Scalable Data Chunk Similarity for Mobile Node Data Using Group Pattern Object

Jayanthi.R[1], John Augustine.P[2]
[1]*PG Scholar, Department of CSE, Sri Eshwar college of Engineering, Coimbatore*
[2]*Associate Professor, Department of CSE, Sri Eshwar college of Engineering, Coimbatore*

Abstract- **Spatial co-location pattern mining is an interesting and important task in spatial data mining which discovers the subsets of spatial features frequently observed together in nearby geographic space. However, the traditional framework of mining prevalent co-location patterns produces numerous redundant co-location patterns, which makes it hard for users to understand or apply. To address this issue, in this paper we study the problem of reducing redundancy in a collection of prevalent co-location patterns by utilizing the spatial distribution information of co-location instances**

In this project considered the redundancy reduction problem of the spatial prevalent co-locations by applying distribution information from co-location instances. It is worth mentioning that the proposed method not only solves the redundancy reduction problem but also provides high efficiency. There are several interesting directions that we are considering for future work: (1) Compression of the spatial prevalent co-locations (to get fewer co-locations but more usability, i.e., a set of representative co-locations); (2) Ordering of the spatial prevalent co-locations; and (3) Reducing the redundancy of prevalent co-locations found in incrementally updated data.

In proposed study analysis a natural phenomena show that many creatures form large social groups and move in regular patterns. However, previous works focus on finding the movement patterns of each single object or all objects. This thesis proposes an efficient distributed mining algorithm to jointly identify a group of moving objects and discover their movement patterns in wireless sensor networks. Afterward, a compression algorithm, called (2 phase and 2D) 2P2D is proposed, which utilizes the discovered group movement patterns shared by the transmitting node and the receiving node to compress data and thereby reduces the amount of delivered data.

Index Terms- Compressive Sensing, Data Gathering, Random Walk, Wireless Sensor Network

I. INTRODUCTION

Sensor networks are deployed to sense, monitor, and report events of interest in a wide range of applications including, but are not limited to, military, health care, and animal tracking. In many applications, such monitoring networks consist of energy constrained nodes that are expected to operate over an extended period of time, making energy efficient monitoring an important feature for unattended networks. In such scenarios, nodes are designed to transmit information only when a relevant event is detected.

Consequently, given the location of an event-triggered node, the location of a real event reported by the node can be approximated within the node's sensing range. In the example depicted in Fig. 1.1, the locations of the combat vehicle at different time intervals can be revealed to an adversary observing nodes transmissions. There are three parameters that can be associated with an event detected and reported by a sensor node: the description of the event, the time of the event, and the location of the event.

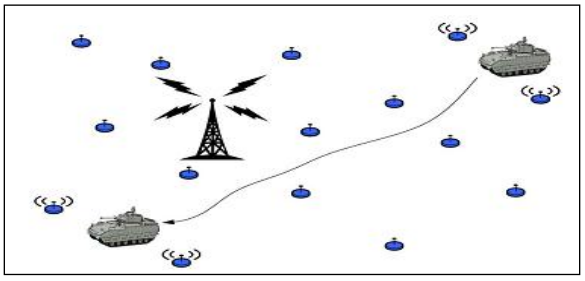

Fig 1.1 Sensor Network

When sensor networks are deployed in untrustworthy environments, protecting the privacy of the three parameters that can be attributed to an event-triggered transmission becomes an important security feature in the design of wireless sensor networks. While transmitting the "description" of a sensed

event in a private manner can be achieved via encryption primitives, hiding the timing and spatial information of reported events cannot be achieved via cryptographic means.

Encrypting a message before transmission, for instance, can hide the context of the message from unauthorized observers, but the mere existence of the cipher text is indicative of information transmission. The source anonymity problem in wireless sensor networks is the problem of studying techniques that provide time and location privacy for events reported by sensor nodes.

II. LITRATURE SURVEY

Giorgio Quer et al [1] describe a addressed the problem of exploiting CS in WSNs taking into account network topology and routing, which is used to transport random projections of the sensed data to the sink. Thus, the main contribution of this paper is the quantification of the benefits of CS in realistic multi-hop WSNs when CS in used in conjunction with routing. In addition, they study the problem of finding good transformations to make real sensed data meet the sparsity requirements of CS and show that widely used transformations are not suitable for a large spectrum of real signals. They also provide a simulation based comparison between the commonly used random sampling (considered here in conjunction with spline interpolation) and CS based data gathering, for synthetic and real sensed data. In this work they studied the behavior of CS when used jointly with a routing scheme for recovering two types of signals: synthetic ones and real sensor data. They showed that for the synthetic signal the reconstruction at the sink node is enhanced when applying CS, whereas the application of CS for real sensor data is not straightforward. Thus, as a next step of their ongoing research, they intend to further investigate which signal representation and routing allows CS to outperform random sampling in realistic WSN deployments. This requires to jointly investigate the design of the two matrices Φ and , since the sparsity requirements and the incoherence between routing and signal representation have to be met.

Liu Xiang et al [3] describe a burgeoning technique for signal processing, compressed sensing (CS) is being increasingly applied to wireless communications. However, little work is done to apply CS to multihop networking scenarios. In this paper [12], they investigate the application of CS to data collection in wireless sensor networks, and they aim at minimizing the network energy consumption through joint routing and compressed aggregation. They first characterize the optimal solution to this optimization problem, then they prove its NP-completeness. They further propose a mixed integer programming formulation along with a greedy heuristic, from which both the optimal (for small scale problems) and the near-optimal (for large scale problems) aggregation trees are obtained.

Energy efficiency of data collection is one of the dominating issues of wireless sensor networks (WSNs). It has been tackled from various aspects since the outset of WSNs, which include, among others, energy conserving sleep scheduling [13], topology control, mobile data collectors and data aggregation [14]. Whereas the first three approaches (and many others) focus on the efficiency of networking techniques that transport the sensory data, data aggregation directly aims at significantly reducing the amount of data to be transported, and it hence complements other approaches and is deemed as the most crucial mechanism to achieve energy efficient data collection for WSNs.

Fatemeh Faze et al [4], describe a power-efficient underwater sensor network scheme employing compressed sensing and random channel access. The proposed scheme is suitable for applications where a large number of sensor nodes are deployed uniformly over a certain area to measure a physical phenomenon. The underlying assumption is that most physical phenomena have sparse representations in the frequency domain. The network is assumed to have a Fusion Center (FC) that collects the observations of sensor nodes and reconstructs the measured field based on the obtained measurements. The proposed method is completely decentralized, i.e., sensor nodes act independently without the need for coordination with each other or with the FC. During each frame, a Bernoulli random generator at each node determines whether the node participates in sampling or stays inactive during that sampling period. If selected, it measures the physical quantity of interest, e.g. temperature. A second random generator with a uniform distribution then picks a (random) delay for the node to send its data to the

FC. The proposed network scheme, referred to as Random Access Compressed Sensing (RACS), results in a simple power-efficient design, for: a) it eliminates the need for duplexing, which requires coordination from the FC b) there is no need for acknowledgment packets and retransmissions in case packets collide; and moreover, c) it is efficient in terms of the communication resources used (only a small fraction of nodes sample and transmit in each sampling period).

Sensor networks consist of a large number of sensor nodes that are deployed over a region of interest to observe the physical environment. Each sensor node communicates its observation of the field to a central node, referred to as the Fusion Center (FC) and the FC retrieves the information about the physical field. In this paper, they are interested in the case where the field of interest is sparse in some domain, noting that most natural phenomena are compressible (sparse) in an appropriate basis. The theory of compressed sensing establishes that under certain conditions on a signal, exact signal recovery is possible with a small number of random measurements [17].

III. COMPRESSION MODEL

A. CLUSTERING FORMATION

Recently, clustering based on objects' movement behavior has attracted more attention. transform the location sequences into a transaction-like data on users and based on which to obtain a valid group, but the proposed AGP and VG growth are still Apriority-like or FP-growth based algorithms that suffer from high computing cost and memory demand. Propose a density-based clustering algorithm, which makes use of an optimal time interval and the average Euclidean distance between each point of two trajectories, to approach the trajectory clustering problem.

Consider a wireless sensor network, where n nodes are randomly and independently distributed in a unit square. At a sampling instant, each sensor node i takes a measurement xi. Let x = {x1….xn}denote the signal vector sampled by all the sensor nodes. For simplicity and tractability of analysis, we assume that the vector x can be represented as a k-sparse signal in the canonical basis (i.e., C = I). The objective of our algorithm is to reconstruct all of the measurements fx1xng by collecting a sufficient number of linear combinations of measurements {x1….xn} via the approach of CS. To accomplish this goal, we propose a random walk algorithm to collect measurements over the network. The above problem can be formulated as follows: Consider a given graph G = (VE) with vertex set V and edge set E. Perform a random walk of length t starting at any vertex, which is chosen uniformly at random from V.

Measurements are aggregated and transmitted along the path visited by each walk. At the end of each walk, the last node visited by the walk will obtain one random projection and then transmit it to the sink through shortest path routing strategy.

B. NETWORK MODEL

As described above, there are n sensor nodes randomly and independently deployed in a wireless sensor network. We model the wireless sensor network as a two-dimensional random geometric graph G(n,r)where two nodes are connected if the euclidean distance between them is smaller than some transmission radius r(n).

Such graphs have been widely used as simplified models for wireless sensor networks, which make it possible for us to analyze our algorithm. It is well known that the transmission radius R(n) should be scale as Ø√log(n/n).plog n=n) to guarantee the connectivity of the network with high probability. It has been shown that such a graph has "nice" properties including the uniformly of node distribution and the regularity of node degree. In particular, each node has the same order of degree Q(log n) with high probability.

C.COMPRESSIVE SENSING WITH RANDOM WALKS FOR DATA GATHERING

The algorithm is proposed to use standard random walk algorithm to collect random measurements. The proposed algorithm is described as follows. At the beginning of the algorithm, m sensor nodes are selected at random to initialize m independent random walks. For each walk, at each step one node chooses one of its neighbors and performs a linear combination with the measurement of the neighbor. At the end of random walks, m random projections are generated from these random walks and then these projections will be sent to the sink through shortest path routing strategy. In practice, all these random walks can be performed simultaneously.

Finally, the sink can reconstruct all the measurements using the recovery algorithm of compressive sensing.

For practical implementation, we can generate a packet containing a time stamp t at the first node for each random walk, and then transmit it to a randomly selected neighbor. The time stamp t decrements when the walk passes a node until the time stamp reaches zero. Note that for each walk revisiting to the same node is allowed in our scheme but the measurement of the node contributes to the computation results only once. On the other hand, to recover the original data, the sink needs to know the measurement matrix A, which is characterized by our random walk based routing algorithm. To avoid the overhead incurred by routing information, we can adopt the following approach. Before invocation of the algorithm, each node generates a random seed and sends it to the sink.

These random seeds can be used through a pseudo-random number generator to determine the next neighbor node that the walk should pass over. Finally, the sink can know which nodes are visited by each random walk according to the seeds received from sensor nodes. Thus, we can only store the length of a walk in the header of a packet. We note that this overhead does not increase with the size of the messages for sensor readings and only depends on the number of random walks that are to be performed simultaneously.

Encapsulated into a large packet, user only needs one of the common information's for the length of random walk. Here do not consider the impact of packet loss during data transmission since it can be handled at the lower layer. Moreover, due to the randomness of the proposed routing strategy, it is possible to forward data to another neighbor node to avoid packet loss in the case of link failure.

To prove that the measurement matrix A constructed from random walks can be used to recover k-sparse signals using '1 minimization decoding algorithm. To do this, we first construct a bipartite graph from the matrix A, and derive the probability that a vertex is visited by a random walk to show the distribution of nonzero elements in A (Lemma 1).

Definition 2: (Uniform Graph).

The undirected graph G = (V E) is called a (D c) uniform graph if for some constant c - 1, the degree of each vertex y 2 V is between D and cD. As stated above, a random geometric graph G has a good property for node degree. From the bounds of the numberof neighbors for any node in graph G [?], the degree of each node is between (1 - m) k log ln and (1 + m) k log n with high probability when the transmission radius r(n) pk log n=n, where m 2 (0 1) and k > 1= log (4=e). Therefore, the random geometric graph is a uniform graph.

Definition 3: (Mixing Time).

Let G = (V E) be a (D c) uniform graph and p be the stationary distribution of a random walk over graph G. The d-mixing time of G is defined as the smallest t0 such that (Π – Π') , where p0 is the distribution that a random walk of length t0 starting at any vertex ends up.

D.CONSTRUCTING MEASUREMENT MATRIX A FROM RANDOM WALKS

Let G=(VE) be an undirected graph with j Vj = n, and A be an m n boolean matrix. By performing m independent random walks on graph G, each row of A can be seen as the characteristic vector of a subset vertices in V. Visited by the walk.

First construct a bipartite graph from the matrix A, where A can be seen as an m n adjacency matrix of the bipartite graph. The nodes on the left hand side corresponds to the signal coefficients set V, and the nodes on the right hand side corresponds to the measurements set M with j M j = m.

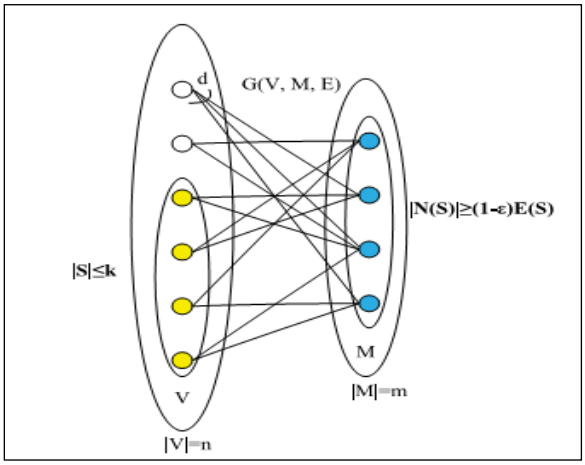

Fig 4.1 Object Random Walk

$$A = \begin{pmatrix} 0 & 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 & 1 \end{pmatrix}.$$

A node i in V connects to a node j in M if and only if the jth random walk visits node i. For simplicity of analysis, assume that any set S has the cardinality jSj = k. E (S) represents the set of links connecting from S to the nodes in M.

IV. DATA COMPRESSIVE PARADIGM

A. NETWORK CONSTRUCTION

In this module, an object is defined as a target, such as an animal or a bird that is recognizable and trackable by the tracking network. To represent the location of an object, geometric models and symbolic models are widely used. A geometric location denotes precise two-dimension coordinates; while a symbolic location represents an area, such as the sensing area of a sensor node or a cluster of sensor nodes, defined by the application.

B. BJECT TRACKING

Object tracking is defined as a task of detecting a moving object's location and reporting the location data to the sink periodically at a time interval. Hence, an observation on an object is defined by the obtained location data. A picture box control is used as the 2D graph area in which circles are drawn such that one circle represents one sensor node and redrawn such that they are moving randomly in all directions with given random speed. The X and Y coordinates are also maintained at regular intervals and updated to the cluster head nodes. It is assumed that sensor nodes wake up periodically to detect objects. Using the timer controls, it is designed such that sensor nodes functions. When a sensor node wakes up on its duty cycle and detects an object of interest, it transmits the location data of the object to its CH. Here, the location data include a times tamp, the ID of an object, and its location. Instead of forwarding the data upward immediately, the CH compresses the data accumulated for a batch period and sends it to the CH of the upper layer. The process is repeated until the sink receives the location data. To learn the significant movement patterns, Probabilistic Suffix Tree (PST) is adopted for it has the lowest storage requirement. PST is also useful and efficient in predicting the next item of a sequence. For a given sequence s and a PST T, the predict_next algorithm is proposed which outputs the most probable next item, denoted by predictnext(T,s).

C. GMP MINE ALGORITHM

The GMP Mine algorithm extracts the movement patterns from the location sequences by learning a PST for each object. In this module, a new similarity measure simp to compare the similarity of two objects is proposed. For each of their significant movement patterns, the new similarity measure considers not merely two probability distributions but also two weight factors, i.e., the significance of the pattern regarding to each PST. The similarity score simp of oi and oj based on their respective PSTs, Ti and Tj. The similarity score is equally divided into a threshold value so that the objects can be grouped in a single cluster if the similarity score falls within that range.

D. COMPRESSION ALGORITHM WITH GROUP MOVEMENT PATTERN

To reduce the amount of delivered data, the 2P2D algorithm is proposed which leverages the group movement patterns derived to compress the location sequences of moving objects elaborately. The algorithm includes the sequence merge phase and the entropy reduction phase to compress location sequences vertically and horizontally. In the sequence merge phase, the Merge algorithm is proposed to compress the location sequences of a group of moving objects.

Since objects with similar movement patterns are identified as a group, their location sequences are similar. The Merge algorithm avoids redundant sending of their locations, and thus, reduces the overall sequence length. It combines the sequences of a group of moving objects by 1) trimming multiple identical symbols at the same time interval into a single symbol or 2) choosing a qualified symbol to represent them when a tolerance of loss of accuracy is specified by the application. Therefore, the algorithm trims and prunes more items when the group size is larger and the group relationships are more distinct. Besides, in the case that only the location center of a group of objects is of interest, the approach can find the aggregated value in the phase, instead of transmitting all location sequences back to the sink for post-processing. To compress the location sequences for a group of moving objects, the proposed system processes the Merge algorithm

V. EXPERIMENTAL RESULTS

The following Table 5.1 describes experimental result for existing and proposed system's probabilities of successful data recovery. The table contains the number of nodes or objects used for the experimental result analysis, the probability of the data compression and recovery of successful data of the Probability Suffix Tree (PST) and Probability Merge Algorithm (2D2PMA) methods.

Number of Objects (Node)	Probability Suffix Tree (PST)	Probability Merge Algorithm (2D2PMA)
N8, N4 [2]	20	40
N7, N6, N9, N10[4]	45	65
N5, N12,N14, N21,N23, N8[6]	58	78
N9, N13, N15, N18,N22, N2, N8, N14 [8]	65	85
N9, N13, N15, N18, N22,N2, N8, N14, N21,N23[10]	75	95

Table 5.1 The probabilities of successful data recovery of PST & 2D2PMA

The following Fig 5.1 & Fig 5.2 describes experimental result for existing and proposed system's probabilities of successful data recovery. The Figures shows the number objects used for the experimental result analysis, the probability of the data compression and recovery of successful data of the Probability Suffix Tree (PST) and Probability Merge Algorithm (2D2PMA) methods.

Compressed Ratio of PST & 2D2PMA Methods		
Number of Nodes [N]	PST [%]	2D2PMA [%]
20	40	46
45	65	72
58	78	83
65	84	89
75	91	97

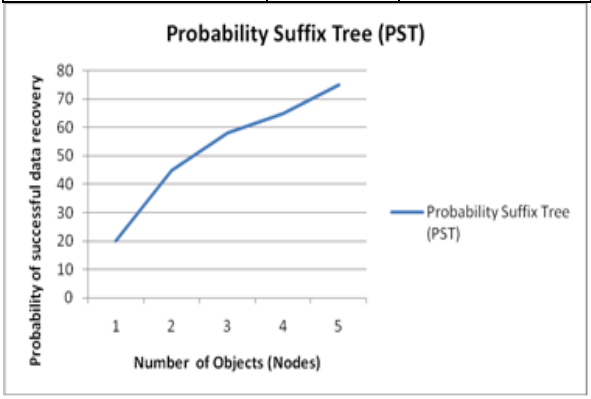

Fig 5.1 Probability Suffix Tree Model

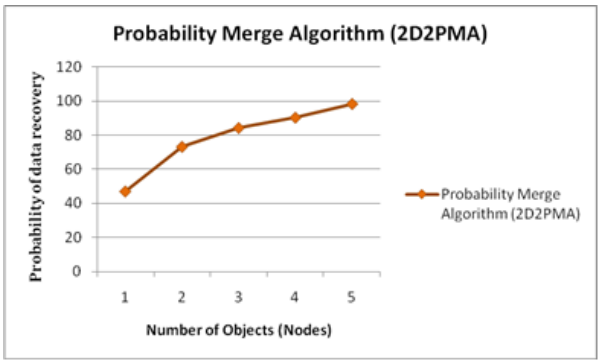

Fig 5.2 Probabilities Merge Algorithm (2D2PMA) Model

The compressive sensing for data gathering in WSNs of existing and proposed approaches performance is evaluated. The following Table 5.2 describes the Compressed Ratio of Probability Suffix Tree (PST) & 2 phase and 2D Probability Merging Algorithm (2D2PMA) Methods. The table shows the details about the number of objects or nodes, ratio of compressed data for gathering of the PST and 2D2PMA approaches.

PST and 2D2PMA approaches.

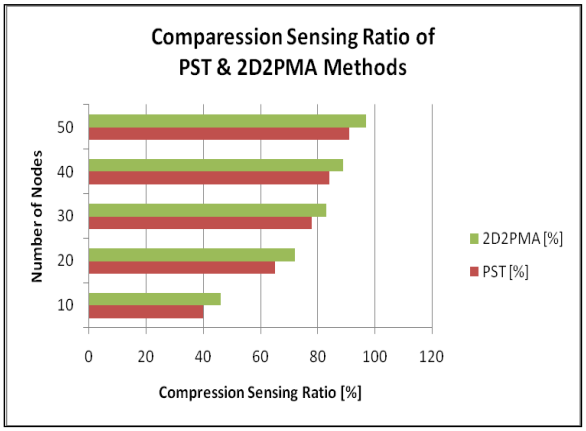

Fig 5.3 Compressed Ratio of PST & 2D2PMA Methods

The following Table 5.3 describes Performances Analysis for PST and 2P2PMA methods. The table contains Number of Objects and Transmitting nodes and Receiving nodes to compress the Data in milliseconds.

Table 5.3 Performances Analysis for PST and 2P2PMA Methods

The following Fig 5.4 describes Performances Analysis for PST and 2P2PMA methods. It shows the details about the Number of Objects and Transmitting

nodes and Receiving nodes to Compress the Data in milliseconds.

Number of Objects (Nodes)	Transmitting and Receiving Nodes to Compress Data (ms)	
	PST	2P2PMA
10	163	161
20	165	163
30	173	170
40	176	174
50	181	177
60	184	182
70	185	183
80	191	188
90	194	190
100	197	195

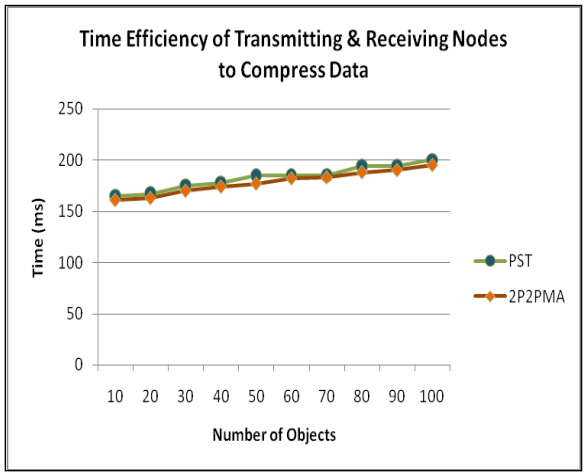

Fig 5.4 Time Efficiency of Transmitting & Receiving Nodes to Compress Data of PST and 2P2PMA Methods

RESULTS

- A noticeable observation from the performance evaluation of the existing and proposed approaches, that the proposed probability merging scheme requires much fewer transmissions than the existing PST schemes when achieving a similar reconstruction quality of compressed sensing data.
- The proposed novel compression algorithm for compress the location data of a group of moving objects with lossless of information and reducing the communication cost of data gathering in terms of number of transmissions.
- The proposed novel compression algorithm effectively reduces the amount of delivered data and also the compression ratio is enhanced.

VI.CONCLUSION AND FUTURE ENHANCEMENT

Through this paper, the Compression algorithm effectively reduces the size of delivered data, storage and consumption expense for data transmission in WSNs. The system is very fast and any transaction can be viewed or retaken at any level. Error messages are given at each level of input in individual stages. The data management process becomes easy. All the day-to-day activities are assigned to them through browser interface. The administrator can view the contents in the server, select a file and download to the system wherever the administrator is working. Several areas to be developed in future, so the application must be upgraded for the new ones required and it is possible to modifications according to new requirements and specifications. The survey provides a best assistance in object tracking and merging two path sequences.

REFERENCES

[1] W. Wang, M. Garofalakis, and K. Ramachandran, “Distributed Sparse Random Projections for Refinable Approximation,”Proc. ACM/IEEE 6th Int. Conf. Inf. Process. Sensor Network. (IPSN ’07), Apr. 2007.

[2] M. Vetterli and J. Kovaˇ cevi´ c. Wavelets and Subband Coding. Prentice Hall, Englewood Cliffs, NJ, 1995.

[3] S. Mallat. A Wavelet Tour of Signal Processing. Academic Press, San Diego, CA, 1999.

[4] E. Candes and T. Tao. Near Optimal Signal Recovery From Random Projections: Universal Encoding Strategies. IEEE Transactions on Information Theory, 52(12), pp. 5406-5425, December 2006.

[5] D. Donoho. Compressed Sensing. IEEE Transactions on Information Theory, 52(4), pp. 1289-1306, April 2006.

[6] G. Quer, R. Masiero, D. Munaretto, M. Rossi, J. Widmer, and M.Zorzi, “On the Interplay between Routing and Signal Representation for Compressive Sensing in Wireless Sensor

Networks,"Proc.Inf. Theory Appl. Workshop (ITA '09), Feb. 2009.

[7] D. Donoho, "Compressed sensing," IEEE Trans. on Information Theory, vol. 52, no. 4, pp. 4036–4048, 2006.

[8] E. Cand`es and T. Tao, "Near optimal signal recovery from random projections: Universal encoding strategies?" IEEE Trans. On Information Theory, vol. 52, no. 12, pp. 5406–5425, 2006.

[9] E. Cand`es, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," IEEE Trans. on Information Theory, vol. 52, no. 2, pp.489–509, 2006.

[10] J. Haupt, W. Bajwa, M. Rabbat, and R. Nowak, "Compressive Sensing for Networked Data: a Different Approach to Decentralized Compression," IEEE Signal Processing Magazine, vol. 25, no. 2, pp. 92–101, 2008.