Received January 5, 2015, accepted January 25, 2015, date of publication February 24, 2015, date of current version March 31, 2015.

Digital Object Identifier 10.1109/ACCESS.2015.2407194

Sparse Signal Processing Concepts for Efficient 5G System Design

GERHARD WUNDER[1,2], HOLGER BOCHE[3], THOMAS STROHMER[4], AND PETER JUNG[1]
[1]Technische Unversität Berlin, Berlin 10623, Germany
[2]Fraunhofer Heinrich-Hertz-Institut, Berlin 10587, Germany
[3]Technische Unversität München, Munich 80333, Germany
[4]University of California at Davis, Davis, CA 95616 USA

Corresponding author: G. Wunder (gerhard.wunder@hhi.fraunhofer.de)

The work of G. Wunder was supported by the European Commission through the 5GNOW Project (www.5gnow.eu) under Grant 318555, by the German Federal Ministry of Education and Research within the PROPHYLAXE project (www.ict-prophylaxe.de) under Grant 16KIS0005K, and the Deutsche Forschungsgemeinschaft (DFG) under Grant WU 598/4-1. The work of T. Strohmer was supported in part by the National Science Foundation under Grant DTRA-DMS 1042939 and in part by the Defense Advanced Research Projects Agency under Grant N66001-11-1-4090. The work of P. Jung was supported by the Deutsche Forschungsgemeinschaft under Grant JU 2795/2-1.

ABSTRACT As it becomes increasingly apparent that 4G will not be able to meet the emerging demands of future mobile communication systems, the question what could make up a 5G system, what are the crucial challenges, and what are the key drivers is part of intensive, ongoing discussions. Partly due to the advent of compressive sensing, methods that can optimally exploit sparsity in signals have received tremendous attention in recent years. In this paper, we will describe a variety of scenarios in which signal sparsity arises naturally in 5G wireless systems. Signal sparsity and the associated rich collection of tools and algorithms will thus be a viable source for innovation in 5G wireless system design. We will also describe applications of this sparse signal processing paradigm in Multiple Input Multiple Output random access, cloud radio access networks, compressive channel-source network coding, and embedded security. We will also emphasize an important open problem that may arise in 5G system design, for which sparsity will potentially play a key role in their solution.

INDEX TERMS Compressed sensing, cloud radio acess networks, massive random access, embedded security, source coding.

I. WHAT DRIVES 5G?

The introduction of 4G clearly marked a first peak of the smartphones' revolution, offering high bandwidth mobile radio access almost anywhere and anytime. However, as it becomes increasingly apparent that 4G will not be able to meet the emerging demands of future mobile communication systems, the question what could make up a 5G system, what are the key drivers, is still open and part of intensive discussions. Actually with 5G and related visions, we believe, mobile communications research is on the brink towards a new innovation cycle [1]–[3]:

- The *Internet of Things* (IoT) will connect billions of devices, i.e., the things of our everyday life, which is far more than 4G can technically and economically accommodate. This will then open up new ways to monitor, assist, secure, control e.g. in the telemedicine area, smart homes, smart factory etc. In fact, the IoT could change the way we see the Internet as a human-to-human interface towards a more general machine-to-machine (M2M) platform.
- *Security, privacy, and data integrity* will be a key issue in the 5G market. Current security solutions e.g. for the IoT fall short due to the sheer number of nodes which must be flexibly managed and distributed in the network.
- Moreover, the *Tactile Internet* (TI) comprises a vast amount of real-time applications with extremely low latency requirements including *industrial wireless applications* such as *Smart Grids*. Motivated by the human tactile sense, which requires round-trip times in the order of 1ms, 5G can then be applied for steering and control scenarios implying a disruptive change from today's content driven communications. This is far shorter than current 4G cellular systems allow for, missing the target by nearly two orders of magnitude.
- *Gigabit wireless connectivity* is required in large crowd gatherings with possibly interactively connected devices

using angle-controlled 3D video streaming, augmented reality, etc.

These examples make it very clear that 5G networks are not only just about providing higher rates for the next smartphone generation (although certainly important!), but more about *enabling, integrating services* and *embedded security* which both implement very different (virtually contradicting) application requirements. From a technical perspective it seems to be utmost challenging to provide such uniform service experience to users under the premises of future heterogeneous networking and small cell scenarios. Consequently, the radio access has to be *flexible, scalable, content aware, robust, reliable and efficient in terms of energy and spectrum*. Actually, with the limitations of current 4G system (i.e., high latency, very bulky control signalling architecture, no embedded security etc.), this would put further pressure on the common value chains on which the operators rely in order to compensate for investment costs for future user services. Hence, there is a clear motivation for an innovative and disruptive re-design of current mobile communication networks from scratch.

Having the short-comings of 4G in mind, we develop the elements of a 5G research agenda based on *sparse signal processing* (also called compressed sensing (CS)). Here, sparsity typically means that only a few samples of the signal are actually non-zero but their locations may not be known. This new paradigm has been an intriguing topic in mathematics and signal processing in recent years. Sparsity-based concepts have also been succesfully applied in specific communication problems, e.g. the *peak power control problem* (see [4]). Moreover, in the context of 5G, [5] has recently identified five disruptive *technology directions* (device-centric architectures, massive Multiple Input Multiple Output (MIMO), mmWave, native support of massive M2M, smarter devices). We will indeed argue here that sparsity in communication signals is a viable innovation source for these technology directions and will, hence, appear as our basic methodology.

II. ENABLING 5G TECHNICAL CONCEPTS

Let us first identify a series of enabling 5G concepts and corresponding research challenges which shall be approached with the new paradigm:

i) *Fast and scalable random access* is one 5G key concept to handle the massive number of *sporadic traffic* generating devices (e.g. IoT devices, but also Smartphones' Apps etc.) which are most of the time inactive but regularly access the network for minor updates with no human interaction. Sporadic traffic will dramatically increase in the 5G market and, obviously, cannot be handled with the bulky 4G random access procedures. Two major challenges must be addressed to leverage successful 5G business models: i) unprecedented number of devices asynchronously access the network over a limited resource, ii) the same resource carries *control signaling and users' payload*. Dimensioning the channel according to classical theory results in a severe waste of resources which, even worse, does not scale towards the requirements of the IoT (low-cost, deep indoor coverage, long life time of devices). Yet, since typically user activity [6], channel profiles [7] and message sizes are compressible within a very large receive space, sparse signal processing methodology is a natural framework to support sporadic traffic, cf. Sec. IV where we discuss a suitable "one shot" approach.
In addition, the TI requires ultra-fast acquisition in the order of $100\mu s$ on the physical layer to enable the 1ms round-trip time [1]. Notably, this implies that even small, say 1kBit data bursts, result in a huge bandwidth requirement. Again, we will argue that classical theory requires that *for each real-time connected device* a significant control signaling overhead is necessary to allow for swift channel estimation, equalization and demodulation. Since, in addition, this traffic class must be also extremely reliable, control signaling must be separated from the data which is very inefficient and can be much better handled by sparse signal processing.

ii) *Densification of cells* together with *cloud-powered baseband processing* and *wireless network virtualization* (so-called *Cloud-RAN*) is another 5G key concept to increase spectrum and energy efficiency and handle the projected traffic growth [1]. It is based on the deployment of many light base stations with overlapping coverage, performing only signal conversion to/from the digital domain, connected through a high capacity link to a cloud of data centers. It is worth emphasizing that such architecture can be efficiently realized in the mmWave frequency bands. Coordinated processing of signals by multiple network nodes is a key design element in such a virtualized cellular network. Yet, it is still mostly overlooked in the literature that existing cooperative designs do not scale in terms of run time requirements and required control information. In Sec. V we analyze existing schemes and discuss a new control signaling architecture thereby efficiently exploiting not only the compressible channels but also the number of effectively coordinated nodes (out of the total number of nodes). We will also discuss the beneficial interacting role of sparse prediction and coordination in this scenario.

iii) *5G source coding concepts* for the massive number of distributed sensors and actuators will be very different as well. Shannon's famous separation theorem states that under appropriate conditions data compression (source coding) and error protection (channel coding) can be performed separately and sequentially, without any performance loss. While the separation theorem has has tremendous impact on the design of communication systems, in several practical scenarios the conditions of the Shannon's separation theorem neither hold nor can be used as a good approximation. Naturally, this raises the question whether and in which form *Shannon's separation theorem* holds true and can serve as a guiding design principle for the communication scenarios we expect to encounter in 5G. In Sec. VI we discuss the related concepts and algorithms.

iv) *Security* will play a central role in 5G networks. In today's communication systems there is an architectural separation between data encryption and error correction. The encryption module is based on cryptographic principles and views the underlying communication channel as an ideal bit pipe. The error correction module is typically implemented at the physical layer. It adds redundancy into the source bits in order to combat channel impairments or multiuser interference and transforms the noisy communication channel into a reliable bit pipe. While such a separation based architecture has long been an obvious solution in most systems, a number of applications have emerged in recent years where encryption mechanisms must be embedded in the physical layer (*called physical layer security/or embedded security* [8]). Embedded security is a relatively new research area exploiting the *stochastic and physically unclonable* nature of the wireless channel including noise *and* the hardware, e.g. for symmetric key and fingerprint generation. It is a further 5G vision to embed security into the concepts of Sec. IV – Sec. VI from scratch and to explore the benefits and tradeoffs in such innovative designs to enable scalable, fast security mechanism implemented without user interaction, please see Sec. VII for a discussion of new approaches in this regard.

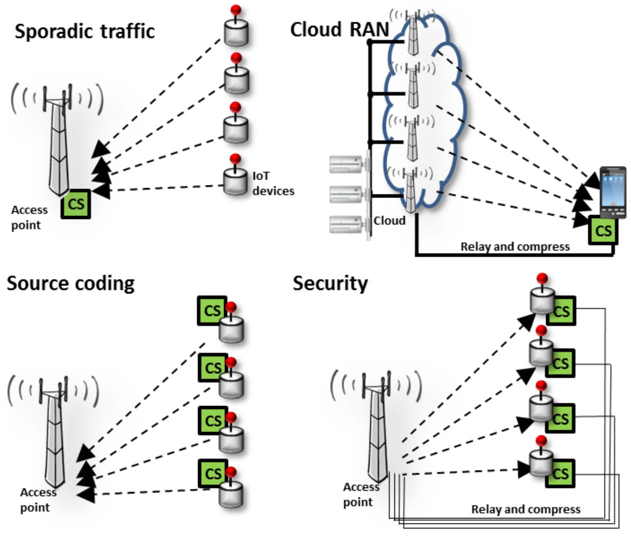

FIGURE 1. Considered 5G deployment scenarios and respective "location" of CS entities.

The scenarios are depicted in Fig. 1. In the figure, we also indicate where we possibly see the CS measurement device. Note, though, that this does not at all imply that all the processing is done in this location, e.g. for the C-RAN "relay and compress" architecture described in Sec. V! Clearly, in general, we have not said much about the complexity yet and it is regarded as a crucial topic for futue research when realizing the agenda outlined in this paper.

Notation: For a vector $x \in \mathbb{C}^n$, $\|x\|_p$ denotes standard ℓ_p–norm and $\|x\|_0$ counts the number of non–zero elements in x. For sets/matrices we use capital/calligraphic letters and $\text{vec}(X)$ denotes the vectorization of the matrix X. For a given vector x, $\text{diag}(x)$ and $\text{circ}(x)$ refer to a diagonal matrix with x on its diagonal and to a circulant matrix (of appropriate size) with x as its first row, respectively. The nuclear norm $\|X\|_*$ of a matrix X is the ℓ_1–norm of the vector of its singular values.

III. THE SPARSE SIGNAL PROCESSING PARADIGM

At the core of compressive sensing (CS) lies the discovery, that it is possible to reconstruct a sparse signal exactly from an underdetermined linear system of equations *and* that this can be done in a computationally efficient manner via convex programming [9]. Consider $\Phi x = y$, where Φ is an $m \times n$ matrix of rank m with $m < n$. Here, Φ models the measurement (or sensing) process, $y \in \mathbb{C}^m$ is the vector of observations and $x \in \mathbb{C}^n$ is the signal of interest. Conventional linear algebra wisdom tells us that in principle the number of measurements m has to be at least as large as the signal length n, otherwise the system would be underdetermined and there would be infinitely many solutions. Most data acquisition devices of current technology obey this principle in one way or another.

Assume now that x is k–sparse, i.e., x satisfies $\|x\|_0 \leq k \ll n$ but we do *not* know the locations of the non-zero entries of x. Due to the sparsity of x one could try to compute x via an exhaustive search, which however is NP-hard. Instead, the CS paradigm tells us that that under certain conditions on the matrix Φ and the sparsity of x we can reconstruct x from its measurements via linear or quadratic programming techniques [9]. While x may not be sparse with respect to the standard basis, in many cases we are dealing *compressible* signals which are well–approximated by signals which are sparse in some specific domain (e.g. in the Fourier or wavelet domain). Thus, let Ψ be a $n \times n$ matrix which sparsifies x and $y = \Phi\Psi x + z$, where z is a noise vector. To recover x we consider the convex optimization problem:

$$\tilde{x} = \arg\min \|x\|_1 \quad \text{s.t. } \|y - \Phi\Psi x\|_2 \leq \epsilon \tag{1}$$

with ϵ depending on variance of z. Here, the ℓ_1–norm appears as the convex relaxation of $\|\cdot\|_0$.

A particularly important subject for the application in multi-node/multi-terminal scenarios is the separation, respectively demixing of multiple sparse signal contributions from the compressed superimposed receive signal (*compressive demixing*). Consider the model:

$$y = \Phi \sum_{p=1}^{P} \Psi_p x_p + e \tag{2}$$

where each contribution $\Psi_p x_p$ is compressible in its own domain. Here, the synthesis matrices $\{\Psi_p\}$ must essentially ensure that there is no common intersecting subspace. Following [10] a condition for exact (and/or stable) recovery is to avoid nontrivial (well-separated) intersections between the per-user ℓ_1-descent cones at the true but unknown signals. See also [11] for a quite general formulation of this problem.

In recent years, the idea of compressive sensing has been extended and generalized in several ways. Adopting such a more general viewpoint (see [12]) we may assume that the objects of interest x (for example n–dimensional vectors or $n \times n$–matrices) can be well–approximated by superpositions $\sum_j c_j \psi_j$ of *a few* fixed atoms taken from an atomic set $\mathcal{A} \equiv \{\psi_j\}$. Compared to the full linear span of *all* atoms $\mathcal{A}$ the *feasible set of objects has then substantially reduced complexity*. In the standard CS problem $\mathcal{A}$ corresponds to Ψ.

A more advanced set $\mathcal{A}$ is given by the superposition of a few rank–one $n \times n$–matrices. The *nuclear norm* $\|\cdot\|_*$ will serve as convex relaxation of the rank–function [13], [14]. Assume that we observe the $n \times n$ matrix x by taking m Hilbert–Schmidt inner products of the form $(\Phi X)_l := \langle \Phi_l, X \rangle$. In analogy to (1), we can attempt to recover x via nuclear norm minimization, i.e., via:

$$\tilde{X} = \arg\min \|X\|_* \quad \text{s.t. } \|y - \Phi X\|_2 \leq \epsilon \tag{3}$$

which can be solved e.g. by semidefinite programming.

While by now we have a fairly good theoretical understanding of such low-rank matrix recovery (LMR) problems as in (3), only preliminary results are known for the practical relevant combination of *simultaneously* low–rank and sparse structures, as present for example in (sparse) *compressive phase retrieval* or *compressive blind deconvolution* discussed in Sec. V and Sec. IV. Multi-objective (ℓ_1 and nuclear norm) convex programs for simultaneously sparse and low-rank matrices are limited by the so called rank-sparsity incoherence and stable recovery cannot be achieved at the theoretically optimal (for non–convex recovery) number of measurements [15].

In several applications we are confronted with nonlinear measurements, such as intensity or quantized measurements. Thus, we may consider the more general case where we get information about the object x by taking m non–adaptive, linear or nonlinear noisy observations of the form $(y)_l = f(\langle \phi_l, x \rangle) + (z)_l$. The famous *phase retrieval problem* falls in this setting, here $f(\langle \phi_l, x \rangle) = |\langle \phi_l, x \rangle|^2$. Recently it has been shown [16] that the phase retrieval problem can be casted as a linear matrix–estimation problem under rank–constraints and exploiting the LMR framework with further cone constraints, i.e., $\mathcal{A}$ is the set of rank–one positive–semidefinite matrices. The approach in [16], called PhaseLift, proceeds by *lifting* the absolute–square map on vectors to a linear map on rank-one matrices. To be explicit, with $\Phi_l = \phi_l \phi_l^*$ and $X = xx^*$ one has $(y)_l = \langle \Phi_l, X \rangle + (z)_l = (\Phi X)_l + (z)_l$ which matches the LMR model, i.e., algorithms related to (3) can be used for recovery. An even more constrained form of (noisy) phase retrieval problem is to recover a *sparse* complex vector from noisy intensity measurements. See [17] for some recent theoretical results and further references.

IV. COMPRESSIVE MULTI-ANTENNA RANDOM ACCESS

In this section we will introduce a general *compressive multiple antenna random access*. In this model each device asynchronously accesses the network thereby carrying overlapping data and control signals. In such a system, data detectability becomes increasingly erroneous the more the control is interfering with the data. Yet, the control must somehow interfere with data in order to allow for (swift) estimation of the channel. This seems to be a contradicting, irresolvable task at first sight. However, we show how we can cope with this task by exploiting sparse signaling principles.

A. RANDOM ACCESS: A KEY APPLICATION FOR SPARSE SIGNAL PROCESSING

A key 5G application of the sparse signal processing paradigm is the evolution of the random access channel (RACH) with asynchronous short–message support which explicitly exploits compressibility. In RACH, sparse structures are present in many directions: (i) due to user activity and the near/far–behavior only a small but unknown subset of users participate in the random access at a particular base station (ii) the mobile channels (its spreading function) from the terminals to base stations are sparse in delay and Doppler and (iii) MIMO channel matrices are often low–rank due to collocation of antennas and (iv) sporadic traffic with short-message type payload is intrinsically sparse.

On the other hand, multiple challenging tasks have to performed simultaneously in such an improved compressive RACH architecture. Firstly, (i) the active user set has to be identified. Then, (ii) the associated channel coefficients for these users have to be estimated. The separate determination of the channel characteristics is important for possible resource assignment for successive high data rate uplink. Finally, (iii) the RACH data payload for each active user has to be reconstructed. Without explicitly exploiting the sparse structures discussed above, it seems to be practically impossible to achieve steps (i), (ii), and (iii) in a single or few transmission steps. Therefore, traditionally, only step (i) is accomplished in the RACH and steps (ii) and (iii) are postponed to a synchronized uplink channel which comes then with the already mentioned control overhead. Thus, *exploiting the system sparsity* can be a key enabler for IoT and, by similar reasons as we pointed out in Sec. II, for the TI on the physical layer.

Sparse reconstruction methods have been used here already in all these steps separately. For example, step (i) user activation—also known as on-off RACH—can be cast as a CS problem [18], [19]. Multipath channel estimation is meanwhile a classical field for CS methods [7]. Compressive demodulation or demixing of superimposed signals, step (iii), is a further field of research [10], [11].

Let us discuss suitable detection strategies in the following. We will distinguish between *coherent concepts* and *incoherent concepts* where the receiver estimates the multi-antenna channel separately prior to data demodulation or not, respectively.

B. COHERENT MULTIANTENNA RECEIVER CONCEPTS

Our target system to be investigated is described as follows: We adopt a general model where multiple receive antennas n_r are incorporated from scratch. Transmission is on a frame by frame basis where the signal space dimension n within the frame can be very large, say several thousands samples due to large bandwidth or large observation times. The time-space (rows-columns) signal to be compressively sensed in some slot is given by the matrix $Y \in \mathbb{C}^{n \times n_r}$. Let us assume that the channel coherence time is essentially larger than the slot time. There are three sources for sparsity or compressibility that should be exploited:

- *Sporadic (Sparse) Traffic:* The communication in random access is sporadic so that out of n_t nodes only an unknown small subset of size k_0 are actually active. Alternatively, we can assert certain probabilities to each node. This is our *primary source for sparsity* within the receive space $\mathbb{C}^{n \times n_r}$.
- *Multipath Channels:* Communication is over a multipath channel with delay spread n_d. For each p-th/ q-th transmit/receive pair, the channel vector h_q^p contains the n_d coefficients of the *channel impulse response* (CIR) which form the matrix:

$$H_p = \begin{bmatrix} \vdots & & \vdots & & \vdots \\ (h_1^p)_i & \cdots & (h_q^p)_i & \cdots & (h_{n_r}^p)_i \\ \vdots & & \vdots & & \vdots \end{bmatrix} \in \mathbb{C}^{n_d \times n_r} \qquad (4)$$

  We assume that out of the n_d channel coefficients in each column, only k_1 in the CIRs are non-zero and the exact positions of the coefficients within H_p are unknown. This sparsity assumption is fulfilled in most wireless communication scenarios and, indeed, channel estimation was one of the first CS applications here [7]. Meanwhile, CS estimation methods have been extended towards sparsity in the delay-Doppler domain (see [20]). Large bandwidth channels tend to exhibit sparsity only in the delay domain where the support of the pathes is invariant though [21]. This is our *second source for sparsity*.
- *Compressible Short Messages:* Each user transmits a sequence $x_p = C_p d_p$ where $d_p \in \mathcal{M} \cup \{0\} \equiv \mathcal{M}_0$ is from some modulation alphabet which, by the activity model includes a zero energy symbol as well. The matrix $C_p \in \mathbb{C}^{n \times n}$ is some designed or random code matrix. We assume that k_2 data symbols out of the n are actually non-zero. This is our *third source of sparsity*. Note that in the random case the transmitter has to be informed about the code matrices. Further reasoning for this assumption follows when taking intrinsic compressibility of the data payload into account as it will be explained, for example, in Sec. VI.
- *Spatial corelations:* Another source of sparsity is the spatial domain particularly when the number of antennas is large (the so-called *massive antenna regime*). In this situation the covariance matrix is sparse but in some domain which is typically unknown; only in some special cases this matric can be decomposed in the Fourier domain, e.g. for the linear array [22]. A natural question is then how to measure this such (huge) covariance matrix or alternatively how to select a proper basis to exploit the sparsity in the process. Additionally, this cannot be done independent of the data detection process.
- *Topology*: Due to sparse connectivity, users will be separated by their received transmission powers. This is e.g. a possible way to inherently distinguish relevance of certain nodes in a cooperative design (see also the Sec. V).

As we outlined before, our goal is to achieve "one shot" transmission, i.e., user detection, channel estimation and data detection in one time slot. Therefore, for each node p, a control sequence $s_p \in \mathbb{C}^n$ drawn from some pool of sequences known to the access, and an unknown data sequence $x_p \in \mathcal{M}_0^n$ is transmitted. Each transmit sequence has some individual power and the power is split between the control and the data. The sampling equation for the receiver q can be mathematically expressed with (4) as[1]:

$$y = \Phi \left(\sum_{p=1}^{n_t} \left(S_p + X_p\right) H_p + Z \right) =: \Phi(Y) \qquad (5)$$

whereby, for the sake of exposition, we use a circular matrix model $S_p = \text{circ}(s_p)$ and $X_p = \text{circ}(x_p)$ to represent all circular convolutions between transmit signals $s_p + s_q$ and channel impulse responses h_q^p. The matrix Z denotes additive white Gaussian noise with variance σ^2 per component. Measurement on this matrix signal Y is performed with a linear mapping Φ giving the sample values $y = \Phi(Y)$. If the same sampling is used independently for all receive antennas, one could take also $\Phi \in \mathbb{C}^{m \times n}$ as the measurement (compression) matrix. Compared with compressive demixing in Sec.III, the dictionary matrices Ψ_p are now related to a circulant model. In CS for a single signal with a random measurement matrix model $\Phi\Psi_p$ where Φ is isotropically–distributed this is sometimes called the anisotropic case. Noteworthy, in the classical setting this model resembles closely the overloaded multiple access channel which is well understood. In the overloaded case, optimal *mean squared error* designs can be achieved [23] which is very different from the compressive case considered here where such problems have not yet been touched.

[1] It is important to note that this "underlay" signal model does not exclude the case where control and data signalling are completely separated, e.g. by FDMA in the frequency domain.

The model in (5) is actually general enough to include several recent models and once H_p is assumed to be known or X_p is zero the determination of the remaining unknowns under sparsity assumptions reduces (up to the anisotropy) to a standard CS problem. For example, if H_p is already known at the receiver, user activity and data demodulation can be performed with CS–based multiuser detection methods [6] and further references in [19]. However, in general H_p itself has to be estimated within the same transmission frame which is a challenging task both from the algorithmic side as well as from the design point of view. Random codebook and/or pilot assignments and spreading, respectively, could be a possible strategy for this problem [19]. The limits of such *compressive demixing* strategies by convex methods has been characterized for the random orientation model in [10] in a quite general setting. According to [10] reliable demixing is possible whenever the number measurements m is sufficiently above the total sum of all contributing statistical dimensions. In the sparse case the statistical dimension of a single signal contribution amounts for the effective dimension of ℓ_1-descent cones at the unknown signal and scales linearly with signal sparsity. A greedy approach to demixing for more general signal manifolds is contained in [11].

A different strategy could be the combination of *illumination and subsampling* as proposed for the random demodulator in [24] and used for random access in [25]. Using the different cancellation properties of stochastic i.i.d. data with deterministic pilots, it is possible (i) to cumulate sufficient pilot power for sparsity–aware channel estimation in a fixed window at the output of the random demodulator and then (ii) to demodulate the whole data payload. To show the potential of such approach let us consider a simple example using LTE-A 4G parameters[2]:

An LTE-A 4G frame consists of a number of subframes with 20MHz bandwidth; the first subframe contains the RACH with one “big” OFDM symbol of $m = 839$ dimensions located around the frequency center of the subframe. The FFT size is $n = 24578 = 24$k corresponding to the 20MHz bandwidth whereby the remainder bandwidth outside PRACH is used for scheduled transmission in LTE-A, so-called PUSCH. The prefix of the OFDM symbol accommodates delays up to $100\mu s$ (or 30km cell radius) which equals 3000 dimensions. In the standard the RACH is responsible for user aquisition by correlating the received signal with preambles from a given set. Here, to mimic a possible 5G situation, we equip the transmitter with the capability of sending information in “one shot” i.e., in addition to user aquisition, channel estimation is performed and the data is detected. For this a fraction of the PUSCH is reserved for data packets of users which are detected in the PRACH. Please note the rather challenging scanrio of only 839 subcarrier in the measurement window versus almost 24k data payload subcarriers.

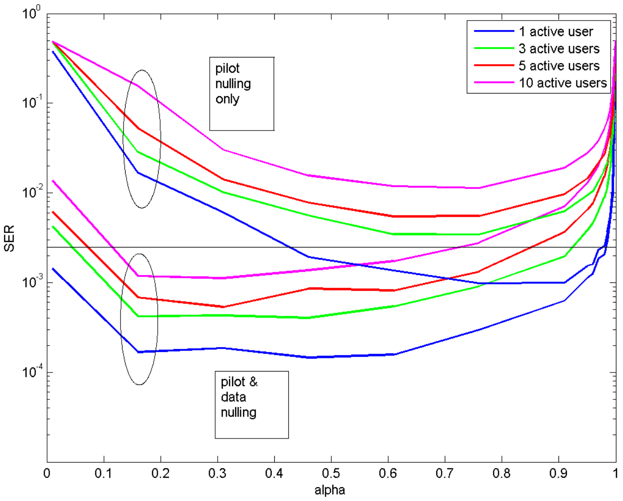

FIGURE 2. Averaged BPSK SER in 5G ”one-shot” random access in a 20MHz LTE-A standard setting at (overall) SNR=20dB. In the first setting (upper curves), $m = 839$ dimension out of $n = 24576$ dimensions are used for CS. This limits the control overhead to below 5%. In the second case, pilots and data are fully separated so that the performance is greatly improved at the expense of a slightly increased control overhead (<14%). Note that the curves do not cross for the same number of active users so the performace is always better with separation.

In our setting, a limited number of users is detected out of a maximum set (here 10 out of 50). We assume that the delay spread is below 300 dimensions of which only a set of 6 pathes are actually relevant. Each active user sends 1000 bits in some predefined frequency slot. This is uniquely achieved by mapping the sequences to a slot. Hence, in the classical Shannon setting 50 users × (300 pathes + 1000 bits) = 65000 dimensions are needed while there are only 24k available! The performance results are depicted in Fig.2 where we show symbol error rates (SER) over the pilot-to-data power ratio α. Moreover, in Fig.3 we depict false detection probability P_{FD} (some user is detected while not active) over missed detection probability P_{MD} (user is active while not detected). We observe that, although the algorithms might not yet capture the full potential of this idea, reasonable detection performance can be achieved by varying α. In the 4G LTE-A standard a minimum $P_{FD} = 10^{-3}$ is required for any number of receive antennas, for all frame structures and for any channel bandwidth. For certain SNRs a minimum $P_{MD} = 10^{-2}$ is required. It can be observed from the simulations that the requirements can be achieved. Actually, compared to 4G LTE-A where the control signalling can be up to 2000% [1] of a single resource element the control overhead is in the CS setting down to to 5% (let alone the huge increase in latency) in the best case!

A great challenge is the link between CS estimation and information-theoretic rates. In [25] we have recently shown that the rate error per subcarrier i is lower bounded by

$$\Delta r_i\,(\alpha) \le \log\left(1+\frac{m\cdot c_2(\delta_{2k_1})^2}{n}\left(\mathrm{SNR}\cdot\frac{(1-\alpha)\,\beta}{\alpha}+\frac{1}{\alpha}\right)\right) \tag{6}$$

[2] Of course this is a simple preliminary example as no 5G setting is available yet.

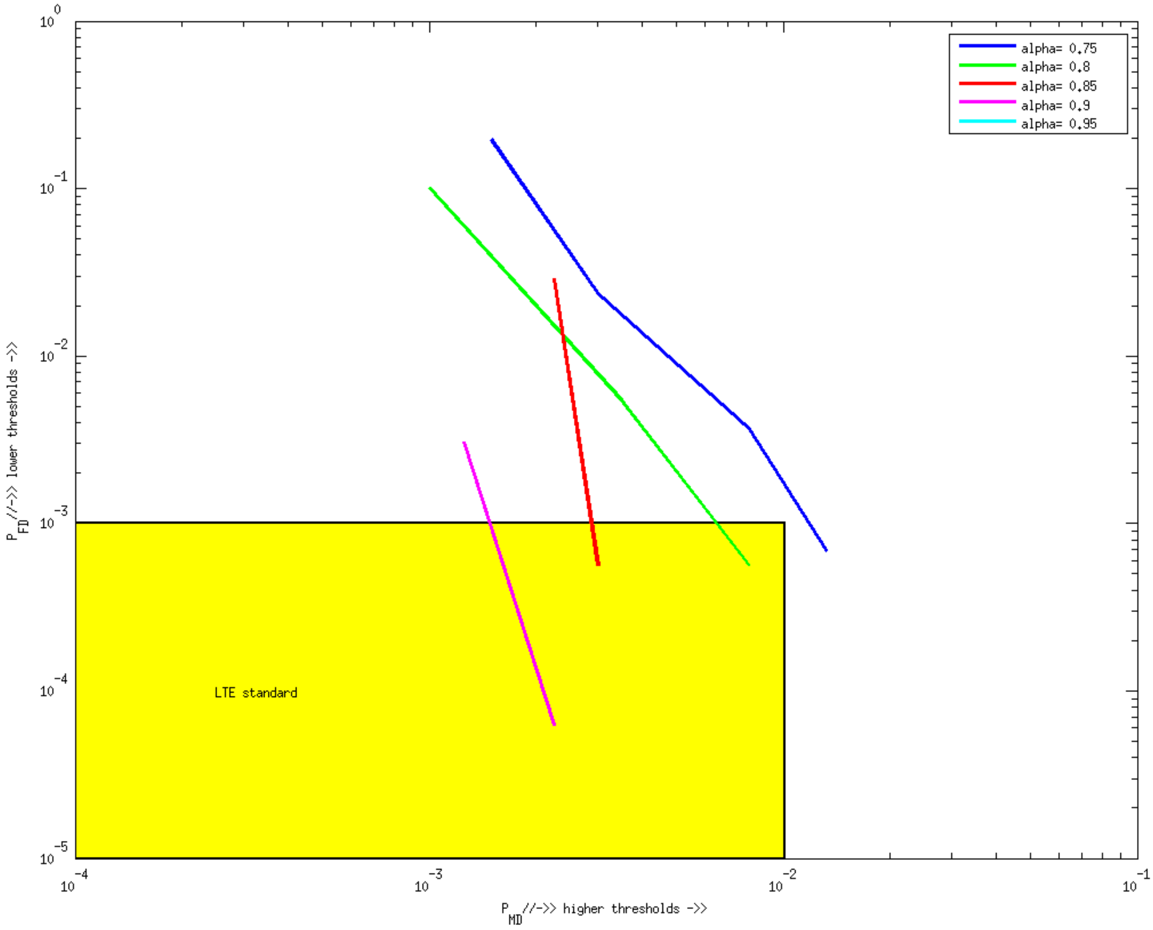

FIGURE 3. $P_{FD} = 10^{-2}$ over $P_{MD} = 10^{-2}$ for the 5G "one-shot" random access with same parameters for the first setting in Fig.2. It is observed that the LTE-A detection (in yellow) can be achieved for certain α. For the second setting in Fig.2 almost no detection errors are observed at overall SNR=20dB.

mimicking the curves' shape in the simulations. Here, the system parameter $c_2(\delta_{2k_1}) = 4\sqrt{1+\delta_{2k_1}}/(1-(1+\sqrt{2})\delta_{2k_1})$ depends only on the *restricted isometry property* δ_{2k_1} of the measurement matrix for k_1 sparse channels which is "small" with high probability provided that $m \geq \mathcal{O}\left(k_1 \log^5(n)\right)$ and $\beta > 0$ is a constant not depending on m, n.

Finally, we mention that we have used a simple FDMA random access strategy for the data which was recently discussed in [26] along with another more sophisticated slotted ALOHA access scheme employing a multiuser detector. Yet, no sparsity is involved. Similarly in [27] a random access scheme called *coded slotted ALOHA* is proposed investigating the interaction of advanced multiuser detection based on CS and successive interference cancellation. In both works the effect of channel/data estimation errors and error propagation in the interference cancellation scheme is crucial [27] and must be carefully considered as it is done in eq. (6).

C. INCOHERENT MULTIANTENNA RECEIVER CONCEPTS

The situation is even more intricate if both the sequence set and the channel are unknown during sampling. For the purpose of exposition we consider the single user/single antenna case and use the short–hand notation $*$ for circular convolution. In such a blind reference model the signals to be sampled are given, for example, as $h * x$ of two unknown but sparse vectors x and h with $\|h\|_0 \leq k_1$ and $\|x\|_0 \leq k_2$. The goals are (i) sample $h * x$ with minimal number of measurements and (ii) determine from $h * x$ the pair (x, h) up to indissoluble ambiguities. Compressive sampling via $\Phi \in \mathbb{C}^{m \times n}$ of such type of bilinear combinations in the strongly undersampled regime (the number of measurements m scales with $k_1 + k_2$ although $\|h * x\|_0 \leq k_1 k_2$) can be achieved under additional stability assumptions [28] and references therein. This bilinear recovery problem can be lifted again to an LMR problem, since circular convolution $h * x = B\left(\text{vec}(hx^T)\right)$ can be expressed as linear mapping B applied on the (vectorized) rank–one matrix $X = hx^T$. More precisely, B is a suitable matrix with elements $(B)_{i,(jk)} = \delta_{i,j \oplus k}$. A convex recovery algorithm not taking sparsity into account follows then from (3). Such a nuclear norm minimization has been considered by some authors for linearly random encoded data x, see for example [29], [46] and further reference therein (also related independent component analysis). Indeed, there it has been shown that this approach is successful with overwhelming probability when certain oversampling is used. It has to be expected that this situation improves if sparsity will be taken into account and even undersampling can then be used. For example, nuclear norm minimization could be ℓ_1–penalized:

$$\min \|X\|_* + \lambda \|\text{vec}(X)\|_1 \quad \text{s.t. } \|y - \Phi B(X)\|_2 \leq \epsilon \tag{7}$$

for some regularization parameter $\lambda > 0$. Yet, such multi-objective convex methods will not scale better then the best separate optimization [15]. Thus, it is of fundamental importance to find recovery algorithms which (i) can operate at the optimal, additive $(k_1 + k_2)$ scaling of measurements and (ii) can be easily extended to the case with multiple interferers.

D. MASSIVE MIMO REGIME

The system in eq. (5) in the non-coherent setting culminates in a many to massive antenna design when the number of receive antennas is scaled up. The standard setting for the massive MIMO case is the flat fading case.

The standard setting for the massive MIMO case is the flat-fading case where all h_q^p contain a non-zero element at their first position only so that we have $H_p = (h_p^1, \ldots, h_p^q) = h_p^T$. The model can be written as

$$Y = \sum_{p=1}^{N_t} s_p h_p^T + Z \tag{8}$$

It can then be shown that an SNR optimal detector is given by a *singular value decomposition* of Y when sparsity is not involved. Very recently, [30] has linked this to the so-called *pilot contamination problem* in a multi-cell scenario using random matrix theory showing that eigenvalue sets related to pilot signals in different cells appear in fact in disjoint intervals so that they can be separated. Hence, by this non-linear detector the pilot contamination problem disappears, however, at the cost of high receiver complexity involving "big" matrix decomposition, of course. Notably, the general case involving sparsity and multipath fading is an important open problem and can be possibly tackled with the incoherent concepts outlined before.

V. CLOUD RADIO ACCESS NETWORKS

In our target architecture, *virtual base stations* located in the data centers control a scalable number of nodes (and terminals) for which transmission/reception is coordinated by sharing control information and/or even messages (so-called *cooperative designs* or *coordinated multipoint (CoMP)*). While it is appealing to exploit existing designs we describe the short-comings of such approach in this section, and how sparse signal processing concepts will enable such designs.

A. STATE-OF-THE-ART COOPERATIVE DESIGNS: A CRITICAL VIEW

A summary of state-of-the-art cooperative designs in multi-cell networks can be found in [31]. However, existing cooperative designs, as they are available at present, are not scalable, i.e., operational regions and switching points depending on the available channel state information (CSI) as well as suitable transitions between different technologies are not clearly defined yet. A prominent coordinated beamforming technique is e.g. interference alignment (IA) [32] which essentially aligns the signal space so that multiple interferer appear in the same subspace. In some scenarios, closed-form solutions are available while in others only iterative solutions exists; rigorous convergence analysis or relevant stopping criteria are still missing. In this context, a scalable control signaling architecture in uplink (feedback) and downlink (feedforward) is a major requirement. In the existing designs, each point-to-point link is treated separately using orthogonal resources. Consequently, with densification of cells it becomes virtually impossible to provide CSI to all coordinated nodes (or antennas). For this small cell scenario, it is an intriguing idea to superpose and compress the control signals suitably, and let the (virtual) base stations recover their own channels by exploiting sparsity of channel profiles and effectively coordinated nodes.

There is another major point which, we think, basically stems from the lack of robustness in existing designs: Industrial field trials show rather disappointing throughput gains of CoMP algorithms [33] far away from the beforehand highlighted information-theoretic limits [31], whereby the major limiting factor is again properly sharing CSI and other overhead among cells. This so-called *limited feedback problem* has been greatly analyzed in [34] (for multiuser MIMO) and recently in [35] (for joint transmission) and [36] (for IA) in terms of the rate distance Δr_i of node i to capacity subject to some offset independent of SNR. Hence, these results essentially provide a systems' degrees-of-freedom analysis, i.e., assuming *infinite SNR regime*. To be specific, let p and $b = b(p)$ denote the SNR and feedback budget (in bits/channel use) as a function of SNR, respectively, then the per-node capacity degradation for any scaling in p is (using order notation $\mathcal{O}(\cdot)$):

$$\Delta r_i(p, b) = \log(1 + p \cdot 2^{-\frac{b}{n_t - 1}}) + \mathcal{O}(1) \tag{9}$$

Very recent studies [37] indicate that these results are fragile and that, in fact, the tradeoffs actually behave very different in more practical regimes. It is shown that for *any finite SNR point* p and for any scaling in b the per-node capacity degradation is:

$$\Delta r_i(p, b) = \mathcal{O}\left(\log(1 + p \cdot 2^{-\frac{b}{2(n_t - 1)}})\right) \tag{10}$$

which actually *doubles the required number of bits* compared to (9). Classical analysis falls short due to several reasons: i) It assumes infinite SNR regime where achieving DoF is optimal. In this operational regime interference mitigation instead of signal enhancement is the primary goal. ii) It asserts that the transmitter can optimally allocate rates while, in practice, the transmitter allocates rates according to the available CSI and corresponding scheduler decisions (real versus ideal link adaption). iii) The optimal scheduling decision is known a priori which is unrealistic since limited feedback not only affects the choice of spatial precoding but also user selection and resource allocation.

Obviously, the problem even worsens if a frequency-selective channel is considered and also persists with alternative time domain quantization. Altogether, the classical analysis renders the performance estimation overly optimistic and it is safe to say that the relevant tradeoffs in a dense C-RAN architecture are not yet well understood. This calls again for a highly efficient control signaling architecture which then considers robustness from scratch as outlined next.

B. TOWARDS A SPARSE ARCHITECTURE: ANALOG RELAY AND LINEARLY COMPRESS

Let us introduce a control signaling architecture where the feedback generating terminals receive the pilot signals from several nodes at different receive antennas at once and act as simple *relay and linearly compress* nodes. Linear compressing means that instead of the complete set of pilot measurements across time and spatial domain only a linear combination of them is (separately) fed back (possibly in analog form). All cooperating nodes collect these compressed control signals from the terminals and makes suitable estimates of the channels.

There are several key features of this scheme: The pilot patterns from the set of cooperating nodes are superposed and can be separated by advanced processing exploiting CIR compressibility. The overhead does not scale with the number of nodes joining, so that every node can join the cooperating set if it wishes. Each cooperating node receives the same signal reflecting the C-RAN architecture where the baseband processing of different nodes is in the same place. Inherently by the linear compressing, only the most relevant subset of nodes are effectively coordinated. Since this step is independent of the used pilot pattern the base station can change it without informing the terminals (so that it can be even random). Altogether, the base station takes over all the processing which is affordable in a centralized C-RAN architecture. From the terminals' point of view, neither do they need to know which nodes are cooperating nor do they need to quantize any information of the channels. By contrast, in simple time domain quantization, it is still assumed that the paths of each antenna of each node is treated the same way which becomes quickly inefficient in a highly dense C-RAN architecture where only the most significant paths across all base stations should be compressed. Moreover, the required communication resources should scale along this number rather than all antennas and nodes. Finally, the terminals could even make use of the same processing principles (only few might be active). Note that our design is efficient and scalable such that it meets the number of essentially unknown parameters (but *not* the sampling theorem) which is typically much smaller.

Mathematically, the scheme is expressed similar to the multiple access scheme in eq. (5). Denoting the number of nodes again by n_t, the received signal at each of the cooperating nodes y is:

$$y = \Phi\left(\sum_{p=1}^{n_t} S_p H_p + Z_1\right) + z_2 \tag{11}$$

Here, the inner bracket is what the terminal receives on multiple antennas similar to eq. (5). The main difference is the additional relay component resulting in two different noise sources Z_1, z_2 which complicates the analysis and performance limits are not known. Another significant challenge is to make such sparse designs robust within the C-RAN. In [37] and [38], we sketch a robust design at finite SNR guided by the intuition to improve the overall performance by suitable metrics (which capture the effects of e.g. scheduling) and transmit/receive strategies (guaranteeing worst-case performance no matter what the scheduling decision is) rather than simply approximating the channel. Combining such robust design with the sparse signal processing paradigm essentially requires to fully understand the tradeoffs between number of measurements and the actual regarded (new) performance metrics. This is actually far away from current achievements in the respective literature. A good starting point is the rate expression in eq. (6) though with suitable estimates for the RIP parameters.

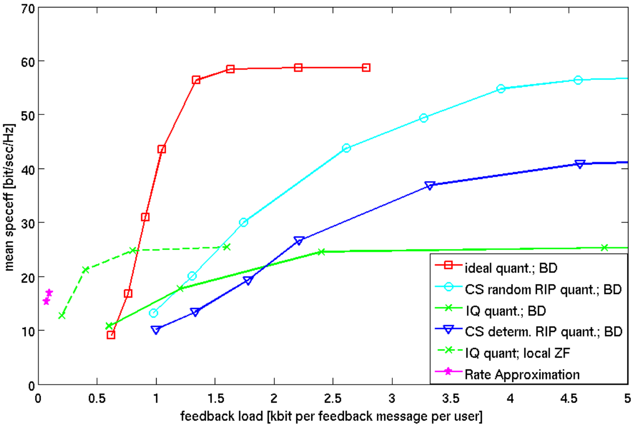

FIGURE 4. Mean spectral efficiency of CoMP schemes over feedback load using IQ quantization or CS based techniques: BD is block diagonalization, ZF is zero forcing and Rate Approximation is the scheme from [38]. Note that the number of users is small but, in principle, only limited by the processing capabilities of the access point.

Let us consider again a simple example with LTE-A 4G parameters in Fig. 4. Here, we compare the analog relay and linearly compress scheme with standard IQ quantization, a genie scheme and the robust schemes described in [38] in a LTE-A 4G setting. We consider 3 base stations with $n_t = 4$ transmit antennas located in 3 adjacent cells with 10 users ($n_r = 1$) uniformly distributed over the network area (radius of 250 meter around the center of the base stations). The physical layer is configured according to 4G. The channels are modeled by the spatial channel model extended (SCME) [15] using the urban macro scenario. It can be clearly observed that the CS provides a much better scaling compared to simple

IQ quantization. Note, that we have not incorporated any sparse topology yet.

C. SPARSE PREDICTION: UNEXPLORED GROUND

CSI aging and corresponding CSI mismatch due to asynchronous signaling is another serious problem in a CRAN. In [34] it was shown that classical prediction is beneficial for zeroforcing beamforming (ZF) in cooperative designs such that the systems are actually not fundamentally interference limited. In case of just stale CSI *retrospective interference alignment* [39] is another technique to be explored within the developed sparse and robust context. Typically it is argued that retrospective IA interference alignment is actually outperformed by ZF over a wide range of velocities [40]. However, under the robustness paradigm this might not be necessarily true as we proved that the performance estimation for ZF is overly optimistic!

So far, the inherent sparsity is not considered in the prediction literature. Clearly, sparsity provides additional structural information which can be explored. The problems are intricate: the system (5) might evolve on a complicated manifold not in a vector space. The simplest approach is then to assume for the time evolution the same support of the path coefficients for all times, which is reminiscent of block- or structured CS, and to iteratively estimate the correct subspace and make a optimal prediction on the estimated support next. Another approach is to let the process truly vary on a sparse manifold where the simplest one is the union of all k-sparse canonical subspaces in $\mathbb{C}^n$. However, so far, neither specific models to enable e.g. advanced Kalman prediction nor any respective algorithms have been developed.

The performance of classical short-time prediction algorithms suffer from the poor granularity in the frequency domain and coarsely quantized complex CSI due to limited feedback problem as well. Another key idea is, hence, to develop a new framework based on sparsity where the sparse CIR is calculated from simple real-valued measurements, cf. Sec.VI for a strong motivation of this assumption. Even though such an approach may not be supported by a physical meaning in all cases, the new process is in any case “slower” and therefore intuitively better predictable. Phase retrieval algorithms can be used for such problems which we outlined in Sec. III. A special case are the Fourier measurements where the problem is related to factorization theory. Alternative methods using so-called *symmetrized Fourier measurements* have been recently proposed in [28] which, subject to the sign, theoretically allow stable reconstruction of the complex-valued CIR. Yet again, suitable algorithms incorporating multiple antennas and nodes exploring the sparsity of the CIRs are not known.

VI. COMPRESSIVE CHANNEL-SOURCE-NETWORK CODING

In this section we carry over the concepts to the terminal side following the idea that sparsity may be already beneficial in the terminals' encoding process.

A. BEYOND THE SEPARATION THEOREM

5G networks, in particular the IoT component, will contain a large number of relatively simple devices that need to operate at a very energy-efficient level and with fairly limited memory. Hence these device can neither afford to employ energy-hungry encoding or compression algorithms, nor can they support high data rates. Thus power budget and communication rates become two main design constraints for 5G networks. These constraints naturally suggest the use of sensors that are built based on the compressive sensing paradigm. Using compressive sensing devices (CSDs), one can avoid the need for power-consuming compression algorithms, while still being able to transmit images of sufficient quality without having to resort to transmitting images pixel by pixel. Instead of recording data at a high resolution (which requires memory and energy) and then throwing away most of the data during the compression step (compression again requires energy), such CSDs collect data directly in the “compressive domain”. The computational burden is shifted from the transmitter to the receiver. Yet, deploying CSDs in a wireless network will impact the entire communication system design. Understanding this impact and taking full advantage of it will be an important aspect for efficient 5G network modeling.

A variety of interesting challenges arise in such “compressive sensor” situations. Unlike in the “classical” communication scenario, where Shannon's separation theorem is one of the guiding principles for designing communication systems (even though the idealized conditions on which the separation theorem is based are hardly met in practice), we are now faced with a different setup. The situation depends strongly on which constraints we put on the complexity of the encoding step. For example, if we equip our simple CSD with a fairly complex encoder, then we find ourselves back in a classical communication scenario. This is due to the fact that in this case we could incorporate both the reconstruction of the original signal from its compressive samples and a standard source coding step into the encoding procedure. But it is clear that such a complicated encoder would be detrimental to our goal of having simple, power-efficient sensing devices. Instead let us look at the other extreme, where we deal with a sensor that does not permit any encoding at all. The only way to add redundancy (and thus to add error protection) in that case is to increase the number of compressive measurements.

Following the CS paradigm such compressive measurements have to be slightly redundant to allow for numerically efficient signal recovery (compared to the number of measurements for an NP-hard reconstruction algorithm). For ideal CS matrices the number of measurements m for a k-sparse signal $x \in \mathbb{C}^n$ has to be at least $\mathcal{O}(k\log(n/k))$. Hence, to increase robustness, we may take $m > \mathcal{O}(k\log(n/k))$. Clearly, m should depend on the SNR. A simple CS-based source coding scheme might look as follows: An instructive numerical simulation illustrates the efficacy of this simple scheme. Similar to [41] we consider the following setup:

(i) *Encoding:* The encoding step consists of taking $m \leq n$ measurements of a signal $x \in C^n$ via $w = \Phi x$, where we assume that x is k-sparse in a known basis Ψ.
(ii) *Transmission:* The measured signal w is transmitted through a noisy channel. The received signal is $y = w + z = \Phi x + z$, where z is AWGN.
(iii) *Decoding:* At the receiver we attempt to recover x via the LASSO, i.e., by solving min $\frac{1}{2}\|\Psi x\|_1 + \lambda\|\Phi x - y\|_2$.

Simulation examples can be found in [41].

B. SPATIALLY CORRELATED NETWORKS

Let us now consider a sensor network where the source signals are sparse with respect to some known basis and in addition are *spatially correlated*. Instead of decoding each signal individually, it is natural to try to exploit the spatial correlation between the sensors. There are various ways to accomplish this. In [41] an elementary sequential reconstruction algorithm is proposed to exploit the spatial correlation between signals, based on the fact that the difference between two sparse signals is again sparse. A different, more flexible and robust approach consists of setting up a low-rank optimization problem. By interpreting the signals as columns of a data matrix X, the spatial correlation suggests that one may attempt to find, among all the matrices consistent with the observed data, that with minimal rank. We can try to improve upon this decoding procedure even further by attempting to combine the powers of compressive sensing with those of matrix completion. Hence we may consider an optimization problem that combines the sparsity and the low-rank structures, as e.g. in (7). While [15] shows (in the noiseless case) that such mixed-convex optimization approaches cannot lead to significant improvements over an approach that optimally exploits one of the individual sparse properties, numerical simulations nevertheless indicate that the improvements can still be worthwhile from a practical viewpoint - in particular in the noisy case. Recovery problems that combine sparsity and low rank as penalty function have been proposed recently in the literature, but their careful use and investigation in wireless communications is unchartered territory.

So far we have not made use of the fact that, unlike in the usual compressive sensing scenarios, the coefficients of the signals we try to reconstruct belong to some finite alphabet, such as e.g. QPSK. Hence, an important open problem is to extend standard decision feedback equalization schemes to the situations described in this section. How can we optimally exploit the sparsity of the signals and the spatial correlation across sensors in a decision feedback scheme? A possible approach might be to try to combine recent advances in mixed integer optimization [42] with sparse recovery methods.

Many variations of the theme are possible. We have ignored the role and effect of quantization so far. Moreover, the sensors may allow for *very simple* encoding in addition to, or instead of, increasing the number of compressive measurements. How does an analog of Shannon's separation theorem look like for such a scenario?

VII. EMBEDDED SECURITY

Embedded security together with sparse signal processing promoting simultaneously secrecy and reliability can fundamentally change the way we approach the design of security, authentication and integrity mechanisms, specifically in the IoT and the TI. In this section, we show that sparse signal processing can be naturally incorporated within the concept of embedded security and exhibits indeed a new degree of freedom in the design of algorithms, naturally entailing new interesting tradeoffs considering compressibility and secrecy.

A. 5G SECURITY CHALLENGES

One example where current security solutions fall far short is the IoT due to the *scalability problem*: Nodes must be flexibly managed and distributed in the network and asymmetric schemes used at the application layer are too complex as well as too computational- and energy intensive for the typical battery driven low-complex wireless transceivers of IoT devices. Symmetric key schemes are 100 to 1000 times less complex ("lightweight security") but assume a common secret key for the nodes so that that there is a *key distribution problem* instead. A detour of this problem is to distribute unique keys already in the chip sets' manufacturing process [43]. However, the overall security architecture still requires Internet server access and lacks some flexibility.

An alternative concept is embedded security aiming at physical layer integration of confidential services in wireless networks [8]. Here, *classical wiretap coding* can achieve provably non-zero secrecy rates within the imposed (often idealistic) channel model, where it is then impossible for an eavesdropper to extract any information about the sent message from the overheard signal. The concept of wiretap coding is somewhat detrimental to the scalability and low complexity requirements in the IoT similar to the argument in Sec. VI. A much simpler method is to exploit the reciprocity and fading nature of CSI and to establish a common secret between sender and transmitter from the CSI measurements [44]. Since keys are then automatically installed, key distribution in these systems is easy to manage and requires no user interaction; moreover beyond the spatial decorrelation length of antenna elements such keys are virtually impossible to recover. Two major practical problems occur though: i) imperfect reciprocity of CSI ii) insufficient entropy of the generated keys due to static channels. In order to handle the imperfect reciprocity, typically a so-called information reconciliation procedure over a public channel is run, careful not to unveil any information about the secret key bits. Imperfect reciprocity and insufficient entropy depend on each other because "coarse" quantization and stronger codes improve on key agreement rate but reduce the keys' entropy. Both aspect affords additional control signaling, yet again bringing up the issue of reliability, complexity, nodes' lifetime, and new security threats in the IoT.

Another example is the ultra-low latency requirement in the TI. Each and every element of the communication and control chain must be optimized and, obviously, fast authentication and secure communication is a "must" for the TI then. We also emphasize the role of data integrity due to the high reliability constraints, e.g. in the context of Industrial Wireless. Applying standard security mechanisms on the application layer is not feasible. Moreover, wireless channel secret key generation is highly limited in the rate of the generated key bits, which is at most 44bit/s by today so that at least roughly 3s are required to generate e.g. a secure 128bit key [44]. Too slow for the TI!

To overcome such limitations, security shall be built in the compressive CSI control signaling architectures developed in Sec. IV and Sec. V from scratch. Here, in contrast to the robust design by taking sufficently many compressive measurements, the opposed direction is taken to disguise the CSI, which shall be discussed next.

B. MAKING SECURITY FAST AND SCALABLE

In our concept, a secret key is periodically generated from CSI and acknowledged between transmitter and legitimate receiver. Let us assume that information reconciliation is part of the "relay and compress" control signaling architecture in Sec. V. Then, since the legitimate receiver can reliably recover the channel, reciprocity and key entropy is preserved which advantageous in case of static channels. In addition, similar to the discussion in Sec. V the transmitter can change the control pattern without informing the receiver. Furthermore, since this is a full duplex scheme the scheme is faster making it a candidate for the TI (standard wireless channel secret key generation schemes typically run in half duplex time division scheme). The catch is though that potentially the eavesdropper can recover the key itself when he/she is able to collect all the control information.

In order to make such schemes workable in practice, additional measures should be taken to ensure that an eavesdropper cannot reveal the message. Actually, this sets a limit to the number of measurements publicly discussed such that slightly erroneous or incomplete information about the control signaling patterns as well as about the compressive measurements make it impossible to extract the original messages. Clearly, one can think of an ocean of possible communication protocols to improve on this line of thinking. Interestingly, by using the wireless channel as the secret key source we have a new interesting trade off between compressibility and secrecy: good compressibility means small entropy in the key, hence longer observation times and vice versa.

To illustrate an example of such "built-in" security we consider the following scenario. Let us assume that the relay and compress scheme from Sec. V is used in a point-to-point link to inform the transmitter (Alice) about the channel of its legitimate receiver (Bob). The purpose is to extract a common key from it in order to conceal the information from an eavesdropper (Eve). Clearly, if Eve knows all the control signalling it can recover the information from the measurements. In practice, by the physical nature of wireless transmission, these measurements are only available subject to some unknown phase shift for each measurement. The simulation in Fig. 5 then clearly indicates that Eve is not able to recover the information while Bob can still get some reasonable performance. The MSE performance is even worse for rank-one distortion. We would like to emphasize that any analytical approach for the such "perturbations" is not known to the best of our knowledge.

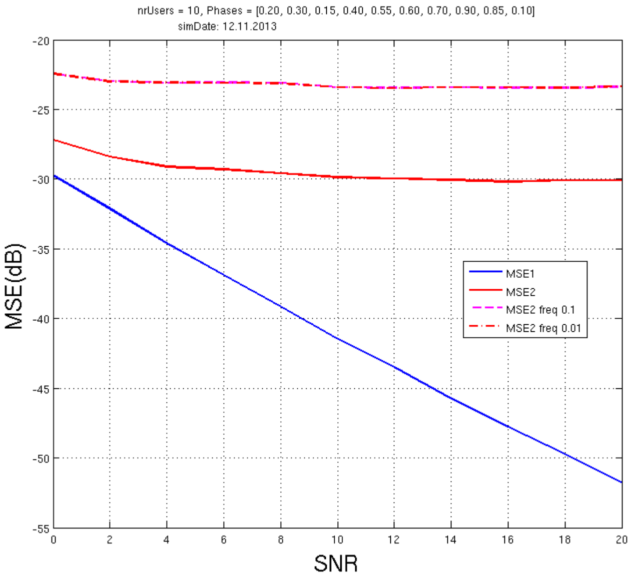

FIGURE 5. MSE performance of the relay and compress scheme under "perturbations" for 1) Bob having ideal measurements and 2) Eve having perturbed measurements with a) phase errors and, worse, b) rank-one distortion.

Allowing fast, efficient, and flexible key distribution are desirable principles in the TI as well. However, to achieve fast authentication even more advanced methods must be used ranging from wireless fingerprinting [45] which can include e.g. the individual sparsity patterns as well as cooperative jamming approaches [44].

VIII. CONCLUSION

We have shown that sparse signal processing is a viable source for an innovative 5G system. To exploit the benefits fundamental research is required addressing the many open questions regarding tradeoffs, performance limits, algorithmic framework etc. We have only touched the surface of this research agenda, where one of the many further fields of exploration is to include new waveforms at the physical layer. We also emphasize that sparsity appears not only as a physical reality, e.g. in the wireless channels, but also by design of the network topology, traffic conditions etc. It is therefore an important future task to measure the degree of sparsity in a system and adapt the signaling architecture acccordingly. We also remark that the simple simulation examples have not been limited by fundamental performance bounds but by the processing capability of the simulation environment. Consequently, lowering the complexity of the CS algorithms is another important furture task.

REFERENCES

[1] G. Wunder *et al.*, "5GNOW: Non-orthogonal, asynchronous waveforms for future mobile applications," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 97–105, Feb. 2014.

[2] FP7 EU Project. (2015). *5th Generation Non-Orthogonal Waveforms*. [Online]. Available: http://www.5gnow.eu

[3] FP7 EU Project. (2015). *METIS2020—Laying the Foundations of 5G*. [Online]. Available: http://www.metis2020.com

[4] G. Wunder, R. F. H. Fischer, H. Boche, S. Litsyn, and J.-S. No, "The PAPR problem in OFDM transmission: New directions for a long-lasting problem," *IEEE Signal Process. Mag.*, vol. 30, no. 6, pp. 130–144, Nov. 2013.

[5] F. Boccardi, R. W. Heath, Jr., A. Lonzano, T. L. Marzetta, and P. Popovski, "Five disruptive technology directions for 5G," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 74–80, Feb. 2014.

[6] H. Zhu and G. B. Giannakis, "Exploiting sparse user activity in multiuser detection," *IEEE Trans. Commun.*, vol. 59, no. 2, pp. 454–465, Feb. 2011.

[7] W. U. Bajwa, J. Haupt, A. M. Sayeed, and R. Nowak, "Compressed channel sensing: A new approach to estimating sparse multipath channels," *Proc. IEEE*, vol. 98, no. 6, pp. 1058–1076, Jun. 2010.

[8] R. F. Schaefer and H. Boche, "Physical layer service integration in wireless networks: Signal processing challenges," *IEEE Signal Process. Mag.*, vol. 31, no. 3, pp. 147–156, May 2014.

[9] S. Foucart and H. Rauhut, *A Mathematical Introduction to Compressive Sensing* (Applied and Numerical Harmonic Analysis). Basel, Switzerland: Birkhäuser, 2013.

[10] M. B. McCoy and J. A. Tropp. (2013). "The achievable performance of convex demixing." [Online]. Available: http://arxiv.org/abs/1309.7478

[11] C. Hegde and R. G. Baraniuk. (2012). "Signal recovery on incoherent manifolds." [Online]. Available: http://arxiv.org/abs/1202.1595

[12] V. Chandrasekaran, B. Recht, P. A. Parrilo, and A. S. Willsky, "The convex geometry of linear inverse problems," *Found. Comput. Math.*, vol. 12, no. 6, pp. 805–849, 2012.

[13] E. J. Candès and B. Recht, "Exact matrix completion via convex optimization," *Found. Comput. Math.*, vol. 9, no. 6, pp. 717–772, 2009.

[14] D. Gross, "Recovering low-rank matrices from few coefficients in any basis," *IEEE Trans. Inf. Theory*, vol. 57, no. 3, pp. 1548–1566, Mar. 2011.

[15] S. Oymak, A. Jalali, M. Fazel, Y. C. Eldar, and B. Hassibi. (2012). "Simultaneously structured models with application to sparse and low-rank matrices." [Online]. Available: http://arxiv.org/abs/1212.3753

[16] E. J. Candès, T. Strohmer, and V. Voroninski. (2011). "PhaseLift: Exact and stable signal recovery from magnitude measurements via convex programming." [Online]. Available: http://arxiv.org/abs/1109.4499

[17] M. Ehler, M. Fornasier, and J. Sigl. (2013). "Quasi-linear compressed sensing." [Online]. Available: http://arxiv.org/abs/1311.1642

[18] A. K. Fletcher, S. Rangan, and V. K. Goyal. (2009). "On–off random access channels: A compressed sensing framework." [Online]. Available: http://arxiv.org/abs/0903.1022

[19] C. Bockelmann, H. F. Schepker, and A. Dekorsy, "Compressive sensing based multi-user detection for machine-to-machine communication," *Trans. Emerg. Telecommun. Technol.*, vol. 24, no. 4, pp. 389–400, 2013.

[20] G. Taubock, F. Hlawatsch, D. Eiwen, and H. Rauhut, "Compressive estimation of doubly selective channels in multicarrier systems: Leakage effects and sparsity-enhancing processing," *IEEE Trans. Sel. Topics Signal Process.*, vol. 4, no. 2, pp. 255–271, Apr. 2010.

[21] P. Cheng *et al.*, "Channel estimation for OFDM systems over doubly selective channels: A distributed compressive sensing based approach," *IEEE Trans. Commun.*, vol. 61, no. 10, pp. 4173–4185, Oct. 2013.

[22] A. Adhikary, J. Nam, J.-Y. Ahn, and G. Caire, "Joint spatial division and multiplexing—The large-scale array regime," *IEEE Trans. Inf. Theory*, vol. 59, no. 10, pp. 6441–6463, Oct. 2013.

[23] S. Stanczak, G. Wunder, and H. Boche, "On pilot-based multipath channel estimation for uplink CDMA systems: An overloaded case," *IEEE Trans. Signal Process.*, vol. 54, no. 2, pp. 512–519, Feb. 2006.

[24] J. A. Tropp, J. N. Laska, M. F. Duarte, J. K. Romberg, and R. G. Baraniuk, "Beyond Nyquist: Efficient sampling of sparse bandlimited signals," *IEEE Trans. Inf. Theory*, vol. 56, no. 1, pp. 520–544, Jan. 2010.

[25] G. Wunder, P. Jung, and C. Wang, "Compressive random access for post-LTE systems," in *Proc. IEEE Int. Conf. Commun. Workshop (ICC)*, Sydney, NSW, Australia, Jun. 2014, pp. 539–544.

[26] H. S. Dhillon, H. C. Huang, H. Viswanathan, and R. A. Valenzuela. (2014). "Fundamentals of throughput maximization with random arrivals for M2M communications," [Online]. Available: http://arxiv.org/abs/1307.0585

[27] Y. Ji, C. Stefanovic, C. Bockelmann, A. Dekorsy, and P. Popovski. (2014). "Characterization of coded random access with compressive sensing based multi-user detection." [Online]. Available: http://arxiv.org/abs/1404.2119

[28] P. Walk and P. Jung. (2013). "A stability result for sparse convolutions." [Online]. Available: http://arxiv.org/abs/1312.2222

[29] A. Ahmed, B. Recht, and J. Romberg. (2012). "Blind deconvolution using convex programming." [Online]. Available: http://arxiv.org/abs/1211.5608

[30] L. Cottatellucci, R. R. Müller, and M. Vehkaperä, "Analysis of pilot decontamination based on power control," in *Proc. IEEE 77th Veh. Technol. Conf. (VTC Spring)*, Dresden, Germany, Jun. 2013, pp. 1–5.

[31] P. Marsch and G. P. Fettweis, Eds., *Coordinated Multi-Point in Mobile Communications: From Theory to Practice*. Cambridge, U.K.: Cambridge Univ. Press, 2011.

[32] S. A. Jafar, "Interference alignment: A new look at signal dimensions in a communication network," *Found. Trends Commun. Inf. Theory*, vol. 7, no. 1, pp. 1–134, 2011.

[33] R. Irmer *et al.*, "Coordinated multipoint: Concepts, performance, and field trial results," *IEEE Commun. Mag.*, vol. 49, no. 2, pp. 102–111, Feb. 2011.

[34] G. Caire, N. Jinal, M. Kobayashi, and N. Ravindran, "Multiuser MIMO achievable rates with downlink training and channel state feedback," *IEEE Trans. Inf. Theory*, vol. 56, no. 6, pp. 2845–2866, Jun. 2010.

[35] P. de Kerret and D. Gesbert. (2013). "The multiplexing gain of the network MIMO channel with distributed CSI." [Online]. Available: http://arXiv:1108.3742

[36] O. El Ayach and R. W. Heath, Jr., "Interference alignment with analog channel state feedback," *IEEE Trans. Wireless Commun.*, vol. 11, no. 2, pp. 626–636, Feb. 2012.

[37] J. Schreck, G. Wunder, and P. Jung, "Robust iterative interference alignment for cellular networks with limited feedback," *IEEE Trans. Wireless Commun.*, vol. 14, no. 2, pp. 882–894, Feb. 2015.

[38] G. Wunder, J. Schreck, and P. Jung, "Nearly doubling the throughput of multiuser MIMO systems using codebook tailored limited feedback protocol," *IEEE Trans. Wireless Commun.*, vol. 11, no. 11, pp. 3921–3931, Nov. 2012.

[39] M. A. Maddah-Ali and D. Tse, "Completely stale transmitter channel state information is still very useful," *IEEE Trans. Inf. Theory*, vol. 58, no. 7, pp. 4418–4431, Jul. 2012.

[40] M. Kobayashi and G. Caire, "On the net DoF comparison between ZF and MAT over time-varying MISO broadcast channels," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Boston, MA, USA, Jul. 2012, pp. 2286–2290.

[41] S. Feizi and M. Médard, "A power efficient sensing/communication scheme: Joint source-channel-network coding by using compressive sensing," in *Proc. 49th Annu. Allerton Conf. Commun., Control, Comput.*, Monticello, IL, USA, Sep. 2011, pp. 1048–1054.

[42] J. De Loera, R. Hemmecke, and M. Köppe, *Algebraic and Geometric Ideas in the Theory of Discrete Optimization* (MPS-SIAM Series on Optimization). Philadelphia, PA, USA: SIAM, 2012.

[43] Special Interest Group. (2015). *WEIGHTLESS—Global Standard for IoT/M2M Connectivity*. [Online]. Available: http://www.weightless.org

[44] S. Gollakota and D. Katabi, "Physical layer wireless security made fast and channel independent," in *Proc. IEEE INFOCOM*, Apr. 2011, pp. 1125–1133.

[45] L. Lai, S.-W. Ho, and H. V. Poor, "Privacy–security trade-offs in biometric security systems—Part I: Single use case," *IEEE Trans. Inf. Forensics Security*, vol. 6, no. 1, pp. 122–139, Mar. 2011.

[46] S. Ling and T. Strohmer. (2015). "Self-calibration and biconvex compressive sensing." [Online]. Available: http://arXiv:1501.06864

GERHARD WUNDER (M'05) received the Dipl.-Ing. (Hons.) degree in electrical engineering from the University of Hannover, Germany, in 1999, the Ph.D. (Dr.Ing.) (Hons.) (*summa cum laude*) degree in communication engineering with a focus on the peak-to-average power ratio (PAPR) problem in OFDM from Technical University Berlin (TUB), Germany, in 2003, and the Habilitation degree, in 2007. He became a Privatdozent with TUB in the field of detection/estimation theory, stochastic processes, and information theory. In 2000 and 2005, he was a Visiting Professor with the Georgia Institute of Technology, Atlanta, GA, USA (Prof. Jayant), and Stanford University, Palo Alto, CA, USA (Prof. Paulraj). In 2009, he was a Consultant with Alcatel—Lucent Bell Labs, (NJ, USA), Murray Hill (S. Stolyar) and Crawford Hill (R. Valenzuela). Since 2003 he is with the Fraunhofer Heinrich-Hertz-Institut. He is a Coordinator and Principal Investigator both in the FP7 Call 8 project 5GNOW by the European Commission and PROPHYLAXE, the largest IoT physical layer security project supported by the German Ministry of Education and Research. He is a recipient of research fellowships from the German National Research Foundation (DFG). He was the General Co-Chair of the 2009 International ITG Workshop on Smart Antennas (2009), and a lead Guest Editor of a special issue of the *Journal of Advances on Signal Processing* regarding the PAPR problem of the European Association for Signal Processing in 2011. Since 2011, he has been an Editor of the IEEE Transactions on Wireless Communications in the area of wireless communications theory and systems. He was also a recipient of the 2011 Award for outstanding scientific publication in the field of communication engineering by the German Communication Engineering Society.

HOLGER BOCHE (M'04–SM'07–F'11) received the Dipl.-Ing. and Dr.Ing. degrees in electrical engineering from the Technische Universität Dresden, Dresden, Germany, in 1990 and 1994, respectively, the degree in mathematics from the Technische Universität Dresden, in 1992, and the Dr.rer.nat. degree in pure mathematics from the Technische Universität Berlin, Berlin, Germany, in 1998. From 1994 to 1997, he did post-graduate studies in mathematics with the Friedrich-Schiller Universität Jena, Jena, Germany. In 1997, he joined the Fraunhofer-Institut für Nachrichtentechnik, Heinrich-Hertz-Institut (HHI), Berlin. In 2002, he was a Full Professor of Mobile Communication Networks with the Institute for Communications Systems, Technische Universität Berlin. In 2003, he became the Director of the Fraunhofer German-Sino Laboratory for Mobile Communications, Berlin, and the Director of HHI in 2004. Since 2010, he has been with the Institute of Theoretical Information Technology, and a Full Professor with the Technische Universität München (TUM), Munich, Germany. Since 2014, he has been a member and an Honorary Fellow of the TUM Institute for Advanced Study, Munich. He was a Visiting Professor with ETH Zurich, Zurich, Switzerland, from 2004 to 2006, and the Royal Institute of Technology, Stockholm, Sweden, in Summer 2005. He is a member of the IEEE Signal Processing Society's Signal Processing and Communications and Signal Processing Theory and Methods Technical Committees. He was elected as a member of the German Academy of Sciences Leopoldina, in 2008, and the Berlin Brandenburg Academy of Sciences and Humanities in 2009. He was a recipient of the Technische Kommunikation Research Award from the Alcatel SEL Foundation in 2003, the Innovation Award from the Vodafone Foundation in 2006, and the Gottfried Wilhelm Leibniz Prize from the Deutsche Forschungsgemeinschaft (German Research Foundation) in 2008. He was a co-recipient of the 2006 IEEE Signal Processing Society Best Paper Award, and a recipient of the 2007 IEEE Signal Processing Society Best Paper Award.

THOMAS STROHMER received the M.S. and Ph.D. degrees in mathematics from the University of Vienna, Austria, in 1991 and 1994, respectively. He was a Research Assistant with the Department of Mathematics, University of Vienna, from 1991 to 1997. He spent one year as an Erwin-Schroedinger Fellow with the Department of Statistics, Stanford University, Stanford, CA, and then joined the Department of Mathematics, University of California in Davis, in 1998, where he is currently a Full Professor. His general research interests are in harmonic analysis, numerical analysis, signal and image processing, information theory, high-dimensional data analysis, and wireless communications. His recent awards include the 2013 IEEE Signal Processing Society Best Paper Award and the 2014 SIAM Outstanding Paper Prize. He is on the Editorial Board of several journals. He also serves as a Consultant to the industry in the areas of telecommunications, bioengineering, and signal and image processing.

PETER JUNG (M'–) received the Dipl.-Phys. degree in high energy physics from Humboldt University, Berlin, Germany, in 2000, in cooperation with DESY, Hamburg, and the Dr.rer.nat. (Ph.D.) degree in Weyl–Heisenberg representations in communication theory from the Technical University of Berlin (TUB), Germany, in 2007. Since 2001, he has been with the Department of Broadband Mobile Communication Networks, Fraunhofer Institute for Telecommunications, Heinrich-Hertz-Institut, and the Fraunhofer German-Sino Laboratory for Mobile Communications since 2004. He is currently working under DFG grants JU 27951&2 with TUB in the field of signal processing and information and communication theory. His current research interests are in the area of compressed sensing, time–frequency analysis, dimension reduction, and randomized algorithms. He is giving lectures in compressed sensing and estimation theory. He is a member of VDE/ITG.