This full text paper was peer reviewed at the direction of IEEE Communications Society subject matter experts for publication in the IEEE Globecom 2010 proceedings.

Efficient Data Collection with Sampling in WSNs: Making Use of Matrix Completion Techniques

[1]Jie Cheng, [1]Hongbo Jiang, [1]Xiaoqiang Ma, [1]Lanchao Liu, [2]Lijun Qian, [1]Chen Tian, and [1]Wenyu Liu
[1]Department of EIE, Wuhan National Laboratory for Optoelectronics
Huazhong University of Science and Technology, Wuhan 430074, China
[2]Department of Electrical & Computer Engineering
Prairie View A&M University, Prairie View, Texas 77446
Email: {jiecheng2009, hongbojiang2004, mxqhust, hustlanchao}@gmail.com,
liqian@pvamu.edu, {tianchen,liuwy}@mail.hust.edu.cn

***Abstract*—Data collection is of paramount importance in many applications of wireless sensor networks (WSNs). Especially, to accommodate ever increasing demands of signal source coding applications, the capacity of processing multi-user data query is crucial in WSNs where the efficiency is one key consideration. To that end, this paper presents EDCA: an Efficient Data Collection Approach for data query in WSNs, which exploits recent matrix completion techniques. Specifically, for the efficiency of energy consumption, we randomly select a part of nodes from the sensor network to sample at each time instance and directly forward the data to the sink. Then, to recover the data precisely, we shift the rank minimization problem, which is NP-hard, to a convex optimization one. Compared with the centralized scheme, energy consumption using EDCA is significantly reduced due to lower sampling rate and fewer packets to transmit. The experimental results demonstrate that EDCA significantly outperforms the existing naive method in terms of energy consumption and the introduced errors are quite trivial.**

I. Introduction

As the important component of aggregation networks, wireless sensor networks (WSNs) aim to achieve the ubiquitous sensing and connect the information world with the physical world. In the past decades, they have provided us with a novel method to acquire data. An example of WSNs system is shown in Fig. 1. WSNs were originally build in military fields and related applications are battlefield surveillance, monitoring friendly forces, equipment and ammunition, targeting, and so on. With rapid development and reduced cost, WSNs quickly extend to civil field [10]. For example, they can be used in forest fire detecting, habitat monitoring, tele-monitoring of human physiciological data and so on. In addition, the emerging technology Internet of Things is mainly based on WSNs and this technology may bring us many demanded products.

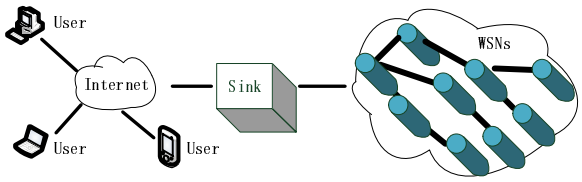

Fig. 1. A simple WSNs data collection system

In wireless sensor networks, one problem should be first addressed is how to sense the physical world efficiently, namely, to achieve the system efficiency. Despite that the energy consumption of sending and receiving packets is one of the most concerned factors in the design of WSNs, the energy consumption of sampling is often ignored by researchers. Previous work has shown that the sensing of physical quantity is usually very power-wasting [11]. Table I lists the power consumption of the transmitting, receiving and sensing components of Mica motes with Mica weather board. It is worth noting that the cost of sensing (I^2C Temperature) is non-negligible and even higher than that of transmission. For this reason, a substantial amount of energy consumption could be reduced if sensor nodes do not sample at every time instance.

TABLE I
Power consumption of Mica motes

Operation	Cost (nAh)
Transmitting a packet	20.00
Receiving a packet	8.00
Radio listening for 1 millisecond	1.25
Operating sensor for 1 sample (digital)	0.35
Thermistor	0.37
Operating sensor for 1 sample (analog)	1.08
Barometric Pressure	1.67
Photoresistor	3.60
Barometric Pressure Temp	1.67
Thermopile	9.67
I^2C Temperature	41.67

Due to the limited resources of sensor nodes, algorithms and protocols designed for WSNs should be of high efficiency. And we describe the efficiency of WSNs from the prospective of network efficiency, compressive efficiency, and universal efficiency.

As to network efficiency, since there is often no single center node which is responsible for routing, scheduling and traffic control of the whole network in WSNs, the sensor nodes are required to be self-organized. Researchers have developed many distributed and adaptive methods for the management of sensor networks. Some of these approaches aim to balance

978-1-4244-5637-6/10/$26.00 ©2010 IEEE

the traffic of sensor nodes in order to prolong the lifetime of the whole network. In terms of compressive efficiency, source coding algorithms often consume a large portion of computational capacity, memory and power resources and thus whether to perform source coding lies heavily on the remaining resources of the nodes. The data collected from large-scale sensor networks is of huge volume but contains much redundancy because it is highly spatial and temporal correlated [8]. As a result, it is desirable to adopt efficient source coding algorithms which can compress data efficiently and consume as few resources of sensor nodes as possible. At last, universal efficiency means that sensor networks should adapt to various work conditions and applications. However, the cost of sensor nodes is still considerably high, especially when the nodes are required to measure different physical quantities. And different physical quantities are not compressible or sparse in the same basis, for example acoustic signals are sparse in the Fourier basis, and image signals are sparse in the wavelet basis, sensors will adopt different compressive algorithms to process the signals respectively before transmitting, which will significantly increase the computational burden and complexity.

To address these above-mentioned impediments, we propose a novel approach to achieving the system efficiency for large-scale WSNs. This paper makes two main contributions. First, we propose a scalable power-saving sampling model to acquire data. Second, we propose to use the nuclear norm optimization for data collection in WSNs and the performance significantly outperforms the naive centralized scheme.

The rest of this paper is organized as follows: Section II discusses the related work of data collection in WSNs. Section III details the approach proposed in this paper. Section IV illustrates the simulation results with both synthetic and real world data. In the last section we come to a conclusion.

II. Related work

In this part, we briefly introduce prior work in traditional source coding, distributed source coding, and compressed sensing.

A. Traditional source coding

Traditional source coding has several critical deficiencies. First of all, the efficiency of data compression is highly correlated with routing protocol of the network. However, the jointly optimization of the two factors has been proved to be NP-hard [1]. What is more, the transmission in wireless networks is lossy and most traditional source coding algorithms don't consider this into consideration and thus lack robustness. Besides, traditional source coding lays too much burden on the encoders to implement the highly complex coding algorithms, so distributed source coding methods are proposed .

B. Distributed source coding

Distributed source coding (DSC) exploits both intra- and inter-signal correlation to compress data and shifts the computational burden in encoders to the joint decoder [2]. The Slepian-Wolf theorem [3] is the conceptual basis for most DSC algorithms, which has proved that several isolated sources can compress data as efficiently as though they can communicate with each other. However, DSC is not satisfactory for system efficiency and robustness when applied to sensor network. First, DSC algorithms incur high time and space complexity, unbalanced nodes transmission load, and needs to know the global correlation structure as a prerequisite to allocate appropriate number of bits to quantify the data transmitted for each sensor, which is a impossible mission in large-scale wireless sensor networks. Second, DSC is only suitable when the correlation among neighbor sensors remains stable. When the correlation changes or the data contain outliers, the decoding accuracy will significantly deteriorate.

C. Compressed sensing

A new concept of signal sensing and compression, compressed sensing (CS), or compressive sensing, has been developed [12]. CS can sample a signal far below the Nyquist rate if the signal has a sparse representation in one basis. In CS, the signal is sampled and compressed simultaneously and accurately reconstructed with high probability. By now, the basic frame of CS has been established, including the compression and recovery of signal, and various applications. However, traditional CS focuses on single-signal sampled by one source and only exploits intra- or inter-signal sparsity to compress and reconstruction signal. Compressive Data Gathering (CDG) proposed in [5] focuses on data collection in WSNs utilizing only the sparsity property on inter-signal. So the entire power consumption of WSNs is high and range of convenience is relatively narrow. Base on compressed sensing theorem, Duarte et al. propose distributed compressed sensing (DCS) [4]. They exploit both intra- and inter-signal sparsity to lower sampling rates, and put forward two joint sparse models as well as related recovery algorithms. However, they have not proved their method works well for large-scale sensor networks.

Recently, researchers discuss low-rank completion problem [6], [7] which is an extension of compressive sensing. Based on nuclear norm minimization, [9] presents a novel approach to estimate the missing values in traffic matrices. In [14], authors compare three recovery methods with noise observations data.

III. EDCA Scheme

In this section, we describe our approach to collect data in WSNs. Rather than traditional information theory, we consider source coding in view of network information theory. Our goal is not only to reduce the resource consumption of sensor nodes, but also to improve the efficiency and robustness of the whole network system.

A. System description

We consider the simple example of sensor networks shown in Fig. 1. Over the Internet, many users want to inquire the data of the whole network. After receiving the inquiry request of a user, the sink node will forward this request to the whole

978-1-4244-5637-6/10/$26.00 ©2010 IEEE

network, and the network will reply to the user through the sink node. In this work, we suppose that WSN has established a routing protocol, for example, the most frequently used tree-based protocol. Packet loss occurs when data packets travel across the sensor network to the sink node. Without loss of generality, we assume all the measurements acquired by sensor nodes are positive real numbers.

B. Algorithm

We consider a sensor network consisting of N nodes. Each node samples at a fixed rate and forwards the data to the sink through a multi-hop way. Then at the sink, we can get an $N \times T$ matrix, in which T means the number of samples of each node, i.e., $X \in \mathbb{R}^{N \times T}$. However, due to the lossy transmission and the failure of sensor nodes, some data are missing and the matrix is incomplete. Thus we can adopt a linear operator $\mathcal{A}(\cdot)$ to represent such effect:

$$\mathcal{A}(X) = B, \tag{1}$$

where X is the original matrix without loss of data and B denotes the data collected by the sink at a certain ratio of missing entries. The operator $\mathcal{A}(\cdot)$ represents the position in which we can acquire readings or not during data collection. For the sake of clarity and simplicity, we can specify $\mathcal{A}(\cdot)$ as a $N \times T$ matrix Q, such that:

$$Q(i,j) = \begin{cases} 0, & \text{if } X(i,j) \text{ is unavailable.} \\ 1, & \text{otherwise.} \end{cases} \tag{2}$$

The word *unavailable* means that we can choose the node i not to sample at time instance j. In addition, this *unavailable* may be also caused by node failing or the packets losing during transmission. Our method is fault-tolerate against these errors above caused by *unavailable*. In this instance, $Q. * X = B$, where .* represents an element-wise product and B maintain the available entries of X. Usually, the probability of numbers in Q is given by binomial distribution and we call the mean of this distribution as "sampling ratio".

Previous studies have shown that data collected from WSNs is highly spatial and temporary correlated, so the matrix X collected at the sink is approximately low rank [8]. E. Candès and B. Recht's recent work on matrix completion has proved that it is highly possible to recover a low-rank matrix from a subset of its entries [7]. Thus we can formulate our problem as follows:

$$\begin{aligned} \text{minimize} \quad & \text{rank}(X), \\ \text{subject to} \quad & \mathcal{A}(X) = B. \end{aligned} \tag{3}$$

However, solving this rank minimization problem is often not practical because it is NP-hard. The time complexity of existing algorithms is at least doubly exponential in the dimension n of the matrix. In this paper, to work out this optimization problem, we propose to use the nuclear norm heuristic [6], which performs rank minimization exactly for a low rank matrix. A rank r matrix has r nonzero singular values, and the sum of them is the nuclear norm, denoted by $\|X\|_*$:

$$\|X\|_* = \sum_{k=1}^{r} \sigma_k(X), \tag{4}$$

where $\sigma_k(X)$ is the k-th largest singular value of X, and the heuristic optimization is given by:

$$\begin{aligned} \text{minimize} \quad & \|X\|_*, \\ \text{subject to} \quad & \mathcal{A}(X) = B. \end{aligned} \tag{5}$$

There have been several effective solutions to nuclear norm minimization [6]. We briefly discuss three methods towards the tradeoff between computational complexity and accuracy of results. Interior point methods are numerically efficient and precise, but the memory requirements for computing significantly increase with the size of the problem; subgradient methods to minimize the nuclear norm problem can be used to solve much larger problems, though it incurs lower accuracy than Interior point methods; for even larger problems, like the matrix in our scenario, we adopt a low-rank semidefinite programming which factorizes the decision variable.

Considering our $N \times T$ matrix X, we intensively decompose it into three elements such that:

$$X = U \Sigma V^T \tag{6}$$

where U is an $N \times N$ unitary matrix and V is a $T \times T$ unitary matrix. Σ is an $N \times T$ diagonal matrix containing the singular values σ_k, which is arranged in a monotonically decreasing order. Then we can factorize matrix X as $X = U\Sigma V^T = LR^T$, where $L = U\Sigma^{1/2}$ and $R = V\Sigma^{1/2}$.

In our algorithm, we strive to find a suitable X to satisfy $X = LR^T$, where L is an $N \times K$ matrix and R an $T \times K$ matrix. There may be more than one such pair of L and R, so we seek for L and R which have Frobenius norm as small as possible:

$$\begin{aligned} \text{minimize} \quad & \|L\|_F^2 + \|R\|_F^2, \\ \text{subject to} \quad & \mathcal{A}(LR^T) = B. \end{aligned} \tag{7}$$

Often the matrix in our scenario is not exactly low-rank, as well as the readings received by the sink are not accurate but contain errors. We therefore solve the following optimization instead:

$$\text{minimize } \|\mathcal{A}(LR^T) - B\|_F^2 + \|L\|_F^2 + \|R\|_F^2. \tag{8}$$

In this paper, achieving low-rank and measurement equations is of equal importance compared to traffic matrix [9], in which a precise fit to the measured data is more critical than the goal of achieving low rank.

The goal of Equ.(8) is two-fold: reducing the error and guaranteeing a low rank solution. This is a convex optimization problem and we could exploit the method proposed by Y. Zhang et al. [9] to recover the missing data. First we randomly initialize L and R. Then we fix L and optimize R with linear least squares method. After that we swap L and R and let L be the optimization variable. This step is iteratively executed with alternating the roles of L and R. Our implementation often converges after a moderate number of iterations and results in an acceptable recovery error. We illustrate our numerical simulation results in the next section.

978-1-4244-5637-6/10/$26.00 ©2010 IEEE

IV. Implementation and evaluation

In our paper we implement simulations on two data-sets to evaluate the efficiency and effectiveness of our proposed algorithm. The first data-set is a real world trace from the Intel Berkeley Research lab and we choose the temperature information collected by 54 sensor nodes on March 1st, 2004 [13]. Another data-set is a synthetic trace which contains matrixes with sizes varying from 100×100 to 3000×3000 with the same rank. After the routing protocol is established as an initial step, the data collected from sensor nodes are transmitted to the sink through this routing tree. Sensor nodes are loosely synchronized and they perform operations of the sampling and data transmission during each time instance.

A. Error evaluation on real world trace

The real world trace forms a 54×2880 matrix which represents the temperature values sampled by 54 nodes in one day. Sink install a uniform probability to each nodes in WSN. We called this probability as "sampling ratio" (from 90% to 20% entries) and the $1 - Sampling\ ratio$ means the dumping ratio in all sensor readings. Then, each node forwards their uniformly random selected readings to sink according to the same probability. At last, sensor readings is collected by sink, based on which we will recover the original matrix using our EDCA algorithm. We calculate an error matrix by comparing the recovered matrix with the original matrix. The estimation of mean and standard deviation of errors are denoted as μ and σ respectively. The mean μ and standard deviation σ on different sampling ratios are shown in Fig. 2. Fig. 2 shows that the errors of means are quite small as the loss rate increases. While the triple mean square roots errors is increased with the increase of the loss rate, it is still considerable.

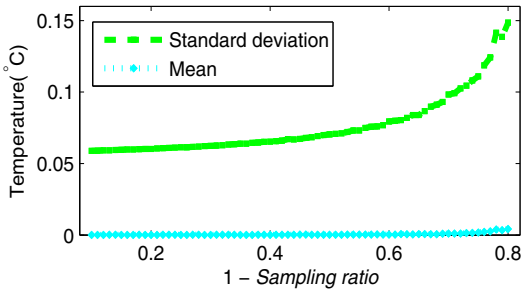

Fig. 2. Accuracy vs. $1 - Sampling\ ratio$

Fig. 3 shows the cumulative distribution function (CDF) of the relative errors as a function of the sampling rate. We can see that despite varying sampling ratios, our recovered sensor readings only have less than 1% relative error at most nodes.

B. Error evaluation on synthetic trace

We then test our EDCA on synthetic data including number of nodes N which grows up from 100 to 3000 and time instances are from 100 to 3000 at the same time. For keeping the rank of the matrix X we generate matrix C which is a square matrix whose rank remains the same, for instance, 10.

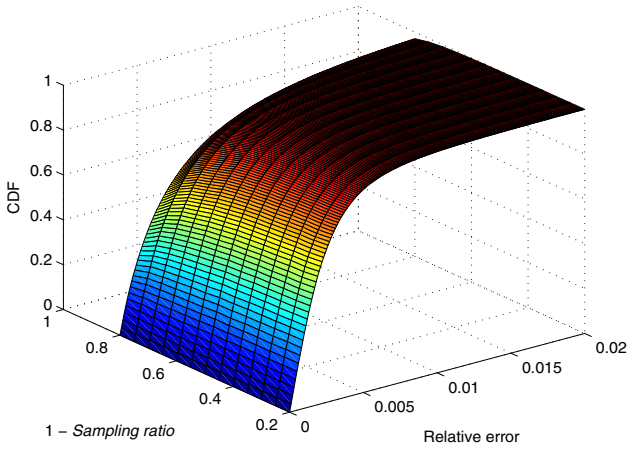

Fig. 3. CDF vs. $1 - Sampling\ ratio$ and Relative error

Then we assign value in X to $X(i, j) = max(C(i, j)) + b + 20, 0)$, where b is a standard norm distributed noise. By doing so, we add a jitter in the original X and make sure that X is satisfied $X \geq 0$. In this simulation, we fix the percent of sampling to 15% and study the errors of our algorithms over the size N of the sensor network. Fig. 4 depicts the mean μ and standard deviation δ of errors of processing data collection on N nodes at N time instances. Obviously, the errors using EDCA is quite small with a variety of network sizes.Interestingly, as the networks grows in size, EDCA even results in the decreased errors. For example, when the number of nodes is more than 1000, the errors are close to zero. This is due to the reason that with fixed sampling rate at nodes, more sensor nodes included in the data processing often result in more enough information to recovery the original information.

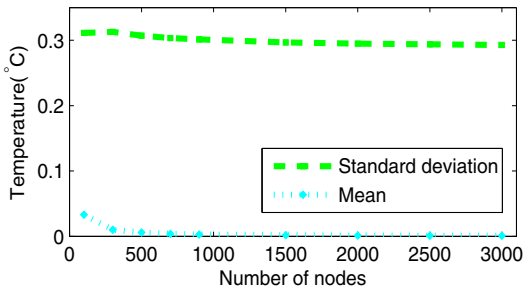

Fig. 4. Accuracy vs. Number of nodes

C. Evaluation of power consumption and lifetime of WSNs

In this scenario,we calculate the power consumption and the lifetime of the WSNs using Intel Berkeley Lab Data. Here for convenience, we suppose that transmitting a 32bit packet as the unit of power consumption. Since the power consumption is in direct proportion to size of packet and function part of node. Since the size of packets sent from each node back to sink is 64 bits: 32 bits used for storing temperature numerical

value ,16 bits for the node ID, and 16 bits for the time instance. We could calculate this power consumption as 2 packets. In this way, receiving a 64bit packet count as 2×0.4 packets and sampling a 32 physical quantity count as 1×2 packets according to Table 1. In Fig. 5, We also compare our results with the naive centralized scheme (denoted by "Centralized exact") where every sampled packet is simply forwarded to the sink. We can see that the EDCA can apparently reduce the total power consumption as the loss rate increases.

And the network lifetime is defined by the time duration of the first sensor node which runs out of its power. So we find the maximum power wasted numerical value M_{max} on every simulation using EDCA. Then we mark the maximum power consumption numerical value using centralized exact as M_0 and it is all the same in every simulation. Finally, we define $(1/M_{max})/(1/M_0)$ as the appraisal value of life time.

We compare our results with the naive centralized scheme. Fig. 6 demonstrates the effectiveness of EDCA on prolonging the lifetime of sensor networks. For example, when the discard rate is 80%, EDCA can prolong the lifetime of WSNs about five times compared with the centralized scheme.

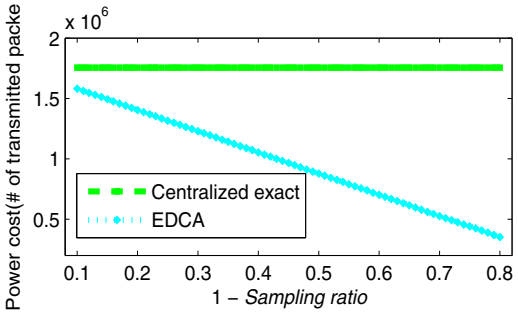

Fig. 5. Power consumption vs. 1 – *Sampling ratio*

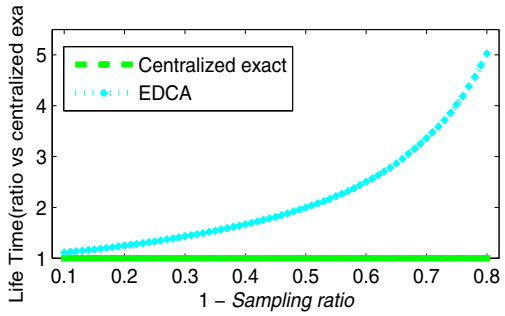

Fig. 6. Life-time vs. 1 – *Sampling ratio*

V. Conclusion

In this paper, we propose an efficient data collection approach so-called EDCA for wireless sensor networks. The EDCA takes advantage of spatial and temporal correlation in WSNs to save energy by randomly choosing the node and time instance to sample data. The key problem is how to recover the original data matrix based on an incomplete one which contains many missing values. To address this problem, matrix completion problem, we exploit the nuclear norm minimization methods. The EDCA can efficiently recover the original data with considerably small errors and simulation results show the efficiency and effectiveness of our proposed approach for data collection in wireless sensor networks. What is more, the EDCA can reduce energy consumption significantly and thus prolong the life time of sensor networks as shown in the simulation part. In the future work, we will further exploit the spatial and temporal correlation to reduce the number of nodes to execute sampling. Besides, the routing and topology information can be utilized to further reduce the transmitted packets.

Acknowledgment

This work was supported in part through Chinese National 863 project (No.2007AA01Z223), National Natural Science Foundation of China (No.60803115, No. 60873127), Chinese National University Basic Research Funding (No.M2009022), and SRF for ROCS, SEM. The corresponding author is Hongbo Jiang.

References

[1] R. Cristescu, B. Beferull-Lozano, M. Vetterli, and R. Wattenhofer, "Network correlated data gathering with explicit communication: Np-completeness and algorithms," *IEEE/ACM Transaction on Networking,* Vol.14, Issue.1, pp. 41-54, 2006.

[2] J. Chou, D. Petrovic, and K. Ramchandran, "A distributed and adaptive signal processing approach to reducing energy consumption in sensor networks", In *Proceedings of IEEE INFOCOM,* 2003.

[3] D. Slepian and J. K. Wolf, "Noiseless encoding of correlated information sources," *IEEE Transaction on Information Theory,* Vol.19, Issue.4, pp. 471-480, 1973.

[4] M. F. Duarte, S. Sarvotham, M. B. Wakin, D. Baron, and R. G. Baraniuk, "Joint Sparsity Models for Distributed Compressed Sensing," In *Proceedings of the Workshop on Signal Processing with Adaptative Sparse Structured Representations,* 2005.

[5] C. Luo, F. Wu, J. Sun, and C. W. Chen, "Compressive data gathering for large-scale wireless sensor networks," In *Proceedings of MobiCom,* 2009.

[6] B. Recht, M. Fazel, and P. Parrilo, "Guaranteed minimum rank solutions of matrix equations via nuclear norm minimization," *To appear in SIAM Review*.

[7] E. Candès and B. Recht, "Exact Matrix Completion via Convex Optimization," *Foundations of Computational Mathematics,* Vol.9, pp.717–772, 2009.

[8] M. C. Vuran, Ö. B. Akan , and I. F. Akyildiz, "Spatio-temporal correlation: theory and applications for wireless sensor networks," *Computer Networks,* Vol.45, Issue.3, pp.245-259, 2004.

[9] Y. Zhang, M. Roughan, W. Willinger, and L. Qiu,"Spatio-temporal compressive sensing and internet traffic matrices," In *Proceedings of the ACM SIGCOMM conference on Data communication,* 2009.

[10] I. F. Akyildiz, W. Su, Y. Sankarasubramaniam , and E. Cayirci, "Wireless sensor networks: a survey," *Computer Networks,* Vol.38, Issue.4, pp. 393-422, 2002.

[11] A. Mainwaring, J. Polastre, R. Szewczyk, and D. Culle, "Wireless sensor networks for habitat monitoring," In *Proceedings of ACM Workshop on Sensor Networks and Applications*, 2002.

[12] D. L. Donoho, "Compressed Sensing", *IEEE Transaction on Information Theory*, Vol. 52, Issue. 4, pp. 1289-1306, 2006.

[13] Intel Lab Data, *URL: http://db.csail.mit.edu/labdata/labdata.html*

[14] R. H. Keshavan, A. Montanari and S. Oh, "Low-rank Matrix Completion with Noisy Observations: a Quantitative Comparison", In *Allerton conference on Communications, Control and Computing,* September 2009.