IEEE TRANSACTIONS ON
EMERGING TOPICS
IN COMPUTING

Received 15 November 2013; revised 25 February 2014; accepted 2 March 2014. Date of publication 16 April 2014;
date of current version 30 October 2014.

Digital Object Identifier 10.1109/TETC.2014.2318177

Toward Energy Efficient Big Data Gathering in Densely Distributed Sensor Networks

DAISUKE TAKAISHI[1], (Student Member, IEEE), HIROKI NISHIYAMA[1], (Senior Member, IEEE), NEI KATO[1], (Fellow, IEEE), AND RYU MIURA[2]

[1]Graduate School of Information Sciences, Tohoku University, Sendai 980-8577, Japan
[2]Wireless Network Research Institute, National Institute of Information and Communications Technology, Tokyo 184-8795, Japan
CORRESPONDING AUTHOR: D. TAKAISHI (takaishi@it.ecei.tohoku.ac.jp)

ABSTRACT Recently, the big data emerged as a hot topic because of the tremendous growth of the information and communication technology. One of the highly anticipated key contributors of the big data in the future networks is the distributed wireless sensor networks (WSNs). Although the data generated by an individual sensor may not appear to be significant, the overall data generated across numerous sensors in the densely distributed WSNs can produce a significant portion of the big data. Energy-efficient big data gathering in the densely distributed sensor networks is, therefore, a challenging research area. One of the most effective solutions to address this challenge is to utilize the sink node's mobility to facilitate the data gathering. While this technique can reduce energy consumption of the sensor nodes, the use of mobile sink presents additional challenges such as determining the sink node's trajectory and cluster formation prior to data collection. In this paper, we propose a new mobile sink routing and data gathering method through network clustering based on modified expectation-maximization technique. In addition, we derive an optimal number of clusters to minimize the energy consumption. The effectiveness of our proposal is verified through numerical results.

INDEX TERMS Big data, wireless sensor networks (WSNs), clustering, optimization, data gathering, and energy efficiency.

I. INTRODUCTION

Recent development of various areas of Information and Communication Technology (ICT) has contributed to an explosive growth in the volume of data. According to a report published by IBM in 2012 [1] , 90 percent of the data in the world was generated in the previous two years. As a consequence, the concept of the big data has emerged as a widely recognized trend, which is currently attracting much attention from government, industry, and academia [2]. As shown in Fig. 1, the big data comprises high volume, high velocity, and high variety information assets [3], which are difficult to gather, store, and process by using the available technologies. The variety indicates that the data is of highly varied structures (e.g. data generated by a wide range of sources such as Machine-to-Machine (M2M), Radio Frequency Identification (RFID), and sensors) while the velocity refers to the high speed processing/analysis (e.g., click-streaming, fast database transactions, and so forth). On the other hand, the volume refers to the fact that a lot of data needs to be gathered for processing and analysis. Although currently used services

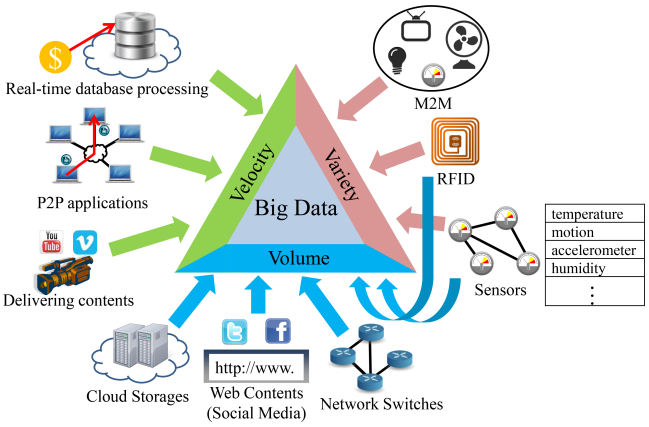


FIGURE 1. Major trends of big data gathering.

(e.g. social networks, cloud storage, network switches, and so forth) are already generating much volume of the big data, it is anticipated that more and more data will be generated

by sensors/RFID devices such as thermometric sensors, atmospheric sensors, motion sensors, accelerometers, and so on. In fact, according to a report by ORACLE [4], the volume of data generated by sensors and RFID devices is expected to reach the order of petabytes. Interestingly as shown in Fig. 1, the sensors are responsible for generation of big data in big volume and also in a wide variety.

Gathering the large volume and wide variety of the sensed data is, indeed, critical as a number of important domains of human endeavor are becoming increasingly reliant on these remotely sensed information. For example, in smart-houses with densely deployed sensors, users can access temperature, humidity, health information, electricity consumption, and so forth by using smart sensing devices. In order to gather these data, the Wireless Sensor Networks (WSNs) are constructed whereby the sensors relay their data to the ''sink''. However, in case of widely and densely distributed WSNs (e.g. in schools, urban areas, mountains, and so forth) [5], [6], there are two problems in gathering the data sensed by millions of sensors. First, the network is divided to some sub-networks because of the limited wireless communication range. For example, sensors deployed in a building may not be able to communicate with the sensors which are distributed in the neighboring buildings. Therefore, limited communication range may pose a challenge for data collection from all sensor nodes. Second, the wireless transmission consumes the energy of the sensors. Even though the volume of data generated by an individual sensor is not significant, each sensor requires a lot of energy to relay the data generated by surrounding sensors. Especially in dense WSNs, the life time of sensors will be very short because each sensor node relays a lot of data generated by tremendous number of surrounding sensors. In order to solve these problems, we need an energy-efficient method to gather huge volume of data from a large number of sensors in the densely distributed WSNs.

To achieve energy-efficient data collection in densely distributed WSNs, there have been many existing approaches. For example, the data compression technology [7] is capable of shrinking the volume of the transmitted data. Although it is easy to be implemented, the data compression technology requires the nodes to be equipped with a big volume of storage and high computational power. In addition, the topology control technology can evaluate the best logical topology and reduce redundant wireless transmissions [8], [9]. When the redundant wireless transmissions are reduced, the required energy for wireless transmissions can be also reduced. Furthermore flow control and routing can choose the path which consists of nodes having high remaining energy [10], [11]. However, these technologies are not able to deal with the divided networks problem.

To deal with both the divided sub-network problem and the energy consumption issue, the mobile sink schemes have received great attention in literature. In such schemes, the data collector, referred to as the ''sink node'' (or simply the sink) is assumed to be mobile such as Vehicle, Unmanned Aerial Vehicle (UAV), and so on. As the sink node moves around the sensing area, the sensor nodes send data to the sink node when the sink node comes in their proximity. Thus, energy consumption can be decreased by reducing the amount of relays in the WSN. Since the mobile sink schemes aim to reduce wireless transmissions, the trajectory of the sink node is decided based on the sensor nodes' information (e.g., location and residual energy). The sink node divides the sensor nodes into a number of clusters based on a certain condition. Then, the sink node roams around in these clusters.

In this paper, we propose an energy minimized clustering algorithm by using the Expectation-Maximization (EM) algorithm for 2-dimensional Gaussian mixture distribution. Our proposal aims to minimize the sum of square of wireless communication distance since the energy consumption is proportional to the square of the wireless communication distance. Moreover, we first focus on the ''data request flooding problem'' to decide the optimal number of clusters. The data request flooding problem refers to the energy inefficiency that occurs when all the nodes broadcast data request messages to their respective neighboring nodes. This problem wastes energy, particularly in the high density WSNs. Previous research work advocates increasing the number of clusters to reduce the data transmission energy. However, in this paper, we point out that an excessive number of clusters can result in performance degradation, and therefore, we propose an adequate method for deriving the optimal number of clusters.

The remainder of the paper is organized as follows. Section II reviews some related works and presents our research motivation. In Section III, we present our proposed clustering algorithm based on a modified EM technique. Section IV illustrates the derivation of the optimal number of clusters. Performance evaluation is presented in Section V. Finally, Section VI concludes the paper.

II. RELATED WORKS AND OUR MOTIVATION

The review conducted by Sagiroglu *et al.* [3] highlighted that big data and its analysis are at the core of modern science and business. Sagiroglu *et al.* identified a number of sources of big data such as online transactions, emails, audios, videos, images, click-streams, logs, posts, search queries, health records, social networking interactions, mobile phones and applications, scientific equipment, and sensors. Also, it was pointed out, in their work, that the big data are difficult to capture, form, store, manage, share, analyze, and visualize via conventional database tools. Furthermore, the three main characteristics of big data, namely variety, volume, and velocity are discussed in that work that were briefly described in Section I.

According to the report by ORACLE in [12], the concept of big data is stimulating a wide range of industry sectors. Specific examples of big data generated by sensors were provided in the report. For instance, manufacturing companies usually embed sensors in their machinery for monitoring usage patterns, predicting maintenance problems, and enhancing the product quality. By studying the data streams generated by the sensors embedded in the machinery allow

the manufacturers to improve their products. The numerous sensors deployed in the supply lines of utility providers generate a huge volume of data, which are consistently monitored for production quality, safety, maintenance, and so forth. Other examples of sensors generating a bulk of the big data consist in electronic sensors monitoring mechanical and atmospheric conditions. In addition, sensors used for healthcare services (to monitor bio-metrics of the human body, patients' conditions, healthcare diagnoses, treatment phases, and so forth) are identified to be a rich source of big data in the report presented in [12]. However, how to gather the sensed data from these numerous sensors in an energy-efficient manner remained beyond the scope of the report.

The work in [13] presented a cloud-based federated framework for sensor services. The main objective of the work was to enable seamless exchange of feeds from large numbers of heterogeneous sensors. Various applications using big data generated by densely distributed WSNs have also emerged in literature. In addition, in [14] and [15], big data in terms of the healthcare information (e.g., blood pressure and heart rate) sensed by numerous sensors are used to realize remote medical care services. Furthermore, patients' location information are used to arrange prompt dispatch of ambulances. Large volume of data gathered from location-sensors attached to animals enabled researchers to observe various animal habitats [16], [17]. Because widely and densely distributed WSNs collect various types of data, the overall data which are gathered is, indeed, overwhelming. To efficiently gather the big data generated by the densely distributed WSNs is, however, not an easy task since the WSNs may be divided into sub-networks because of the limited wireless communication range of the sensors.

In conventional research works, data gathering using the mobile sink in WSNs has been widely studied in literature. Data Mobile Ubiquitous LAN Extensions (MULEs) [18] is the one of the most prominent and earliest studies on the mobile sink scheme. Data MULEs follow the basic steps of all the mobile sink schemes. First, it divides sensor nodes into clusters. Second, it decides the route for patrolling each cluster. The work in [18] assumes a simple data collection scheme whereby the mobile sink node divides sensor nodes into grids regardless of the sensor nodes' location, and patrols the grids by using random walk between the neighboring grids. However, this type of clustering, which is not based on the nodes' location, might result in inefficient data gathering. If there is no sensor node remaining in the cluster, patrolling the empty cluster results in waste of time and degraded efficiency. Also, patrolling based on randomness might result in unbalanced visits to clusters with different numbers of sensor nodes. Thus, the mobile sink might fail to collect information.

Low-Energy Adaptive Clustering Hierarchy (LEACH) [19] is one of the most famous clustering algorithms in WSNs using the static sink node. In LEACH, the clustering algorithm is executed by the each sensor node. Sensor nodes exchange information on their residual energies, and the nodes with higher residual energy are given a higher probability of becoming a cluster head. By doing periodical re-clustering, energy consumption of each node becomes eventually equal. However, LEACH still has several shortcomings. For example, because LEACH is based on the assumption that each node can communicate with all other nodes, the WSNs deployed in wide areas are not able to use the algorithm. Most of the distributed algorithms like LEACH naturally consider the limitation of the node's communication range. K-hop Overlapping Clustering Algorithm (KOCA) [20] and k-hop connectivity ID (k-CONID) [21] are examples of the distributed clustering algorithms. Authors of KOCA focused on multiple overlapping clusters, and designed the KOCA algorithm based on a probabilistic cluster head selection and nodes' location. The k-CONID algorithm is also a probabilistic algorithm. The nodes exchange their random IDs with each other, and the node that has the minimum ID within k-hop is selected as a cluster head.

In WSNs, minimizing data transmission is difficult for a distributed clustering algorithm. If a WSN is physically divided into sub-networks, a node cannot possess information about all the nodes in the WSNs. Thus, the algorithm cannot achieve optimization. To realize minimum energy clustering, we need to use the centralized clustering algorithm. Moreover, the centralized clustering algorithm, which is conducted by a super node, is suitable for the mobile sink scheme. Power-Efficient Gathering in Sensor Information Systems (PEGASIS) [22] and KAT mobility (K-means And TSP mobility) [23] are one of the centralized clustering algorithms. PEGASIS algorithm constructs chain clusters of nodes based on location, and repeats cluster head selection. PEGASIS algorithm considers the limitation of the communication range, and achieves uniform energy consumption. However, the algorithm still does not achieve minimization of energy consumption because the clustering algorithm uses greedy algorithm. KAT mobility divides the nodes into clusters by using k-means algorithm. Because k-means algorithm is the centralized clustering algorithm based on the node's location, the clustering result is closer to the total optimization. While the result is the optimal cluster that reduces energy consumption, the KAT mobility algorithm is designed without considering the communication range limitation. Therefore, the mobile sink might fail to collect information from all nodes.

Contemporary research on the sensor node clustering algorithm can be classified into three types, namely centralized algorithms without considering nodes' information (i.e., location or communication range), distributed algorithms without considering nodes' information, and distributed algorithms that consider the nodes' location and communication range. However, to achieve both minimization of data transmission and data collection from all the nodes, we need to use a centralized algorithm, which considers the nodes' location and communication range. Unlike existing algorithms, our proposed clustering algorithm achieves both minimization of data transmission and data collection.

Earlier research works on sensor node clustering algorithms demonstrates that the increasing number of clusters reduces energy consumption for data transmission. Certainly, the idea holds since increasing the number of clusters decreases the cluster-sizes and shortens the transmission length. Some researchers consider that certain limitations on the number of cluster can be decided by other factors. For example, in [24], the limitation is the maximum acceptable latency of data collection. The authors of [24] also defined the limitation by a node's buffer size. While these limitations are realistic assumptions, they do not consider the energy consumption for data requests. In our paper, we first focus on the effect of data request messages by increasing the number of clusters. Based on a simple and common data gathering model of the densely distributed WSNs, we demonstrate that the number of data request messages has a noticeable impact on the energy consumption of the sensor nodes. When the connectivity of the nodes becomes bigger, the impact becomes larger also. In this paper, we present how to evaluate the optimal number of clusters to minimize the energy consumption of the sensor nodes.

III. CLUSTERING-BASED BIG DATA GATHERING IN DENSELY DISTRIBUTED WSN

In this section, we first outline the clustering problem in WSN using mobile sink and the challenges in solving this problem. After that, we introduce the considered network model and the overview of EM algorithm for clustering. Based on EM algorithm, we proposed our clustering method and the procedure to gather data using the proposed method.

A. CLUSTERING PROBLEM

When considering the scheme of data gathering in WSN using mobile sink, the biggest challenge in reducing energy consumption is how to decide the location where data gathering is conducted. In other words, this problem has same meaning as answering the following two questions. 1) What is the best algorithm for dividing nodes into clusters? 2) How many clusters is optimal in terms of reducing energy consumption? As we assume that required energy for data transmission of node is proportional to the square of transmission distance, the best clustering algorithm to minimize energy consumption for data transmission must minimize the sum of square of data transmission distance in a network. EM algorithm is powerful and well-known tool to solve the clustering problem by repeatedly calculate the simple math formula. Since the EM algorithm can minimize the sum of square of distance between every node and cluster centroid, we adopt EM algorithm over the 2-dimensional Gaussian mixture distribution. However, there is a limitation of the maximum communication range in the realistic situation. Not all nodes can connect to each other and also to the cluster centroid. Nodes that cannot directly communicate with the cluster centroid need to communicate in a multi-hop manner. In multi-hop communication, communication distance is a sum of distance between nodes in multi-hop path. Therefore, as shown in Fig. 2, communication distance is different from direct distance. However, the EM algorithm minimize the sum of square of direct distance, not communication distance. Thus, we need to adapt the EM algorithm to the situation of limited maximum communication range and improve it such as to minimize the sum of square of communication distance.

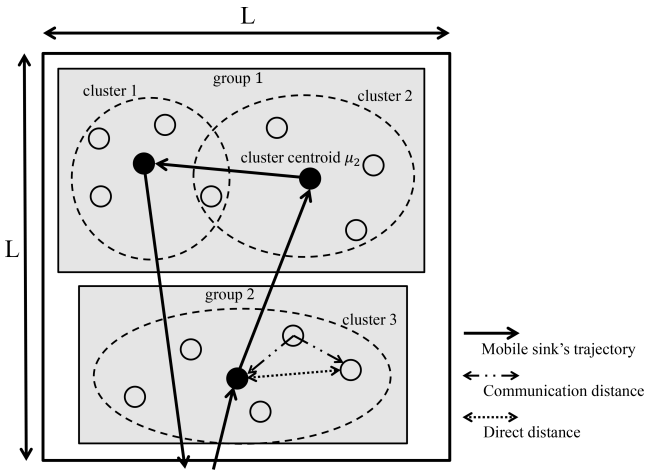


FIGURE 2. **An example of the considered network.**

B. CONSIDERED NETWORK MODEL

In this paper, we consider a network which consists of a mobile sink and many sensor nodes spread within a limited field. Every sensor node knows its location by using localization technology, and the mobile sink knows all nodes' locations. Regardless of being a sink or the sensor, a node has a limited communication range R and communication is always successful if it is within R. The mobile sink node patrols the cluster centroids that are calculated to minimize energy consumption for data transmission, and collects data from sensor nodes. Sensor nodes are equipped with a buffer memory and store sensed information until mobile sink approaches the cluster centroid. The information is transferred to the sink node by multi-hop fashion. In this paper, we assume a densely distributed WSN in a large area such as schools, urban areas, mountains, and so forth and thus WSNs are divided into sub-networks. Fig. 2 shows a simple example of the assumed network. N sensor nodes illustrated by circles are distributed in the target $L \times L$ area. K centers of clusters illustrated by filled circle is to be visited by mobile sink. A solid-fill area and a dotted circle means ''group'' of nodes and cluster, respectively. In this paper, ''group'' means a set of nodes that can communicate with each other. The nodes that belong to different groups cannot communicate with each other due to being far away. There are G groups in the field, and N_g and K_g refers to the number of nodes and number of clusters in the gth group, respectively. The number of groups is calculated by the nodes' location and communication range R. In case of Fig. 2, $N_1 = 7$ and $K_1 = 2$ because there are 7 nodes and 2 clusters in group 1.

C. OVERVIEW OF EM ALGORITHM FOR CLUSTERING

The EM algorithm is a classical clustering algorithm, which assumes that nodes are distributed according to Gaussian mixture distribution,

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right), \tag{1}$$

where K and π_k indicate the total number of clusters and the mixing coefficient of the kth cluster, respectively. $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is defined as follows,

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)\,|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}, \tag{2}$$

where $\mathbf{x}$ is the position vectors of all nodes. Cluster parameters, μ_k and Σ_k, are the position vector of centroid of cluster k and 2×2 covariance matrix of the kth cluster, respectively.

At the first step, EM algorithm calculates each node's value of degree of dependence that is referred to as responsibility. The responsibility shows how much a node depends on a cluster. The nth node's value of degree of dependence on kth cluster is given by following equation.

$$\gamma_{nk} = \frac{\pi_k \mathcal{N}\left(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right)}{\sum_{j=1}^{K} \pi_j \mathcal{N}\left(\mathbf{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j\right)}. \tag{3}$$

Because of its definition, the responsibility takes values between 0 and 1. At the second step, the EM algorithm evaluates K weighted center of gravity of a 2-dimensional location vector of nodes. This evaluation uses the responsibility value as weight of nodes. At the third step, the locations of the cluster centroids are changed to the weighted centers of gravity evaluated in the second step. And EM algorithm evaluates the value of the log likelihood as shown below.

$$\mathcal{P} = \ln p(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) = \sum_{n=1}^{N} \ln\left\{\sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\right\}. \tag{4}$$

Until the value of log likelihood converge, the EM algorithm repeats all steps. This value of log likelihood is monotonously decreasing, and the EM algorithm always terminates. Because the EM algorithm repeatedly updates cluster centroids' position vector, μ_k, and nodes' responsibility to kth cluster, γ_{n_k}, the sum of square of distances of each node to cluster gradually decreases and finally becomes optimal.

D. PROPOSED CLUSTERING METHOD

Our objective is to propose a clustering method based on the EM algorithm. In supposed widely and densely deployed WSNs, which have high variety and high volume of data, we need to consider "groups", which refer to sets of nodes that can communicate with each other. Therefore, nodes that cannot communicate with each other belongs to different groups. To collect data from all nodes, the number of clusters must be set to more than the number of groups. Our proposed clustering method can be summarized in Algorithm 1.

Algorithm 1 Proposed Clustering Algorithm

```
Initialize cluster centroids, μ, to random locations.
Calculate clusters' parameters, π and Σ.
Calculate D_nk and P.
while |P − P^new| < ε do
  Select a group g which has the biggest value v_g.
  for k ∈ K_g do
    for n ∈ N_g do
      Calculate nth node's responsibility value, γ_nk.
    end for
    Calculate number of nodes belong to cluster, N_k.
    Update the clusters' parameters, π, μ and Σ, by using
    N_k.
  end for
  Evaluate the log likelihood P^new.
end while
Return cluster centroids, μ, covariance matrix, Σ, and the
number of nodes that belongs to each cluster.
```

At first, the mobile sink sets the cluster centroids, μ, to random locations. By using a random position vector of cluster centroids, communication distances of each node to cluster centroids, D_{nk}, are calculated. Thereafter, the mixing coefficient, π, and covariance matrix, Σ, are calculated.

After the cluster initialization phase, our proposed method selects a group g that has the largest value of proportion of number of nodes to the number of clusters in group g, shown as follows,

$$v_g = \frac{K_g}{N_g}. \tag{5}$$

In the selected group that has the highest value of v_g, our proposed method picks up all nodes that belong to group g and updates these node's responsibility value, γ_{nk}. This responsibility value reflects how much node n belongs to cluster k. By using the updated responsibility, γ_{nk}, cluster centroids, μ, and covariance matrix, Σ, are re-calculated, and the number of nodes which belongs to kth cluster is calculated as shown in the following equation,

$$N_k = \sum_{x_n \in \boldsymbol{X}} \gamma_{nk}. \tag{6}$$

These calculation are repeatedly executed until the difference between the newly calculated $\mathcal{P}$ and previously calculated $\mathcal{P}$ becomes smaller than small number, ϵ.

E. DATA GATHERING PROCEDURE USING THE PROPOSED CLUSTERING TECHNIQUE

After clustering, the mobile sink patrols every cluster centroid and collects the data from the nodes in the cluster. It is easy to see that delay is a main problem of using mobile sink in WSNs. This delay is the waiting time from data generation to data sending. Because the mobile sink moves relatively slow

compared with electrical communication between nodes, the mobile sink scheme causes long delay. To shorten this delay, we need to minimize total patrolling path length. Thus, in our scheme the mobile sink patrols along Traveling Salesman Problem (TSP) path of all cluster centroids.

Once the mobile sink arrives at the cluster centroids, it collects data from sensor nodes. Directed Diffusion [25] is one of the most famous data collection schemes in WSNs. In our method, we consider using a typical example of them, i.e., ''One Phase Pull [26] where the mobile sink node sends data request message at the cluster centroids. When a sensor node receives a data request message from cluster k, the node re-broadcasts the data request message and replies data to the neighboring node, which is the parent node in the data request tree of cluster k. Then, the node relays data messages to the sink.

To minimize the total required energy to send data, all nodes send the sensed information according to the value of responsibility of the cluster. The responsibility value is calculated based on the given parameters, $\boldsymbol{\mu}$, $\boldsymbol{\pi}$, and $\boldsymbol{\Sigma}$, according to (3). These parameters are added to data request message and sent by the sink. Only after the sensor deployment, each node exchanges its own position vector, x, with sensor nodes belonging to same groups. Because the exchange of position vector is executed only one time after the sensor deployment, the energy consumption is not significant. As a result, when a node belongs to only one cluster, the node can send all data to the sink node. And when a node belongs to more than one clusters, the node sends data according to the responsibility of each cluster. In case of $\gamma_{n1} = 0.6$ and $\gamma_{n2} = 0.4$, if the nth node receives a data request message that is sent by the sink node at the centroid of cluster 1, the node replies 60% of data. And if the node receives data request message that is sent from cluster 2, the node sends 40% of data to the sink node at the centroid of cluster 2. By sending data using this cluster adapted Directed Diffusion scheme, we can minimize total required energy to send data.

IV. DERIVING THE OPTIMAL NUMBER OF CLUSTERS IN THE PROPOSED CLUSTERING METHOD

The data gathering method presented in the previous section aims to minimize energy consumed by gathering data. However, it still has a remaining issue, which is to find the optimal number of clusters. Previous researches in literature often consider increasing the number of clusters lead to the decrease of energy consumption for data transmission. However, such researches do not take into consideration the energy consumption of data request message. In this section, we point out this problem, and show an analysis to derive the optimal number of clusters.

A. DEFINITION OF CONNECTIVITY

To analyze the correlation between energy consumption and connectivity, we formulate the connectivity of nodes. In this paper, we define the connectivity as the portion of nodes that can communicate with each other.

$$C = \frac{\sum_{g=1}^{G} N_g(N_g - 1)}{N(N-1)}. \tag{7}$$

This metric takes a value between 0 and 1. When all nodes can communicate with each other, the value of connectivity is 1. If every node is isolated, the value is 0. When the mobile sink starts computing the optimal number of clusters, the mobile sink node knows every sensor nodes' location. Therefore, the mobile sink can calculate the connectivity value C based on nodes' location.

B. DATA REQUEST FLOODING PROBLEM

In WSN using mobile sink, the sink node sends data request message to invoke data transmission from sensor nodes when it arrives at the cluster centroids. The nodes that receive data request message send the data to the sink node and broadcast data request message to their neighboring nodes. That data request message is repeatedly broadcasted until all nodes that belong to the same group receive the message. Although some nodes may receive data request message more than 2 times, they only send data and broadcast the data request message once after the first time of receiving the message. These broadcasts of data request message cause high energy consumption because the network will be flooded with redundant wireless communication. Thus, reducing data request transmission is also important for mobile sink scheme.

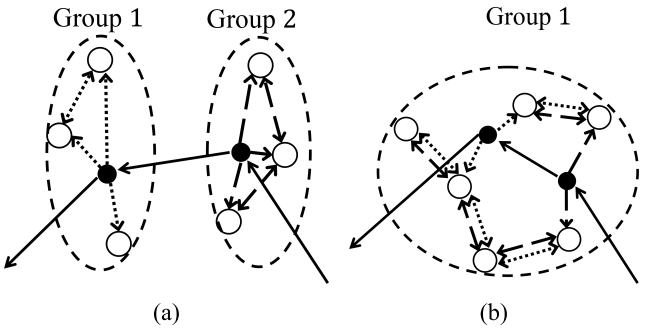


FIGURE 3. Data request flooding in low and high connectivity network. (a) Low connectivity network. (b) High connectivity network.

The impact of data request flooding issues becomes significant when connectivity becomes larger as an example shown in Fig. 3. In Fig. 3(a) and Fig. 3(b), there are two groups and one group, respectively. Six sensor nodes are scattered on the ground. Furthermore, the sink nodes traverse the two cluster centroids, and broadcasts the data request message. In the case of Fig. 3(a), nodes can only communicate with the nodes that belongs to the same group. Sink node broadcasts data request message to each node at cluster 1, and these nodes broadcast the data request message. Therefore, the sum of the transmission of data request message of both cluster 1 and cluster 2 is 6. On the other hand, in Fig. 3(b), where nodes can communicate with all nodes, the data request message

sent at the cluster 1 is transferred to all nodes. Furthermore, all nodes broadcast the data request message. Therefore, sum of the transmission of data request message of both cluster 1 and cluster 2 is 12.

Even if number of nodes and clusters stay the same, the data request flooding problem becomes more serious with higher connectivity. Moreover, it is clearly understood that the total number of transmitted data request messages increase when the numbers of clusters increases. Because of this problem, it is necessary to find the optimal number of clusters in terms of connectivity and energy consumption.

C. COMPUTING THE OPTIMAL NUMBER OF CLUSTERS

To decide the optimal number of clusters, we need to define objective function. The objective function is defined as the sum of required energy of data and data request message transmissions. Thus, the objective function, $W(K)$, can be defined as the sum of energy consumption in one cycle of mobile sink patrol as follows.

$$W(K) = D_{\mathrm{Req}} E_{\mathrm{Req}}(K) + D_{\mathrm{Dat}} E_{\mathrm{Dat}}(K), \tag{8}$$

where $E_{\mathrm{Req}}(K)$ and $E_{\mathrm{Dat}}(K)$ are the sums of the square of transmission distance of data requests and data messages, respectively. D_{Req} and D_{Dat} indicate the data size of data and data request messages, respectively. $E_{\mathrm{Dat}}(K)$ is evaluated according to the following equation:

$$E_{\mathrm{Dat}} = \sum_{n=1}^{N} \sum_{k=1}^{K} \sum_{h=1}^{H_{nk}} \gamma_{nk} \cdot l_h^2, \tag{9}$$

where H_{nk} is the hop count from nth node to kth cluster centroid and l_h is communication distance of each hop. When nth node cannot communicate with kth centroid, we set the value of H_{nk} to 0 and the value of required energy to 0. Moreover, each node re-broadcasts each data request message one time with the maximum transmission power. Since the data transmission energy, $E_{\mathrm{Dat}}(K)$, is a decreasing function of K while data request transmission energy, $E_{\mathrm{Req}}(K)$, is an increasing function of K, there is a trade-off relationship between the first and second terms in the right side of (8). By considering the condition that the number of clusters, K, must be greater than the number of groups, G, the optimal number of clusters, K_{opt}, is defined by the following equation.

$$K_{opt} = \max(G, \arg\min_{K}(W(K))). \tag{10}$$

In order to calculate the required energy to transmit data request messages, we consider one group of node that has N_g nodes and K_g cluster centroids. Data request message is sent from every cluster and every node re-broadcasts it one time. Thus, the total required energy to transmit data request massage is formulated as follows:

$$E_{\mathrm{Req}} = \sum_{g=1}^{G} K_g N_g R^2, \tag{11}$$

where R is the maximum transmission range of sensor nodes. For simplicity, constant variables are omitted. If there is no imbalance of location of cluster centroids, the number of nodes that belongs to each cluster is the same.

$$\frac{K_g}{N_g} = \frac{K}{N} \tag{12}$$

Here, if the number of nodes is larger than 1, the connectivity, C, can be approximated as follows:

$$C = \frac{\sum_{g=1}^{G} N_g(N_g - 1)}{N(N-1)} \doteqdot \frac{\sum_{g=1}^{G} N_g^2}{N^2}. \tag{13}$$

Therefore, from (11), (12) and (13), E_{Req} can be calculated as follows:

$$E_{\mathrm{Req}} = KNR^2 C. \tag{14}$$

This analysis says that the required energy for data request transmission is proportional to connectivity. Thus, it can be seen that the number of clusters has a significant effect on connectivity. Moreover, the function is a monotonically increasing function of K which indicates that a lower number of clusters is better for reducing energy of data request transmissions.

By calculating the required energy for data transmission as in (14), and data request transmission as in (9), the optimal number of clusters, (10), can be calculated by using (8) and (10).

V. PERFORMANCE EVALUATION

We conducted performance evaluation by using a clustering simulator built by C++ programming language. In this section, we first evaluate the clustering efficiency. Then we evaluate total energy consumption to evaluate our proposed methd of optimizing number of clusters.

A. EFFICIENT DATA COLLECTION

In this experiment, we measure the energy consumption for data transmissions, E_{Dat}, and the efficiency of our proposed clustering algorithm by varying the number of nodes. Table 1 shows simulation parameters used in the first experiment. Sensors are uniformly deployed in a 5000 × 5000 square meters area. The nodes' communication range is set to 438.57 meters, and we measure E_{Dat} and efficiency of our proposal clustering by varying the number of sensor nodes. E_{Dat} represented in (9) simply shows how much energy is needed for data transmissions from sensor nodes to the mobile sink. However, if locations of every centroid is far away from nodes and they cannot establish connection,

TABLE 1. Environments of 1st experiment.

Node distribution	Uniformly random
Number of cluster, K	10
Number of node, N	20 - 100
Communication range, R	438.57m
Length of one side of field, L	5000m

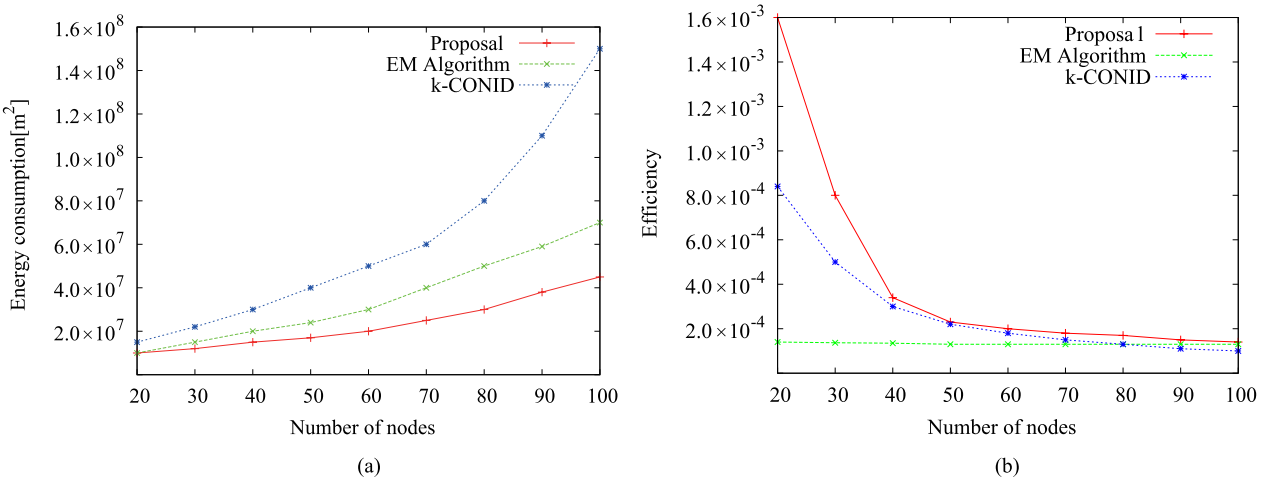


FIGURE 4. Energy consumption for data transmission and efficiency. (a) Required transmission energy E_{Dat}. (b) Efficiency of each clustering algorithm.

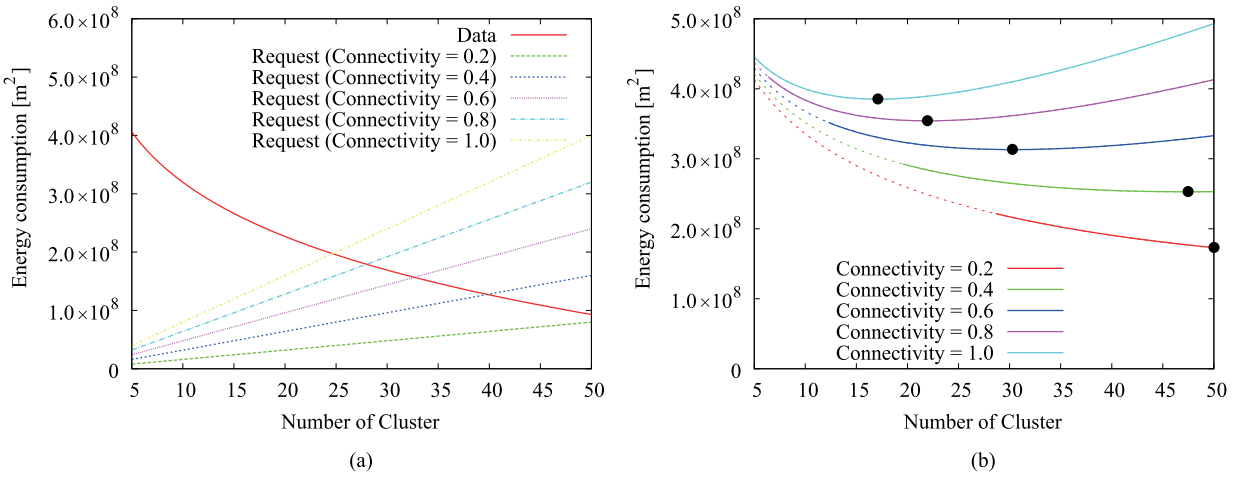


FIGURE 5. Effect of the number of cluster. (a) Energy consumption for data transmissions and data request messages. (b) Sum of energy consumption of data transmissions and data request messages.

E_{Dat} value is calculated as 0 according to (9). This value does not reflect energy saving, but it only indicates failure of clustering. The clustering algorithm that do not consider connectivity suffer from this failure (e.g. pure EM algorithm). Thus, we also use a second metric, referred to as efficiency, which combines E_{Dat} value and number of connected nodes.

$$\text{Efficiency} = \frac{(\text{Number of connected node})}{E_{\mathrm{Dat}}}. \qquad (15)$$

We compare our clustering algorithm with EM algorithm and k-COIND algorithm, which are centralized clustering algorithm and distributed clustering algorithm, respectively. Our clustering algorithm is a centralized algorithm considering connectivity, unlike EM algorithm.

Figure 4(a) and 4(b) are experimental results of required energy and efficiency respectively. As can be seen from (9), (11), the energy consumption is proportional to the square of data transmission range, we measure energy in units of m^2, i.e., omitting constant variables. Figure 4(a) shows the proposed scheme and pure EM algorithm can reduce required energy significantly compared with k-CONID algorithm. The reason of this difference is based on difference between centralized and distributed cluster establishment. The centralized algorithm can calculate more efficient clustering than the distributed one. Our proposed scheme behaves similar to the EM algorithm, but has less energy consumption. This improvement occurs from considering connectivity and communication distance. Fig. 4(b) shows that EM algorithm is the worst clustering algorithm when node density is low. Since EM algorithm does not consider node connectivity and is centralized algorithm, when the number of nodes is low and node density is small, centroids of EM algorithm can connect only to a small number of nodes. Our proposed scheme succeeds to adapt to node density variation and minimizes transmission energy.

TABLE 2. Environment of 2nd experiment.

Number of cluster, K	5 - 50
Number of node, N	50 - 100
Communication range, R	200m - 600m
Length of one side of field, L	1000m - 5000m
Clustering algorithm	Proposed algorithm

B. OPTIMAL NUMBER OF CLUSTERS

To evaluate our proposed method of optimizing number of clusters, we measure the energy consumption by varying the number of clusters. Energy consumption is defined as the sum of energy consumption of data transmissions and data requests. Given parameters are enumerated in Table 2. We set parameter $D_{\mathrm{Dat}}/D_{\mathrm{Req}}$ to 512.

Figure 5(a) shows the required energy for data transmissions and data request transmissions, and Fig. 5(b) shows the objective function. As described in the previous subsection, energy is measured in units of m^2. Black dots are the optimal number of cluster computed by using our method. Dash lines in Fig. 5(b) are the area where the number of clusters is smaller than the number of groups. In those areas, a mobile sink cannot collect all data. By using our method, the optimal number of cluster is decided as 17, 23, 32, 47, and 50 when connectivity is 1.0, 0.8, 0.6, 0.4, and 0.2, respectively. These results show that ''traditional method'' which increases the number of clusters is not always the best solution to reduce energy consumption.

VI. CONCLUSION

In this paper, we investigated the challenging issues pertaining to the collection of the ''big data'' generated by densely distributed WSNs. Our investigation suggested that energy-efficient big data gathering in such networks is, indeed, necessary. While the conventional mobile sink schemes can reduce energy consumption of the sensor nodes, they lead to a number of additional challenges such as determining the sink node's trajectory and cluster formation prior to data collection. To address these challenges, we proposed a mobile sink based data collection method by introducing a new clustering method. Our clustering method is based upon a modified Expectation-Maximization technique. Furthermore, an optimal number of clusters to minimize the energy consumption was evaluated. Numerical results were presented to verify the effectiveness of our proposal.

ACKNOWLEDGMENT

Part of this work was conducted under the national project, Research and Development on Cooperative Technologies and Frequency Sharing Between Unmanned Aircraft Systems (UAS) Based Wireless Relay Systems and Terrestrial Networks, supported by the Ministry of Internal Affairs and Communications (MIC), Japan.

REFERENCES

[1] IBM, Armonk, NY, USA. (2012, Jan.). *Four Vendor Views on Big Data and Big Data Analytics: IBM* [Online]. Available: http://www-01.ibm.com/software/in/data/bigdata/

[2] D. Agrawal *et al.*, (2012). *Challenges and Opportunities With Big Data* [Online]. Available: http://cra.org/ccc/docs/init/bigdatawhitepaper.pdf

[3] S. Sagiroglu and D. Sinanc, ''Big data: A review,'' in *Proc. Int. Conf. CTS*, 2013.

[4] *Big Data: Business Opportunities, Requirements and Oracle's Approach*, Winter Corporation, Cambridge, MA, USA, 2011, pp. 1–8.

[5] I. Bisio and M. Marchese, ''Efficient satellite-based sensor networks for information retrieval,'' *IEEE Syst. J.*, vol. 2, no. 4, pp. 464–475, Dec. 2008.

[6] I. Bisio *et al.*, ''A survey of architectures and scenarios in satellite-based wireless sensor networks: System design aspects,'' *Int. J. Satellite Commun. Netw.*, vol. 31, no. 1, pp. 1–38, 2013.

[7] S. Katti, H. Rahul, W. Hu, D. Katabi, M. Medard, and J. Crowcroft, ''XORs in the air: Practical wireless network coding,'' *IEEE/ACM Trans. Netw.*, vol. 16, no. 3, pp. 497–510, Jun. 2008.

[8] K. Miyao, H. Nakayama, N. Ansari, and N. Kato, ''LTRT: An efficient and reliable topology control algorithm for ad-hoc networks,'' *IEEE Trans. Wireless Commun.*, vol. 8, no. 12, pp. 6050–6058, Dec. 2009.

[9] N. Li, J. Hou, and L. Sha, ''Design and analysis of an MST-based topology control algorithm,'' *IEEE Trans. Wireless Commun.*, vol. 4, no. 3, pp. 1195–1206, May 2005.

[10] S. He, J. Chen, D. Yau, and Y. Sun, ''Cross-layer optimization of correlated data gathering in wireless sensor networks,'' in *Proc. 7th Annu. IEEE Commun. Soc. Conf. Sensor Mesh and Ad Hoc Commun. and Netw. (SECON)*, Jun. 2010, pp. 1–9.

[11] C. Jiming, X. Weiqiang, H. Shibo, S. Youxian, P. Thulasiraman, and S. Xuemin, ''Utility-based asynchronous flow control algorithm for wireless sensor networks,'' *IEEE J. Sel. Areas Commun.*, vol. 28, no. 7, pp. 1116–1126, Sep. 2010.

[12] D. Baum and CIO Information Matters. (2013). *Big Data, Big Opportunity* [Online]. Available: http://www.oracle.com/us/c-central/cio-solutions/information-matters/big-dat%a-big-opportunity/index.html

[13] L. Ramaswamy, V. Lawson, and S. Gogineni, ''Towards a quality-centric big data architecture for federated sensor services,'' in *Proc. IEEE Int. BigData Congr.*, Jul. 2013, pp. 86–93.

[14] C.-C. Lin, M.-J. Chiu, C.-C. Hsiao, R.-G. Lee, and Y.-S. Tsai, ''Wireless health care service system for elderly with dementia,'' *IEEE Trans. Inf. Technol. Biomed.*, vol. 10, no. 4, pp. 696–704, Oct. 2006.

[15] P. E. Ross, ''Managing care through the air [remote health monitoring],'' *IEEE Spectr.*, vol. 41, no. 12, pp. 26–31, Dec. 2004.

[16] S. Wen-Zhan, H. Renjie, X. Mingsen, B. A. Shirazi, and R. Lahusen, ''Design and deployment of sensor network for real-time high-fidelity volcano monitoring,'' *IEEE Trans. Parallel Distrib. Syst.*, vol. 21, no. 11, pp. 1658–1674, Nov. 2010.

[17] K. Baumgartner, S. Ferrari, and A. V. Rao, ''Optimal control of an underwater sensor network for cooperative target tracking,'' *IEEE J. Ocean. Eng.*, vol. 34, no. 4, pp. 678–697, Oct. 2009.

[18] R. C. Shah, S. Roy, S. Jain, and W. Brunette, ''Data MULEs: Modeling and analysis of a three-tier architecture for sparse sensor networks,'' *Ad Hoc Netw.*, vol. 1, nos. 2–3, pp. 215–233, 2003.

[19] W. Heinzelman, A. Chandrakasan, and H. Balakrishnan, ''Energy-efficient communication protocol for wireless microsensor networks,'' in *Proc. 33rd Annu. Hawaii Int. Conf. Syst. Sci.*, vol. 2. Jan. 2000.

[20] M. Youssef, A. Youssef, and M. Younis, ''Overlapping multihop clustering for wireless sensor networks,'' *IEEE Trans. Parallel Distrib. Syst.*, vol. 20, no. 12, pp. 1844–1856, Dec. 2009.

[21] T. Khac and C. Hyunseung, ''Connectivity-based clustering scheme for mobile ad hoc networks,'' in *Proc. IEEE Int. Conf. RIVF*, Jul. 2008, pp. 191–197.

[22] S. Lindsey and C. Raghavendra, ''PEGASIS: Power-efficient gathering in sensor information systems,'' in *Proc. IEEE Aerosp. Conf.*, Mar. 2002, pp. 1125–1130.

[23] H. Nakayama, N. Ansari, A. Jamalipour, and N. Kato, ''Fault-resilient sensing in wireless sensor networks,'' *Comput. Commun.*, vol. 30, nos. 11–12, pp. 2375–2384, Sep. 2007.

[24] L. He, Z. Yang, J. Pan, L. Cai, J. Xu, and Y. Gu, "Evaluating service disciplines for on-demand mobile data collection in sensor networks," *IEEE Trans. Mobile Comput.*, vol. 13, no. 4, pp. 797–810, Apr. 2014.

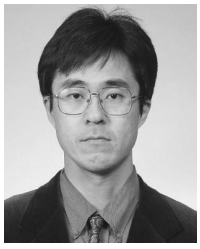
[25] C. Intanagonwiwat, R. Govindan, and D. Estrin, "Directed diffusion: A scalable and robust communication paradigm for sensor networks," in *Proc. 6th Annu. Int. Conf. Mobile Comput. Netw.*, 2000, pp. 56–67.

[26] M. Chen, T. Kwon, and Y. Choi, "Energy-efficient differentiated directed diffusion (EDDD) in wireless sensor networks," *Comput. Commun.*, vol. 29, no. 2, pp. 231–245, 2006.

DAISUKE TAKAISHI (S'13) received the B.E. degree in Information engineering from Tohoku University, Sendai, Japan, in 2013, where he is currently pursuing the M.S. degree with the Graduate School of Information Sciences. He was a recipient of the 2013 IEEE VTS Japan Student Paper Award at the 78th Vehicular Technology Conference. His current research interests include UAS networks and ad hoc networks.

HIROKI NISHIYAMA (SM'13) is an Associate Professor with the Graduate School of Information Sciences, Tohoku University, Sendai, Japan, where he received the M.S. and Ph.D. degrees in information science in 2007 and 2008, respectively. He has authored more than 100 peer-reviewed papers, including many high-quality publications in prestigious IEEE journals and conferences. He was a recipient of the Best Paper Awards from many international conferences, including the IEEE's flagship events, such as the 2013 IEEE Global Communications Conference (GLOBECOM), the 2010 GLOBECOM, and the 2012 IEEE Wireless Communications and Networking Conference. He was also a recipient of the 2013 IEEE Communications Society Asia-Pacific Board Outstanding Young Researcher Award, the 2011 IEICE Communications Society Academic Encouragement Award, and the 2009 FUNAI Foundation's Research Incentive Award for Information Technology. He has served as the Co-Chair for the Cognitive Radio and Networks Symposium of the 2015 IEEE International Conference on Communications (ICC), and the Selected Areas in Communications Symposium of 2014 IEEE ICC, an Associate Editor of the *IEEE Transactions on Vehicular Technology* and the *Springer Journal of Peer-to-Peer Networking and Applications*, and the Secretary of the IEEE ComSoc Sendai Chapter. His research interests cover a wide range of areas, including satellite communications, unmanned aircraft system networks, wireless and mobile networks, ad hoc and sensor networks, green networking, and network security. One of his outstanding achievements is Relay-by-Smartphone, which makes it possible to share information among many people by using only WiFi functionality of smartphones. He is a member of the Institute of Electronics, Information and Communication Engineers.

NEI KATO (F'13) received the bachelor's degree from Tokyo Polytechnic University, Tokyo, Japan, in 1986, and the M.S. and Ph.D. degrees in information engineering from Tohoku University, Sendai, Japan, in 1988 and 1991, respectively. He joined the Computer Center, Tohoku University, as an Assistant Professor in 1991, where he was promoted to Full Professor with the Graduate School of Information Sciences in 2003. He was a Strategic Adviser to the President of Tohoku University in 2013. He has been involved in research on computer networking, wireless mobile communications, satellite communications, ad hoc, sensor, and mesh networks, smart grid, and pattern recognition. He has authored more than 300 papers in peer-reviewed journals and conference proceedings. He currently serves as a Member-at-Large on the Board of Governors, the IEEE Communications Society, the Chair of the IEEE Ad Hoc and Sensor Networks Technical Committee, the Chair of the IEEE ComSoc Sendai Chapter, an Associate Editor-in-Chief of the *IEEE Internet of Things Journal*, an Area Editor of the *IEEE Transactions on Vehicular Technology*, an Editor of *IEEE Wireless Communications Magazine* and the *IEEE Network Magazine*. He has served as the Chair of the IEEE ComSoc Satellite and Space Communications Technical Committee from 2010 to 2012 and the IEICE Satellite Communications Technical Committee from 2011 to 2012), a Guest-Editor of many IEEE transactions/journals/magazines, a Symposium Co-Chair of GLOBECOM'07, ICC'10, ICC'11, ICC'12, the Vice Chair of the IEEE WCNC'10, WCNC'11, ChinaCom'08, ChinaCom'09, a Symposia Co-Chair of GLOBECOM'12, the TPC Vice Chair of ICC'14, and the Workshop Co-Chair of VTC2010. He was a recipient of the Minoru Ishida Foundation Research Encouragement Prize (2003), the Distinguished Contributions to Satellite Communications Award from the IEEE ComSoc, Satellite and Space Communications Technical Committee (2005), the FUNAI Information Science Award (2007), the TELCOM System Technology Award from the Foundation for Electrical Communications Diffusion (2008), the IEICE Network System Research Award (2009), the IEICE Satellite Communications Research Award (2011), the KDDI Foundation Excellent Research Award (2012), the IEICE Communications Society Distinguished Service Award (2012), five Best Paper Awards from the IEEE GLOBECOM/WCNC/VTC, and the IEICE Communications Society Best Paper Award (2012). In addition to his academic activities, he also serves on the expert committee of Telecommunications Council, Ministry of Internal Affairs and Communications, and as the Chairperson of ITU-R SG4 and SG7, Japan. He is a Distinguished Lecturer of the IEEE Communications Society, and a fellow of IEICE.



RYU MIURA received the B.E., M.E., and Dr. Eng. degrees in electrical engineering from Yokohama National University, Yokohama, Japan, in 1982, 1984, and 2000, respectively. He joined the National Institute of Information and Communications Technology (NICT, reorganized from CRL), Tokyo, Japan, in 1984. Since then, he has been involved in research and development on communication systems using satellites and high-altitude/long-endurance aerial platforms. From 1991 to 1992, he was a Visiting Researcher with AUSSAT Pty. Ltd., Sydney, Australia, where he served for the prototyping of mobile satellite communication systems. From 1993 to 1996 and from 2009 to 2011, he was with the Advanced Telecommunications Research Institute, Kyoto, Japan, where he was involved in research and development on digital beamforming antennas and intelligent transport systems for driving safety support, respectively. He is currently the Director of the Dependable Wireless Laboratory, Wireless Network Research Institute, NICT, where he is involved in the wireless systems for disaster-tolerant and body area networks. He is a member of IEICE.