International Journal of Future Generation Communication and Networking
Vol. 9, No. 11 (2016), pp. 275-286
http://dx.doi.org/10.14257/ijfgcn.2016.9.11.25

An Efficient Data Collection Protocol for Maximum Sensor Network Data Persistence

Jian Wan[1, 2, 3, a], Li Yang[1, b], Wei Zhang*[1, 3, c], Huayou Si [1, 3, d] and Jin Feng [4, e]

[1]*Department of Computer Science and Technology, Hangzhou Dianzi University, Hangzhou 310037, China*
[2]*Zhejiang University of Science and Technology, Hangzhou 310023, China*
[3]*Key Laboratory of Complex Systems Modeling and Simulation, Ministry of Education, Hangzhou 310023, China*
[4]*PLA the Rocket Force Command College，Wuhan 430000，China*
[a]*wanjian@hdu.edu.cn,* [b]*cyfsdm_yangli@163.com,* [c]*magherozhw@hdu.edu.cn,* [d]*sihy@hdu.edu.cn* [e]*fengjin8236@126.com*

Abstract

Sensor network has lot applications in the early warning and assistant of disaster environment such as debris flows, floods and forest fires. However, such disaster environment pose an interesting challenge for data collection since sensor nodes may be destroyed unpredictably and centrally, resulting in the decrease of data persistence in the network. Growth Codes Protocol (GCP) first focuses on increase sensor network data persistent in the disaster. However, the completely random data transmission way in GCP may cause a large number of invalid data transmissions and therefore, the efficiency of data collection of the protocol is not ideal in the late stage of data collection. In this paper, we propose an efficient data collection protocol (DGCP) to maximize sensor network data persistence by changing the completely random data transmission way. Packet classification mechanism and a novel dynamic probability model of data transmission in DGCP are proposed to control the effective direction of data flow. Furthermore, we found that the parameter optimization problem of the probabilistic model is a problem of searching the optimal solution in a mathematical view. Based on this property, we propose a genetic algorithm to optimize the dynamic probability model. The performance of the proposed DGCP is shown by a comparative experimental study. When compared with GCP, our DGCP has better performance in a variety of environments

Keywords*: Packet Classification Mechanism; dynamic probability model of data transmission; Genetic Algorithm; data collection; Growth Codes*

1. Introduction

Wireless sensor networks (WSN) have the inherent characteristics of self-organization, without human duty and rapid deployment. One of the motivating uses of WSN is the monitoring of disaster scenarios, such as debris flows, forest fires and volcanic eruptions. However, the disaster scenarios offer new challenges for data collection since sensor nodes may be destroyed unpredictably and centrally. These failure nodes are especially troubling because they directly lead to the loss of data collected by the sensor nodes. In addition, node failures will result in frequent changes of routing configuration.

WSN often expected to be deployed in zero-configuration networking, where nodes have no opportunity to know specifics of the topology, aside from some information describing their neighbor nodes. Hence, how to recover lost data already in the zero-

*Corresponding author: Wei Zhang; E-mail: magherozhw@hdu.edu.cn
ISSN: 2233-7857 IJFGCN

configuration networking needs to be paid attention and to be considered in the design process of the data collection protocol.

In such disaster scenes, it is obvious that the number of failure nodes is increasing with the passage of time. Hence, we need to complete the collection of data as soon as possible to avoid more nodes being destroyed. However, the networks have a feature described as a funnel effect [1], where there is a large number of congestions and delays in the neighborhood of the sink. And the amount of data generated in a typical wireless sensor networks greatly exceeds the amount of data collected by the sink node, which causes the significant transmission delay. Due to the significant transmission delay, the risk of data loss is greatly increased. Therefore, how to increase the efficiency of data transmission also need to be considered in the design process of the data collection protocol.

Data persistence is one of the most important performance metrics for data collection since it reflects the proportion of the generating data and the collected data in disaster. To build an efficient data collection, we need to improve data persistence of zero configuration networks by improving the efficiency of data collection and the recovery of lost data.

Growth Codes is the first coding technology focused on improving data persistence in the zero-configuration networking. The code uses a simple encoding and random data transmission way to improve the likelihood that data will survive when some storage nodes fail. To recover the maximum amount of data at any time point, GCP is designed to increase the degree of codeword monotonically with time, and found the optimal time points when the degree of the codeword is increased.

However, due to the use of a completely random data transmission way in GCP, the sink frequently collects redundant data and seriously affected the performance of the data collection. As the figure 1, the efficiency of data collection protocol has dropped sharply as the network becomes sparse. Simultaneously, GCP may also do no good to the data collection in the later period.

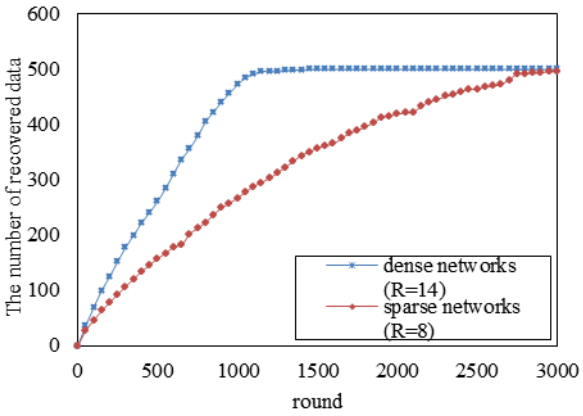

Figure 1. The Performance of GC in Different Density Network

When using completely random data transmission in GCP, the different data flows to be random in all directions. Only when the direction of the data flow and the direction of the sink node are consistent, the data can be collected by the sink node. While nodes transmit data to sink nodes with different hops, which leads that different data have the different probability to be collected by the sink node.

In this paper, we change the flow of data to improve data persistence, and propose an efficient data collection protocol for maximum sensor network data persistence (DGCP). Packet classification mechanism in DGCP allows each data packet to know information about the distance between the data packet and the sink node. A novel dynamic

probability model of data transmission in DGCP is proposed to control the direction of data flow.

The trend is as follows: initially the data packets take approximately a random way to transfer data in order to make the data evenly distributed. But over time, the bigger series' data packet has a greater probability to flow to the direction of the sink node. And the smaller series' data packet has a smaller probability to flow to the direction of the sink node. A well-designed probability model can not only make the data evenly distributed, but also reduce the acceptance of redundant data.

To maximize the data persistence, the parameters of the probability model need to be optimized. From a mathematical point of view, we found that the parameter optimization problem of the probabilistic model is a problem of searching the optimal solution in an interval.

Based on the property, we introduce genetic algorithm to optimize the data transmission probability model. The performance of the proposed DGCP is shown by simulation experiments. When compared with GCP, DGCP shows significant performance improvement.

Furthermore, we use canonical genetic algorithm to optimize the dynamic probability model. The performance of the proposed DGCP is shown by simulation experiments. When compared with GCP, DGCP shows significant performance improvement.

2. Related Works

A lot of research work in sensor networks targeted towards increasing the performance of data collection, which takes advantage of the routing technology, such as [3], [4], and[5], and takes advantage of the network coding [6] to improve the efficiency of data collection [7] and reliability [8].

To improve the efficiency of data collection, routing technology is very straightforward means. It finds the optimal path to reduce the transmission overhead. However, this technique is not applicable to the disaster scene since the technology requires nodes to establish a routing tree but the failure nodes will frequently interrupt routing establishment.

Therefore, some scholars introduce the network encoding [2] to improve the performance of data collection in the disaster. Data collections based on network encoding use the feature that network coding can increase the data transfer information entropy [9] and network data backup [10] to improve the data persistence. However, before the accumulation of a large number of data its decoding rate is relatively low in [11], which affects the performance of data collection.

Growth Codes is specifically designed to improve data persistence in a disaster environment. However, the efficiency of data collection of the protocol is not ideal in the late stage of data collection and sparse networks due to the use of a completely random data transmission way.

An extensive researches based on Growth codes are described in [12], where the problem of poor data collection efficiency in sparse networks is addressed. In recent years, some scholars have focused on the use of intelligent algorithms to optimize network coding technology [13] and improve data collection persistence.

3. Network Model

Figure 2 depicts network model of DGCP. DGCP is based on the series information of data packets to control the flow of data, therefore the biggest difference between network model of DGCP and the network model of GCP model is that each data package adds a flags of series in network model of DGCP. In order to facilitate the description, we use the round [2] to represent time and use degree [2] of the codeword to represent the number of symbols which are encoded together.

N sensor nodes are randomly deployed to sense data in the monitoring region. Node can communicate with other nodes in the circular area centered at the node with the radius of R. A codeword in GCP is called a data packet in DGCP. Each node contains a number of data packets, and each data packet contains a series.

The first data packet's series from left to right can represent the distance between the node and the sink node due to the coding scheme, so the series is called the node's series. Additional series in the node can represent the distance between the data packet and the sink node. The sink node is placed in a safe region to collect and restore the sensing data. Select a proxy node, and all sensing data in network are transmitted through the proxy node to the sink node.

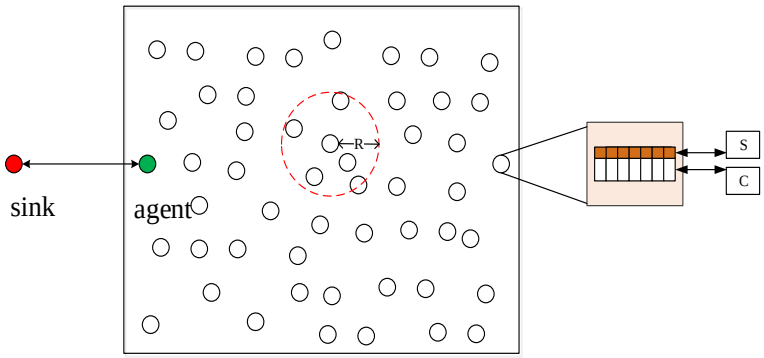

Figure 2. Network Model of DGCP

4. About GA Parameters

The evaluation function f_i of the genetic algorithms corresponds to measure the degree of improving data persistence. That is, the evaluation value is affected by the efficiency and reliability of the data collection.

In a large scale wireless sensor network, we use the number of recovered data and the data recovery time to affect the evaluation function. The evaluation function is inversely proportional to data recovery time, and it is proportional to the total amount of data recovered. The evaluation function in the i -th time is given as follows：

$$f_i = \begin{cases} data' & if\ \ t_c == 0 \\ (s - t_{0.5*c} - \partial(t_c - t_{0.5*c})) + data & other \end{cases}$$

where the index t_c represents the number of rounds needed to collect all data. When $t_c == 0$, it indicates that the data are not fully collected , at that time ,the evaluation function only is determined by the total number of recovered data $data'$. The constant s slightly larger than the total number of rounds. The index $t_{0.5*c}$ represents the number of rounds needed to collect half the data. The constant $data$ represents the total number of data that needs to be recovered. Regulatory factor ∂ is defined in order to distinguish the difficulty of data collection at different time periods.

5. DGCP

5.1. Data Packet Classification Mechanism

In this section, we introduce the design principle of data packet classification mechanism. The purpose of data packet classification mechanism is to allow each data packet to know information about the distance between the data packet and the sink node.

We define a series for each data packet to reflect the distance between the data packet and the sink node. Furthermore, we use the number of hops required to transmit a data packet to the sink node is used to describe the series of the data packet.

When nodes occur failure in a disaster environment, the series does not need to perform the update operation since it reflects the data packet itself. However, when increasing the degree of a data packet, the series needs to be updated as two old packets are encoded and a new data packet is generated. Identifying the new packet's series is simple, the series is equal to the bigger series in the two coded data packets.

Data packet classification mechanism is divided into establishing phase and updating phase. Establishing phase is required to complete before data collection begins. We need to update the series at the transition points at which the data packet switches to higher degree of the data packet.

1) Establishing phase

- The initial value of each data packet's series is set to 0.
- The sink node first broadcasts a data packet *count* containing a series flag bit to its neighbors and the a series flag bit *s* is set to 2.
- For a neighbor node that receives the *count* sent by sink node, if the series of the data packets stored in the node are 0, the series of all the data packets in the node is set to *s*, otherwise, series is unchanged. Then *s* is changed to 3 and the node is sent *count* to its neighbor node.
- For any node receiving the *count*, if the series of the data packets stored in the node are 0, the series of all the data packets in the node is set to *s*, otherwise, series unchanged. Then, the *s* is changed to *s*+1, and the node sends *count* to its neighbor nodes.
- Until all packets' series are not zero in the network, stop.

2) Updating phase

The new data packet's series is equal to the bigger series in the two coded data packets.

Initially each sensor node placed to sense its data has only multiple backups of its own data. After the establishing phase is completed, each data packet is added to the series that represents its distance between the data packet and the sink node, obviously, data packets' series (in the figure, series 3) in same node are the same at the moment. When the data packet switches to higher degree of data packet, the series is updated. At this point, data packets' series in same node are different. The new series reflects the new data packet contains the maximum value of encoding data packets' series.

5.2. A Novel Data Transmission Probability Model

Wireless sensor network data transmission involves a problem: how to select the data packets in a node. In order to further facilitate the use of the series of information, we need to sort the data packets according to the series.

This paper uses the function to measure the urgency of selecting the k-th packet in the node that has a greater distance between the node and the sink.

When the value of is larger, the node is more likely to select the k-th packet. The function is given by Eq. 1:

$$f_{ik} = 1 + \lambda * k^{\alpha} + \eta * t^{\delta}, k = 1, 2, ..., cache - 1 \quad (1)$$

where the variable k is influenced by the series and the variable t is influenced by the time of data collection. The parameters λ and α are series factors of adjusting the probability by series information. The parameters η and δ are time factors of adjusting the probability by time information. In order to guarantee the trend of data flow, we set up $\alpha \geq 1$、 $\delta \geq 1$、 $\lambda > 0$ and $\eta > 0$.

The probability model of selecting the k-th packet in the node that has a greater distance between the node and the sink in a certain round is given as Eq. 2:

$$p_{ik}^{greater} = f_{ik} / \sum_{j=1}^{j=cache-1} f_{ij}, \quad k = 1,2,...,cache\text{-}1 \tag{2}$$

This paper uses the function $f_{i(cache-s-1)}^{small}$ to measure the urgency of selecting the $(cache-s-1)$-th packet in the node that has a smaller distance from the sink. When the value of $f_{i(cache-s-1)}^{small}$ is larger, the node i is more likely to select the $(catch-s-1)$-th packet. The function is given by Eq. 3:

$$f_{i(catch-s-1)}^{small} = 1+\lambda * s^{\alpha} + \eta * t^{\delta}, \ s = 0,1,2,...,cache\text{-}2 \tag{3}$$

where the variable t is influence by the time of data collection. The parameters λ and α are series factors. The parameters η and δ are time factors.The probability model of selecting the k-th packet in the node that has a smaller distance from the sink in a certain round is given as Eq. 4:

$$p_{ik}^{smaller} = f_{ik}^{small} / \sum_{j=1}^{j=cache-1} f_{ij}^{small}, \quad k = 0,1,2,...,cache\text{-}1 \tag{4}$$

We can use the probability models $p_{ik}^{greater}$ and $p_{ik}^{smaller}$ to control the direction of data flow, and the probability models contain four parameters (λ、 α、 η and δ) are used to adjust the growth rate of the two variables. To maximize the persistence of data collection, the parameters need to be optimized. From a mathematical point of view, we find that the parameters' optimization problem is to a problem of searching the optimal solution in an interval. Based on the property, we use a canonical genetic algorithm to optimize the Data transmission probability model.

5.3. DGCP

DGCP based on the degree distribution of GCP proposes a novel probability model of data transmission to control the flow of data, which is elaborated as follows.

1) Chromosome encoding and initialization: Four parameters (λ、 α、 η and δ) are converted into a binary string and the first gene pool is $S_0 = (s_1, s_2, ...s_L)$.

2) Wireless sensor network initialization: Monitoring region are randomly deployed sensor nodes, and each node stores the sensed data. Meanwhile, sets the series of data packets to 0.

3) Establishing series phase

4) Node communication options for the i-th round:

- If the degree of data packet *x* is less than maxdegree, the data packet *x* and the first data packet *y* in the node are encoded. At the same time, series of all data packets need to be updated.
- If the round is greater than or equal to the codewords conversion time $k_{\max}$, the maxdegree need to add 1;
- A data exchange of communication nodes: The node that has a smaller distance between the node and the sink uses probability $p_{ik}^{smaller}$ to select the k-th packet, and the node that has a greater distance between the node and the sink uses probability $p_{ik}^{greater}$ to select the k-th packet, then exchange the two packets.

5) Calculation fitness for the k-th chromosome: we use the number of recovered data and the data recovery time to affect the fitness function. The fitness function is inversely proportional to data recovery time, and it is proportional to the total amount of data recovered. Parameter values are put into the fitness function. Then again jump to step 2 and calculate the fitness of the next chromosome.

6) Termination condition: Judge whether the algorithm is terminated.

7) New populations are created: New population is generated from current population by selection, mutation and crossover. Return step 2.

6. Experimental Results

In this section, we evaluate the performance of DGCP in different scenarios. Herein, a description of the sensor network is given. 500 sensor nodes are deployed in a square field with dimensions of 100*100, and the number of data packets stored in a sensor node is 10. The series flag bits of a data packet are small and can be ignored in the packet header.

Setting the time interval for the completion of a data acquisition is 3000 rounds. Node communication radius (R=8) simulates sparse networks. Node communication radius (R=14) simulates dense networks. The following experimental data are the average of several simulation experiment. On the X-axis, the number of rounds shows the time of data collection while the Y-axis is the number of data recovered at the sink. Experimental environments simulate the centralized disaster networks by setting all nodes to fail in the circular area centered at any node with the radius of 20 at time t=500. Set every 10 rounds will have a node fails, which simulates in discrete disaster networks.

In the stable networks, Figure. 3 and Figure. 4 how much data is recovered using GCP or DGCP at each time point. In the centralized disaster networks, Figure. 5 and Figure. 6 show the data collection process. Reduced time ratios using DGCP are collected in the same data are shown in table 1.

To a certain extent, it represents the improvement of the efficiency of data collection using DGCP. Table 2 shows total number of recovered data by GCP and DGCP. Through the comparison of GCP and DGCP, we can know the reliability of the data collection in various scenarios. The simulation results show that our proposed DGCP is an efficient method for improving data persistence in different scenarios.

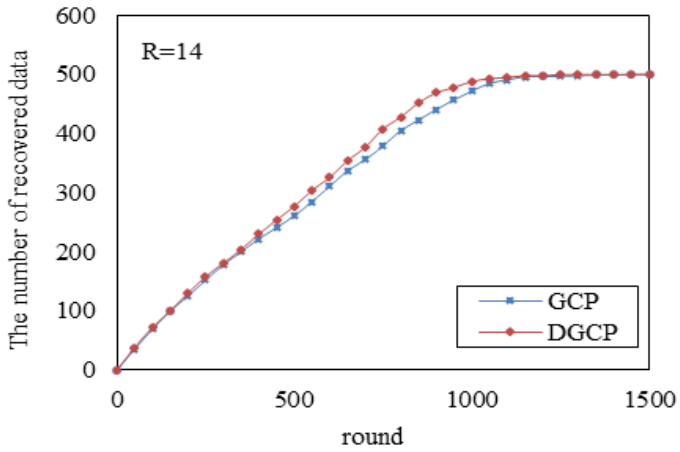

Figure 3. Data Collection Process using GCP and DGCP in Dense Networks (Stable Networks)

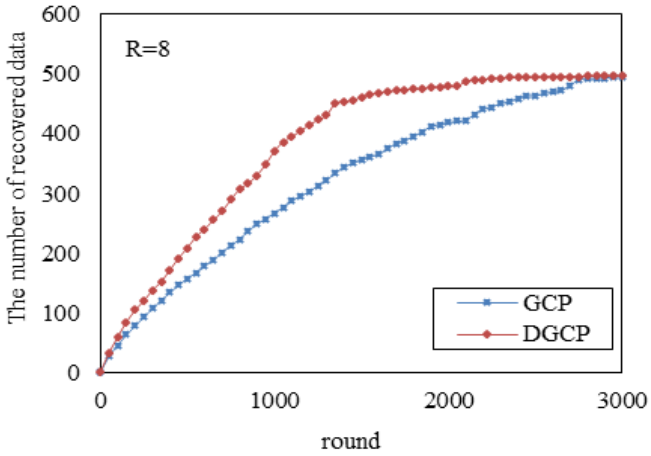

Figure 4. Data Collection Process using GCP and DGCP in Sparse Networks (Stable Networks)

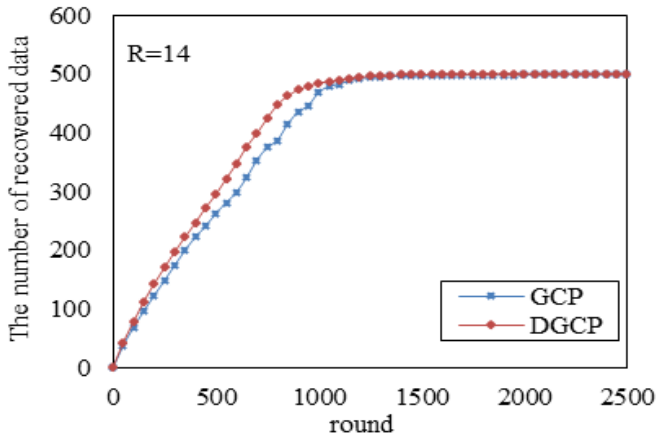

Figure 5. Data Collection Process using GCP and DGCP in Dense Networks (Centralized Disaster Networks)

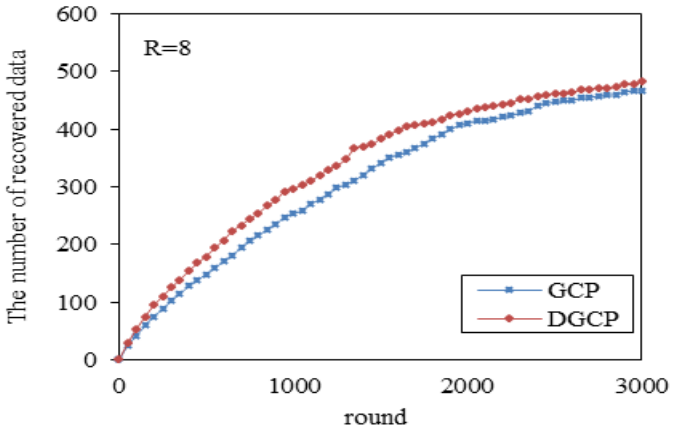

Figure 6. Data Collection Process using GCP and DGCP in Sparse Networks (Centralized Disaster Networks)

Table 1. Reduced Time Ratios using DGCP are Collected in the Same Data

Scenes	Dense Networks	Sparse Networks
Stable networks	13.79%	15.30%
Centralized Disaster Networks	29.70%	11.70%
Discrete Disaster Networks	11.67%	23.30%

Table 2. Total Number of Recovered Data by GCP

Scenes	(GCP)Dense Networks	(GCP)Sparse Networks	(DGCP)Dense Networks	(DGCP)Sparse Networks
Stable networks	500	496	500	498
Centralized Disaster Networks	500	467	500	483
Discrete Disaster Networks	499	429	500	463

7. Conclusions

In this paper, we propose a data collection protocol (DGCP) to improve data persistence in the disaster. To change the direction of data flow, we propose packet classification mechanism to distinguish data packets according to the distance between data packets and the sink node. Then a novel dynamic probability model of data transmission is proposed to control the direction of data flow. To maximize the data persistence of data collection, we introduce a canonical genetic algorithm to optimize the data transmission probability model. Finally, extensive simulations are conducted to validate that DGCP can improve data persistence by improving the efficiency of data collection and reliability.

Acknowledgments

This work is supported by the Zhejiang Provincial Natural Science Foundation of China (No.LY14F020044), the National Natural Science Foundation of China (No. J1524009, 61472112), the opening fund of Key Laboratory of Complex Systems Modeling and Simulation, Ministry of Education, and the National Key Technology Research and Development Program of the Ministry of Science and Technology of China (No. 2014BAK14B00).

References

[1] C. Y. Wan, S. B. Eisenman, A. T. Campbell and J. Crowcroft, "Siphon: overload traffic management using multi-radio virtual sinks in sensor networks", Proceedings of the 3rd ACM International Conference on Embedded Networked Sensor Systems, San Diego, USA, **(2005)**.
[2] A. Kamra, V. Misra, J. Feldman and D. Rubenstein, "Growth codes: maximizing sensor network data persistence", ACM SIGCOMM Computer Communication Review, vol. 36, no.4, **(2006)**, pp. 255-266.
[3] C. Intanagonwiwat, R. Govindan, D. Estrin, J. Heidemann and F. Silva, "Directed diffusion for wireless sensor networking", IEEE/ACM Transactions on Networking (ToN), vol. 11, no. 1, **(2003)**, pp. 2-16.
[4] R. Sugihara and R. K. Gupta, "Optimal Speed Control of Mobile Node for Data Collection in Sensor Networks", IEEE Transactions on Mobile Computing, vol. 9, no. 1, **(2010)**, pp. 127-139.
[5] Y. Yao, Q. Cao and A. V. Vasilakos, "EDAL: an energy-efficient, delay-aware and lifetime-balancing data collection protocol for heterogeneous wireless sensor networks", IEEE/ACM Transactions on Networking, vol. 23, no. 3, **(2013)**, pp. 182-190.

[6] C. Gkantsidis, P. R. Rodriguez, "Network coding for large scale content distribution", Proceedings IEEE 24th Annual Joint Conference of the IEEE Computer and Communications Societies, Miami, USA, **(2005)** March 13-17.
[7] S. Katti, H. Rahul, W. Hu, D. Katabi, M. Médard and J. Crowcroft, "XORs in the air: practical wireless network coding", IEEE/ACM Transactions on Networking (ToN), vol. 16, no. 3, **(2008)**, pp. 497-510.
[8] A. Dâmaso, N. Rosa and P. Maciel, "Reliability of wireless sensor networks", Sensors, vol. 14, no. 1 **(2014)**, pp. 15760-15785.
[9] W. C. Kuo and C. C. Wang, "Robust and optimal opportunistic scheduling for downlink 2-flow inter-session network coding with varying channel quality", IEEE INFOCOM 2014-IEEE Conference on Computer Communications, Toronto, Ontario, Canada, **(2014)**27 April-2 May.
[10] A. G. Dimakis, P. B. Godfrey, Y. Wu, M. J. Wainwright and K. Ramchandran, "Network coding for distributed storage systems", IEEE Transactions on Information Theory, vol. 56, no. 9, **(2007)**, pp. 2000-2008.
[11] M. G. Luby, M. Mitzenmacher, M. A. Shokrollahi and D. A. Spielman, "Efficient erasure correcting codes", IEEE Transactions on Information Theory, vol. 47, no. 2, **(2001)**, pp. 569-584.
[12] X. Xu, C. Shen and J. Wan, "Regulative Growth Codes: Enhancing Data Persistence in Sparse Sensor Networks", APSCC '10 Proceedings of the 2010 IEEE Asia-Pacific Services Computing Conference, **(2010)** December 6-10.
[13] L. U. Wen-Wei, Y. H. Zhu and G. H. Chen, "Energy-efficient routing algorithms based on linear network coding in wireless sensor networks", Acta Electronica Sinica, vol. 38, no. 10, **(2010)**, pp. 2309-2314.

Authors

Wan Jian, he received his Ph.D. Degree in Computer Science from Zhejiang University, China, in 1996. He is now a professor of Zhejiang University of Science and Technology, Hangzhou, China. His research interest includes virtualization, grid computing, services computing and wireless sensor networks.

Wei Zhang, he received the BE degree in School of Information Science and Engineering of Wuhan University of Science and Technology in China in 2000 and he received the MEc and PhD degree in Computer School of Wuhan University in China in 2004 and 2008, respectively. He is currently an associate professor with School of Computer Science and Technology, Hangzhou Dianzi University, China. His research interests include wireless sensor network and Intelligent Computing. He is a member of Association for Computing Machinery (ACM) and China Computer Federation (CCF).

Li Yang, she received bachelor's degree from Zhejiang Ocean University, Zhejiang, China and master's degree from Hangzhou Dianzi University, Zhejiang, in 2013 and 2016, respectively

Huayou Si, he is a lecturer in School of Computer Science and Technology, Hangzhou Dianzi University. He received M.S. and Ph.D. in Computer Science from Peking University in 2004 and 2012 respectively. During the past several years, His research interests include P2P network, service-oriented computing and Semantic Web. In the related research field, he has published more than 20 academic

papers. In addition, he has served in the Technical Program Committee of several international conferences.

Jin Feng, she is a lecturer of PLA the Rocket Force Command College.She gains Law Juris Master degree of Central China Normal University in 2013.Her research areas are Cyber Crimes and Cyber Security Law.