Information Centric Sensor Network Management Via Community Structure

Tzu-Yu Chuang, Kwang-Cheng Chen, *Fellow, IEEE*, and H. Vincent Poor, *Fellow, IEEE*

***Abstract*—Self-organizing dense sensor networks are expected to resolve the challenges of spectral efficiency, energy efficiency, and device management. This paper explores the use of information dependence among sensors to form a community structure of sensor data, and uses this structure to develop an information–centric processing methodology to achieve self-organizing dense sensor networks. Moreover, a data aggregator adopts ℓ_1 regularization to easily enhance energy efficiency with uncertain sensor availability. Combining this community structure and these data recovery algorithms, a novel self-organizing device management scheme is proposed to mitigate the sensor maintenance costs. This approach is justified through simulations and analysis, and is envisioned to have future application in the Internet of Things.**

***Index Terms*—Information-centric processing, energy efficiency, WSNs, LASSO, community structure, self-organization, device management, IoT, social network analysis.**

I. Introduction

SENSOR network management, addressing issues such as spectral and energy efficiency and device management, is essential in dense *wireless sensor networks* (WSNs) and the *Internet of Things* (IoT) [1]. Such networks usually consist of large numbers of spatially scattered sensors that organize themselves into hierarchical network structures, while each battery-operated or energy harvesting sensor has limited computing power to support transmissions of measurements to/from a *Data Aggregator* (DA). Because of highly limited spectral and energy resources, it is critical to exploit the tradeoff between transmissions and accuracy of measurements.

In practice, spatial signals exhibit statistical dependence, and they are captured by sensors in the form of correlated data, which provides significant opportunities for data processing and sensor management. Statistical decision theory indicates that, given statistical properties of states, variables can often be estimated by the realizations (data) of other dependent variables with vanishing errors. Therefore, if a portion of data can be accurately reconstructed from other highly correlated data at the DA, their transmissions are redundant, and thus can be reduced to achieve spectral efficiency, while the data accuracy is maintained by data recovery algorithms. Furthermore, missing data from transmissions or sensor failures can be properly reconstructed without re-transmissions or human intervention.

A realization of the above idea relies on the *discovery of community* in networks [2], in which all sensors can be grouped into communities such that desired information can be inferred with relatively small and tolerable error by using only a portion of the sensor data in the same group. We might determine communities in the fashion of (almost) isolated clusters to simplify the access of such information, but this approach is not desirable in dense WSNs because of the high complexity of clustering algorithms. Even worse, the inference of data often requires full access to statistical parameters, and thus leads to heavy loads of data accumulation and computation in large scale systems. To address these challenges, an approximation algorithm based on the *Least Absolute Shrinkage and Selection Operator* (LASSO) principle provides a fast algorithm to determine community members associated with each sensor, instead of finding the exact isolated cluster structure.

In this paper, given the community structure of sensor data, an algorithm to identify an energy-efficient transmission pattern is proposed to save resources by reducing the number of necessary transmissions. Furthermore, we develop an algorithm for *device management* to maintain the time-varying WSN operation given unknown malfunctioning sensors. Our proposed algorithms are justified via numerical experiments that demonstrate simultaneous achievements in measurement accuracy and energy savings. Also, a theoretical framework for further analysis of device management is provided.

There have been various sensor management schemes proposed in the literature, including those addressing the lifetime and organization of WSNs [3], [4], protocols [5], and cluster-based efficient transmission schemes [6]. This letter presents the first effort to enable sensor management by leveraging community structures, with a theoretical framework for optimal operations. This approach exploits the tradeoff between the accuracy of measurements and costs of maintenance, which is critical for future development and management of WSNs and the IoT.

II. Problem Formulation

We consider a system of p distributed sensors that takes measurements from a physical field as illustrated in Fig. 1. These sensors are represented by a point set $\mathcal{V} = \{1, \dots, p\}$, and their measurements are characterized by a p-variate $X = [X_1, \dots, X_p]^\top \in \mathbb{R}^p$ with a non-singular $p \times p$ covariance matrix $\mathbf{\Sigma}$. A data aggregator can collect the sensors' measurements, and coordinate sensor operations. Inspired by social network analysis [7], we introduce an information-centric principle that forms a network of data via the dependence of these measurements.

Manuscript received September 17, 2014; revised February 15, 2015; accepted February 19, 2015. Date of publication March 5, 2015; date of current version May 7, 2015. This work was supported in part by the Ministry of Science and Technology of Taiwan, National Taiwan University and Intel Corporation under Grants MOST 103-2221-E-002-022-MY3, MOST 103-2911-I-002-001 and NTU-ICRP-104R7501, and in part by the U.S. Army Research Office under MURI Grant W911NF-11-1-0036. The associate editor coordinating the review of this paper and approving it for publication was D. Qiao.

T.-Y. Chuang and K.-C. Chen are with the Graduate Institute of Communication Engineering, National Taiwan University, Taipei 10617, Taiwan (e-mail: d99942022@ntu.edu.tw; ckc@ntu.edu.tw).

H. V. Poor is with the Department of Electrical Engineering, Princeton University, Princeton, NJ 08544 USA (e-mail: poor@princeton.edu).

Digital Object Identifier 10.1109/LCOMM.2015.2409070

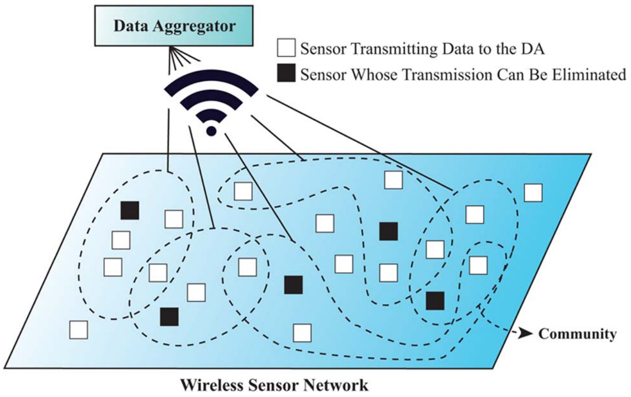
Data Aggregator
□ Sensor Transmitting Data to the DA
■ Sensor Whose Transmission Can Be Eliminated
Community
Wireless Sensor Network

Fig. 1. Ideally, we can group sensors into almost disjoint clusters to simplify algorithms for efficient transmission and device management. In practice, this situation does not always arise, and optimal algorithms to solve such problems are NP-hard. Instead, we choose the most dependent nodes of a given sensor to form its community and control the size of the community and accuracy of data recovery by ℓ_1 optimization. A fraction of the sensors' measurement can thus be eliminated to save transmission energy.

Given limited resources for transmissions, the DA shall determine an appropriate transmission pattern for sensors to obtain knowledge of the physical field as accurately as possible and to save spectrum and energy. From the information-centric viewpoint, a straightforward idea is that the DA can suppress some sensors whose data can be recovered with vanishing error by a fraction of other sensors' data. This operation requires the discovery of the community of an arbitrary sensor, which is defined as follows.

Definition 1 (Community): The community of node $i \in \mathcal{V}$, denoted by $\mathcal{D}_i$, is a subset of $\mathcal{V} \setminus \{i\}$. A node j is in $\mathcal{D}_i$ if its data can be used in the reconstruction of the data of node i, *i.e.*, x_i. The size of $\mathcal{D}_i$ is its cardinality, $|\mathcal{D}_i|$

Furthermore, we can impose certain constraints such that the loss of accuracy in the recovery of x_i using the data collected from $\mathcal{D}_i$, *i.e.*, $\{x_j\}_{j\in\mathcal{D}_i}$, is minimized. With this definition of community, one can easily see that an efficient transmission scheme can be realized by the knowledge of the entire community structure of the network. For simplicity, we use the notation x_A to denote the vector composed of $\{x_i\}_{i\in A}$. We formulate community discovery in the following problem.

Problem 1 (Community Discovery): Given a data point $x_i, i \in \mathcal{V}$ to be recovered, determine a regression f and a community of node i, $\mathcal{D}_i$, such that the expected value of the loss function, $g(x; f) = \mathsf{Loss}(x_i, f(x_{\mathcal{D}_i}))$, is minimized while the size of $\mathcal{D}_i$ is kept as small as possible.

Theoretically, constructing a large community for data recovery to achieve high accuracy is possible, but not practical, because including too many nodes in one community can result in the problem of *overfitting*, and it also introduces a heavy computational load.

Once the community structure of WSNs is discovered, the DA can determine a transmission pattern by suppressing a fraction of sensors for which data can be recovered by its community with little cost in the accuracy of measurements, in order to save transmissions and thus energy. The determination of such a transmission pattern can be stated as the following procedure:

Problem 2 (Efficient Transmission): Given an error tolerance level ϵ, and the community structure, $\{\mathcal{D}_i\}_{i\in\mathcal{V}}$, discovered, determine a subset $\mathcal{C} \subset \mathcal{V}$ such that the data collected by nodes in $\mathcal{C}$, $x_{\mathcal{C}}$, can be recovered by data in $\cup_{i\in\mathcal{C}}\mathcal{D}_i$ with error less than ϵ.

Furthermore, the community structure can be used for device management. Consider that, because of the extremely large scale of WSNs, even with a small probability of sensor failure, there will be frequent malfunctions, thus making the timely replacement of broken sensors expensive in terms of maintenance cost. In addition, such maintenance is not practical if the locations of failed sensors are not tracked precisely. However, the discovery of community structure serves the purpose of device management because the network can maintain a time-varying WSN in a self-organizing way. Given an error level specified by the service provider, when any sensor breaks, the DA can update the community structure immediately, and then recover the missing data caused by malfunctions if necessary. In the meantime, the service provider can track the accuracy of measurements by the real-time error level as a metric for device management, and thus determine an appropriate strategy for operations, such as the timing of maintenance, or replacements of sensors.

Problem 3 (Device Management): Given a current error level ϵ, community structure $\{\mathcal{D}_i\}_{i\in\mathcal{V}}$, and a malfunctioning sensor $i \in \mathcal{V}$, determine a procedure to recover the data of i, x_i, and update the error level and the community structure.

In summary, we shall construct an efficient algorithm to discover the community structure based on the accuracy of data recovery, and then use it to select a proper subset of data to transmit, in order to save spectrum and energy.

III. Main Results

A. Community Discovery

In **Problem 1**, assume that the loss function g is mean square error (MSE) and the regression f is constrained to be linear. For a node $i \in \mathcal{V}$, we want to select a subset of nodes, $\mathcal{D}_i \subseteq \mathcal{V} \setminus \{i\}$, such that there is a decision $\mathbf{u} \in \mathbb{R}^{|\mathcal{D}_i|}$ minimizing

$$\mathbb{E}\left[g(X,\mathbf{u})\right] = \mathbb{E}\left[\left(X_i - \mathbf{u}^\top X_{\mathcal{D}_i}\right)^2\right] \tag{1}$$

where X_A is the random vector composed of $\{X_i\}_{i\in A}$. Given n samples of historical data $\mathbb{X} = [\mathbf{x}_1\mathbf{x}_2\cdots\mathbf{x}_p] \in \mathbb{R}^{n\times p}$, a heuristic solution is to find a decision $\mathbf{u}_i$ over all data other than $\mathbf{x}_i$, i.e, $\mathbf{x}_{[-i]} \triangleq [\mathbf{x}_j]_{j\in\mathcal{V}, j\neq i}$, and to keep the number of zero entries of $\mathbf{u}$ as large as possible. That is, we want to select a sparse decision $\mathbf{u}_i$ such that the mean square error and the ℓ_1 norm of $\mathbf{u}_i$ are minimized simultaneously. This is luckily a typical LASSO problem and can be easily solved by introducing a LASSO parameter τ:

$$\mathsf{minimize}\ \left[\tau\|\mathbf{u}_i\|_1 + \frac{1}{2}\left\|\mathbf{x}_{[-i]}\mathbf{u}_i - \mathbf{x}_i\right\|_2^2\right]. \tag{2}$$

The nodes corresponding to non-zero entries of $\mathbf{u}_i$ are collected together to form the $\mathcal{D}_i$. The parameter τ controls the expected MSE of prediction and the sparsity of $\mathbf{u}_i$; *i.e.*, the size of the community. With proper setting of τ, we can form a small community of node i, from which data reconstructs x_i with expected error less than the pre-defined error level.

Furthermore, we can find the community of i for an arbitrary subset $\mathcal{C} \subseteq \mathcal{V} \setminus \{i\}$ in the same fashion if necessary.

B. Efficient Transmission

It is of interest to use the community structure, $\{\mathcal{D}_i\}_{i\in\mathcal{V}}$, to determine a transmission pattern to save spectrum and energy. A simple way to do this is for the DA to select a subset of nodes, $\mathcal{C} \subset \mathcal{V}$, in which sensors can switch into sleep mode with their data being reconstructed from other data at the DA. Given a tolerance of error level ϵ, we aim to save as many transmissions as possible. This leads to finding an algorithm that optimizes the following objective:

$$\begin{aligned} &\mathsf{maximize}\ |\mathcal{C}| \\ &\mathsf{subject\ to}\ \{\cup_{i\in\mathcal{C}}\mathcal{D}_i^\epsilon\} \subseteq \mathcal{C}^c,\ \mathcal{C} \subseteq \mathcal{V} \end{aligned} \tag{3}$$

where $\{\mathcal{D}_i^\epsilon\}_{i\in\mathcal{V}}$ is the community discovered with error bound ϵ, and $\mathcal{C}^c$ is the complement of $\mathcal{C}$. Generally, the complexity of finding an optimal solution to (3) is $\mathcal{O}(2^p)$, which is impractical when p is large. Furthermore, to the best of our knowledge, there is no efficient algorithm to solve this problem with low complexity. To mitigate the computational issue, we propose that, given $\mathcal{T}$ to be the set of sensors that must transmit, a node i is added to $\mathcal{C}$ if the size of $\mathcal{T}$ increased by adding i to $\mathcal{C}$ is minimal. It is guaranteed that we can reconstruct as much data as possible with minimum marginal increase in transmissions. This algorithm is of complexity $\mathcal{O}(p^2 \log p)$, and stated as **Algorithm 1**.

Algorithm 1 Efficient Transmission

```
Input: $\mathbb{X}$, $\epsilon$ **Initial Sets**: $\mathcal{C} \leftarrow \varnothing$, $\mathcal{T} \leftarrow \varnothing$
Find $\{\mathcal{D}_i^\epsilon\}_{i\in\mathcal{V}}$ by LASSO with proper LASSO parameter.
while $\mathcal{C} \cup \mathcal{T} \neq \mathcal{V}$ do
    Check $\mathcal{V} \setminus \{\mathcal{C} \cup \mathcal{T}\}$. ▷ If $D_i^\epsilon$ involves any member in $\mathcal{C}$,
      $i$ should also be included in $\mathcal{T}$
    Find the $i \in \mathcal{V} \setminus \{\mathcal{C} \cup \mathcal{T}\}$ that has minimum $|\mathcal{D}_i^\epsilon \setminus \mathcal{T}|$
    $\mathcal{C} \leftarrow \mathcal{C} \cup \{i\}$, $\mathcal{T} \leftarrow \mathcal{T} \cup \mathcal{D}_i^\epsilon$
```

C. Device Management

We have shown that, once the community structure $\{\mathcal{D}_i\}_{i\in\mathcal{V}}$ has been discovered, it is easy to reconstruct any missing data of malfunctioning sensors with little loss in the accuracy of measurements. In addition, thanks to the fast and easy algorithm for community discovery, the WSN can update the community structure dynamically, and calculate the error level anytime. Therefore, the service provider need not conduct any maintenance until the sensor measurements cannot meet the requirement for accuracy. This procedure enables the WSN to run device management in a self-organizing fashion such that most errors can be fixed automatically, while the best timing of human intervention can be determined by further self-organizing optimization. In particular, given a pre-defined error level ϵ, if sensor $i \in \mathcal{V}$ is broken, then given its community $\mathcal{D}_i$, and the set of current effective sensors, $\mathcal{T} \subset \mathcal{V}$, we conduct the procedure for device management in **Algorithm 2**.

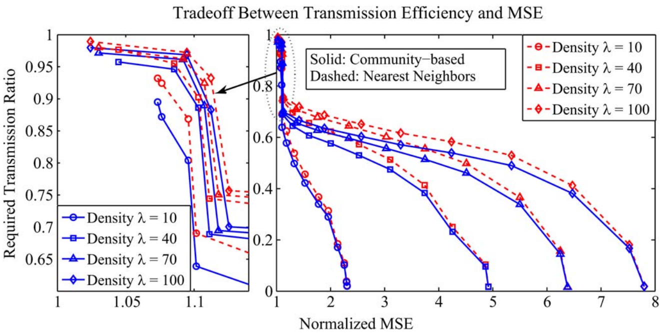

Fig. 2. This figure demonstrates the saving of transmissions by **Algorithm 1** (blue solid lines). It is also compared with the Nearest Neighbor method (red dashed lines).

Algorithm 2 Device Management

```
Input: Broken sensor $i$, $\mathcal{D}_i^\epsilon$, and $\mathcal{T}$
if we can use $\mathcal{T}$ to recover $x_i$ then    ▷ Check if $\mathcal{D}_i^\epsilon \subseteq \mathcal{T}$
    Recover $x_i$ using data from $\mathcal{D}_i^\epsilon$ (using $\mathbf{u}_i^\top x_{\mathcal{D}_i}$)
else
    Restructure $\mathcal{D}_i^\epsilon$ by LASSO with candidates in $\mathcal{T} \setminus \{i\}$
    Recover $x_i$ using the updated $\mathcal{D}_i^\epsilon$
Calculate the MSE of $\hat{x}_i$.
DA reports the new error level to the service provider.
```

There are two advantages of this proposed self-organizing management scheme. First, the DA uses the community structure of WSNs such that they can mitigate the loss of data without creating too much of a burden on computing, and the restructuring of communities can extend the life of WSNs. Secondly, the DA reports the error level to the service provider in terms of MSE, which is a critical measurement of performance, and can be used directly in future planning of device management. The proposed approach not only provides a framework for quantitative analysis of device management other than most intuitive approaches, but also can be extended to many important applications, such as change point detection in a dynamic environment. To the best of our knowledge, there is no other satisfactory framework addressing the device management of WSNs in such a systematic way and providing a framework for quantitative optimization.

IV. Simulation and Conclusion

We simulate a WSN distributed in a unit square area according to a Poisson point process with density λ. The data $\mathbb{X}$ are generated according to $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$ with $\rho_{ij} \overset{\Delta}{=} \exp(-d_{ij}/\theta)$, where d_{ij} is the distance between node i and j, and θ is a randomly selected parameter in [1, 10]. Fig. 2 shows that, with a 10% increase in MSE, we can save 10% to 20% of transmissions depending on the density of WSNs. This is a significant saving in spectrum and energy. Please further note a critical point at which the MSE increases rapidly around 12% with only 70% of total transmissions. This point provides a reference for a tolerable error for device management. In Fig. 3, we show that WSNs with high density are not superior against the malfunctioning sensors under our approach,

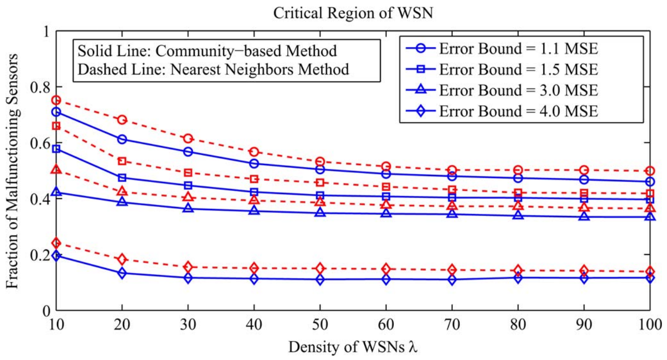

Fig. 3. The fraction of malfunctioning sensors for different densities of WSNs, constrained at different error tolerance levels. It can be seen that, the denser the network is the smaller fraction we can lose against error. Our method (blue solid line) is compared with Nearest Neighbor method (red dashed line).

because they often include too many nodes in one community, making data recovery less efficient. Obviously there is a tradeoff between the density of WSNs and their robustness in device management.

We adopt the *Nearest Neighbor* method for comparison with our proposed methodology. In Nearest Neighbor, each node selects k nodes closest to itself, *i.e.*, the ones with k highest correlation coefficients, as its community, with k varying to meet the error constraint. The simulation shows that the Nearest Neighbor method often yields a larger community than the community selected by our proposed method (LASSO) because it does not optimize the community size, and thus results in worse performance.

A. *Necessary Condition for Community*

Here we take a closer look at the condition defining community. Given $\mathbf{u}^*$ to be a solution to (2), $\mathbf{u}^*$ must obey the following conditions [8]:

$$|\mathbf{x}_{\mathcal{D}_i}^\top(\mathbb{X}\mathbf{u}^* - \mathbf{x}_i)| = \tau\mathbf{1}_{|\mathcal{D}_i|};\ \|\mathbf{x}_{\bar{\mathcal{D}}_i}^\top(\mathbb{X}\mathbf{u}^* - \mathbf{x}_i)\|_\infty < \tau.$$

By the law of large numbers, an asymptotic version of this condition is that, $u_j = 0$ if $|\sigma_{ij} - \sum_{k=1}^{p}\sigma_{jk}u_k| \leq \tau$. That is, node j is excluded from i's community if the estimator of σ_{ij} is close enough with the absence of data $\mathbf{x}_j$, implying either that i and j are almost independent, or that j can be replaced by other highly correlated nodes.

Consider a degenerate case: Suppose i's location is fixed, and all other sensors are separated enough such that they are mutually independent. A necessary condition for the solution of LASSO provides a bound on σ_{ij} that depends on the LASSO parameter, and thus we can determine a region with radius r_i based on the correlation model such that the nodes with distances to i greater than r_i are excluded from $\mathcal{D}_i$.

B. *Probability of Community Overlap*

The performance of efficient transmission depends on two key factors: the WSN density and the fraction of community overlap. A huge community is often discovered in a WSN with high density, which is not favored in efficient data recovery. Furthermore, the more sensors that are responsible for multiple nodes in data recovery, the more transmissions we can save.

We can derive an upper bound on the probability of community overlap from a necessary condition for forming a community. Assume there is a realization of the network distributed according to a homogeneous spatial point process. Given an area with volume a, let the number of points n of a process falling within this area be distributed according $\rho(n, a)$. Consider two arbitrary nodes i and j in the network and a necessary condition for community, $d(i, j) \leq r_f$. The condition that they can both use another node k as their community simultaneously is equivalent to the condition that k belongs to both i's and j's communities while i and j are not in each other's communities. The probability of such an event is upper bounded by the probability that there are at least two nodes within k's area of radius r_f, and these two nodes are sufficiently far apart. That is

$$\mathbb{P}(\text{overlapping}) \leq \sum_{n=2}^{\infty} \rho\left(n, \pi r_f^2\right)\left(1 - \mathbb{P}(D < r_f)^{\binom{n}{2}}\right). \quad (4)$$

This probability provides a baseline for further analytical work on the performance of sensor management, but is not the focus of this letter.

C. *Conclusion*

In this letter, we have discussed two common problems encountered in dense wireless sensor network management: spectrum and energy efficient transmission and device management. We have proposed an approach to sensor management using community structure, with a consequent framework for reliable and efficient algorithms. Future directions of this research include use of this framework to develop fast and efficient parameter estimates in large scale WSNs, and the detection of change points of statistical properties of WSNs.

References

[1] L. Atzori, A. Iera, and G. Morabito, "SIOT: Giving a social structure to the Internet of things," *IEEE Commun. Lett.*, vol. 15, no. 11, pp. 1193–1195, Nov. 2011.

[2] S. Fortunato, "Community detection in graphs," *Phys. Rep.*, vol. 486, no. 3–5, pp. 75–174, Feb. 2010.

[3] J. Zhu and S. Papavassiliou, "On the energy-efficient organization and the lifetime of multi-hop sensor networks," *IEEE Commun. Lett.*, vol. 7, no. 11, pp. 537–539, Nov. 2003.

[4] Y. Chen and Q. Zhao, "On the lifetime of wireless sensor networks," *IEEE Commun. Lett.*, vol. 9, no. 11, pp. 976–978, Nov. 2005.

[5] 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. 2004.

[6] W. Ye, J. Heidemann, and D. Estrin, "Medium access control with coordinated adaptive sleeping for wireless sensor networks," *IEEE/ACM Trans. Netw.*, vol. 12, no. 3, pp. 493–506, Jun. 2004.

[7] K.-C. Chen, M. Chiang, and H. V. Poor, "From technological networks to social networks," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 9, pp. 548–572, Sep. 2013.

[8] J.-J. Fuchs, "On sparse representations in arbitrary redundant bases," *IEEE Trans. Inf. Theory*, vol. 50, no. 6, pp. 1341–1344, Jun. 2004.