TSINGHUA SCIENCE AND TECHNOLOGY
ISSN 1007-0214 05/14 pp44–55
DOI: 10.26599/TST.2019.9010023
Volume 25, Number 1, February 2020

Approximate Data Aggregation in Sensor Equipped IoT Networks

Ji Li, Madhuri Siddula, Xiuzhen Cheng, Wei Cheng, Zhi Tian, and Yingshu Li*

Abstract: As Internet-of-Things (IoT) networks provide efficient ways to transfer data, they are used widely in data sensing applications. These applications can further include wireless sensor networks. One of the critical problems in sensor-equipped IoT networks is to design energy efficient data aggregation algorithms that address the issues of maximum value and distinct set query. In this paper, we propose an algorithm based on uniform sampling and Bernoulli sampling to address these issues. We have provided logical proofs to show that the proposed algorithms return accurate results with a given probability. Simulation results show that these algorithms have high performance compared with a simple distributed algorithm in terms of energy consumption.

Key words: data aggregation; sampling; Internet-of-Things (IoT) networks

1 Introduction

As the urban population snowballs, the smart city has become inevitable to solve many day-to-day problems. These problems include power supply, disaster prediction, and traffic maintenance[1–4]. Some of the smart city applications that are already being used are parking services, intelligent light systems, and water conservation. For the better utilization of natural resources, we should incorporate these applications even in rural areas. The fundamental working principle of a smart city application is that there are various sensors deployed all over the city that are used for collecting data. This data helps us understand information at a city level and hence the data should be well-spread.

Similar to the smart city, we are also focusing on smart home applications[5–10]. These applications are based on the fact that in today's world all the home electronics are connected to the internet. A network with such connected devices is called Internet-of-Things (IoT). Recent devices like Alexa and Google home are built on such a network. These devices interact with all the other devices connected to the same network. Since not all devices are the same, we need a way to collect sensory data from different sensors. The primary objective of any IoT network is to reduce cost and provide faster access to data. One of the distinct challenges in these applications is the deployment of a considerable number of sensing devices.

It is clear that sensors are the building blocks of any IoT network. However, a network with sensors has some drawbacks such as the issue of dynamic traffic, adding new service, adaptive to channel condition, and ever-changing user requirements. Having self-configurable sensors helps address some of these issues. Additionally, many algorithms have been proposed to solve the issues of routing, topology control, and time synchronization[11–24]. Using sensors in IoT networks reduces the communication cost but increases the processing cost. We deploy sensors in our network because they collect data over a long period and could be placed over a long period of time and could be

• Ji Li is with Kennesaw State University, Marietta, GA 30060, USA. E-mail: jli38@kennesaw.edu.
• Madhuri Siddula and Yingshu Li are with Georgia State University, Atlanta, GA 30303, USA. E-mail: msiddula1@student.gsu.edu; yili@gsu.edu.
• Xiuzhen Cheng is with the George Washington University, Washington, DC 20052, USA. E-mail: cheng@gwu.edu.
• Wei Cheng is with the Department of Computer Science, Virginia Commonwealth University, Richmond, VA 23284, USA. E-mail: wcheng3@vcu.edu.
• Zhi Tian is with the Department of Electrical & Computer Engineering, George Mason University, Fairfax, VA 22030, USA. E-mail: ztian1@gmu.edu.
* To whom correspondence should be addressed.
Manuscript received: 2019-05-22; accepted: 2019-05-27

placed over a vast network. Hence, the data collected from these sensors is huge and requires high processing power to aggregate and analyze the data. Hence, if the data aggregation problem is addressed at the sensor level, we do not have to deal with extensive data. However, adding data aggregation functionality to the sensor might consume a lot of sensor's energy. This further raises the energy consumption issue as the aggregation costs much energy and the sensors are not equipped with huge amounts of power supply. According to Ref. [25], cost of transmitting one bit of data is equivalent to the energy cost of executing 1000 instructions. Therefore, reducing data transmission is one of the major ways to decrease the energy cost in IoT. Hence, it is critical to design energy efficient data aggregation methods for sensor equipped IoT networks.

In this paper, we study two kinds of aggregation queries: maximum query and distinct set query. The maximum query is to calculate the maximum of all the sensory data. The distinct set query is to calculate the unique values in the sensory data. Both the queries are critical for an IoT. These two queries can be widely used in practice. For example, in the field of environmental monitoring, the maximum value query can be used to acquire the most serious level of pollution. While the user may get all the pollution levels in the monitored area using the distinct-set query. Therefore, the energy efficient data aggregation model should accommodate both queries in its development.

In practice, exact query results are not always necessary. Approximate query results may also be acceptable for some applications[26,27]. Therefore, in this paper, we propose two algorithms to process approximate maximum queries and distinct-set queries. These algorithms are based on uniform sampling and Bernoulli sampling, respectively. Proposed algorithms will return the exact query results with probability not less than $1-\delta$, where δ is a real number and its value can be arbitrarily small. In summary, the main contributions of our paper can be summarized are as follows:

(1) Mathematical estimators for the two aggregation operations are provided.

(2) The mathematical methods to determine the required sample size and sample probability for calculating approximate maximum value and approximate distinct-set are designed.

(3) Distributed algorithms for approximate maximum value and approximate distinct-set are provided. The energy costs of these algorithms are analyzed.

(4) Extensive simulation results are presented which show the proposed algorithms perform significantly better than a simple distributed algorithm in the aspect of energy consumption.

The rest of the paper is organized as follows. Section 2 defines the problem. Section 3 provides the mathematical proof for the δ-approximate aggregation algorithms. Section 4 explains the proposed δ-approximate aggregation algorithms. Section 5 shows the simulation results and the related works are discussed in Section 6. Section 7 concludes the paper.

2 Problem Definition

Suppose we have an IoT network with n sensor nodes and s_{ti} is the sensory data of node i at time t. $S_t = \{s_{t1}, s_{t2}, \ldots, s_{tn}\}$ is the set of all the sensory data at time t and $Dis(S_t) = \{s^d_{t1}, s^d_{t2}, \ldots, s^d_{t|Dis(S_t)|}\}$ contains the distinct values in S_t. For example, if $S_t = \{s_{t1}, s_{t2}, s_{t3}, s_{t4}, s_{t5}\}$ and $s_{t1} = 1, s_{t2} = 1, s_{t3} = 2, s_{t4} = 3, s_{t5} = 3$, then $Dis(S_t) = \{1, 2, 3\}$.

In this paper, we address maximum and distinct set queries by performing *max* and *distinct set* operations, respectively. The definition of these operations are as follows:

(1) The exact maximum value denoted by $Max(S_t)$ satisfies $Max(S_t) = \max\{s_{ti} \in S_t | 1 \leqslant i \leqslant n\}$.

(2) The exact distinct-set of S_t denoted by $Dis(S_t)$ satisfies $\forall s \in S_t, \exists s^d \in Dis(S_t), s = s^d$, and $\forall s^d_x, s^d_y \in Dis(S_t), x \neq y \Rightarrow s^d_x \neq s^d_y$.

Obviously, the following steps can be used to solve the max and distinct set aggregation problems.

(1) Arrange all the nodes in the network in the form of an aggregation tree where the sink node broadcasts the aggregation operation.

(2) All the nodes submit their sensory data to the sink node along the aggregation tree.

(3) The intermediate nodes in the aggregation tree calculate the partial results during the data transmission.

Although this method results in accurate aggregation results, it will also lead to huge communication and computation cost. Hence, we propose a δ-approximation to the results that can be achieved by the above said aggregation operations. Let I_t and $\widehat{I}_t$ are the accurate and approximate aggregation results at the time "t", respectively. The definition of the δ-estimator is as follows:

Definition 1 (δ-estimator) For any δ ($0 \leqslant \delta \leqslant 1$), $\widehat{I}_t$ is called the δ-estimator of I_t if $\Pr(\widehat{I}_t \neq I_t) \leqslant \delta$.

According to Definition 1, the problem of computing δ-approximate maximum value and δ-approximate distinct-set is defined as follows.

Input: (1) A sensor equipped IoT network; (2) The sensory data set S_t; and (3) Aggregation operator $Agg \in \{Max, Dis\}$ and δ $(0 \leqslant \delta \leqslant 1)$.

Output: δ-approximate aggregation result of Agg.

3 Preliminaries

In this paper, we use two sampling techniques to sample the raw data in the network, which are uniform sampling and Bernoulli sampling, respectively. The preliminaries of computing δ-approximate maximum value and δ-approximate distinct-set are presented in the following subsections.

3.1 Uniform sampling-based approximate aggregation

Let $u_1, u_2, ..., u_m$ to denote m simple random samplings with replacement from sensory data set S_t, $U(m) = \{u_1, u_2, ..., u_m\}$ is a uniform sample of S_t with sample size m, we have the following conclusions.

(1) u_i and u_j are independent with each other for all $1 \leqslant i \neq j \leqslant m$.

(2) $\Pr(u_i = s_{tj}) = \dfrac{1}{n}$ for any $1 \leqslant i \leqslant m$, $1 \leqslant j \leqslant n$.

Based on the above two conclusions, we have the following lemma.

Lemma 1 For any given value $x \in Dis(S_t)$, we have

$$\Pr(x \notin U(m)) = \left(1 - \frac{n_x}{n}\right)^m,$$

where n_x is the number of appearance of value x in S_t.

Proof $\Pr(x \notin U(m)) = \Pr(u_1 \neq x \wedge u_2 \neq x \wedge \ldots \wedge u_m \neq x)$. Since all the samples $u_1, u_2, \ldots, u_m$ are independent with each other, we have

$$\Pr(x \notin U(m)) = \prod_{i=1}^{m} \Pr(u_i \neq x) = (\Pr(u_1 \neq x))^m.$$

Moreover, we have

$$\Pr(u_1 \neq x) = 1 - \Pr(u_1 = x) = 1 - \frac{n_x}{n}.$$

Then this lemma is proved. ■

To obtain δ-approximate maximum value, the mathematical estimator needs to be defined firstly. Let $\widehat{Max(S_t)_u}$ denote the uniform sampling-based estimator of $Max(S_t)$. Then $\widehat{Max(S_t)_u}$ is defined as

$$\widehat{Max(S_t)_u} = Max(U(m)).$$

Based on Lemma 1, we have the following theorem.

Theorem 1 $\widehat{Max(S_t)_u}$ is a δ-estimator of $Max(S_t)$ if

$$m \geqslant \frac{\ln \delta}{\ln\left(1 - \dfrac{n_{min}}{n}\right)},$$

where n_{min} is the number of appearances for the least appearing data.

Proof Based on the condition, we have

$$m \ln\left(1 - \frac{n_{min}}{n}\right) \leqslant \ln \delta,$$

$$\left(1 - \frac{n_{min}}{n}\right)^m \leqslant \delta.$$

According to Lemma 1, we have

$$\Pr(Max(S_t) \notin U(m)) = \left(1 - \frac{n_{Max(S_t)}}{n}\right)^m,$$

where $n_{Max(S_t)}$ is the number of appearance for the maximum value in S_t. Since $n_{Max(S_t)} \geqslant n_{min}$, we have

$$\Pr(Max(S_t) \notin U(m)) \leqslant \left(1 - \frac{n_{min}}{n}\right)^m \leqslant \delta.$$

Then this theorem is proved. ■

Let $\widehat{Dis(S_t)_u}$ denote the uniform sampling-based estimator of exact result $Dis(S_t)$. Then $\widehat{Dis(S_t)_u}$ is defined as

$$\widehat{Dis(S_t)_u} = Dis(U(m)).$$

Based on Lemma 1, we also have the following theorem.

Theorem 2 $\widehat{Dis(S_t)_u}$ is a δ-estimator of $Dis(S_t)$ if

$$m \geqslant \frac{\ln(1 - (1-\delta)^{n_{min}/n})}{\ln\left(1 - \dfrac{n_{min}}{n}\right)}.$$

Proof First, we have

$$\left(1 - \frac{n_{min}}{n}\right)^m \leqslant 1 - (1-\delta)^{n_{min}/n},$$

$$\left(1 - \left(1 - \frac{n_{min}}{n}\right)^m\right)^{n/n_{min}} \geqslant 1 - \delta,$$

$$1 - \prod_{i=1}^{|Dis(S_t)|} \left(1 - \left(1 - \frac{n_{min}}{n}\right)^m\right) \leqslant \delta.$$

Let $n_{s_{ti}^d}$ to denote the number of appearance for s_{ti}^d, then we have

$$1 - \prod_{i=1}^{|Dis(S_t)|} \left(1 - \left(1 - \frac{n_{s_{ti}^d}}{n}\right)^m\right) \leqslant \delta,$$

since $n_{min} \leqslant n_{s_{ti}^d}$. Moreover, according to Lemma 1, we have

$$1 - \prod_{i=1}^{|Dis(S_t)|} (1 - \Pr(s_{ti}^d \notin U(m))) \leqslant \delta,$$

$$1 - \prod_{i=1}^{|Dis(S_t)|} \Pr(s_{ti}^d \in U(m)) \leqslant \delta,$$

$$1 - \Pr(\widehat{Dis(S_t)_u} = Dis(S_t)) \leqslant \delta,$$

$$\Pr(\widehat{Dis(S_t)_u} \neq Dis(S_t)) \leqslant \delta.$$

Then this theorem is proved. ■

3.2 Bernoulli sampling-based approximate aggregation

Let $B(q) = \{b_1, b_2, \ldots, b_{|B(q)|}\}$ denote a Bernoulli sample of data set S_t with sample probability q. Then we have the following lemma.

Lemma 2 For any given value $x \in Dis(S_t)$, we have

$$\Pr(x \notin B(q)) = (1-q)^{n_x},$$

where n_x is the number of appearance of value x in S_t.

Proof Without loss of generality, we assume $s_{t1} = s_{t2} = \cdots = s_{tn_x} = x$, then we have $\Pr(x \notin B(q)) = \Pr(s_{t1} \notin B(q) \wedge s_{t2} \notin B(q) \wedge \cdots \wedge s_{tn_x} \notin B(q))$. Therefore, we have

$$\Pr(x \notin B(q)) = (\Pr(s_{t1} \notin B(q)))^{n_x}.$$

According to the definition of Bernoulli sampling, we have

$$\Pr(s_{t1} \notin B(q)) = 1 - \Pr(s_{t1} \in B(q)) = 1 - q.$$

Then this lemma is proved. ■

Let $\widehat{Max(S_t)_b}$ denote the Bernoulli sampling-based estimator of exact value $Max(S_t)$. $\widehat{Max(S_t)_b}$ is defined as

$$\widehat{Max(S_t)_b} = Max(B(q)).$$

Based on Lemma 2, we have the following theorem.

Theorem 3 $\widehat{Max(S_t)_b}$ is a δ-estimator of $Max(S_t)$ if

$$q \geqslant 1 - (\delta)^{1/n_{min}}.$$

Proof Based on the condition, we have

$$(1-q)^{n_{min}} \leqslant \delta.$$

According to Lemma 2, we have

$$\Pr(Max(S_t) \notin B(q)) = (1-q)^{n_{Max(S_t)}},$$

where $n_{Max(S_t)}$ is the number of appearance for the maximum value in S_t. Since $n_{Max(S_t)} \geqslant n_{min}$, we have

$$\Pr(Max(S_t) \notin B(q)) \leqslant (1-q)^{n_{min}} \leqslant \delta.$$

Then this theorem is proved. ■

Let $\widehat{Dis(S_t)_b}$ denote the Bernoulli sampling-based estimator of exact result $Dis(S_t)$. Then $\widehat{Dis(S_t)_b}$ is defined as

$$\widehat{Dis(S_t)_b} = Dis(B(q)).$$

Based on Lemma 1, we have the following theorem.

Theorem 4 $\widehat{Dis(S_t)_b}$ is a δ-estimator of $Dis(S_t)$ if

$$q \geqslant 1 - (1 - (1-\delta)^{n_{min}/n})^{1/n_{min}}.$$

Proof According to the condition, we have

$$(1-q)^{n_{min}} \leqslant 1 - (1-\delta)^{n_{min}/n},$$

$$(1 - (1-q)^{n_{min}})^{n/n_{min}} \geqslant 1 - \delta,$$

$$1 - \prod_{i=1}^{|Dis(S_t)|} (1 - (1-q)^{n_{min}}) \leqslant \delta.$$

Let $n_{s_{ti}^d}$ denote the number of appearance for s_{ti}^d, since $n_{min} \leqslant n_{s_{ti}^d}$, we have

$$1 - \prod_{i=1}^{|Dis(S_t)|} (1 - (1-q)^{n_{s_{ti}^d}}) \leqslant \delta.$$

Moreover, according to Lemma 2, we have

$$1 - \prod_{i=1}^{|Dis(S_t)|} (1 - \Pr(s_{ti}^d \notin B(q))) \leqslant \delta,$$

$$1 - \prod_{i=1}^{|Dis(S_t)|} \Pr(s_{ti}^d \in B(q)) \leqslant \delta,$$

$$1 - \Pr(\widehat{Dis(S_t)_b} = Dis(S_t)) \leqslant \delta,$$

$$\Pr(\widehat{Dis(S_t)_b} \neq Dis(S_t)) \leqslant \delta.$$

Then this theorem is proved. ■

4 δ-Approximate Aggregation Algorithm

Theorems in Section 3 describe the calculation methods required for sampling size and probability. However, we need to address the following problems:

(1) Broadcasting the sampling information by the sink node to the whole network.

(2) Sampling the sensory data.

(3) Transmission and aggregation of the partial aggregation results.

4.1 Uniform sampling-based aggregation algorithm

One of the naive methods to calculate sample size m can be described as follows:

(1) The sink nodes generate and broadcast m random numbers $\{1, 2, 3, \ldots, n\}$ into the network.

(2) A sensor node identifies itself by the random number sent by the sink node, thereby receiving the sensory data.

This procedure needs huge energy cost due to the broadcasting information transmitted through out network sensors. Hence, we need to develop a mechanism to reduce the energy cost for broadcasting. Therefore, to reduce the energy cost, we cluster the network into "k" clusters $\{C_1, C_2, \ldots, C_k\}$ that are disjoint. By using the method proposed in Ref. [28], we organize the cluster heads in the network as a minimum hop-count spanning tree that has sink node as the root. We then perform uniform sampling algorithm proposed in Ref. [29]. We describe the algorithm as follows:

(1) The sink node generates random numbers Y_i with the probability $\Pr(Y_i = l) = \dfrac{|C_l|}{n}$ $(1 \leqslant i \leqslant m)$.

(2) Let m_l be the sample size of C_l. Then m_l is calculated by $m_l = |\{Y_i | Y_i = l\}|$.

(3) The sink node sends the sample size $\{m_l \mid 1 \leqslant l \leqslant k\}$ to each cluster head. Each cluster head samples the sensory data in its own cluster using the above naive sampling algorithm.

If the sensory data received by the l-th cluster head is $U(m_l)$, it then calculates the partial aggregation result $R(U(m_l))$ based on the aggregation operation Agg by using the following method:

$$R(U(m_l)) = \begin{cases} Max(U(m_l)), \text{ if } Agg = Max; \\ Dis(U(m_l)), \text{ elsewhere.} \end{cases}$$

Then $R(U(m_l))$ is transmitted along the spanning tree. To further reduce the transmission cost, the intermediate nodes also aggregate the received partial result while transmitting the sensory data. The whole process is explained in Algorithm 1.

According to the content in Section 3.1, we have

$$m = \begin{cases} \left\lceil \dfrac{\ln \delta}{\ln(1 - \frac{n_{min}}{n})} \right\rceil, & \text{if } Agg = Max; \\ \left\lceil \dfrac{\ln(1 - (1-\delta)^{n_{min}/n})}{\ln(1 - \frac{n_{min}}{n})} \right\rceil, & \text{if } Agg = Dis. \end{cases}$$

Therefore, we have

$$m = \begin{cases} O\left(\ln \dfrac{1}{\delta}\right), & \text{if } Agg = Max; \\ O\left(\ln\left(\dfrac{1}{1 - (1-\delta)^{n_{min}/n}}\right)\right), & \text{if } Agg = Dis. \end{cases}$$

In practice, $|R_j|$ can be regarded as a constant. According to Ref. [29], the communication cost and the energy cost of the above algorithm is $O(\ln \dfrac{1}{\delta})$ if $Agg = Max$, while the cost is $O\left(\ln\left(\dfrac{1}{1 - (1-\delta)^{n_{min}/n}}\right)\right)$ if $Agg = Dis$. In practice, the value of n_{min} can be acquired by the background knowledge of the specific applications. For example, in the field of environmental monitoring, the user can get the value of n_{min} according to the historical data.

Algorithm 1 Uniform sampling-based aggregation algorithm

Input: δ, aggregation operator $Agg \in \{Max, DistinctSet\}$
Output: δ-approximate aggregation results

if $Agg = Max$ **then**
  $m = \lceil \frac{\ln \delta}{\ln(1 - \frac{n_{min}}{n})} \rceil$
else
  $m = \lceil \frac{\ln(1-(1-\delta)^{n_{min}/n})}{\ln(1 - \frac{n_{min}}{n})} \rceil$
end if
generate Y_i following $\Pr(Y_i = l) = \frac{|C_l|}{n}$,
$m_l = |\{Y_i \mid Y_i = l\}|$ $(1 \leqslant i \leqslant m, 1 \leqslant l \leqslant k)$, the sink sends m_l to each cluster head by multi-hop communication
for each cluster head of the clusters C_l $(1 \leqslant l \leqslant k)$ **do**
  generate random numbers $k_1, k_2, \ldots, k_{m_l}$ then broadcast inside the cluster
end for
for each cluster member of C_l $(1 \leqslant l \leqslant k)$ **do**
  send sensory value to cluster head if its id belongs to $\{k_1, k_2, \ldots, k_{m_l}\}$
end for
for each cluster head of the clusters C_l $(1 \leqslant l \leqslant k)$ **do**
  receive sample data $U(m_l)$ and calculate partial result $R(U(m_l))$
end for
for each node j in the spanning tree **do**
  if j is the leaf node **then**
    Send R_j to its parent node
  else
    Receive partial results $R_{j1}, R_{j2}, \ldots, R_{jc}$ from its children
    if $Agg = Max$ **then**
      $R_j = \max(R_{j1}, R_{j2}, \ldots, R_{jc})$
    else
      $R_j = \bigcup_{i=1}^{c} R_{ji}$
    end if
    if j is the sink node **then**
      return R_j
    else
      Send R_j to its parent node
    end if
  end if
end for

4.2 Bernoulli sampling-based aggregation algorithm

Unlike the uniform sampling-based aggregation algorithm, the sampling information of Bernoulli sampling-based aggregation algorithm utilizes only the sampling probability q. Additionally, Bernoulli-based method provides a mechanism for each node in the network to do the sampling independently. Therefore, the following steps are used in the Bernoulli sampling-based aggregation algorithm to perform sampling and the network need not be divided into clusters.

(1) Sink node broadcasts the sampling probability q in the network.

(2) Each node generates a random number *rand* in the range of [0,1], submit its sensory data to the parent node if $rand < q$.

When the intermediate nodes in the spanning tree receive the submitted sensory data, they will calculate the partial aggregation results using a similar method introduced in Section 4.1. These nodes then transmit the

partial results along the spanning tree. Similarly, during the process of transmitting partial aggregation results to the sink node along the spanning tree, the intermediate nodes in the spanning tree aggregate the received partial results. The process mentioned above is explained in detail in Algorithm 2.

According to the analysis in Section 3.2, for the sample probability q, we have

$$q = \begin{cases} 1-(\delta)^{1/n_{min}}, & \text{if } Agg = Max; \\ 1-(1-(1-\delta)^{n_{min}/n})^{1/n_{min}}, & \text{if } Agg = Dis. \end{cases}$$

Similarly, the communication cost and the energy cost of the Bernoulli sampling-based δ-approximate aggregation algorithm is $O(n - n(\delta)^{1/n_{min}})$ if $Agg = Max$, while the cost is $O(n - n(1-(1-\delta)^{n_{min}/n})^{1/n_{min}})$ if $Agg = Dis$.

Algorithm 2 Bernoulli sampling-based aggregation algorithm

Input: δ, aggregation operator $Agg \in \{Max, Dis\}$
Output: δ-approximate aggregation results

if $Agg = Max$ **then**
  $q = 1-(\delta)^{1/n_{min}}$
else
  $q = 1-(1-(1-\delta)^{n_{min}/n})^{1/n_{min}}$
end if
Sink node broadcasts q in the network
for each leaf node j in the spanning tree **do**
  if $rand < q$ **then**
    Send its own sensory data to its parent node;
  end if
end for
for each non-leaf node j in the spanning tree **do**
  Receive partial results $R_{j1}, R_{j2}, \ldots, R_{jc}$ from its children
  if $Agg = Max$ **then**
    $R_j = \max(R_{j1}, R_{j2}, \ldots, R_{jc})$
  else
    $R_j = \bigcup_{i=1}^{c} R_{ji}$
  end if
  if $rand < q$ **then**
    if $Agg = Max$ **then**
      $R_j = max(R_j, j.data)$
    else
      $R_j = R_j \cup \{j.data\}$
    end if
  end if
  if j is the sink node **then**
    return R_j
  else
    Send R_j to its parent node
  end if
end for

5 Simulation Results

In order to evaluate the proposed algorithms, we simulated an IoT network with 1000 nodes. All nodes are randomly distributed in a 300 m × 300 m rectangular region and the sink node is in the center of the region. The following steps are used to define the clusters.

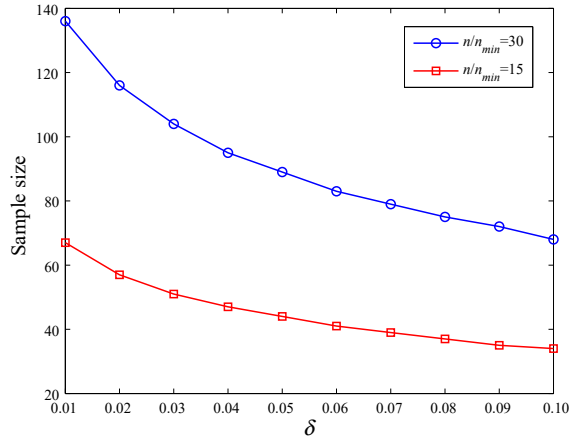
(1) Divide the region into 10×10 grids.
(2) Group the nodes in the same gird into one cluster.
(3) The cluster head is randomly chosen.

For each node, the energy cost to send and receive one byte is defined as 0.0144 mJ and 0.0057 mJ, respectively[30]. The communication range of each sensor node is set to be $30\sqrt{2}$ m in our simulation[31]. By these simulation settings, we ensure that each sensor node at a one-hop distance from its corresponding cluster head.

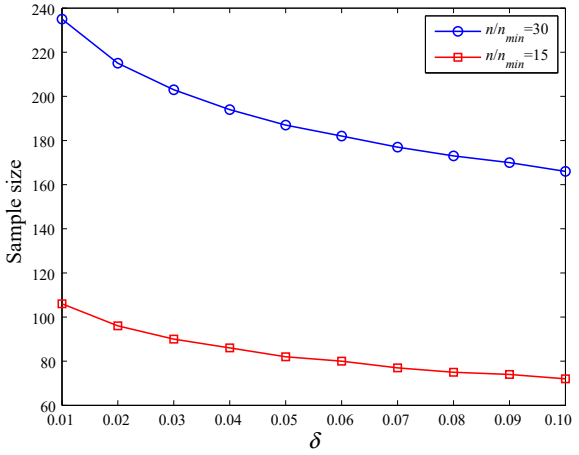
5.1 Uniform sampling-based aggregation algorithm

The first group of simulations is to study the relationship between δ and the sample size. The results are shown in Fig. 1. The results for both the maximum value aggregation and the distinct-set aggregation are listed. Additionally, two groups of results with different $\frac{n}{n_{min}}$ are also listed for comparison. These results indicate that the sample size increases with a decline of δ. Moreover, the sample sizes are much smaller than that of the network. For example, if we have $\delta = 0.01$ and $\frac{n}{n_{min}} = 15$, the sample size is about 67, which indicates that we only need to sample 6.7% sensory data to guarantee that the estimated maximum value being equal to the actual maximum value with the probability greater than 99%. Hence, the proposed algorithm based on uniform sampling preserves a tremendous amount of energy as the amount of sensory data sampled is little. Additionally, in the same condition, the sample size for the distinct-set aggregation is greater than the sample size for maximum value aggregation. Hence, we have to ensure that the distinct-set aggregation has all distinct values that are sampled.

The second group of simulations is to study the relationship between δ and the energy cost. The results are listed in Fig. 2. These results indicate that the energy cost increases with the decrease of δ. It can be observed that the energy cost for the distinct-set aggregation is higher than that of the maximum value aggregation as the distinct-set aggregation requires sample size.

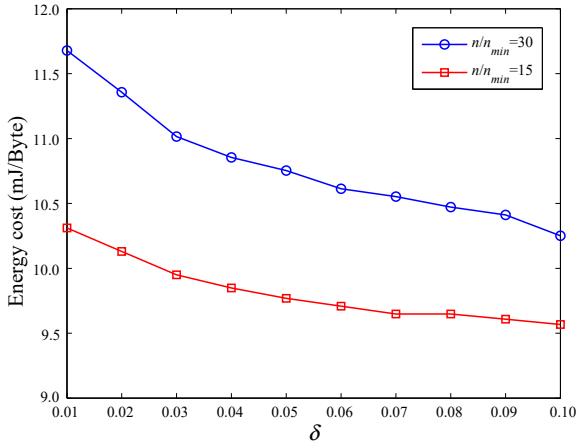


(a) Maximum value aggregation

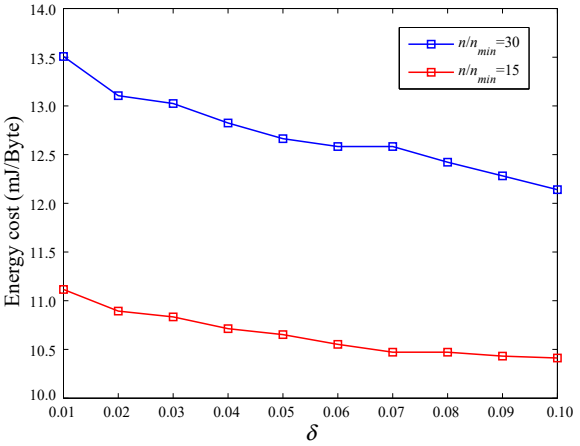


(b) Distinct-set aggregation

Fig. 1 Relationship between δ and the sample size.



(a) Maximum value aggregation



(b) Distinct-set aggregation

Fig. 2 Relationship between δ and the energy cost for the uniform sampling-based aggregation algorithm.

The third group of simulation is to compare the energy cost between the uniform sampling-based aggregation algorithm and the simple distributed algorithm. The steps of the simple distributed algorithm are as follows.

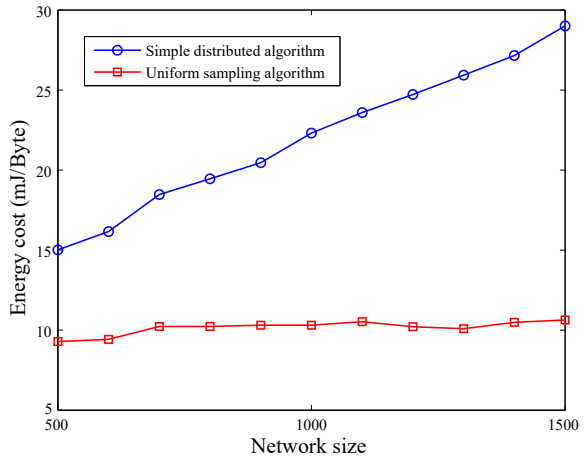
(1) Collect all the raw sensory data.

(2) Aggregate the partial results during the transmission.

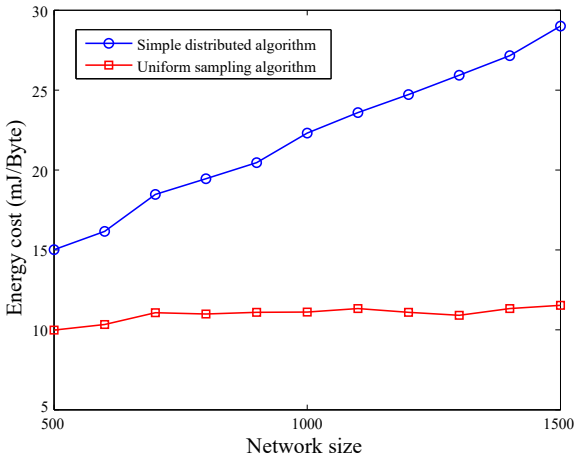
We can see that the simple distributed algorithm can always return accurate aggregation results. For the uniform sampling-based aggregation algorithm, we set $\delta = 0.1$, $\frac{n}{n_{min}} = 15$, and the network size changes from 500 to 1500. The results are listed in Fig. 3. We can see that for all these two algorithms, the energy cost increases with the increase of the network size. Moreover, the energy cost of the uniform sampling-based aggregation algorithm is much lower than that of the naive distributed algorithm since only a small number of nodes need to transmit their sensory data. These results indicate that the uniform sampling-based aggregation algorithm performs much better in energy cost although it may return wrong aggregation results. Finally, with the increase in the network size, the energy cost of the simple distributed algorithm proliferates, while the energy cost for the uniform sampling-based aggregation algorithm almost remains the same. That is because the uniform sampling algorithm's required sample size depends on the value of δ and $\frac{n_{min}}{n}$ rather than the network size n itself. This phenomenon also indicates that the uniform sampling algorithm is appropriate for large scale networks, which is verified by the results shown in Fig. 4.

5.2 Bernoulli sampling-based aggregation algorithm

The first group of simulations is about the relationship between δ and the sample probability. The results are presented in Fig. 5. The results show that the

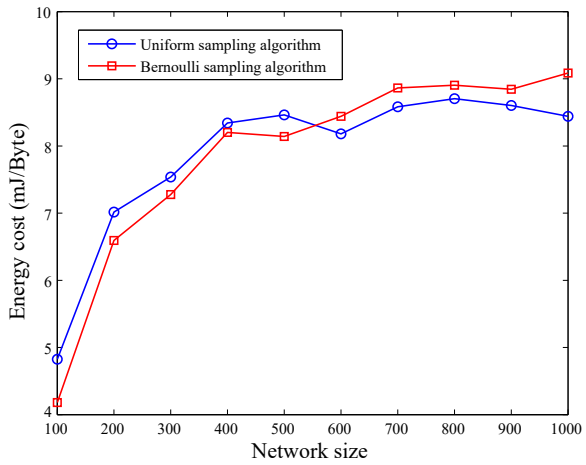


(a) Maximum value aggregation

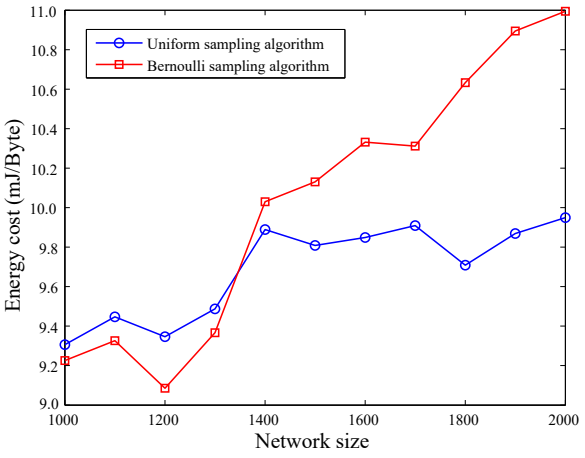


(b) Distinct-set aggregation

Fig. 3 Energy cost comparison between the uniform sampling-based aggregation algorithm and the simple distributed algorithm.



(a) Maximum value aggregation



(b) Distinct-set aggregation

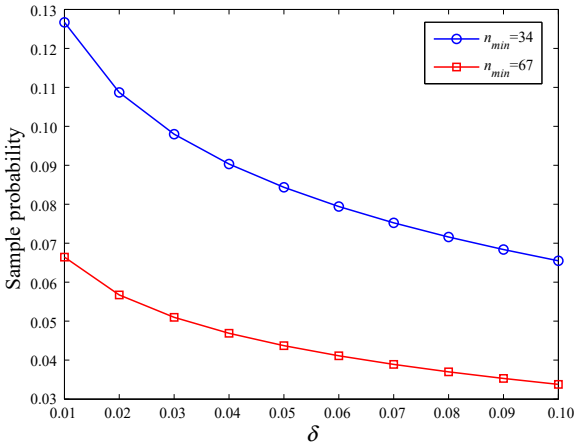
Fig. 4 Energy cost comparison between the uniform sampling-based aggregation algorithm and Bernoulli sampling-based aggregation algorithm.

sample probability increases with the decline of δ. Moreover, the sample probabilities are much smaller than 1. For example, when $\delta = 0.01$, the sample probability is about 0.066 for deriving δ-approximate maximum value. Therefore, our Bernoulli sampling-based algorithm also saves a great deal of energy. Similarly, the required sample size for the distinct-set aggregation is greater than that of the maximum value aggregation in the same condition.

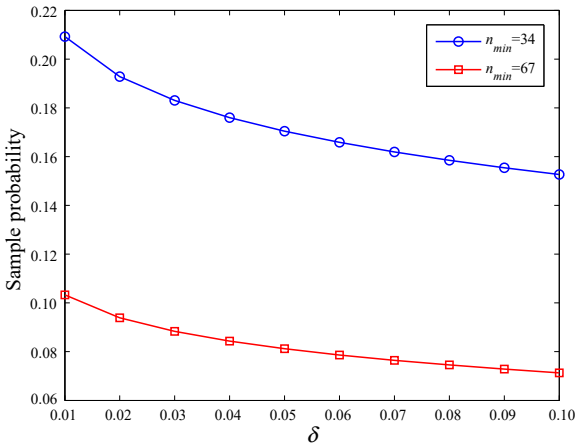
The second group of simulations is about the relationship between δ and the energy cost. The results are shown in Fig. 6. Similarly, we can see that the energy cost increases with the decline of δ and the energy cost for the distinct-set aggregation is greater than that of the maximum value aggregation.

The third group of simulation is to compare the energy cost between the Bernoulli sampling-based aggregation algorithm and the simple distributed algorithm. For the Bernoulli sampling-based aggregation algorithm, we set $\delta = 0.1$ and $n_{min} = 67$. The network size varies from 500 to 1500. The results are listed in Fig. 7. Similarly, we can see for the same network size, the energy cost of the Bernoulli sampling-based aggregation algorithm is much lower than that of the simple distributed algorithm which indicates that Bernoulli sampling-based aggregation algorithm has high performance on energy consumption. Moreover, we can also see that the Bernoulli sampling-based aggregation algorithm has even better performance on large scale networks.

The fourth group of simulation is to compare the energy cost between the Bernoulli sampling-based aggregation algorithm and the uniform sampling-based aggregation algorithm. We set $\delta = 0.1$ and $\frac{n}{n_{min}} = 15$.

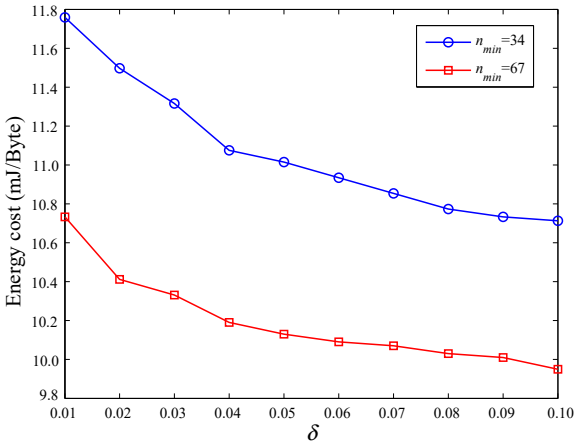


(a) Maximum value aggregation

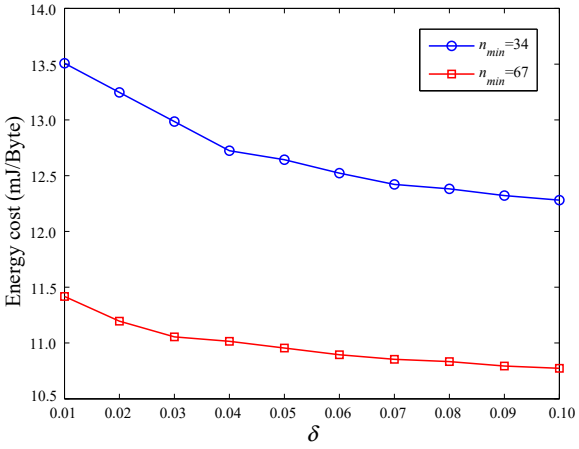


(b) Distinct-set aggregation

Fig. 5 Relationship between δ and the sample probability.



(a) Maximum value aggregation



(b) Distinct-set aggregation

Fig. 6 Relationship between δ and the energy cost for the Bernoulli sampling-based aggregation algorithm.

In order to ensure the network connectivity when the network size is small, we set node's communication to 60 m for this group of simulation. The results are shown in Fig. 4. We can see that for both the uniform sampling-based aggregation algorithm and the Bernoulli sampling-based aggregation algorithm, the energy cost increases with the increase of network size. Moreover, the Bernoulli sampling-based aggregation algorithm has lower energy cost when the network size is small, while the uniform sampling-based aggregation algorithm has lower energy cost when the network size is large. From the above results, we can see the Bernoulli sampling-based aggregation algorithm has the following advantages.

(1) The Bernoulli sampling-based aggregation algorithm can be used in unclustered networks.

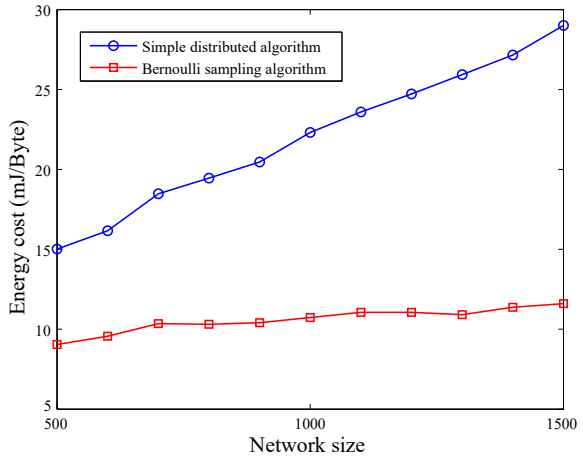
(2) The Bernoulli sampling-based aggregation algorithm has lower energy cost in small scale networks.

While on the other hand, the uniform sampling algorithm is appropriate for large scale clustered networks.

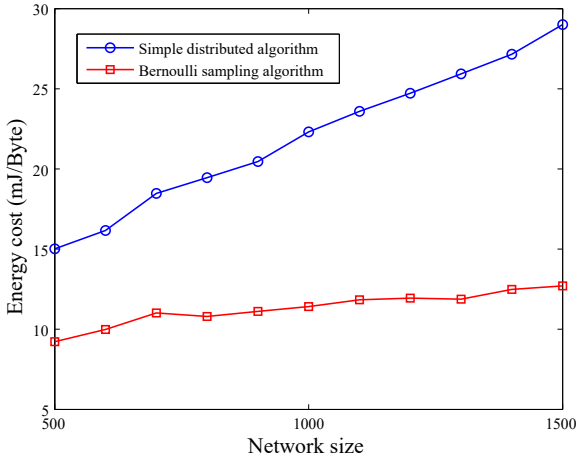
6 Related Work

The sampling technique has been widely used, such as quantile calculation, data collection, and top-k query. For example, Ref. [32] is about an approximate algorithm to calculate the quantiles in wireless sensor networks. By using the sampling technique, Ref. [33] develops ASAP, which is an adaptive sampling-based method to do energy-efficient data collection in sensor networks. Reference [34] uses samples of past sensory data to define the problem of optimizing approximate top-k queries. However, all these techniques cannot be used in our problem because these operations differ a lot with the maximum query and distinct-set query.

The distinct-count query has been widely studied in many works, such as Refs. [35, 36]. Reference



(a) Maximum value aggregation



(b) Distinct-set aggregation

Fig. 7 Energy cost comparison between the Bernoulli sampling-based aggregation algorithm and simple distributed algorithm.

[35] introduces an algorithm to calculate approximate distinct-count based on approximate frequency query results. Reference [37] is about range count queries in big IoT data. Reference [36] is about an algorithm to compute the approximate distinct-count. However, this algorithm is centralized and not appropriate for IoT networks. Moreover, all these works are for the distinct-count query, which is about the size of the distinct set rather than the content of the distinct set. Therefore, the above works still cannot be used in our problem.

7 Conclusion

In this paper, the approximate algorithms for the maximum value aggregation and distinct-set aggregation operations in sensor equipped IoT networks are proposed. These algorithms are based on the uniform sampling and Bernoulli sampling, respectively. We have proposed mathematical estimators for the two algorithms. Moreover, we have derived the values for the required sample size and the required sample probability for any given δ. Finally, an algorithm based on uniform sampling and an algorithm based on Bernoulli sampling are provided. Simulation results are shown for various δ values and the network sizes. These simulation results indicate that the proposed algorithms have high performance in terms of the energy cost.

Acknowledgment

This work was partly supported by the National Science Foundation (NSF) (Nos. 1741277, 1741287, 1741279, 1851197, and 1741338).

References

[1] Z. P. Cai, X. Zheng, and J. G. Yu, A differential-private framework for urban traffic flows estimation via taxi companies, *IEEE Trans. Ind. Inf.*, doi: 10.1109/TII.2019.2911697.

[2] Z. P. Cai and X. Zheng, A private and efficient mechanism for data uploading in smart cyber-physical systems, *IEEE Trans. Netw. Sci. Eng.*, doi: 10.1109/TNSE.2018.2830307.

[3] X. Zheng, Z. P. Cai, and Y. S. Li, Data linkage in smart internet of things systems: A consideration from a privacy perspective, *IEEE Commun. Mag.*, vol. 56, no. 9, pp. 55–61, 2018.

[4] Y. Liang, Z. P. Cai, J. G. Yu, Q. L. Han, and Y. S. Li, Deep learning based inference of private information using embedded sensors in smart devices, *IEEE Netw. Mag.*, vol. 32, no. 4, pp. 8–14, 2018.

[5] Y. Huo, C. Q. Hu, X. W. Qi, and T. Jing, LoDPD: A location difference-based proximity detection protocol for fog computing, *IEEE Internet Things J.*, vol. 4, no. 5, pp. 1117–1124, 2017.

[6] Y. Huo, C. T. Yong, and Y. F. Lu, Re-ADP: Real-time data aggregation with adaptive ω-event differential privacy for fog computing, *Wirel. Commun. Mobile Comput.*, vol. 2018, pp. 1–13, 2018.

[7] Y. K. Wen, Y. Huo, L. R. Ma, T. Jing, and Q. H. Gao, A scheme for trustworthy friendly jammer selection in cooperative cognitive radio networks, *IEEE Trans. Veh. Technol.*, vol. 68, no. 4, pp. 3500–3512, 2019.

[8] Y. Huo, M. Xu, X. Fan, and T. Jing, A novel secure relay selection strategy for energy-harvesting-enabled internet of things, *EURASIP J. Wirel. Comm.*, vol. 2018, p. 264, 2018.

[9] Y. Q. Jia, Y. Chen, X. S. Dong, P. Saxena, J. Mao, and Z. K. Liang, Man-in-the-browser-cache: Persisting https attacks via browser cache poisoning, *Comput. Secur.*, vol. 55, pp. 62–80, 2015.

[10] J. Mao, S. S. Zhu, J. D. Bian, Q. X. Lin, and J. W. Liu, Anomalous power-usage behavior detection from smart home wireless communications, *J. Commun. Inf. Netw.*, vol. 4, no. 1, pp. 13–23, 2019.

[11] C. Schurgers and M. B. Srivastava, Energy efficient routing in wireless sensor networks, in *2001 MILCOM Proc. Communications for Network-Centric Operations: Creating the Information Force*, McLean, VA, USA, 2001, pp. 357–361.

[12] S. Y. Cheng, Z. P. Cai, J. Z. Li, and H. Gao, Extracting kernel dataset from big sensory data in wireless sensor networks, *IEEE Trans. Knowl. Data Eng.*, vol. 29, no. 4, pp. 813–827, 2017.

[13] S. Y. Cheng, Z. P. Cai, J. Z. Li, and X. L. Fang, Drawing dominant dataset from big sensory data in wireless sensor networks, in *Proc. 2015 IEEE Conf. Computer Communications*, Kowloon, China, 2015, pp. 531–539.

[14] S. Y. Cheng, Z. P. Cai, and J. Z. Li, Curve query processing in wireless sensor networks, *IEEE Trans. Veh. Technol.*, vol. 64, no. 11, pp. 5198–5209, 2015.

[15] S. Y. Cheng, J. Z. Li, and Z. P. Cai, $O(\varepsilon)$-approximation to physical world by sensor networks, in *Proc. 32nd Ann. IEEE Int. Conf. Computer Communications*, Turin, Italy, 2013, pp. 3084–3092.

[16] Z. B. He, Z. P. Cai, S. Y. Cheng, and X. M. Wang, Approximate aggregation for tracking quantiles and range countings in wireless sensor networks, *Theor. Comput. Sci.*, vol. 607, pp. 381–390, 2015.

[17] X. Zheng and Z. P. Cai, Real-time big data delivery in wireless networks: A case study on video delivery, *IEEE Trans. Ind. Inf.*, vol. 13, no. 4, pp. 2048–2057, 2017.

[18] X. Zheng, Z. P. Cai, J. Z. Li, and H. Gao, A study on application-aware scheduling in wireless networks, *IEEE Trans. Mobile Comput.*, vol. 16, no. 7, pp. 1787–1801, 2017.

[19] J. G. Yu, Q. B. Zhang, D. X. Yu, C. C. Chen, and G. H. Wang, Domatic partition in homogeneous wireless sensor networks, *J. Netw. Comput. Appl.*, vol. 37, pp. 186–193, 2014.

[20] J. G. Yu, X. L. Ning, Y. C. Sun, S. L. Wang, and Y. W. Wang, Constructing a self-stabilizing CDS with bounded diameter in wireless networks under SINR, in *Proc. IEEE INFOCOM 2017-IEEE Conf. Computer Communications*, Atlanta, GA, USA, 2017, pp. 1–9.

[21] J. G. Yu, B. G. Huang, X. Z. Cheng, and M. Atiquzzaman, Shortest link scheduling algorithms in wireless networks under the SINR model, *IEEE Trans. Veh. Technol.*, vol. 66, no. 3, pp. 2643–2657, 2017.

[22] S. L. Wang, X. Wang, X. Z. Cheng, J. H. Huang, R. F. Bie, and F. Zhao, Fundamental analysis on data dissemination in mobile opportunistic networks with levy mobility, *IEEE Trans. Veh. Technol.*, vol. 66, no. 5, pp. 4173–4187, 2017.

[23] Y. Wang, Topology control for wireless sensor networks, in *Wireless Sensor Networks and Applications*, Y. S. Li, M. T. Thai, and W. L. Wu, eds. Springer, 2008, pp. 113–147.

[24] J. Elson and D. Estrin, Time synchronization for wireless sensor networks, in *Proc. 15th Int. Parallel and Distributed Processing Symp.*, San Francisco, CA, USA, 2001.

[25] J. B. Li and J. Z. Li, Data sampling control, compression and query in sensor networks, *Int. J. Sens. Netw.*, vol. 2, nos. 1&2, pp. 53–61, 2007.

[26] J. Considine, F. Li, G. Kollios, and J. Byers, Approximate aggregation techniques for sensor databases, in *Proc. 20th Int. Conf. Data Engineering*, Boston, MA, USA, 2004, pp. 449–460.

[27] G. Hartl and B. C. Li, infer: A Bayesian inference approach towards energy efficient data collection in dense sensor networks, in *Proc. 25th IEEE Int. Conf. Distributed Computing Systems*, Columbus, OH, USA, 2005, pp. 371–380.

[28] R. Lachowski, M. E. Pellenz, M. C. Penna, E. Jamhour, and R. D. Souza, An efficient distributed algorithm for constructing spanning trees in wireless sensor networks, *Sensors*, vol. 15, no. 1, pp. 1518–1536, 2015.

[29] S. Y. Cheng and J. Z. Li, Sampling based (epsilon, delta)-approximate aggregation algorithm in sensor networks, in *Proc. 29th IEEE Int. Conf. Distributed Computing Systems*, Montreal, Canada, 2009, pp. 273–280.

[30] Crossbow, *MPR-Mote Processor Radio Board User's Manual*. San Jose, CA, USA: Crossbow Technology Inc, 2003.

[31] G. Anastasi, A. Falchi, A. Passarella, M. Conti, and E. Gregori, Performance measurements of motes sensor networks, in *Proc. 7th ACM Int. Symp. Modeling, Analysis and Simulation of Wireless and Mobile Systems*, Venice, Italy, 2004, pp. 174–181.

[32] Z. F. Huang, L. Wang, K. Yi, and Y. H. Liu, Sampling based algorithms for quantile computation in sensor networks, in *Proc. 2011 ACM SIGMOD Int. Conf. Management of Data*, Athens, Greece, 2011, pp. 745–756.

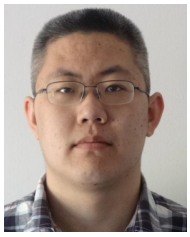
[33] B. Gedik, L. Liu, and P. S. Yu, ASAP: An adaptive sampling approach to data collection in sensor networks, *IEEE Trans. Parallel Distrib. Syst.*, vol. 18, no. 12, pp. 1766–1783, 2007.

[34] A. S. Silberstein, R. Braynard, C. Ellis, K. Munagala, and J. Yang, A sampling-based approach to optimizing top-k queries in sensor networks, in *Proc. 22nd Int. Conf. Data Engineering*, Atlanta, GA, USA, 2006, p. 68.

[35] J. Li, S. Y. Cheng, Z. P. Cai, J. G. Yu, C. K. Wang, and Y. S. Li, Approximate holistic aggregation in wireless sensor networks, *ACM Trans. Sens. Netw.*, vol. 13, no. 2, p. 11, 2017.

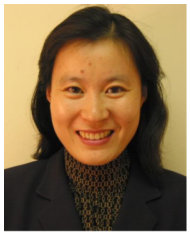
[36] K. Beyer, P. J. Haas, B. Reinwald, Y. Sismanis, and R. Gemulla, On synopses for distinct-value estimation under multiset operations, in *Proc. 2007 ACM SIGMOD Int. Conf. Management of Data*, Beijing, China, 2007, pp. 199–210.

[37] Z. Cai and Z. He, Trading private range counting over big iot data. in *Proc. 39th IEEE Int. Conf. Distributed Computing Systems*, Dallas, TX, USA, 2019.

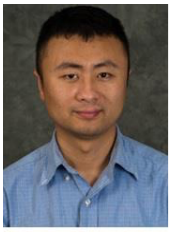


Ji Li received the BS degree from Heilongjiang University, China in 2012, and the PhD degree from Georgia State University in 2018. He is currently an assistant professor in the College of Computing and Software Engineering at Kennesaw State University. His research focuses on mobile crowdsensing and big data management in IoT networks.

Madhuri Siddula received the BS degree from Osmania University, India in 2010 and the MS degree from Indraprastha Institute of Information Technology, India in 2012. She is currently a PhD student at Georgia State University, USA. Her research interests include social networks, privacy and security, and big data mining.



Xiuzhen Cheng received the MS and PhD degrees from University of Minnesota Twin Cities, Minneapolis, MN, USA, in 2000 and 2002, respectively. She is a professor with the Department of Computer Science, the George Washington University, Washington, DC, USA. She was a program director for the National Science Foundation from April to October in 2006 (full time) and from April 2008 to May 2010 (part time). She has published more than 170 peer-reviewed papers. Her current research interests include privacy-aware computing, wireless and mobile security, dynamic spectrum access, mobile handset networking systems (mobile health and safety), cognitive radio networks, and algorithm design and analysis. She has served on the Editorial Board of several technical journals (e.g., *IEEE Transactions on Parallel and Distributed Systems* and *IEEE Wireless Communications*) and the Technical Program Committee of various professional conferences/workshops (e.g., IEEE Conference on Computer Communications, IEEE International Conference on Distributed Computing Systems, IEEE International Conference on Communications, and IEEE/ACM International Symposium on Quality of Service). She also has chaired several international conferences (e.g., IEEE Conference on Communications and Network Security, and International Conference on Wireless Algorithms, Systems, and Applications).



Wei Cheng received the BS and MS degrees from the National University of Defense Technology, Changsha, China, in 2002 and 2004, respectively, and the PhD degree from the George Washington University, Washington, DC, USA, in 2010. He is currently an assistant professor with Virginia Commonwealth University, Richmond, VA, USA. He was a post-doctoral scholar with University of California, Davis, CA, USA. His current research interests include wireless networks, cyber-physical networking systems, and algorithm design and analysis. In particular, he is interested in localization, security, fog computing, and smart cities. He is a member of the ACM.

Zhi Tian is a professor in the Electrical and Computer Engineering Department of George Mason University, Fairfax, VA, USA, since 2015. Prior to that, she was on the faculty of Michigan Technological University from 2000 to 2014. Her research interests lie in statistical signal processing, wireless communications, and wireless sensor networks. She is an IEEE fellow. She is an elected member of the IEEE Signal Processing for Communications and Networking Technical Committee and a member of the Big Data Special Interest Group IEEE Signal Processing Society. She served as an associate editor for *IEEE Transactions on Wireless Communications* and *IEEE Transactions on Signal Processing*. She is a distinguished lecturer of the IEEE Vehicular Technology Society from 2013 to 2017 and the IEEE Communications Society from 2015 to 2016.

Yingshu Li received the BS degree from the Department of Computer Science and Engineering, Beijing Institute of Technology, Beijing, China in 2001, and the MS and PhD degrees from the Department of Computer Science and Engineering, University of Minnesota Twin Cities, Minneapolis, MN, USA in 2003 and 2005, respectively. She is currently an associate professor with the Department of Computer Science, Georgia State University, Atlanta, GA, USA. Her research interests include wireless networking, sensor networks, sensory data management, social networks, and optimization. She received the National Science Foundation CAREER Award in 2006.