application area. The number of nodes is set to 10 while the number of participants is varied between 1 and 4.

In Fig. 4, we employ a single participant and vary the number of nodes and the size of the application area. The results are relative to the node count. Fig. 4 denotes that the number of connected nodes is proportional to the number of nodes in the network even when the application area is varying in size. As expected, network connectivity ratio improves when the network size declines. This is due to the decreased average distance between the nodes which leads to increased chance of visit by a participant.

Total number of messages successfully delivered to the data center are given in Figs. 5 and 5 for varying number of participants and nodes respectively. Figs. 5 reveals that the number of delivered messages increases when the number of participants increases. Size of the application area is inversely propor-

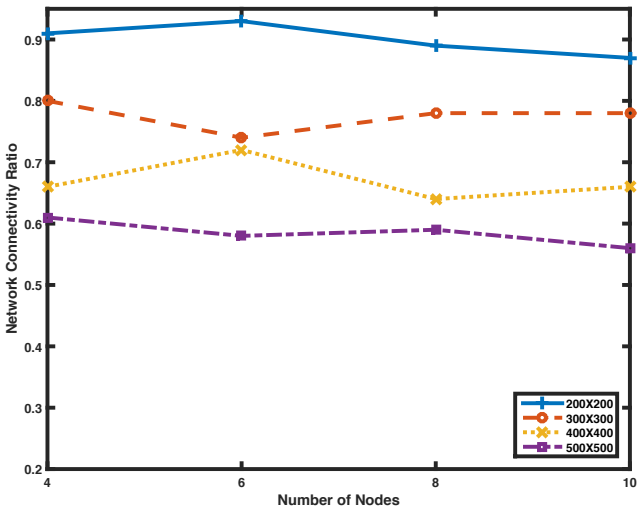


Figure 4: Network connectivity with respect to the number of nodes and the size of the monitored application area. The number of participants is set to 1 while the number of nodes is varied between 4 and 10.

tional to the message count. This can be attributed to the increased average distance between the nodes in sparse networks and the decreased probability of participant-node encounter. As expected, the highest number of message count is attained when the application area is set to 200 meters $\times$ 200 meters.

Figs. 6 suggests that the number of messages successfully delivered to the data center increases with the increased node count. This is expected due to the increased message count generated by the sensor nodes and the improved likelihood of participant-node encounter considering the increased node density when the size of the application area is fixed. On the other hand, improvement in the message count diminishes in larger application areas. If the number of nodes is increased from 4 to 10 when the size of the application area is set to 200 meters $\times$ 200 meters, message count increases 156%. On the other hand, if the size of the application area is set to 500 meters $\times$ 500 meters then the increase in the message count declines to 51%.

We evaluate accuracy and integrity of the collected data next. The number of nodes is set to 10 and the number of participants is set to 4. The size of the application area is 500 meters $\times$ 500 meters. We introduce the concept of malicious participants for the rest of the experiments. Malicious participants are assumed to always manipulate the data they collect from sensors and forward the altered data to the data center. In the following experiments, we considered a sample space with 5 possible values for the sensed data as given in Table 3. Depending on the application, sample space may denote incremental or particular values as discussed earlier. We have employed $AADV$ and FDV approaches to resolve conflicting data and validate the data for corresponding nodes. Success

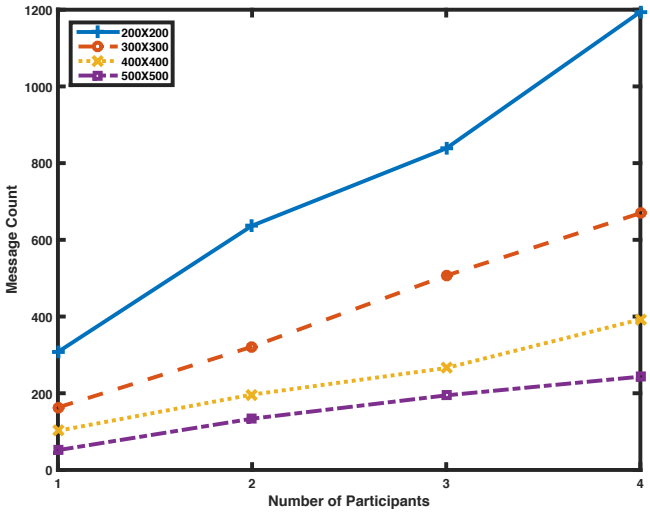


Figure 5: The number of delivered messages with respect to the number of participants and the size of the monitored application area. The number of sensor nodes is set to 10 while the number of participants is varied between 1 and 4.

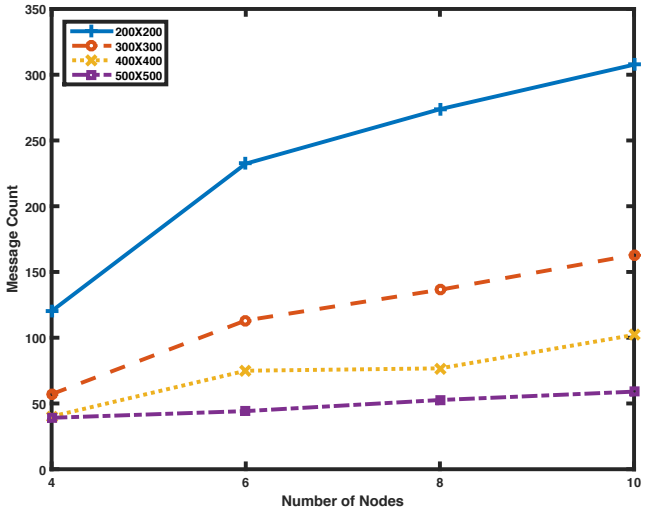


Figure 6: The number of delivered messages with respect to the number of nodes and the size of the monitored application area. The number of participants is set to 1 while the number of nodes is varied between 4 and 10.

rates for the mentioned approaches are illustrated in Figs. 7 and 8 in terms of accuracy and integrity respectively. Integrity evaluates whether the precise data can be obtained. This metric is primarily useful when the data values are not

incremental but particular. On the other hand, accuracy evaluates how close the obtained data is to the original data. For instance, let us assume that the original data is "very light". In terms of integrity, there is no difference whether the validated data is "light" or "heavy" and both will be marked as failure. However, we assess the validated data based on its distance to the original value in accuracy.

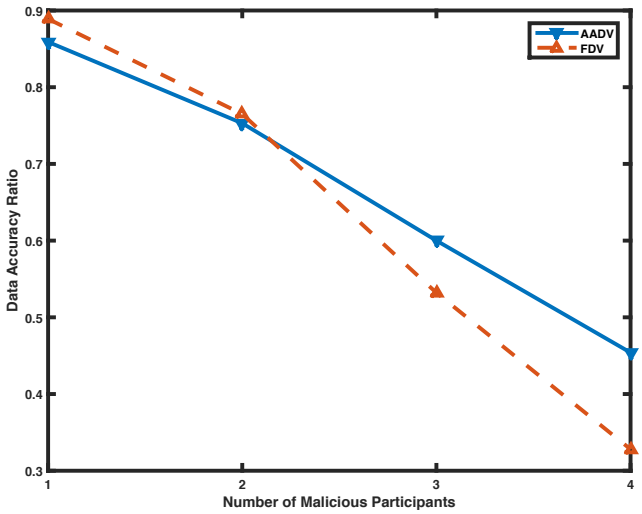


Figure 7: Data accuracy ratio with respect to the number of malicious mobiles when the number of mobiles is set to 4.

Fig. 7 denotes that both approaches perform worse when the malicious activity is increased. Both $AADV$ and FDV provide almost 90% success in validating the data when the number of malicious participants is 1. FDV performs better initially but it is outperformed by $AADV$ when the number of malicious participants is 3 or more. The decline in the performance of the FDV can be attributed to its working principle in validating the data. Recall that FDV considers frequency of the collected data for validation. When the number of honest participants is more than or equal to malicious participants, FDV performs better. However, when the number of malicious participants is increased further, frequency becomes misleading.

Fig. 8 demonstrates the results in terms of integrity of the collected data when malicious participants exist in the network. Initially, FDV reaches a success rate of 85% while $AADV$ provides 71% integrity. However, performance of the FDV declines rapidly and it is outperformed by $AADV$ when the number of malicious participants is more than the half of the total participants. If all the participants are malicious, success rate drops to 0% for FDV. $AADV$, on the other hand, provides 16% integrity.

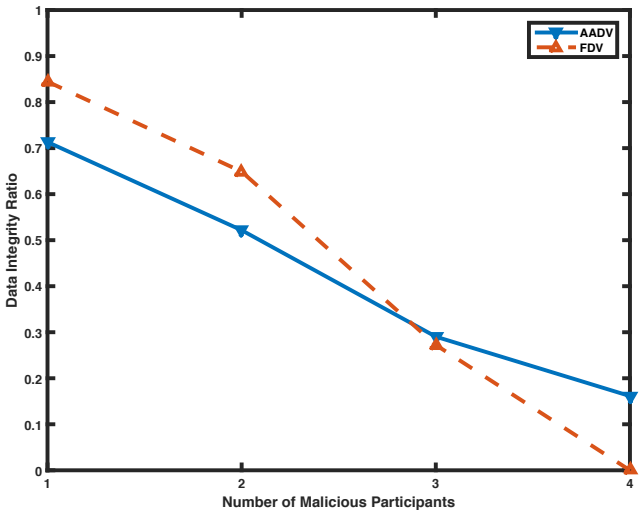


Figure 8: Data integrity ratio with respect to the number of malicious mobiles when the number of mobiles is set to 4.

6 Conclusion and Future Issues

Availability of several sensors on ubiquitous devices such as smartphones, smartwatches, and intelligent vehicles enabled large scale sensing tasks to be carried out by public crowd. On the one hand, this sensing model avoids installation of custom hardware and its maintenance. On the other hand, reliability of the application is highly dependent on the participants. Considering inherent human mobility, participants provide ample opportunities to improve some of the network performance metrics such as connectivity and coverage. Participants may even be employed to provide connectivity to an otherwise disconnected network of sensor nodes and enable a crowd assisted network. However, in the best case, nodes can be intermittently connected with the data center due to random mobility. Depending on the number of participants and their movement trajectories, some of the nodes may never be able to send their data to the data center. Therefore, the number of participants should be increased through incentive mechanisms. Another major challenge is reliability of the participants. In case of malicious participants, the sampled data can be altered. Consequently, the data center will likely receive conflicting data from different participants. In order to provide a certain level of reliability, data conflicts must be resolved and the actual data must be identified. Considering malicious participants, we defined two different metrics, namely accuracy and integrity in order to assess the data quality. While accuracy metric evaluates the disparity between the obtained value and the expected value, integrity metric expects the exact value. Accuracy is useful when sensor data is incremental (e.g. brightness obtained from a light sensor). Integrity, on the other hand, can be employed

when precise measurement is required (e.g. directions obtained from a gesture sensor). To resolve conflicting data, we present two novel approaches based on arithmetic average and frequency considering the defined metrics.

Data quality in CrAN can be assessed in terms of accuracy, integrity, and latency. In this chapter, we investigated accuracy and integrity metrics and we are planning to consider latency as a future work. Latency is a major issue when participants employ only low-rate short range wireless communication means. In such a network, a base station must be present within the network to provide internet connection. In order to deliver sampled data to the data center, participants should visit the base station. This scheme introduces additional delay for the data delivery. Another issue that warrants additional investigation is the mobility model. Besides synthetic mobility models which simulate human behavior, real traces can be obtained and considered in future studies.

References

[1] Vandrico's wearable devices database. `http://vandrico.com/wearables/wearable-technology-database`. Accessed: 2017-09-11.

[2] Vandrico's wearable devices database. `http://www.gartner.com/newsroom/id/3790965`. Accessed: 2017-09-11.

[3] iphone 7 specifications. `https://www.apple.com/iphone-8/specs/`. Accessed: 2017-09-11.

[4] Galaxy s8 specifications. `http://www.samsung.com/global/galaxy/galaxy-s8/specs/`. Accessed: 2017-09-11.

[5] Garmin vivoactive 3 specifications. `https://buy.garmin.com/en-US/US/p/571520#specs`. Accessed: 2017-09-11.

[6] Tesla model x specifications. `https://www.tesla.com/sites/default/files/model_x_owners_manual_north_america_en.pdf`. Accessed: 2017-09-11.

[7] M. Ma and Y. Yang. Data gathering in wireless sensor networks with mobile collectors. In *2008 IEEE International Symposium on Parallel and Distributed Processing*, pages 1–9, April 2008.

[8] I. F. Senturk, K. Akkaya, F. Senel, and M. Younis. Connectivity restoration in disjoint wireless sensor networks using limited number of mobile relays. In *2013 IEEE International Conference on Communications (ICC)*, pages 1630–1634, June 2013.

[9] S. Janansefat, K. Akkaya, I. F. Senturk, and M. Gloff. Rethinking connectivity restoration in wsns using feedback from a low-cost mobile sensor network testbed. In *38th Annual IEEE Conference on Local Computer Networks - Workshops*, pages 108–115, Oct 2013.

[10] G. Wang, G. Cao, T. La Porta, and W. Zhang. Sensor relocation in mobile sensor networks. In *Proceedings IEEE 24th Annual Joint Conference of the IEEE Computer and Communications Societies.*, volume 4, pages 2302–2312 vol. 4, March 2005.

[11] Amitabha Ghosh and Sajal K. Das. Coverage and connectivity issues in wireless sensor networks: A survey. *Pervasive and Mobile Computing*, 4(3):303 – 334, 2008.

[12] Bang Wang, Hock Beng Lim, and Di Ma. A survey of movement strategies for improving network coverage in wireless sensor networks. *Computer Communications*, 32(13):1427 – 1436, 2009.

[13] Stefano Basagni, Alessio Carosi, Emanuel Melachrinoudis, Chiara Petrioli, and Z. Maria Wang. Controlled sink mobility for prolonging wireless sensor networks lifetime. *Wirel. Netw.*, 14(6):831–858, December 2008.

[14] Wei Wang, Vikram Srinivasan, and Kee-Chaing Chua. Using mobile relays to prolong the lifetime of wireless sensor networks. In *Proceedings of the 11th Annual International Conference on Mobile Computing and Networking*, MobiCom '05, pages 270–283, New York, NY, USA, 2005. ACM.

[15] Mohamed Younis, Izzet F. Senturk, Kemal Akkaya, Sookyoung Lee, and Fatih Senel. Topology management techniques for tolerating node failures in wireless sensor networks: A survey. *Computer Networks*, 58:254 – 283, 2014.

[16] H. Ma, D. Zhao, and P. Yuan. Opportunities in mobile crowd sensing. *IEEE Communications Magazine*, 52(8):29–35, Aug 2014.

[17] Low-rate wireless personal area networks. `https://en.wikipedia.org/wiki/IEEE_802.15.4`. Accessed: 2017-09-11.

[18] Long-term evolution. `https://en.wikipedia.org/wiki/LTE_(telecommunication)`. Accessed: 2017-09-11.

[19] D. He, S. Chan, and M. Guizani. User privacy and data trustworthiness in mobile crowd sensing. *IEEE Wireless Communications*, 22(1):28–34, February 2015.

[20] X. Zhang, Z. Yang, W. Sun, Y. Liu, S. Tang, K. Xing, and X. Mao. Incentives for mobile crowd sensing: A survey. *IEEE Communications Surveys Tutorials*, 18(1):54–67, Firstquarter 2016.