# Joint Clustering and Routing Design for Reliable and Efficient Data Collection in Large-Scale Wireless Sensor Networks

Zhezhuang Xu, *Member, IEEE*, Liquan Chen, Cailian Chen, *Member, IEEE*, and Xinping Guan, *Senior Member, IEEE*

***Abstract*—For data collection in large-scale wireless sensor networks (WSNs), dynamic clustering provides a scalable and energy-efficient solution, which uses cluster head (CH) rotation and cluster range assignment algorithms to balance the energy consumption. Nevertheless, most existing works consider the clustering and routing as two isolated issues, which is harmful to the connectivity and energy efficiency of the network. In this paper, we provide a detailed analysis on the relations between clustering and routing, and then propose a joint clustering and routing (JCR) protocol for reliable and efficient data collection in large-scale WSN. JCR adopts the backoff timer and gradient routing to generate connected and efficient intercluster topology with the constraint of maximum transmission range. The relations between clustering and routing in JCR are further exploited by theoretical and numerical analyses. The results show that the multihop routing in JCR may lead to the unbalanced CH selection. Then, the solution is provided to optimize the network lifetime by considering the gradient of one-hop neighbor nodes in the setting of backoff timer. Theoretical analysis and simulation results prove the connectivity and efficiency of the network topology generated by JCR.**

***Index Terms*—Connectivity, dynamic clustering, energy efficiency, gradient routing, Internet of Things (IoT), large-scale wireless sensor networks (WSNs).**

## I. Introduction

WIRELESS sensor networks (WSNs) play important roles in the Internet of Things (IoT) [1] for collecting large-scale physical data that can be further utilized in various fields such as environment monitoring [2], intelligent transportation system [3], [4], and industrial control [5], [6]. In WSN, a great number of sensor nodes with limited energy are self-organized in a vast region; thus, the data collection in large-scale network meets the challenges of scalability and energy efficiency [7], [8].

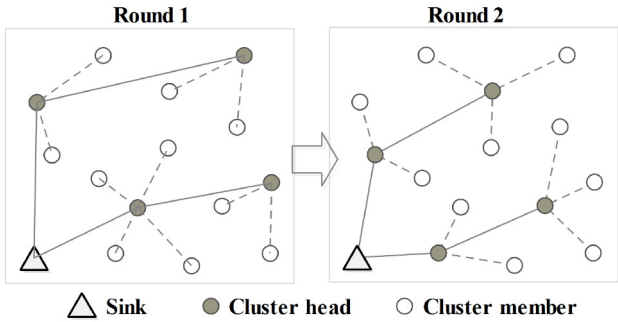


Fig. 1. Sample of dynamic clustering.

Dynamic clustering provides a promising solution for data collection in large-scale WSNs [9]. Sensor nodes are grouped into clusters which have a leader called cluster head (CH) and a number of cluster members (CMs). As shown in Fig. 1, the data are first collected by each CH from its cluster, and then forwarded to the sink through multihops routing. Moreover, the cluster-based topology is reorganized periodically in order to balance the heavy traffic load of CHs.

Due to the advantages of scalability and energy efficiency, dynamic clustering has attracted considerable attentions from different research communities [10]–[16], and has been implemented in various applications [17]–[19]. However, most previous works deal with the dynamic clustering without in-depth consideration on the impact of intercluster routing. We argue that there are tight relations between clustering and routing that may influence the performance of data collection in large-scale WSN.

1) *Energy efficiency* of intercluster communication: The intercluster communication is built among CHs; therefore, the efficiency of intercluster communication depends on not only the node placement but also the CHs selection. Take an example shown in Fig. 1, for the unreasonable CHs selection, the edges among CHs in round 1 is much longer than that in round 2.
2) *Network connectivity* with limited transmission range: Most dynamic clustering protocols assume that the nodes

Manuscript received April 01, 2015; revised July 11, 2015; accepted September 10, 2015. Date of publication September 25, 2015; date of current version July 27, 2016. This work was supported in part by the National Science Foundation (NSF) of China under Grant U1405251, Grant 61304260, Grant 61301096, Grant 61221003, Grant 61290322, and Grant 61273181, in part by the NSF of the Fujian Province of China under Grant 2014J05072, in part by the Ministry of Education of China under Grant NCET-13-0358, and in part by the Science and Technology Commission of Shanghai Municipality (STCSM), China under Grant 13QA1401900.

Z. Xu and L. Chen are with the School of Electrical Engineering and Automation, Fuzhou University, Fuzhou 350000, China, and also with the Fujian Key Laboratory of Industrial Control and Information Security Technology, Fuzhou 350000, China (e-mail: zzxu@fzu.edu.cn; lqchen_fzu@qq.com).

C. Chen and X. Guan are with the Department of Automation, Shanghai Jiao Tong University, Shanghai 200240, China, and also with the Key Laboratory of System Control and Information Processing, Ministry of Education of China, Shanghai 200240, China (e-mail: cailianchen@sjtu.edu.cn; xpguan@sjtu.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JIOT.2015.2482363

have enough transmission power to keep the network connected [9], [10], [13]. For example, in LEACH [9], it is assumed that every node should have enough power to communicate with sink directly. In HEED [10], which has considered multihops intercluster communication, the intercluster transmission range $R_t$ is assumed to be $R_t > 6R_c$, where $R_c$ is the transmission range for intracluster communication. These assumption are hard to be satisfied in large-scale WSN. The node has maximum transmission range $R_{\max}$ which is bounded by its hardware capability. If the network exists in any edge that is longer than $R_{\max}$, the network will be disconnected.

Motivated by these observations, we propose a joint clustering and routing (JCR) protocol which adopts backoff timer [12] and gradient routing [20], [21] to generate connected and efficient network topology for data collection in large-scale WSN. Specifically, the major contributions are summarized as follows.

1) Detailed analysis on the relations between clustering and routing is provided based on several typical dynamic clustering algorithms. The results show that, if the clustering and routing decisions are made separately, the clustering range $R_c$ affects both the connectivity and the energy efficiency of the network. In this case, the energy efficiency can hardly be achieved with the constraint of network connectivity.
2) The JCR protocol is proposed to build the framework of JCR design. Taking advantages of the traffic pattern in data collection scenario, the JCR protocol establishes the gradient field to present the direction through which the sink can be reached. Then, the *gradient* can be exchanged among one-hop neighbors for making the decisions of CH selection and routing discovery. With the help of gradient field, JCR ensures the network connectivity with any value of $R_c$. Therefore, in JCR, the $R_c$ can be freely adjusted to achieve energy efficiency.
3) The relations between clustering and routing in JCR are further exploited by theoretical and numerical analysis. The results show that the multihop routing in JCR may lead to unbalanced CH selection. Then, the solution is provided to optimize the network lifetime by considering the gradient of neighbor nodes in the setting of backoff timer.

This paper is organized as follows. Section II provides a brief survey of related works. Section III provides the network model and evaluates the clustering and routing problem addressed in this work. The details of the JCR protocol are provided in Section IV. Then, we evaluate the network topology generated by JCR in Section V, and the energy efficiency of JCR is further analyzed in Section VI. Finally, Section VII concludes this paper.

## II. Related Works

Dynamic clustering is first proposed in LEACH [9]. The basic idea is the periodic reclustering with a randomized CHs rotation such that energy consumption of CHs can be balanced all over the network. Due to the advantages of scalability and energy efficiency, dynamic clustering has considered to be a promising solution for large-scale data collection in WSNs, and it has attracted considerable attentions from various research communities.

Since the network topology is built based on the set of CHs, the CHs selection is one of the fundamental problems in dynamic clustering. HEED [10] proposes an iteration-based algorithm which considers both residual energy and communication cost in CHs selection. The algorithm improves the CHs distribution in the network and hence has better efficiency than LEACH. Observing that HEED suffers from considerable overhead in the iteration, backoff strategy clustering (BSC) [12] uses the random backoff timer to control the process of CHs selection. The node with smaller backoff time has greater probability to be CH. BSC can generate well-distributed CHs, whereas the overhead in CHs selection is greatly reduced. LEACH-SWDN [22] sets up a sliding window to ensure that the number of CHs is maintained in an optimum range. It improves the energy efficiency after some nodes runs out of energy. Different from the distributed algorithms given above, the research work in [17] uses harmony search algorithm (HSA) to select the CHs with centralized optimization. It is expected to minimize the intracluster communication cost and optimize the energy distribution of the network.

Cluster formation is another important issue in dynamic clustering. A reasonable cluster formation algorithm can balance the traffic load in different clusters, and eventually prolong the network lifetime. EECS [23] considers the difference of the distance between CHs and sink, and then proposes a novel cost function for CMs to select proper CHs based on their locations. The unbalanced energy consumption of intercluster communication is balanced by intracluster communication in EECS, and the network lifetime is significantly improved. A similar cluster formation algorithm is proposed in [11], where the cluster size is balanced by two threshold: 1) the distance of intracluster communication and 2) the number of CMs. Different to the design of cluster formation algorithms, many research works have provided mathematical analysis on how to determine the clustering range in different scenarios [13], [14]. EC [13] considers the multihop data collection scenario, and formulates an optimization problem that determines suitable clustering ranges depending on the hop distance to the sink. The research work in [14] provides a mathematical framework to determine the optimal number of clusters by minimizing the energy consumption in both single-hop and multihop scenarios.

Almost all the aforementioned works deal with the dynamic clustering without in-depth consideration on the impact of intercluster routing. Our work in [24] has a preliminary study on the relations between clustering and routing, and proposes a solution to keep the network connectivity in multihop data collection. However, it does not consider the impact of intercluster routing on the balance of energy consumption, and hence has poor performance in network lifetime.

Compared with related works, this paper distinguishes them in two aspects. 1) We emphasize the relations between clustering and routing in the large-scale data collection. The JCR protocol is designed to select the CHs with joint considerations on the intercluster routing efficiency. On the other hand,

we also analyze how does the multihop routing leads to the unbalanced CHs selection, and provides the guideline to balance the energy consumption. 2) We consider the limitation of transmission range in the design of JCR. Most related works, such as [9], [10], and [13], assume that all sensor nodes have enough transmission power to keep the network connected. However, the limitation of transmission range becomes nontrivial that may impact both the connectivity and the energy efficiency in large-scale WSN.

## III. Preliminaries

### A. Network Model

Considering a set of nodes and a sink deployed in a sensing area, we assume that the sensor network has the following properties.

1) The mission of the sensor network is to collect sensing data from sensor nodes to the sink, such that the destination of every sensing data is the sink.
2) The network is organized into clusters. The CM sends sensing data to its CH *directly* with the clustering range $R_c$, and the CH forwards packets to the sink via multihop relays with the intercluster transmission range $R_t$. The connectivity problem discussed in this paper focuses on the connectivity among CHs.
3) Every node has a maximum transmission range denoted as $R_{\max}$. This motivates the requirement of keeping network connectivity with the limitation of $R_{\max}$.
4) The critical transmitting range (CTR) is less than $R_{\max}$. The CTR is defined as the length of the longest edge of the Euclidean minimum spanning tree [25] built on the nodes. If this assumption is not satisfied, the network will be doomed to be disconnected even if every node is selected as CH and uses its maximum radio power for data transmission. This assumption can be satisfied by reasonable node deployment [26].

It is worth noting that there are *no* assumptions on the following:

- physical location-awareness;
- distance estimation among nodes;
- specific node placement.

The energy consumption follows the popular model given in [9]. To transmit an $l$-bit message within distance $d$, the radio expends

$$E_{\text{Tx}}(l,d) = l \cdot E_{\text{elec}} + l \cdot \epsilon_{fs} \cdot d^2 \tag{1}$$

and to receive this message, the radio expends

$$E_{\text{Rx}}(l) = l \cdot E_{\text{elec}} \tag{2}$$

The CH executes data aggregation after collecting data from its CMs. The size of aggregated data is formulated as [27]

$$l_{\text{CH}} = l(\beta \cdot n_{\text{CM}} + c) \tag{3}$$

where $n_{\text{CM}}$ denotes the number of CMs. $c$ corresponds to the overhead of aggregation, while $\beta$ is the compression ratio. In most related works on dynamic clustering, the aggregation model is set as $\beta = 0, c = 1$. Thus, we use it as the default setting.

TABLE I
Simulation Parameters

| Type | Parameter | Value |
|---|---|---|
| Application | Initial energy | 2 J |
| | Maximum Tx range ($R_{\max}$) | 70 m |
| | Data packet size ($l$) | 125 Bytes |
| | Round | 20 frames |
| Radio model | $E_{\text{elec}}$ | 50 nJ/bit |
| | $\epsilon_{fs}$ | 10 pJ/bit/m$^2$ |
| | $E_{\text{fusion}}$ | 5 nJ/bit/signal |

### B. Problem Statement

As we introduced in Section I, most related works have not considered multihop routing issues in CHs selection which may lead to the disconnection and inefficiency of the intercluster topology. To verify these problems, we use some simulations to study how does the cluster range assignment impact the network topology and energy efficiency in LEACH [9] and BSC [12].

There are 400 nodes randomly deployed in an area of $200 * 200\,\text{m}^2$, and the sink locates at coordinate $[0, 0]$. The setting of other parameters is given in Table I. In BSC, the cluster size can be controlled by tuning $R_c$, whereas in LEACH, the cluster size can only be adjusted by setting the expected number of CHs $N_{\text{CH}}$. To clarify the comparison, all simulation results are demonstrated with respect to $R_c$. When $R_c$ grows from 20 to 70 m, the expected numbers of CHs in LEACH are set as $N_{\text{CH}} = [58.3, 30.3, 18.6, 12.9, 9.6, 7.7]$, respectively. The value of $N_{\text{CH}}$ is obtained by calculating the average number of CHs in BSC when $R_c$ grows from 20 to 70 m.

Since both LEACH and BSC do not include multihop routing protocol, a greedy routing algorithm is adopted in LEACH and BSC. In the greedy routing algorithm, every node is assumed to know its distance to the sink. The CH $i$ selects the closest CH which is closer to the sink than itself as its relay. The algorithm can be formulated as

$$r(i) = \min\{d(i,j) | d(i,0) > d(j,0), j \in \mathcal{H}\} \tag{4}$$

where $d(i,j)$ is the distance between node $i$ and $j$, and $d(i,0)$ is the distance between node $i$ and the sink. $\mathcal{H}$ denotes the set of CHs.

At first, we evaluate how does the cluster range assignment impact the intercluster topology. The length of edges among CHs (LEH for short) is used as the metric to measure the intercluster topology. Each simulation runs 50 times with different node deployments. At each deployment, the protocols run 20 rounds for selecting different sets of CHs. The results are given in Fig. 2. When $R_c$ grows from 20 to 70 m, in BSC, the average LEH grows from 24.92 to 75.96 and the maximum LEH grows from 53.87 to 138.31, while in LEACH, the average LEH grows from 19.35 to 59.34, and the maximum LEH grows from 91.68 to 211.05.

Then, we study the impact of $R_c$ on network lifetime. Each simulation runs 50 times with different node deployments. At each deployment, the network runs until the first node exhausts out its energy. The results are shown in Fig. 3. In LEACH, the

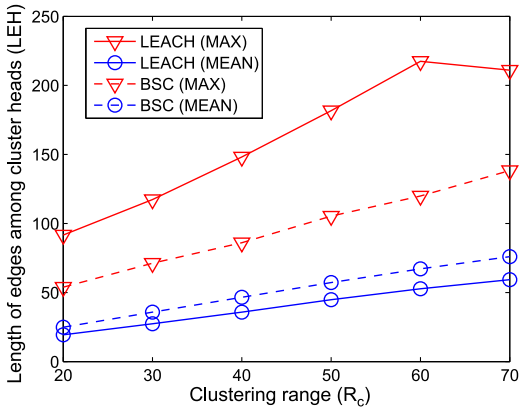


Fig. 2. Length of edges among CHs (LEH) in LEACH and BSC.

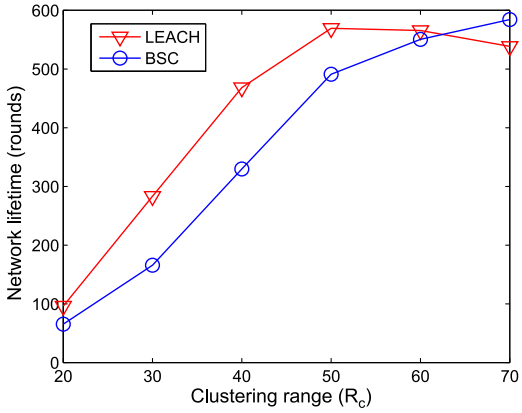


Fig. 3. Network lifetime in LEACH and BSC.

network lifetime grows from 96 rounds when $R_c = 20$ m and achieves maximum at 569 rounds when $R_c = 50$ m. Then, the network lifetime slightly drops to 539 rounds at $R_c = 70$ m. On the other hand, in BSC, the network lifetime monotonically increases from 65 rounds to 584 rounds when $R_c$ grows from 20 to 70 m.

There are two important results that can be derived from the simulation results.

1) The average and maximum values of LEH have great difference. It demonstrates that the intercluster topology is variable due to the randomness of CHs selection. The variance of LEH in BSC is smaller than that in LEACH, which proves that the CHs selection of BSC is more reasonable than that of LEACH. Nevertheless, in BSC, the maximum LEH is still over 150% compared with the average LEH.
2) Both LEH and network lifetime have tight relations with the clustering range $R_c$. For the network connectivity, $R_c$ should be small enough to ensure that the maximum LEH is smaller than $R_{\max}$. However, the maximum network lifetime can hardly be achieved when $R_c$ is not large enough.

To the best of our knowledge, few works have considered the routing issues in the CHs selection, which leads to the variance of LEH shown in Fig. 2. It is harmful to the connectivity and energy efficiency of the network. In this paper, our goal is to design a JCR protocol which can generate connected and efficient intercluster topology. Specifically, the design goals are given as follows.

1) The clustering and routing are hybrid, i.e., after executing the JCR, each node is either a CH or a CM, and it has selected a node as its next hop node.
2) The LEH should be well controlled to ensure both the energy efficiency and the network connectivity with the constraint of maximum transmission range $R_{\max}$.
3) The algorithm should be completely distributed with low overhead.

TABLE II
VARIABLES USED IN THIS PAPER

| Symbol | Description |
|---|---|
| $R_c$ | Clustering range |
| $R_t$ | Intercluster transmission range |
| LEH | Length of edges among cluster heads |
| $R_{\max}$ | Maximum transmission range, determined by hardware |
| $g$ | Gradient, which denotes the minimum hop count to the sink |
| $\mathcal{R}(k)$ | $k$th ring, the set of nodes with $g = k$ |
| $F$ | Set of forward nodes |
| $B$ | Set of backward nodes |
| $Q$ | Set of equal nodes |
| $N_f$ | Number of forward nodes |
| $N_b$ | Number of backward nodes |
| $N_q$ | Number of equal nodes |
| $d(i, j)$ | Distance between node $i$ and $j$ |
| $d(i)$ | Distance from node $i$ to the sink |
| $\Delta d(i)$ | Shortest distance from the node $i$ ($g(i) = k$) to the edge of $\mathcal{R}(k-1)$ |

## IV. JCR PROTOCOL DESCRIPTION

The JCR protocol is described in a typical dynamic clustering scenario [9] where the network operation is divided into rounds. Each round is composed of setup phase and data transmission phase. In the setup phase, every node begins with the status of candidate (CA) and then chooses to be a CM or a CH. In the data transmission phase, each CH periodically collects the sensing data in its cluster, and then forwards the aggregated result to the sink by intercluster communication. After that, each node retires to CA status, and a new round begins.

The goal of JCR is to select the set of CHs that not only organizes well-balanced clusters, but also builds up the connected and efficient backbone for intercluster routing. The basic idea of gradient routing [20] is adopted in JCR to provide a simple and efficient solution for distributed routing among CHs. In the gradient field establishment algorithm, every node obtains its *gradient* to establish the gradient field that presents the direction through which the sink can be reached. Then, the gradient information is exchanged distributively among the nodes for making the decision of *CH selection* and *routing discovery*.

In order to control the topology, each node has two transmission ranges: 1) the clustering range $R_c$ and 2) the intercluster transmission range $R_t$. How to determine the value of $R_c$ and $R_t$ will be discussed in Section V. To clarify the statement, the variables used in this paper are summarized in Table II.

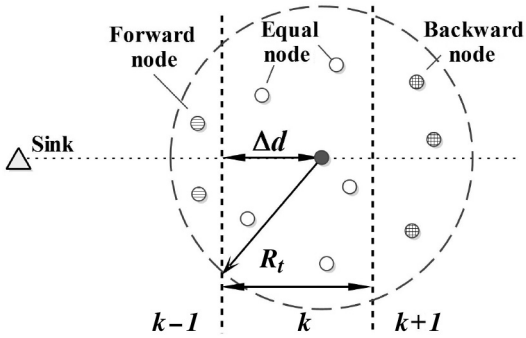


Fig. 4. Analysis of neighbor nodes in JCR.

### A. Gradient Field Establishment

To realize the joint design of clustering and routing, some information should be obtained at the beginning of network operation. To clarify the description, we provide several definitions as follows.

*Definition 1:* (Gradient) Given a transmission range $R_t$, the gradient of node $i$, which is denoted as $g(i)$, is the minimum hop count by forwarding a packet from node $i$ to the sink with $R_t$.

*Definition 2:* (Ring) The set of nodes, which have the same gradient $k$, is defined as Ring $k$ which is denoted by

$$\mathcal{R}(k) = \{i : g(i) = k, i \in \mathcal{S}\} \tag{5}$$

where $\mathcal{S}$ is the set of all sensor nodes.

*Definition 3:* (Distance) Given node $i$ with gradient $k$, the distance from the node $i$ to the sink is denoted as $d(i)$. The shortest distance from the node $i$ to the edge of $\mathcal{R}(k-1)$ can be formulated as $\Delta d(i)$

$$\Delta d(i) = d(i) - (k-1)R_t \tag{6}$$

Fig. 4 depicts an example of neighbor nodes. Based on the definitions of the gradient, the neighbor nodes can be classified into *forward nodes*, *equal nodes*, and *backward nodes* according to their gradient. The definitions are given as follows.

*Definition 4:* (Forward node) Given node $i$ with gradient $k$, the node which locates within the range $R_t$ of the node $i$ and has gradient $k-1$ is the forward node of node $i$. The set of forward nodes of node $i$ is denoted by

$$\mathcal{F}(i) = \{j : d(i,j) < R_t \cap g(j) = k-1\} \tag{7}$$

Similarly, we have the definition of backward node and equal node.

*Definition 5:* (Backward node) Given node $i$ with gradient $k$, the node which locates within the range $R_t$ of the node $i$ and has gradient $k+1$ is the backward node of node $i$. The set of backward nodes of node $i$ is denoted by

$$\mathcal{B}(i) = \{j : d(i,j) < R_t \cap g(j) = k+1\} \tag{8}$$

*Definition 6:* (Equal node) Given node $i$ with gradient $k$, the node which locates within the range $R_t$ of the node $i$ and has gradient $k$ is the equal node of node $i$. The set of equal nodes of node $i$ is denoted by

$$Q(i) = \{j : d(i,j) < R_t \cap g(j) = k\} \tag{9}$$

Based on the definition of equal node, we define *equal clustering nodes* which will be used in the analysis in Section VI.

*Definition 7:* (Equal clustering node) Given node $i$ with gradient $k$, the node which locates within the clustering range $R_c$ of the node $i$ and has gradient $k$ is the equal clustering node of node $i$. The set of equal clustering nodes of node $i$ is denoted by

$$Q_c(i) = \{j : d(i,j) < R_c \cap g(j) = k\} \tag{10}$$

In JCR, the goal of gradient field establishment algorithm is to obtain the gradient $g$, the number of forward nodes $N_f$, and backward nodes $N_b$ for every node. These information will be used in the CHs selection algorithm.

The initial value of $g$, $N_f$, and $N_b$ is set as 0. According to the definitions, the gradient field can be established by flooding a GFE packet generated by the sink. The process of gradient field establishment is divided into frames: At frame 0, the sink generates the GFE packet message with its gradient $g = 0$, and then broadcast it in the range $R_t$. The node received this GFE packet reads the $g_{\text{FE}}$ and sets its gradient $g = g_{\text{FE}} + 1$, i.e., $g(i) = 0 + 1$. Then, it broadcasts the GFE packet with its gradient $g = 1$ in $R_t$ at frame 1. The nodes with the same $g$ should broadcast GFE packets in the same frame; hence, the length of the frame should be long enough to ensure that the nodes have enough time to broadcast their GFE packets. The broadcasting process repeats until all nodes have received GFE packet and set their own gradient.

A node may receive more than one GFE packets. When it receives the GFE packet at the $first$ time, it sets its gradient $g = g_{\text{FE}} + 1$ and set its $N_f = 1$. After that, at each time, it receives a GFE packet, it compares $g_{\text{FE}}$ with its own $g$. If $g_{\text{FE}} = g - 1$, a forward node is found, and the $N_f$ should plus 1. Similarly, if $g_{\text{FE}} = g + 1$, a backward node is found, and the $N_b$ should plus 1.

If a node broadcasts GFE packet in frame $k$ and does not receive any GFE packet in the frame $k+1$, the node will transmit a report (RPT) packet back to the sink. By receiving the RPT packets, the sink can obtain the maximum gradient $G$ in the network. The value of $G$ depends on the node deployment and $R_t$. At the end of gradient field establishment, the sink floods another GFE packet with maximum gradient $G$ over the network, such that every node can obtain $G$.

The packet collision during the gradient field establishment may lead to the miscount of $N_f$ and $N_b$. The CSMA is used to reduce packet collision in the GFE. Moreover, in JCR, each packet has to carry the gradient of its transmitter. The node can use them to upgrade its $N_f$ and $N_b$ during data transmission.

It is worthy to note that the gradient field establishment operates only *once* at the beginning of the network operation; thus, it does not impact the complexity of JCR.

### B. CH Selection

The CH selection algorithm is driven by the backoff scheme [12]. The node decides whether to be CH or CM according to its backoff timer $T_b$ and the advertisement (ADV) packets received before $T_b$ terminates.

At the beginning, the CA sets up a backoff timer $T_b$ according to its gradient $g$, the number of its neighbor nodes $N_f$,$N_b$, and its residual energy $e$

$$T_b = \begin{cases} [(G-g)+\alpha(\frac{E-e}{E})+(1-\alpha)(\frac{N_F-N_f}{N_F})] \times \tau & g > 1 \\ [(G-1)+\alpha(\frac{E-e}{E})+(1-\alpha)\frac{N_b}{N_B}] \times \tau & g = 1 \end{cases} \tag{11}$$

where $G$ is the maximum gradient in the network and $E$ is the initial energy of the node. $N_F$ is the maximum number of forward nodes and $N_B$ is the maximum number of backward nodes. Both $N_F$ and $N_B$ can be approximated by

$$N_F = N_B = \rho \cdot \frac{\pi R_t^2}{2} \tag{12}$$

where $\rho$ is the node density of the network.

According to (11), the process of CHs selection breaks into $G$ frames whose length is $\tau$. The set of CHs in Ring $k$ ($\mathcal{R}(k)$) is selected in the $(G-k)$th frame. The first set of CHs is selected from $\mathcal{R}(k)$, and then go through the gradient field until reaching the sink.

For the nodes in the same *Ring*, the node which has more forward nodes or more residual energy has smaller backoff time $T_b$. The $\alpha$ which is set in range $[0,1]$ is the parameter to adjust the weight between the number of forward nodes $N_f$ and the residual energy $E_r$. The impact of $\alpha$ on the network performance will be studied in Section IV-C.

For the nodes with $g = 1$, all their forward node is the sink; thus, the factor of $N_f$ cannot be used to adjust $T_b$ in $\mathcal{R}(1)$. Therefore, for the node in $\mathcal{R}(1)$, their $T_b$ depends on the number of their backward nodes $N_b$.

After setting $T_b$, the CA listens to the channel for incoming packets. There are two kinds of ADV packets used for CHs selection: 1) the ADV packet for clustering (AC) and 2) the ADV packet for routing (AR). When the CA is determined to be CH, it broadcasts AC packet in $R_c$ and AR packet in $R_t$. Both AC and AR carry the following information of the transmitter CH: identifier (ID), gradient ($g$), residual energy ($E_r$), and NEXT flag, which denotes if the transmitter CH has found its next hop relay. Based on received packets, every node updates the following data lists:

- JOIN list: records other CHs with clustering range $R_c$;
- SERV list: records the requests for relay from CHs.

When the CA receives AC packet, it records the information of the transmitter CH to the JOIN list. The JOIN list will be used for the node to select its CH when it is determined to be CM. When the CA receives AR packet with $g_{\text{AR}} > g$, it will check the NEXT flag in the AR packet. If NEXT $= 0$, which means that the transmitter CH has not found a forward CH, the CA will record the transmitter information to the SERV list. If NEXT $= 1$ and the transmitter information has been recorded in the SERV list, the CA will delete the transmitter from the SERV list. The setting of $T_b$ (11) ensures that the CA cannot receive any AR packet with $g_{\text{AR}} < g$.

When the backoff timer $T_b$ terminates, the node checks the JOIN and SERV lists. The CA will be selected as CH if at least one of the following conditions is satisfied.

- The JOIN list is empty. It means that the node has the smallest backoff time compared with its neighbor nodes (i.e., more forward nodes or more residual energy).
- The SERV list is *not* empty. It means that the node is required to be a relay for another CH.

Otherwise, if the JOIN list is not empty *and* the SERV list is empty, the CA sets its status as CM.

**Algorithm 1.** Cluster Head Selection

```
Set up backoff timer T_b;
repeat
  if Receive AC packet then
    Record transmitter to the JOIN list;
  end if
  if Receive AR packet AND g < k_AR then
    if NEXT=0 then
      Record transmitter to the SERV list;
    else if Transmitter has recorded in the SERV list then
      Delete transmitter from the SERV list;
    end if
  end if
until T_b terminates
if SERV list is not empty OR JOIN list is empty then
  status ← CH;
else
  status ← CM;
end if
```

### C. *Routing Discovery*

The goal of routing discovery algorithm is to find a path to the sink for every node. For simplicity, in this section, we only describe how does every node select its next-hop node. We will prove that this algorithm ensures the network connectivity in Section V.

For the CM, the next-hop node is its CH. The CM with $g(i)$ should continue to receive AC packets until the end of $G - g(i) + 1$th frame. Then it selects the CH which has the highest residual energy in the JOIN list as its own CH.

For the CH with $g = 1$, its next-hop node is destined to be the sink. For other CHs, the relays are selected based on the following algorithm: According to (11), the CH $i$ can not find its relay in the $(G - g(i))$th frame, hence the first AR packet is broadcasted with NEXT=0. Its forward nodes will receive the AR packet and record it to the SERV list. According to the CHs selection algorithm, the forward node $j$ with the smallest $T_b$ becomes CH when its timer terminates. Then it broadcasts AR packet in $R_t$. In this case, the CH $i$ receives the AR packet with $g_{\text{AR}} < g(i)$. Thus, the CH $i$ selects CH $j$ as its relay and broadcasts the second AR packet with NEXT=1 to announce that it finds a relay.

## V. Network Topology Analysis

The JCR protocol organizes the network into clusters whose size is determined by the clustering range $R_c$, and the CH uses the transmission range $R_t$ to forward its data. According to the

results obtained in [28], to minimize the energy consumption, every CH should use its maximum transmission range for intercluster communication. With joint consideration of network connectivity, the value of $R_c$ and $R_t$ should be constrained by

$$R_c \leq R_t = R_{\max} \tag{13}$$

Assume that the (13) holds, the network topology generated by JCR has the following properties.

*Lemma 1:* Each node has at least one forward node.

*Proof:* In the gradient field establishment, the GFE packet is flooding over the network with range $R_t$. Since the CTR is less than $R_t$ (Assumption 4 in Section III), every node can receive the GFE packet. The gradient of the node is set as the $g_{FE}+1$ when it receives GFE packet at the first time, hence the node has at least one forward node which is the transmitter of its first received GFE packet. ■

*Lemma 2:* At the end of the setup phase, each node is either a CH or a CM.

*Proof:* A similar theorem has been proved in [10]. Therefore, the proof is omitted. ■

*Lemma 3:* At the end of the setup phase, each CH has selected a forward CH or the sink as its next-hop node.

*Proof:* For the CH with $g=1$, its next-hop node is the sink. For the CH with $g>1$, it first broadcasts AR packet with NEXT $=0$ in $R_t$. According to Lemma 1 and (11), there is at least one forward CA that can receive the AR packet. The CA records transmitter information to the SERV list. Since the SERV list is not empty, the CA with the lowest $T_b$ will become CH and serves as the relay for the transmitter CH. ■

*Theorem 1:* In JCR, a CH with gradient $k$ transmits its data to the sink in $k$ hops.

*Proof:* According to Lemma 3 and the definition of forward node (Definition 4), the data from CH with $g=k$ can be transmitted to the CH with $g=k-h$ in $h$ hops. The gradient of the sink is 0; thus, the CH with $g=k$ can transmit its data to the sink in $k$ hops. ■

*Theorem 2:* In JCR, a CM with gradient $k$ transmits its data to the sink in $k$ to $k+2$ hops.

*Proof:* In JCR, the CM selects its CH which has the highest residual energy within range $R_c$. Since $R_c$ is constrained by $R_c \leq R_t = R_{\max}$ (13), the gradient of its CH can be $k-1$, $k$, or $k+1$. Combing with Theorem 1, the CM with $g=k$ can transmit its data to the sink in $k$ to $k+2$ hops. ■

*Theorem 3:* The JCR protocol generates a connected network topology.

*Proof:* According to Lemma 2, each node is either a CH or a CM by running JCR protocol. According to Theorems Theorem 1 and 2, every node can transmit its data to the sink with limited hops. Therefore, the network is connected. ■

Based on the analytical results given above, the JCR protocol can generate connected and efficient intercluster topology. Then, we use simulations to analyze the network topology.

### A. Connectivity

The performance of JCR are compared with BSC [12] which uses similar strategy to select CHs. BSC uses the greedy routing, which has been stated in Section III-B, to generate intercluster topology. In JCR, $R_{\max}$ and $R_t$ are both set to be 70 m, while the BSC does not have limitation on the transmission range.

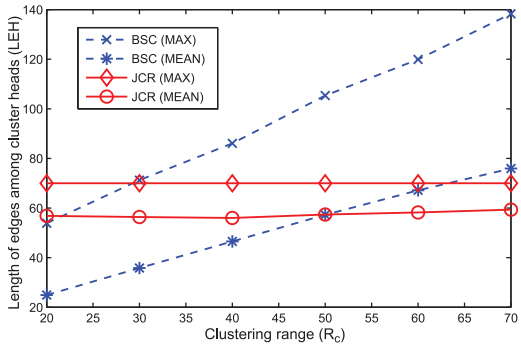


Fig. 5. Length of edges among CHs (LEH).

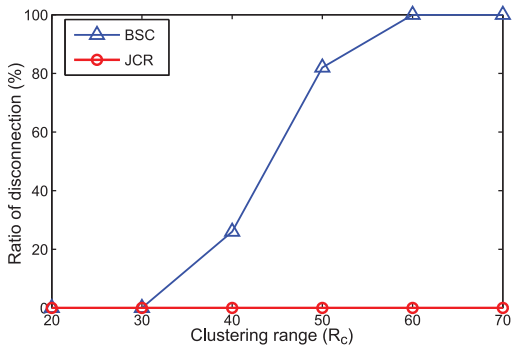


Fig. 6. Ratio of disconnection.

There are 400 nodes randomly deployed in an area of $200*200\,\mathrm{m}^2$, and the sink locates at $[0,0]$. Both BSC and JCR are running in 50 different deployment which runs 20 rounds to select different sets of CHs. The cluster-based topology is reorganized at the beginning of every round. The clustering range $R_c$ is adjusted from 20 to 70 m.

At first, we compare the LEH to evaluate the intercluster topology in Fig. 5. In BSC, when $R_c$ is set from 20 to 70 m, the average LEH grows from 25 to 82.9 m and the maximum LEH grows from 46 to 127.7 m. The LEH of BSC increases almost linearly with the growth of $R_c$. When $R_c$ is larger than 30 m, the maximum LEH of BSC grows larger than $R_{\max}$, which leads to the disconnection of the network. On the other hand, in JCR, the maximum LEH is tightly bounded by $R_t=70\,\mathrm{m}$, and the average LEH ranges from 56 to 59 m.

Fig. 6 shows the ratio of disconnection in the network. The ratio of disconnection is calculated by the percentage of rounds in which the network topology exists any edge that is longer than $R_{\max}$. In BSC, the ratio of disconnection is 26% when $R_c=40\,\mathrm{m}$, and it grows sharply to 82% when $R_c=50\,\mathrm{m}$. When $R_c$ grows larger than 60 m, the network is doomed to be disconnected at every round. Therefore, in BSC, the $R_c$ has to be smaller than 40 m in order to keep the network connectivity. However, the energy efficiency can hardly be achieved in this case [10], [12]. On the other hand, the topology generated by JCR is ensured to be connected with any value of $R_c$.

The reason that JCR provides connected topology can be found in Fig. 7. When $R_c$ grows from 20 to 70 m, the number

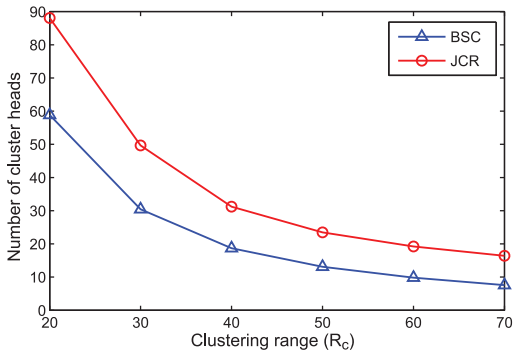


Fig. 7. Number of CHs.

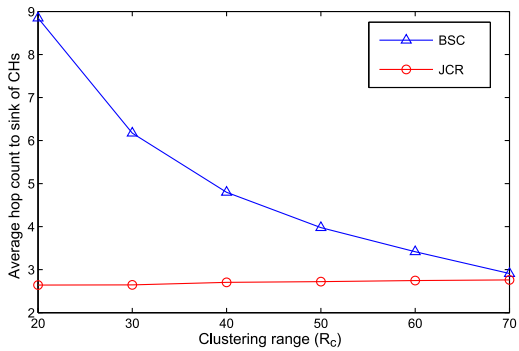


Fig. 8. Average hop count from cluster heads to sink.

TABLE III
HOP COUNT TO SINK

| Gradient | $g=1$ | $g=2$ | $g=3$ | $g=4$ |
|---|---|---|---|---|
| Maximum hop (CH) | 1 | 2 | 3 | 4 |
| Average hop (CH) | 1 | 2 | 3 | 4 |
| Maximum hop (CM) | 1 | 4 | 5 | 5 |
| Average hop (CM) | 1 | 2.89 | 4.13 | 4.78 |

of CHs in JCR drops from 88 to 16, while the number of CHs in BSC drops from 59 to 8. JCR protocol generates more CHs than BSC to guarantee the connectivity among CHs.

### B. *Efficiency of the Topology*

The hop count from the node to the sink is used as the metric to evaluate the efficiency of network topology generated by JCR. To clarify the relations between the hop count and the gradient, the hop count is recorded separately with different $g$. $R_c$ is set to be 70 m. The results are given in Table III.

For CHs, the hop count is the same as the gradient $g$, which proves the results derived in Theorem 1. It is because the CHs selection algorithm ensures that the relay of a CH has to be its forward node. Therefore, the transmission path from CHs to the sink is ensured to be forward. For CMs, the average hop count is larger than $g$; the results depend on the CH that the CM selected. Nevertheless, the hop count of CM is tightly bounded by $k+2$, which has been proved in Theorem 2.

Then, we compare the average hop count of CHs between BSC and JCR. As shown in Fig. 8, when $R_c$ grows from 20 to 70 m, the average hop count of CHs in BSC drops from 8.8 to 2.9. In BSC, the hop count is related to the number of CHs. In contrast, in JCR, $R_c$ has little impacts on the average hop count of CHs. When $R_c$ ranges from 20 to 70 m, the average hop count of CHs grows from 2.64 to 2.76. A larger $R_c$ generates larger area of clusters; thus, the CM with $g=k$ have higher probability to select the CH with $g=k$ or $g=k+1$. It leads to the slight growth of hop count when $R_c$ increases. Nevertheless, the hop count in JCR is still smaller than that of BSC when $R_c=R_{\max}$. It proves that the transmission path generated by JCR is more efficient than BSC, and the advantage is greater when $R_c$ is smaller.

To summarize, the analytical and numerical results demonstrate that the JCR protocol generates connected and efficient intercluster topology. It is worth to note that the gradient field is built by flooding GFE packet with $R_{\max}$; thus, the value of $g$ is equal to the minimum hop count from the node to the sink. Therefore, the JCR can provide optimal transmission path for CHs, and suboptimal path for CMs. Moreover, the analysis results show that the JCR protocol decouples the relations between $R_c$ and the intercluster topology. Thus, the research works of cluster range assignment [13] can be used in JCR with maximum transmission range constraint.

## VI. ENERGY CONSUMPTION ANALYSIS

The CHs selection algorithm of JCR ensures that the CH selects one of its forward nodes as its relay. As we discussed in Section V, this scheme ensures the connectivity and the efficiency of intercluster topology. In this section, we will provide analytical and numerical results to show that this scheme may lead to the imbalance of energy consumption among CHs, and the number of neighbor nodes plays an important role in this problem. Furthermore, we also prove that the energy imbalance problem can be solved by tuning the factor of $\alpha$ in the backoff timer $T_b$.

### A. *Probability to be CHs*

According to the CHs selection algorithm described in Section IV-B, in JCR, the node will be selected as CH in two conditions when its backoff timer $T_b$ terminates.

1) SERV list is not empty. It means that at least one of its backward nodes has been selected as CH, and the node will be selected as CH and serve as its relay. In this case, the probability in this condition can be written as

$$p_1(i)=1-\prod_j [1-p_b(i,j)], j\in\mathcal{B}(i) \qquad (14)$$

where $p_b(i,j)$ is the probability that the node $j$ is selected as CH and chooses the node $i$ as its relay. In JCR, the relay of node $j$ is one of its forward nodes which has the smallest $T_b$. Given $p_f(i,j)$ as the probability that the node $i$ has the smallest $T_b$ in $\mathcal{F}(j)$, $p_b(i,j)$ can be formulated as

$$p_b(i,j)=P_{\mathrm{CH}}(j)\cdot p_f(i,j) \qquad (15)$$

where $P_{\mathrm{CH}}(j)$ is the probability that node $j$ is selected as CH.

2) Both SERV list and JOIN list is empty. It means that, if the first condition given above is not satisfied, the node which has the smallest $T_b$ in $N_e$ will be the CH. Therefore, the probability in this condition is

$$p_2(i) = p_e(i,m) \cdot \prod_j [1 - p_b(i,j)], m \in Q_c(i), j \in \mathcal{B}(i) \tag{16}$$

where $p_e(i,m)$ denotes the probability that the node $i$ has the smallest $T_b$ in $Q_c(i)$ (Definition 7). If the node has no backward node (e.g., the nodes with $g = G$), the probability will be determined by the $T_b$ competition in $Q_c(i)$. The probability can be written as

$$p_2(i) = p_e(i,m), N_b(i) = 0, m \in Q_c(i) \tag{17}$$

Combining (14), (16), and (17), the probability that the node $i$ is selected as CH can be formulated as

$$P_{\text{CH}}(i) = \begin{cases} 1 - (1 - p_e(i,m)) \cdot \prod_j [1 - p_b(i,j)] & N_b(i) \neq 0 \\ p_e(i,m) & N_b(i) = 0. \end{cases} \tag{18}$$

Then, we use simulations to prove the correctness of (18). In the simulations, there are 400 nodes randomly deployed in an area of $200 * 200\,\text{m}^2$, and the sink locates at $[0,0]$. To clarify the analysis, all the nodes set its $R_c = R_t = 70\,\text{m}$, and they use a simple backoff timer

$$T_b = (G - g) + \{\text{rand}\} \tag{19}$$

where rand is a random variable whose value is selected from $[0,1]$ with uniform distribution. In this case, $p_f(i,j)$ can be simplified as $\frac{1}{N_f(j)}$.

At first, every node uses (18) to estimate its probability of being CH. The algorithm is given as follows: the calculation starts from the nodes with $g = G$. Since all the nodes with $g = G$ have no backward node and $R_c = R_t$, all their CH probability can be estimated by $P_{\text{CH}}(i) = p_e(i,m) = \frac{1}{N_e(i)}$. After that, the CH probability of nodes with $g = G - 1$ can be calculated by (18). The calculation repeats until $g = 1$.

On the other hand, the network runs 1000 rounds to obtain different sets of CHs. We can obtain the probability of being CH by recording the times to be CH for every node. The results are given in Fig. 9. To make the results clear to be read, we only demonstrate the nodes whose ID ranges from 1 to 50. There are two important results that we can derive from Fig. 9.

1) Equation (18) can estimate the probability of being CH with acceptable error.
2) The probability of being CH is imbalanced when $T_b$ is set as (19), i.e., without the consideration of residual energy and neighbor nodes.

The results inspire us to look into (18) to exploit the factors that influence $P_{\text{CH}}(i)$. There are two parts in (18): 1) $p_e(i,m)$ and 2) $\prod_j [1 - p_b(i,j)]$. $p_e(i,m)$ is the probability that node $i$ has the shortest $T_b$ in $Q_c(i)$. Therefore, it is determined by the backoff timer $T_b$ and the clustering range $R_c$. $p_e(i,m)$ increases with smaller $R_c$ or larger $T_b$, and so the $P_{\text{CH}}(i)$.

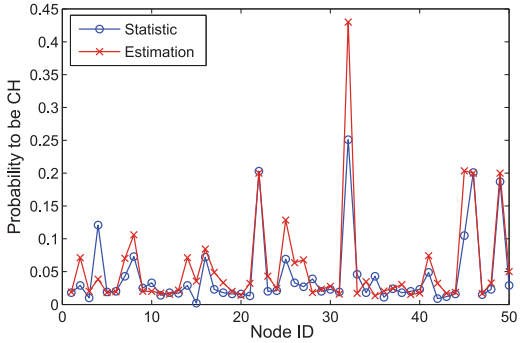


Fig. 9. Estimation of the CH probability.

For $\prod_j [1 - p_b(i,j)]$, it depends on the number of backward nodes $N_b$, and $p_b(i,j)$. According to (15), $p_b(i,j)$ is determined by the CH probability of backward nodes $P_{\text{CH}}(\mathcal{B}(i))$ and the number of their forward nodes $N_f(\mathcal{B}(i))$. Since $p_b(i,j) <= 1$, when $N_b$ grows, $\prod_j [1 - p_b(i,j)]$ reduces and $P_{\text{CH}}(i)$ grows up.

To sum up, we have the following results.

*Remark 1:* The probability of being CH $P_{\text{CH}}(i)$ in JCR depends on the following factors.

1) The backoff timer $T_b$ and the clustering range $R_c$. $P_{\text{CH}}(i)$ grows when $T_b$ and/or $R_c$ is getting smaller.
2) The number of neighbor nodes. $P_{\text{CH}}(i)$ grows with larger $N_b(i)$.
3) The CH probability of backward nodes, i.e., $P_{\text{CH}}(\mathcal{B}(i))$. $P_{\text{CH}}(i)$ increases with the growth of $P_{\text{CH}}(\mathcal{B}(i))$.

Based on the analysis given above, in Section VI-B, we study how does neighbor nodes leads to imbalanced CHs selection shown in Fig. 9. Then, we provide the guideline of setting the backoff timer to balance the energy consumption in Section VI-C.

### B. Impact of Neighbor Nodes

At first, we study the relations between the node's location and its neighbor nodes. As shown in Fig. 4, given $\Delta d(i)$ and node density $\rho$, the number of forward nodes $N_f$ can be estimated by euclidean geometry

$$N_f(i) = \rho \cdot \left( R_t^2 \cdot \arccos \frac{\Delta d(i)}{R_t} - \Delta d(i) \cdot \sqrt{R_t^2 - \Delta d(i)^2} \right) \tag{20}$$

Similarly, the number of backward nodes $N_b$ is

$$N_b(i) = \rho \cdot \left( R_t^2 \cdot \arccos \frac{R_t - \Delta d(i)}{R_t} - \left( R_t - \Delta d(i) \right) \cdot \sqrt{R_t^2 - \left( R_t - \Delta d(i) \right)^2} \right) \tag{21}$$

and the number of equal nodes $N_e$ is

$$N_e(i) = \rho \cdot \pi R_t^2 - N_b - N_f \tag{22}$$

Then, we use simulations to study the relations between $\Delta d$ and neighbor nodes. To reduce the boundary effect, the network area is set as $500 * 100\,\text{m}^2$. There are 500 nodes randomly

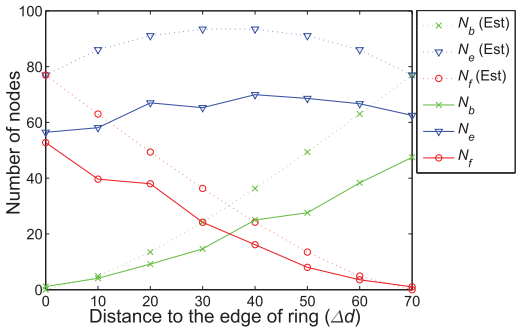


Fig. 10. Number of neighbor nodes, (Est) is estimated by (20)–(22).

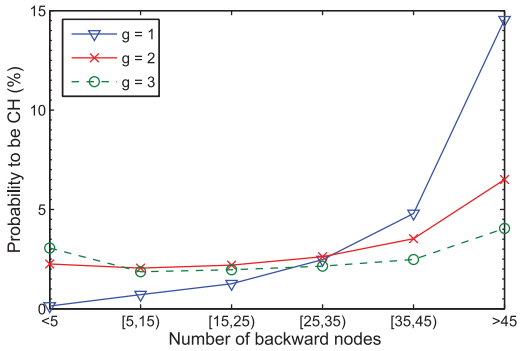


Fig. 11. Probability to be CHs with respect to $g$ and $N_b$ (no $N_b$ in $T_b$).

deployed in the area and the sink locates at $[0, 50]$. At first, we focus on the nodes located at $\mathcal{R}(3)$. The results are given in Fig. 10. For comparison, Fig. 10 also presents the results calculated by (20)–(22).

Due to the boundary effect, the number of neighbor nodes in simulations are smaller than that calculated by (20)–(22). Nevertheless, (20)–(22) are sufficient to demonstrate the relations between $\Delta d$ and neighbor nodes. As shown in Fig. 10, $N_b$ increases with the growth of $\Delta d$. On the other hand, $N_f$ reduces when $\Delta d$ grows. The variance of $N_e$ is relatively small, and its maximum value appears when the node locates at the middle of the *Ring*.

The variance of $P_{\text{CH}}$ and neighbor nodes shown in Figs. 9 and 10 motivate us to study their relations. Fig. 11 shows $P_{\text{CH}}$ with respect to $g$ and $N_b$, and we have three important results as follows.

1) With any value of $g$, $P_{\text{CH}}$ increases sharply when $N_b > 35$. The reason can be found in (18) and Fig. 10. When $N_b$ grows larger than 35, $N_f$ is monotonically increasing and $N_e$ is monotonically decreasing. All these factors lead to the growth of $P_{\text{CH}}$ according to (18).
2) The variance of $P_{\text{CH}}$ is relatively small when $N_b < 35$. It is because when $\Delta d$ grows from 0 to $R_t/2$, $N_e$ increases that reduces $P_{\text{CH}}$. On the other hand, the decrease in $N_f$ and the increase in $N_b$ leads to the growth of $P_{\text{CH}}$. These two opposite efforts make the variance of $P_{\text{CH}}$ small.
3) When $N_b < 35$, $P_{\text{CH}}$ increases with the reduction of $g$. The reason can be found in the $p_b(i, j)$ part in (18). When $\Delta d > R_t/2$, $P_{\text{CH}}(i)$ increases and $N_f$ decreases with the growth of $N_b$. All these factors lead to the growth of $p_b(i, j)$ and so the $P_{\text{CH}}(i)$. On the other hand, when $\Delta d < R_t/2$, this effect can be neglect with small $N_b$ or large $N_f$.

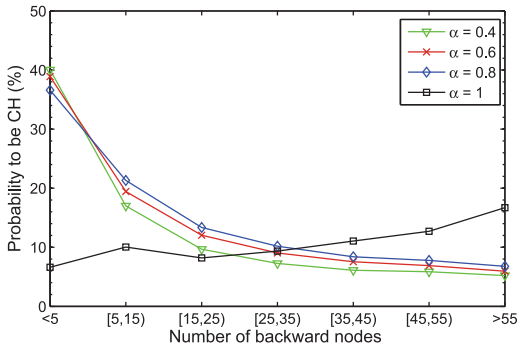


Fig. 12. Impact of $\alpha$ on the probability to be CH in $g = 1$ (with $N_b$ in $T_b$).

In multihop data collection, the nodes in $\mathcal{R}(1)$ have higher energy consumption due to the relay burden, i.e., the energy hole problem. This problem is more severe with imbalanced CHs selection described above. Fortunately, this problem can be solved by considering the number of neighbor nodes in the backoff timer $T_b$. In the next section, we will study how does the backoff timer $T_b$ impact the energy consumption in JCR.

### C. Impact of Backoff Timer

According to the analysis given in Sections VI-A and VI-B, the location and the backoff timer determines the node's probability to be CH. Since the node location is uncontrollable in this paper, the setting of the backoff timer is critical to balance the energy consumption. Let us revisit the backoff timer setting of JCR (11) given in Section IV-B

$$T_b = \begin{cases} [(G-g) + \alpha(\frac{E-e}{E}) + (1-\alpha)(\frac{N_F - N_f}{N_F})] \times \tau & g > 1 \\ [(G-1) + \alpha(\frac{E-e}{E}) + (1-\alpha)\frac{N_b}{N_B}] \times \tau & g = 1. \end{cases}$$

The backoff timer of JCR considers both the residual energy $E_r$ and the number of neighbor nodes ($N_b$ or $N_f$). According to (11), in the same *Ring*, the nodes with higher residual energy and more forward nodes (or less backward node) have higher possibility to be CH. The parameter $\alpha$ is used to balance the impact of $E_r$ and $N_b$ (or $N_f$).

In order to study the efficiency of considering neighbor nodes in $T_b$, simulations are executed to provide a closer view at $\mathcal{R}(1)$. There are 400 nodes randomly deployed in a network area of $200 \times 200$ m$^2$, and the sink locates at $[0, 0]$. The maximum transmission range $R_{\max}$ is set as 70 m, and so $R_t$ and $R_c$. Fig. 12 shows the probability to be CH at $\mathcal{R}(1)$ with different $N_b$ and $\alpha$. When $\alpha = 1$, which means $T_b$ does not consider the impact of $N_b$, the probability grows from 6.6% to 16.7% with the growth of $N_b$. It is similar to the results given in Fig. 11.

On the contrary, when $\alpha = 0.8$, the probability decreases from 36.6% to 6.8%. The results prove that the backoff timer given in (11) effectively adjusts the node's probability to be CH with different neighbor nodes. When the $\alpha$ is smaller, the impact of neighbor nodes is higher, and so the $P_{\text{CH}}$ of the nodes with smaller $N_b$.

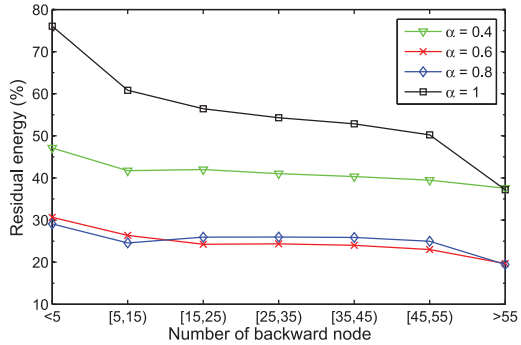


Fig. 13. Impact of $\alpha$ on the residual energy in $g = 1$. (with $N_b$ in $T_b$).

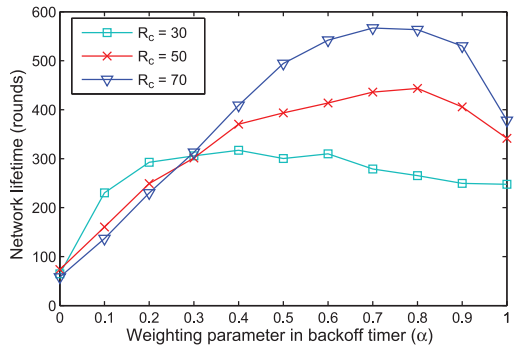


Fig. 14. Impact of $\alpha$ on network lifetime with clustering range $R_c = 30, 50, 70$ m.

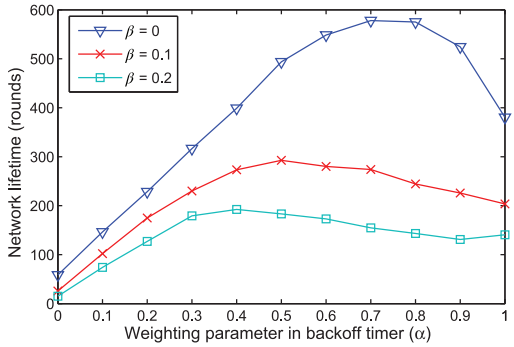


Fig. 15. Impact of $\alpha$ on network lifetime with aggregation rate $\beta = 0, 0.1, 0.2$.

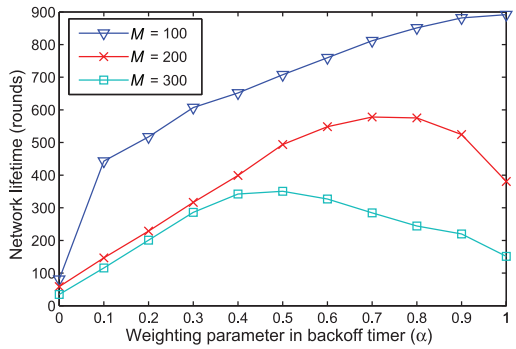


Fig. 16. Impact of $\alpha$ on network lifetime when side length of network area $M = 100, 200, 300$.

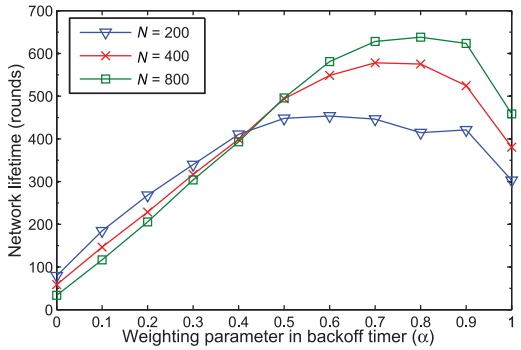


Fig. 17. Impact of $\alpha$ on network lifetime when the number of nodes $N = 200, 400, 800$.

Then, we study the network lifetime to evaluate the energy efficiency of JCR. The network lifetime is defined as the number of the round at which the first dead node appears in the network [9]. The initial energy of every node is $2J$, and the node is dead when it spends all its energy. Fig. 13 shows the residual energy of the nodes when the first dead node appears. In $\alpha = 1$, over 70% energy are left in the nodes with $N_b < 5$. The residual energy decreases with the growth of $N_b$. For the nodes with $N_b > 55$, the residual energy is about 37%. The variance of residual energy is large due to the imbalanced CHs selection and traffic load.

On the other hand, the variance of residual energy is much more smaller when $\alpha < 1$. Specifically, when $\alpha = 0.8$, the residual energy drops sharply to lower than 30%. The imbalance of energy consumption is effectively alleviated by considering the number of neighbors in the backoff timer. Moreover, the result also demonstrates that the value of $\alpha$ impacts the network performance.

In the rest of this section, we use simulations to study how does $\alpha$ impact the network lifetime with different values of clustering range $R_c$ (Fig. 14), aggregation rate (Fig. 15), network area (Fig. 16), and network density (Fig. 17).

Fig. 14 shows the relations between network lifetime and $\alpha$ when the clustering range $R_c$ is set as 30, 50, and 70 m. It is worth to note that when $\alpha = 1$, the $T_b$ does not consider the impact of $N_f$. In $R_c = 70$ m, the network lifetime reaches maximum as 566 rounds when $\alpha = 0.8$, and the network lifetime as 378 rounds when $\alpha = 1$. The network lifetime is prolonged over 50% by considering neighbor nodes in $T_b$. Moreover, the network lifetime reduces when $R_c$ decreases. The result is the same as that given in [10] and [24]. It is because the growth of $R_c$ leads to the decreasing number of CHs and so the number of intercluster packets.

Then, we study the impact of aggregation rate $\beta$ on the network lifetime. In this scenario, $R_c$ is set as 70 m and the $\beta$ is set as 0, 0.1, and 0.2. As shown in Fig. 15, the network lifetime decreases with the growth of $\beta$ because of the growth of packet size. To maximize the network lifetime, the $\alpha$ should be set as 0.8, 0.5, and 0.4, respectively, when $\beta = 0, 0.1, 0.2$. The optimal value of $\alpha$ is smaller when $\beta$ is larger.

Fig. 16 shows the results when the network area grows from $100 \times 100\,\text{m}^2$, $200 \times 200\,\text{m}^2$, to $300 \times 300\,\text{m}^2$. The node density is fixed at $0.01\,\text{node/m}^2$. The $R_c$ is set as 70 m and $\beta = 0$. As shown in Fig. 16, the network lifetime is shorter in a larger area due to the growth of multihop relay burden. In

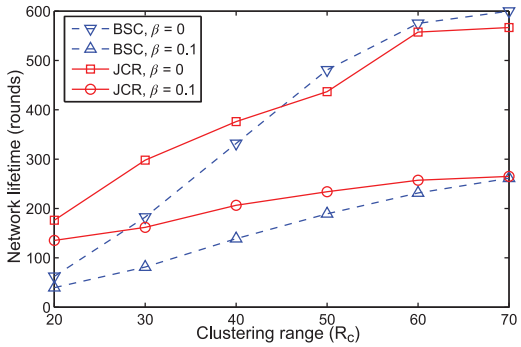


Fig. 18. Network lifetime comparison with different clustering range and aggregation rate.

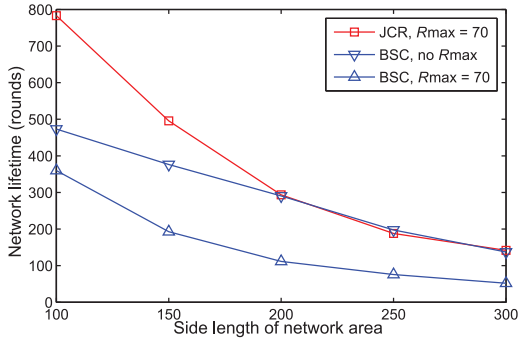


Fig. 19. Network lifetime comparison with different network area.

$100 \times 100\,\mathrm{m}^2$ network, the maximum lifetime is 893 rounds when $\alpha = 1$, whereas the maximum lifetime is 315 rounds in $300 \times 300\,\mathrm{m}^2$ network when $\alpha = 0.5$. The reason is that the multihop relay burden is higher in a larger area. To balance the energy consumption, the $\alpha$ should be smaller such that the node with higher $N_b$ has lower probability to be CH.

At last, we study the network lifetime with different node densities. In a fixed network area of $200 \times 200\,\mathrm{m}^2$, the number of nodes $N$ is set as 200, 400, and 800, respectively. As shown in Fig. 17, the network with higher node density has longer network lifetime. The growth of the node density increases the number of CMs and so as the burden for CH. However, according to (18), the growth of node density leads to the growth of neighbor nodes, and hence reduces the $P_{\mathrm{CH}}$ for each node. Moreover, a higher node density is helpful to load balance. When $N = 800$, the maximum network lifetime is achieved at $\alpha = 0.8$. The optimal $\alpha$ reduces to 0.7 and 0.5 when $N = 400$ and $N = 200$, respectively. It is because the variance of neighbor nodes is larger in a sparse network which leads to more severe imbalance of CHs selection. Reducing $\alpha$ increases the impact of neighbor nodes in CHs selection, and it is helpful to balance the energy consumption.

To summarize, considering residual energy and neighbor nodes in the backoff timer $T_b$ is helpful to improve the performance of JCR. Moreover, the optimal value of $\alpha$ is smaller when the intercluster traffic load grows up.

### *D. Network Lifetime Comparison*

To prove the energy efficiency of JCR, we compare the network lifetime of JCR with that of BSC. The first simulation is run in the network with 400 nodes randomly deployed in an area of $200 \times 200\ \mathrm{m}^2$. The intercluster transmission range $R_t$ in JCR is set as 70 m, whereas the transmission range in BSC has no limitation. Fig. 18 shows the network lifetime when $R_c$ and $\beta$ varies and the value of $\alpha$ is set to be optimal, respectively.

JCR performs better than BSC when $R_c < 40\mathrm{m}$ due to the efficient network topology. The advantage become small when $R_c$ grows larger than 40 m, because JCR generates more CHs to ensure network connectivity that leads to the growth of intercluster traffic load. On the other hand, when $\beta = 0.1$, which means the intercluster traffic load depends on not only the number of CHs but also the number of CMs, the advantage of JCR is larger than that in $\beta = 0$.

Then, we compare the network lifetime when the side length of the network grows from 100 to 300 m. The node density is fixed at $0.01\,\mathrm{node/m}^2$. The $R_c$ is set as 70 m and $\beta = 0.1$. As shown in Fig. 19, JCR performs better in a small area due to the efficient network topology. The advantage is small when $M \geq 200\,\mathrm{m}$, since JCR has to generate more CHs than BSC to ensure network connectivity that leads to the growth of intercluster traffic load.

It is worth to note that the results of BSC given above are based on the assumption that there is no limitation on the maximum transmission range. For completeness, we also study the network lifetime of BSC with the limitation of $R_{\max} = 70\,\mathrm{m}$. According to Fig. 5, the $R_c$ should be set as 35 m to ensure that the maximum LEH is smaller than 70 m. As shown in Fig. 19, the performance of BSC is greatly reduced due to the limitation of $R_{\max}$. The network lifetime of JCR is over 117.6% longer than that of BSC, and the improvement increases from 117.6% to 172.7% with the growth of the network area.

## VII. Conclusion

In this paper, we propose a JCR protocol to provide reliable and efficient data collection in large-scale WSN. The random backoff and gradient routing schemes are adopted in JCR to execute the CH selection and multihop routing simultaneously with low overhead. Theoretical analysis and simulation results prove that JCR can provide connected and efficient intercluster topology with limited transmission range. Moreover, the backoff timer in JCR can be carefully tuned to balance the energy consumption and prolong the network lifetime.

## References

[1] J. Stankovic, "Research directions for the Internet of Things," *IEEE Internet Things J.*, vol. 1, no. 1, pp. 3–9, Feb. 2014.

[2] Y. Liu, X. Mao, Y. He, K. Liu, W. Gong, and J. Wang, "CitySee: Not only a wireless sensor network," *IEEE Netw.*, vol. 27, no. 5, pp. 42–47, Sep. 2013.

[3] T. H. Luan, L. X. Cai, J. Chen, X. Shen, and F. Bai, "Engineering a distributed infrastructure for large-scale cost-effective content dissemination over urban vehicular networks," *IEEE Trans. Veh. Technol.*, vol. 63, no. 3, pp. 1419–1435, Mar. 2014.

[4] R. Du, C. Chen, B. Yang, N. Lu, X. Guan, and X. Shen, "Effective urban traffic monitoring by vehicular sensor networks," *IEEE Trans. Veh. Technol.*, vol. 64, no. 1, pp. 273–286, Jan. 2015.

[5] X. Tian, Y. Zhu, K. Chi, J. Liu, and D. Zhang, "Reliable and energy-efficient data forwarding in industrial wireless sensor networks," *IEEE Syst. J.*, 2015, to be published.
[6] C. Chen, J. Yan, N. Lu, Y. Wang, X. Yang, and X. Guan, "Ubiquitous monitoring for industrial cyber-physical systems over relay assisted wireless sensor networks," *IEEE Trans. Emerg. Topics Comput.*, vol. 3, no. 3, pp. 352–362, Sep. 2015.
[7] Y. Yao, Q. Cao, and A. Vasilakos, "EDAL: An energy-efficient, delay-aware, and lifetime-balancing data collection protocol for heterogeneous wireless sensor networks," *IEEE/ACM Trans. Netw.*, vol. 23, no. 3, pp. 810–823, Jun. 2015.
[8] X.-Y. Liu *et al.*, "CDC: Compressive data collection for wireless sensor networks," *IEEE Trans. Parallel Distrib. Syst.*, vol. 26, no. 8, pp. 2188–2197, Aug. 2015.
[9] W. B. Heinzelman, A. P. Chandrakasan, and H. Balakrishnan, "An application-specific protocol architecture for wireless microsensor networks," *IEEE Trans. Wireless Commun.*, vol. 1, no. 4, pp. 660–670, Oct. 2002.
[10] O. Younis and S. Fahmy, "Heed: A hybrid, energy-efficient, distributed clustering approach for ad hoc sensor networks," *IEEE Trans. Mobile Comput.*, vol. 3, no. 4, pp. 366–379, Oct./Dec. 2004.
[11] V. Pal, G. Singh, and R. P. Yadav, "Balanced cluster size solution to extend lifetime of wireless sensor networks," *IEEE Internet Things J.*, vol. 2, no. 5, pp. 399–401, Oct. 2015.
[12] S. Fang, S. Berber, and A. Swain, "An overhead free clustering algorithm for wireless sensor networks," in *Proc. Global Telecommun. Conf. (GLOBECOM'07)*, Nov. 2007, pp. 1144–1148.
[13] D. Wei, Y. Jin, S. Vural, K. Moessner, and R. Tafazolli, "An energy-efficient clustering solution for wireless sensor networks," *IEEE Trans. Wireless Commun.*, vol. 10, no. 11, pp. 3973–3983, Nov. 2011.
[14] N. Amini, A. Vahdatpour, W. Xu, M. Gerla, and M. Sarrafzadeh, "Cluster size optimization in sensor networks with decentralized cluster-based protocols," *Comput. Commun.*, vol. 35, no. 2, pp. 207–220, 2012.
[15] C. Chen, S. Zhu, X. Guan, and X. S. Shen, *Wireless Sensor Networks: Distributed Consensus Estimation*. New York, NY, USA: Springer, 2014.
[16] R. Zhang, J. Pan, J. Liu, and D. Xie, "A hybrid approach using mobile element and hierarchical clustering for data collection in WSNs," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC'15)*, 2015, pp. 1566–1571.
[17] D. C. Hoang, P. Yadav, R. Kumar, and S. K. Panda, "Real-time implementation of a harmony search algorithm-based clustering protocol for energy-efficient wireless sensor networks," *IEEE Trans. Ind. Informat.*, vol. 10, no. 1, pp. 774–783, Feb. 2014.
[18] H. Zhou *et al.*, "Chaincluster: Engineering a cooperative content distribution framework for highway vehicular communications," *IEEE Trans. Intell. Transp. Syst.*, vol. 15, no. 6, pp. 2644–2657, Dec. 2014.
[19] O. Demigha, W.-K. Hidouci, and T. Ahmed, "On energy efficiency in collaborative target tracking in wireless sensor network: A review," *IEEE Commun. Surv. Tuts.*, vol. 15, no. 3, pp. 1210–1222, Third Quart., 2013.
[20] F. Ye, G. Zhong, S. Lu, and L. Zhang, "Gradient broadcast: A robust data delivery protocol for large scale sensor networks," *Wireless Netw.*, vol. 11, no. 3, pp. 285–298, 2005.
[21] P. Huang, H. Chen, G. Xing, and Y. Tan, "SGF: A state-free gradient-based forwarding protocol for wireless sensor networks," *ACM Trans. Sensor Netw.*, vol. 5, no. 2, pp. 14:1–14:25, 2009.
[22] A. Wang, D. Yang, and D. Sun, "A clustering algorithm based on energy information and cluster heads expectation for wireless sensor networks," *Comput. Electr. Eng.*, vol. 38, no. 3, pp. 662–671, 2012.
[23] M. Ye, C. Li, G. Chen, and J. Wu, "EECS: An energy efficient clustering scheme in wireless sensor networks," in *Proc. IEEE Int. Perform. Comput. Commun. Conf. (IPCCC'05)*, 2005, pp. 535–540.
[24] Z. Xu, C. Long, C. Chen, and X. Guan, "Hybrid clustering and routing strategy with low overhead for wireless sensor networks," in *Proc. IEEE Int. Conf. Commun. (ICC'10)*, 2010, pp. 1–5.
[25] M. Penrose, "The longest edge of the random minimal spanning tree," *Ann. Appl. Probab.*, vol. 7, no. 2, pp. 340–361, 1997.
[26] S. He, J. Chen, and Y. Sun, "Coverage and connectivity in duty-cycled wireless sensor networks for event monitoring," *IEEE Trans. Parallel Distrib. Syst.*, vol. 23, no. 3, pp. 475–482, Mar. 2012.
[27] V. Mhatre and C. Rosenberg, "Design guidelines for wireless sensor networks: Communication, clustering and aggregation," *Ad Hoc Netw.*, vol. 2, no. 1, pp. 45–63, Jan. 2004.
[28] S. Olariu and I. Stojmenovic, "Design guidelines for maximizing lifetime and avoiding energy holes in sensor networks with uniform distribution and uniform reporting," in *Proc. IEEE Int. Conf. Comput. Commun. (INFOCOM'06)*, Apr. 2006, pp. 1–12.

**Zhezhuang Xu** (M'14) received the Ph.D. degree in control and systems from Shanghai Jiao Tong University, Shanghai, China, in 2012.

He is currently an Assistant Professor with the School of Electrical Engineering and Automation, Fuzhou University, Fuzhou, China. He has authored and/or coauthored over 20 referred international journal and conference papers. His research interests include wireless sensor/actuator network and its applications in industrial Internet of Things.

Dr. Xu serves as a Reviewer for several journals including the IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, the IEEE TRANSACTIONS ON INDUSTRIAL ELECTRONICS, and the *IEEE Communications Letters*.

**Liquan Chen** is currently an undergraduate student with the School of Electrical Engineering and Automation, Fuzhou University, Fuzhou, China.

He has authored one paper in the 34th Chinese Control Conference (CCC2015) and one paper in the 27th Chinese Control and Decision Conference (CCDC2015). His research interests include wireless sensor/actuator networks and Internet of Things.

**Cailian Chen** (S'03–M'07) received the B.Eng. and M.Eng. degrees in automatic control from Yanshan University, Qinhuangdao, China, in 2000 and 2002, respectively, and the Ph.D. degree in control and systems from the City University of Hong Kong, Kowloon Tong, Hong Kong, in 2006.

She joined the Department of Automation, Shanghai Jiao Tong University, Shanghai, China, in 2008, as an Associate Professor and is currently a Full Professor. She has authored and/or coauthored 2 research monographs and over 80 referred international journal and conference papers. She holds 20 patents. Her research interests include distributed estimation and control of network systems, wireless sensor and actuator network, multiagent systems, and intelligent control systems.

Dr. Chen serves as Associate Editor of four journals including the IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY and *Peer-to-Peer Networking and Applications* (Springer). She serves as a TPC member of several flagship conferences including IEEE Global Communications Conference (Globecom), IEEE International Conference on Communications (ICC), and IEEE World Congress on Computational Intelligence (WCCI). She was the recipient of the IEEE Transactions on Fuzzy Systems Outstanding Paper Award in 2008. She was one of the First Prize recipients of the Natural Science Award from the Ministry of Education of China in 2007. She was honored as New Century Excellent Talents in University of China and Shanghai Rising Star in 2013, Shanghai Pujiang Scholar, Chenguang Scholar, and SMC Outstanding Young Staff of Shanghai Jiao Tong University in 2009.

**Xinping Guan** (M'03–SM'04) received the Ph.D. degree in control and systems from the Harbin Institute of Technology, Harbin, China, in 1999.

He was a Professor and the Dean of Electrical Engineering with Yanshan University, Qinhuangdao, China. In 2007, he joined the Department of Automation, Shanghai Jiao Tong University, Shanghai, China, where he is currently a Chair Professor, the Deputy Director of the University Research Management Office, and the Director of the Key Laboratory of Systems Control and Information Processing, Ministry of Education of China, Beijing, China. He has authored and/or coauthored 4 research monographs, more than 200 papers in IEEE TRANSACTIONS and other peer-reviewed journals, and numerous conference papers. His research interests include industrial cyber–physical systems, wireless networking and applications in smart city and smart factory, and underwater sensor networks. He is the leader of the prestigious Innovative Research Team of the National Natural Science Foundation of China (NSFC).

Dr. Guan is an Executive Committee Member of the Chinese Automation Association Council and the Chinese Artificial Intelligence Association Council. He was the recipient of the First Prize of the Natural Science Award from the Ministry of Education of China in 2006 and the Second Prize of the National Natural Science Award of China in 2008. He was also the recipient of the IEEE Transactions on Fuzzy Systems Outstanding Paper Award in 2008. He is a National Outstanding Youth honored by the NSF of China, Changjiang Scholar by the Ministry of Education of China, and State-level Scholar of the New Century Bai Qianwan Talent Program of China.