On the Fundamental Limits of Massive Connectivity

Wei Yu
Electrical and Computer Engineering Department
University of Toronto
weiyu@comm.utoronto.ca

***Abstract*—This paper aims to provide an information theoretical analysis of massive device connectivity scenario in which a large number of devices with sporadic traffic communicate in the uplink to a base-station (BS). In each coherence time interval, the BS needs to identify the active devices, to estimate their channels, and to decode the transmitted messages from the devices. This paper first derives an information theoretic upper bound on the overall transmission rate. We then provide a degree-of-freedom (DoF) analysis that illustrates the cost of device identification for massive connectivity. We show that the optimal number of active devices is strictly less than half of the coherence time slots, and the achievable DoF decreases linearly with the number of active devices when it exceeds the number of receive antennas. This paper further presents a two-phase practical framework in which device identification and channel estimation are performed jointly using compressed sensing techniques in the first phase, with data transmission taking place in the second phase. We outline the opportunities in utilizing compressed sensing results to analyze the performance of the overall framework and to optimize the system parameters.**

I. Introduction

Massive connectivity is envisioned to be a key requirement for future cellular networks, in which millions of devices are expected to be connected to the cellular network while performing machine-centric functions such as environment sensing, event detection, surveillance, and control. Differing from the human-centric traffic dominating the wireless cellular Internet of today, these future machine-type communications have the following key features:

- Large Number of Devices: The number of devices connected to each cellular base-station (BS) can potentially be in the order of $10^5 - 10^6$. Thus, significant network resources need to be devoted to identify and to keep track of these potentially active connections.
- Sporadic Traffic: Most devices engaged in machine-type communications do not transmit data constantly. Sensors may report observations periodically, or in response to emerging events. At any given time, only a small subset of potential devices have data to transmit.
- Low-Latency: Machine-type traffic is typically delay sensitive, especially for sensing and control applications. The required latency can be in the millisecond range.

It is apparent that the traffic pattern for mass connectivity is very different from the video dominated cellular traffic for which today's network is designed. Specifically, the current cellular system can only support a small number of devices. Further, scheduling overhead alone in today's network can already overwhelm the latency budget for machine communications.

The paper aims to provide a framework for communication protocol design for massive connectivity. We adopt a system design in which a massive number of connected devices can be connected in an expeditious fashion to the BS by taking advantage of the sparsity in the activity pattern for device communications and by utilizing compressed sensing techniques for active device detection. The main objective of the paper is to reveal the information theoretical limit of such a system by presenting a degree-of-freedom (DoF) analysis. This paper is primarily focused on the uplink, where the major challenge is that of uncoordinated user transmission and activity pattern. The downlink counterpart can be designed using more conventional techniques.

The notation used in the paper is as follows. Lower case letters, e.g, x, are used to denote scalars. Lower case bold-faced letters, e.g., $\mathbf{x}$, are used to denote vectors. Upper case bold-faced letters, e.g., $\mathbf{X}$, are used to denote matrices. Matrix transpose is denoted as $(\cdot)^\tau$ and conjugate transpose as $(\cdot)^\dagger$.

II. Device Communication Framework

Consider a cellular network designed to allow a large number of devices with sporadic traffic to communicate with the BS. We consider a single cell in which the BS is equipped with M antennas and the devices are equipped with a single antenna each. The channels between the BS and the devices are modeled as flat-fading channel with a distance dependent pathloss component, a shadowing component, and a fast-fading component, uncorrelated across the antennas. We further assume a block-fading model (which can be in either time or frequency), where the channel stays constant for a duration of T slots. (Note that if T is in the time domain, we can also interpret T as the latency constraint in delay limited communication.) In this case, the overall channel can be written as

$$\mathbf{y}(i) = \sum_{n=1}^{N} \mathbf{h}_n x_n(i) + \mathbf{z}(i), \qquad i = 1, \cdots, T, \tag{1}$$

where i is the time index, n is the device index, and N is the total number of potential devices in the pool, among which K are assumed to be active at any given time. Here, $\mathbf{y}(i) \in \mathbb{C}^M$ is the received signal across the M BS antennas in the ith time slot, $\mathbf{z}(i) \in \mathbb{C}^M$ is the background noise assumed to be additive white Gaussian, $\mathbf{h}_n \in \mathbb{C}^M$ is the vector channel from the nth device to the M receive antennas at the BS assumed

to be a constant over the coherence time of T time slots, and $x_n(i) \in \mathbb{C}$ is the transmit signal by the nth device in the ith time slot. Thus, the vector $\mathbf{x}(i) = [x_1(i) \cdots x_N(i)]^\tau \in \mathbb{C}^N$ is sparse.

In matrix form, we can define $\mathbf{Y} = [\mathbf{y}(1) \cdots \mathbf{y}(T)] \in \mathbb{C}^{M \times T}$, $\mathbf{H} = [\mathbf{h}_1 \cdots \mathbf{h}_N] \in \mathbb{C}^{M \times N}$, and $\mathbf{X} = [\mathbf{x}(1) \cdots \mathbf{x}(T)] \in \mathbb{C}^{N \times T}$. Further, since only a subset of the total of N potential devices are active, we use a diagonal matrix $\mathbf{A} \in \mathbb{R}^{N \times N}$ to denote user activity, in which the ith diagonal of $\mathbf{A}$ is 1 if and only if the device i is active, and zero otherwise. Then, the overall channel model becomes

$$\mathbf{Y} = \mathbf{HAX} + \mathbf{Z} \tag{2}$$

where the combined matrix $\mathbf{AX}$ is row sparse.

The conventional cellular networks design, in which each device is scheduled on a specific time-frequency tone, is not the most suitable for massive device communication, because not all devices have data to transmit all the time. This paper advocates a contention-based system, in which each device is assigned a signature sequence. The devices contend for the channel whenever they have data to communicate. The BS detects the activity pattern of the devices first, then subsequently the data.

Contention-based system has two key advantages. First, it can accommodate variable system load. Second, it avoids the latency associated with scheduling overhead. However, a contention-based system must also address the issue of how the active devices are identified and subsequently how data communications take place. We address these two issues separately below.

A. Phase I: Device Identification

To allow multiple devices to access a common channel, we must assign a signature sequence to each device n and use a device identification phase to identify the devices based on these pilot sequences. Let the length of these pilot sequences be L, where $L < T$. Denote the pilot sequence for device n as $\mathbf{s}_n = [s_n(1) \cdots s_n(L)]^\tau \in \mathbb{C}^L$. This paper assumes a simplified model in which time is slotted and the devices are synchronized so that all the active devices transmit their respective pilots at the beginning of the slot simultaneously.

The first phase of the proposed framework consists of the joint detection of active users and the estimation of their channels. Let $\mathbf{S} = [\mathbf{s}_1 \cdots \mathbf{s}_N]$ be the matrix of pilot signatures for devices 1 to N, the transmit signal over slots 1 to L is now $\mathbf{X}_{1:L} = \mathbf{S}^\tau$, and the overall channel model becomes:

$$\mathbf{Y}_{1:L} = \mathbf{HAS}^\tau + \mathbf{Z}_{1:L}, \tag{3}$$

where the subscript $(\cdot)_{1:L}$ denote the first L columns of the matrix. It is instructive to rewrite the above as

$$\mathbf{Y}_{1:L}^\tau = \mathbf{S}(\mathbf{HA})^\tau + \mathbf{Z}_{1:L}^\tau. \tag{4}$$

Observe that $\mathbf{S}$ is a known matrix and $(\mathbf{HA})^\tau$ is row sparse, so that the problem of jointly detecting the sparse user activity pattern $\mathbf{A}$ and estimating the channel matrix $\mathbf{H}$ is now a compressed sensing problem. For example, in the case where the BS is equipped with a single antenna, the problem amounts to detecting the non-zero pattern in the vector of channel values below:

$$\begin{bmatrix} y(1) \\ y(2) \\ \vdots \\ y(L) \end{bmatrix} = \begin{bmatrix} | & | & & | \\ \mathbf{s}_1 & \mathbf{s}_2 & \cdots & \mathbf{s}_N \\ | & | & & | \end{bmatrix} \begin{bmatrix} h_1 \\ h_2 \\ \vdots \\ \vdots \\ h_N \end{bmatrix} + \begin{bmatrix} z(1) \\ z(2) \\ \vdots \\ z(L) \end{bmatrix} \tag{5}$$

This is known as the single-measurement compressed sensing problem. When the BS is equipped with multiple antennas, each of the y_i and h_i and z_i above becomes a vector of size M. The resulting problem becomes that of detecting the sparse set of non-zero rows in the channel matrix $\mathbf{H}^\tau$ below:

$$\begin{bmatrix} \mathbf{y}^\tau(1) \\ \mathbf{y}^\tau(2) \\ \vdots \\ \mathbf{y}^\tau(L) \end{bmatrix} = \begin{bmatrix} | & | & & | \\ \mathbf{s}_1 & \mathbf{s}_2 & \cdots & \mathbf{s}_N \\ | & | & & | \end{bmatrix} \begin{bmatrix} \mathbf{h}_1^\tau \\ \mathbf{h}_2^\tau \\ \vdots \\ \vdots \\ \mathbf{h}_N^\tau \end{bmatrix} + \begin{bmatrix} \mathbf{z}^\tau(1) \\ \mathbf{z}^\tau(2) \\ \vdots \\ \mathbf{z}^\tau(L) \end{bmatrix} \tag{6}$$

This matrix version of the problem is known as the multiple-measurement compressed sensing problem.

To summarize, the joint user activity detection and channel estimation problem can now be stated simply as the estimation of non-zero rows in the row-sparse matrix $\mathbf{H}^\tau$ based on the measurement matrix $\mathbf{Y}_{1:L}^\tau$ in (6). Since we assume that only K out of N devices are active, we can proceed to use sparse optimization technique to identify the non-zero rows.

B. Phase II: Device Communication

Assuming successful identification of the devices, the second phase consists of data transmission. This phase is a conventional multiple-access channel with some fixed number of transmitters and a single receiver as modeled in (1), or in matrix form as:

$$\mathbf{Y}_{L+1:T} = \mathbf{HAX}_{L+1:T} + \mathbf{Z}_{L+1:T}, \tag{7}$$

for which the characterization of capacity region is well established. This paper will mostly focus on the achievable sum rate R.

As mentioned earlier, latency is a key requirement for device communication scenario of interest. Phase I and phase II transmissions combined must be completed within T slots. Thus, there is a tradeoff between the L time slots used in device identification and channel estimation versus the $T - L$ slots used for data transmission.

We note that there are two possible system implementations of this phase, depending on the latency and rate requirements. First, the BS can schedule the successfully identified users to the time-frequency resource blocks for transmission. The transmit signals from different users can be assured to be orthogonal. However, to perform scheduling, the BS must inform the users that they are correctly identified. Further, the

BS must provide the index of the time-frequency resources to the users. This not only puts certain capacity requirement on the downlink channel, but also causes considerable additional delay, although in general this scheduling approach is more spectrally efficient.

Alternatively, for delay-sensitive traffic, we can engineer the system so that the BS rarely makes an identification error, and the devices can transmit data immediately after the pilot sequence. In this case, device can transmit using spread spectrum schemes such as code-division multiple-access (CDMA). For this scheme to work, accurate identification of active devices is essential.

C. Problem Statement

The central questions this paper aims to answer are the following. Suppose that we have a latency constrained massive device communication scenario with latency or channel coherence constraint T. What is the fundamental limit in term of how many active devices K can be accommodated in the pool of potential devices N, and what is the maximum data rate R they can transmit?

Related to the above is the question of whether the two-phase transmission strategy involving a device detection and channel estimation stage followed by data transmission stage can achieve the fundamental limit above. If so, what is the optimal division between device/channel estimation stage L and transmission stage $T - L$? Finally, what are the practical algorithms for device detection and for data communication that achieves the optimal tradeoff between (N, K, R)?

D. Related Work

Massive devices detection and communications have been studied previously by several groups of researchers. In the 5G context, [1]–[3] propose compressed sensing based algorithms for joint device detection and data transmission. Compressed sensing approach has also been taken by [4]–[6], and [7], [8] for device detection problems. An earlier work exploiting sparsity in user activities is [9]. More recently, user detection and channel estimation are considered in [10], [11]. The approach taken in this paper is similar to these earlier lines of work. The emphasis here is on information theoretical upper bound and the resulting DoF analysis, which shed light on the optimized operating parameters at the system level. The information theoretical approach taken in this paper is related to the work of [12], [13] that studies the capacity of the multiuser channels in the limit of large number of users, although these works assume perfect channel knowledge while the present paper also accounts for channel estimation cost.

III. Fundamental Limits

A. Capacity Upper Bound

One of the main goals of this paper is to investigate the information theoretical limit of the tradeoff between (N, K, R). Toward this end, we derive the following upper bound on the achievable sum rate of the massive device connectivity.

Proposition 1. *Consider a massive device communications scenario with N potential devices, out of which K are active, communicating to a BS in the uplink. The BS detects the active devices and also decodes the messages from the active devices. Let the overall communication channel be modelled as $\mathbf{Y} = \mathbf{HAX} + \mathbf{Z}$ as in (2) with fixed distributions for $\mathbf{H}$, $\mathbf{A}$, $\mathbf{X}$, and $\mathbf{Z}$. Then the achievable sum rate of data transmission across all the users is approximately bounded by*

$$R \lesssim I(\mathbf{X}; \mathbf{Y}|\mathbf{HA}) - H(\mathbf{A}) - I(\mathbf{HA}; \mathbf{Y}|\mathbf{X}). \tag{8}$$

Proof. (Sketch) Expand the mutual information $I(\mathbf{HA}, \mathbf{X}; \mathbf{Y})$ as follows:

$$\begin{aligned} I(\mathbf{HA}, \mathbf{X}; \mathbf{Y}) &= I(\mathbf{X}; \mathbf{Y}) + I(\mathbf{HA}; \mathbf{Y}|\mathbf{X}) &(9)\\ &= I(\mathbf{HA}; \mathbf{Y}) + I(\mathbf{X}; \mathbf{Y}|\mathbf{HA}) &(10)\end{aligned}$$

We argue that the $I(\mathbf{HA}; \mathbf{Y})$ term is negligible, because it corresponds to the information transmission from the channel coefficients while not knowing $\mathbf{X}$. As $\mathbf{X}$ changes from symbol to symbol while $\mathbf{HA}$ is fixed within the coherence time T, this term must be small. A detailed evaluation requires expansion of $I(\mathbf{HA}; \mathbf{Y}) = h(\mathbf{Y}) - h(\mathbf{Y}|\mathbf{HA})$ according to their respective statistical distribution. For the rest of this paper, we ignore this term.

Assuming the above, the overall information transfer from the input of the channel $\mathbf{X}$ to the output $\mathbf{Y}$ is bounded by

$$I(\mathbf{X}; \mathbf{Y}) \lesssim I(\mathbf{X}; \mathbf{Y}|\mathbf{HA}) - I(\mathbf{HA}; \mathbf{Y}|\mathbf{X}). \tag{11}$$

Now, from the receiver's perspective, the overall information transfer consists of the sum data rate transmitted by the users, plus the information contained in the user activity pattern. Thus, we have:

$$R + H(\mathbf{A}) \leq I(\mathbf{X}; \mathbf{Y}). \tag{12}$$

Note that this bound is similar to a result contained in [12]. Combining the two inequalities above, we arrive at (8). □

The outer bound in Theorem 1 has very intuitive interpretations. The data transmission rate from the user devices to the BS is bounded by the data rate if the channel and activity patterns are known (the $I(\mathbf{X}; \mathbf{Y}|\mathbf{HA})$ term), subtracting from which the information contained in activity pattern (the $H(\mathbf{A})$ term) together with the cost of estimating the channels of the active users (the $I(\mathbf{HA}; \mathbf{Y}|\mathbf{X})$ term).

The outer bound clearly illustrates the crucial roles of user activity detection and channel estimation. For example, if a random set of K users can be active among the N potential users at any given time, the information contained in user activity pattern can be seen as:

$$H(\mathbf{A}) = Nh(p) \approx \log \begin{pmatrix} N \\ K \end{pmatrix}, \tag{13}$$

where $h(\cdot)$ is the binary entropy function, and $p = K/N$ is the user activity probability. When N is large, this term can be comparable to the user data transmission rate and cannot be ignored. Likewise, the need for channel estimation is clearly illustrated in the term $I(\mathbf{HA}; \mathbf{Y}|\mathbf{X})$. The impact of the two terms together can be clearly seen in a DoF analysis.

B. Degree-of-Freedom Analysis

This section asks the question of how the user data rate scales with the signal-to-noise ratio (SNR) and the number of devices by providing a DoF analysis for the massive connectivity scenario. The analysis is modeled in a similar fashion as in traditional MIMO channel [14], [15], but with the crucial difference of accounting for user activity detection. In particular, in traditional MIMO channels, the number of active users is typically smaller than the number of receiver antennas. But in massive connectivity application, we also need to consider the possibility of having a large number of devices as well. We present the analysis for two separate cases below.

1) Case $K \geq M$: In the massive connectivity scenario, there can be a large number of active users in the system. Thus, we can have $K \geq M$. In this case, over a coherence block of T time slots, the data rate assuming known channel scales with SNR as:

$$I(\mathbf{X}; \mathbf{Y}|\mathbf{HA}) \approx MT \log(\mathsf{snr}). \tag{14}$$

However, channel estimation incurs considerable cost. As there are a total of KM channel entries to estimate, the scaling of the channel estimation term can be up to:

$$I(\mathbf{HA}; \mathbf{Y}|\mathbf{X}) \approx KM \log(\mathsf{snr}). \tag{15}$$

Thus, as in a conclusion already reached in [15], for the overall data rate R to scale with SNR with non-zero DoF, we must at least have $K < T$. The channel estimation cost already puts restriction on the number of active devices that can be accommodated, while being able to communicate with significant data rate at the same time.

In the scenario in which the active devices need to be detected among a large set of potential users, there is further restriction on K in order to ensure non-zero DoF. Consider a system in which the number of active devices K, the number of receive antennas M and the coherence time T are fixed, but the number of potential users N is large. The large-system asymptotic analysis below reveals that K must be bounded by a fraction of T in order for the overall data rate to scale with SNR.

Lemma 1. *Consider a massive device communications scenario with K active users among N potential single-antenna users, communicating to a BS with M receive antennas over coherence time interval T. Let K, M, T be fixed. Assume $K \geq M$, but let N and SNR go to infinity as $N \approx (\mathsf{snr})^\eta$. The achievable sum rate scaling across all the users over T time slots is asymptotically bounded by*

$$R \leq (MT - K\eta - KM)^+ \log(\mathsf{snr}), \tag{16}$$

where $(\cdot)^+$ denote the positive part of a number. In particular, the DoF for the overall sum rate is non-zero only if

$$K < \left(\frac{M}{M+\eta}\right) T. \tag{17}$$

Proof. We use the outer bound (8) in Theorem 1. Evaluate the asymptotic scaling of $H(\mathbf{A})$ as follows:

$$\begin{aligned} H(\mathbf{A}) &= N(-p\log p - (1-p)\log(1-p)) \\ &\leq K\log(N/K) \\ &\leq K\eta \log(\mathsf{snr}) \end{aligned} \tag{18}$$

Combining with (14) and (15), we get the DoF result (21). It is easy to see that the DoF is non-zero only if K is bounded as in (17). □

The above result implies that when there are uncoordinated massive number of devices transmitting simultaneously, they can easily overwhelm the number of receive dimensions, defined by the coherence time T, to render high-rate transmission impossible. This result is the counterpart of the result by Lozano, Heath, and Andrews [15], which shows that for the usual multiple-access channel with known user activity, the DoF vanishes if $K \geq T$. The result of this paper shows that when user activities are unknown and need to be detected, then a further factor η, related to the total number of potential devices, needs to be accounted for.

2) Case $K < M$: The conclusion above motivates us to also consider the case where K is smaller, either by engineering the device network to allow only a small number of simultaneous transmissions at a time, or to implement receiver with massive MIMO, so that $M \gg K$. In this case, the relevant mutual information expressions become

$$I(\mathbf{X}; \mathbf{Y}|\mathbf{HA}) \approx KT \log(\mathsf{snr}) \tag{19}$$

and

$$I(\mathbf{HA}; \mathbf{Y}|\mathbf{X}) \approx K^2 \log(\mathsf{snr}), \tag{20}$$

where the latter is due to the fact that the transmit signal is rank K, so the effective receive signal dimension is only K, despite having M antennas at the receiver. In other words, it is not necessary to estimate KM channel elements; the number of effective channel coefficients is only K^2. This allows us to characterize the DoF upper bound as follows:

Lemma 2. *Consider a scenario with K active users among N potential single-antenna users, communicating to a BS with M receive antennas over coherence time interval T. Let K, M, T be fixed. Assume $K < M$, and let N and SNR go to infinity as $N \approx (\mathsf{snr})^\eta$. The achievable sum rate scaling across all the users over T slots is asymptotically bounded by*

$$R \leq \left(KT - K\eta - K^2\right)^+ \log(\mathsf{snr}). \tag{21}$$

Proof. The result follows directly from (19), (20), (18) and the outer bound (8). □

We now summarize the DoF outer bound results of this section by stating the following theorem.

Proposition 2. *Consider a massive device communications scenario with K active users among N potential single-antenna users, communicating to a BS with M receive antennas over coherence time interval T. If K, M, T are fixed,*

and N and SNR go to infinity as $N \approx (\mathsf{snr})^{\eta}$, then a DoF outer bound on the sum rate over all users, on a per time slot basis, is:

$$DoF \leq \begin{cases} \left(1 - \frac{K(M+\eta)}{TM}\right)^{+} M & \text{if} \quad K \geq M, \\ \left(1 - \frac{K+\eta}{T}\right)^{+} K & \text{if} \quad K < M. \end{cases} \tag{22}$$

Further, if K can be optimized, then the above DoF outer bound is maximized when

$$K^* = \min\left\{M, \left\lfloor \frac{T-\eta}{2} \right\rfloor^{+}\right\}, \tag{23}$$

in which case the DoF upper bound per time slot is $\left(1 - \frac{K^*+\eta}{T}\right)^{+} K^*$.

It is instructive to compare the above result with the DoF characterization of conventional noncoherent MIMO system by Zheng and Tse [14], which states that for a MIMO channel with K transmit antennas, M receive antennas, and channel coherence time T, we should use $K^* = \min\{K, M, T/2\}$ transmit antennas to achieve a maximum overall DoF of $(1 - K^*/T)K^*$. It is clear that the result of this paper is an analogous analysis with the key difference of accounting for the effect of device identification. Massive device identification brings a cost to the overall DoF bound; it also reduces the optimal number of transmit antennas.

In [15], the practical value of T for a typical cellular network is calculated to be in the order of 10^4 for Doppler value corresponding to pedestrian speed. For devices that are stationary, the value of coherence time T can be even larger, making the accommodation of simultaneous transmission by thousands of devices per cell a feasible goal, especially when device transmission can be coordinated so that the number of simultaneous transmitting devices at any given time is no more than the number of receive antennas.

However, for systems where coherence time is limited and with uncoordinated transmission across massive number of devices, controlling the number of simultaneous transmissions is not always possible. The result of this paper shows that there are achievable data rate implication for such uncoordinated massive device communications. When the number of simultaneous transmissions goes beyond the number of receive antennas, the system DoF starts to decrease linearly with the number of active devices. When the number of active devices is almost at the coherence time, the achievable DoF becomes zero, making high-rate transmission effectively impossible.

IV. Achievable Rates

The information theoretical analysis of the previous section provides an outer bound for the data rate across all the users in a massive connectivity scenario. The outer bound can be evaluated numerically for comparison with practically achievable rates. This section aims to present the achievable rates of the device communication framework proposed in this paper for comparison with the outer bound.

A. DoF Comparison

Fixing the latency or channel coherence constraint T, the proposed achievable scheme uses L timeslots for device identification, and $T - L$ slots for data transmission. Assuming that K out of N devices are active at any given time, we must choose L sufficiently large so that K devices can be detected with negligible probability of error and their channels are estimated accurately in the first phase. Subsequently, the second phase of $T - L$ time slots then provides the following achievable data rate

$$R = (T - L) \log |\mathbf{H}\mathbf{A}\mathbf{\Sigma}_{\mathbf{X}}\mathbf{A}\mathbf{H}^{\dagger} + \mathbf{I}|, \tag{24}$$

where $\mathbf{\Sigma}_{\mathbf{X}}$ is a $N \times N$ transmit covariance matrix across all the devices. With independent transmission with fixed power P for each device, Σ_X is just a diagonal matrix with diagonal value P in entries corresponding to the active devices.

The above rate expression scales as

$$R \approx \min\{M, K\} \left(1 - \frac{L}{T}\right) T \log(\mathsf{snr}). \tag{25}$$

Comparing the achievable DoF in (25) with the DoF upper bound (22), we see that the bounds match if the following choices of $L = K\left(1 + \frac{\eta}{M}\right)$ for the $K \geq M$ case and $L = K + \eta$ for the $K < M$ case are sufficient for device detection and channel estimation. In particular, the outer bound indicates that if L can be made to scale linearly with K, then we would have been able to achieve the largest possible DoF.

Interestingly, the information theoretical analysis of compressed sensing indeed suggests that the linear scaling with K is all that is needed for sparse recovery (in the $M = 1$ case), i.e., $L = O(K)$ is sufficient under suitable conditions [16]–[19]. However, the existing results in literature are often asymptotic in somewhat different regimes and are not directly comparable to the DoF outer bound derived earlier. For example, the fundamental result of compressed sensing [20], [21] states that if the sensing matrix satisfies a restricted isometry (RIP) property, then $L = O(K \log(N/K))$ suffices for sparse recovery, but results of this type apply only at high SNR. The scaling of the required L as function of SNR is typically not known. Further, results for the $M > 1$ cases are harder to obtain. The important question of the impact of the channel estimation error also have not yet been fully explored.

The computational complexity of the detection algorithm is another important question. Broadly speaking, there are three classes of algorithms of practical interests classified according to complexity: convex optimization based algorithms intended to solve a convex relaxation of the sparse recovery algorithm (e.g., [20]–[23]), combinatorial optimization based algorithms (e.g., [24]; see [25] and references therein), and iterative decoding algorithms based on message passing [26]–[29]. For large-scale sparse recovery problems, iterative decoding based algorithms are the only feasible ones, as the complexity of convex optimization or combinatorial optimization becomes prohibitive when the problem size is large.

Of particular interest is a class of algorithms known as approximate message passing (AMP) [26], which represents a

complexity-performance tradeoff attractive to communications applications. The AMP algorithm for compressed sensing has been extensively studied in the literature. A key advantage of AMP is that it admits an analysis based on state evolution that allows an accurate prediction of successful recovery as function of problem parameters [30], [31]. Further, as shown in [27], the noise-sensitivity phase transition of AMP in fact coincides with that of convex optimization approach. We mention here our recent work applying the AMP algorithm to the massive device detection problem [32], [33].

Finally, we remark that the discussions in this paper are predicated on the standard assumptions in compressed sensing with random sensing matrices, typically chosen from some i.i.d. distribution (e.g., Gaussian). There is additional space for designing the sensing matrix, for example, by choosing the sensing matrices to be sparse.

V. Conclusion

Massive device connectivity presents a new set of challenges and opportunities for communication engineering. This paper provides a framework for the design and analysis of massively connected device network. The main contributions of this paper include an information theoretical upper bound on the transmission rate of the massive device network and a corresponding DoF analysis. Further, this paper presents a two-phase scheme, with joint device identification and channel estimation in the first phase followed by data transmission in the second phase, as a possible way of engineering such a system.

Acknowledgment

This work is supported by Huawei Technologies Canada Co., Ltd. The author wishes to thank Peiying Zhu and Jianglei Ma for many very useful discussions.

References

[1] G. Wunder, H. Boche, T. Strohmer, and P. Jung, "Sparse signal processing concepts for efficient 5G system design," *IEEE Access*, vol. 3, pp. 195–208, 2015.

[2] G. Wunder, P. Jung, and C. Wang, "Compressive random access for post-LTE systems," in *IEEE Inter. Conf. Commun. Workshops (ICC)*, Jun. 2014, pp. 539–544.

[3] G. Wunder, P. Jung, and M. Ramadan, "Compressive random access using a common overloaded control channel," in *IEEE Inter. Conf. Commun. Workshops (ICC)*, Jun. 2015.

[4] H. F. Schepker and A. Dekorsy, "Compressive sensing multi-user detection with block-wise orthogonal least squares," in *IEEE Veh. Tech. Conf. (VTC Spring)*, May 2012, pp. 1–5.

[5] H. F. Schepker, C. Bockelmann, and A. Dekorsy, "Exploiting sparsity in channel and data estimation for sporadic multi-user communication," in *Inter. Symp. Wireless Commun. Sys. (ISWCS)*, Aug 2013, pp. 1–5.

[6] F. Monsees, M. Woltering, C. Bockelmann, and A. Dekorsy, "Compressive sensing multi-user detection for multicarrier systems in sporadic machine type communication," in *IEEE Veh. Tech. Conf. (VTC Spring)*, May 2015, pp. 1–5.

[7] A. K. Fletcher, S. Rangan, and V. K. Goyal, "A sparsity detection framework for on-off random access channels," in *IEEE Inter. Symp. Inf. Theory (ISIT)*, Jun. 2009, pp. 169–173.

[8] ——, "On-off random access channels: A compressed sensing framework," Mar. 2009, [Online] Available: arXiv:0903.1022v2.

[9] 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.

[10] G. Hannak, M. Mayer, A. Jung, G. Matz, and N. Goertz, "Joint channel estimation and activity detection for multiuser communication systems," in *IEEE International Conference on Communication Workshop (ICCW)*, Jun. 2015, pp. 2086–2091.

[11] X. Xu, X. Rao, and V. K. N. Lau, "Active user detection and channel estimation in uplink cran systems," in *IEEE Inter. Conf. Commun. (ICC)*, Jun. 2015, pp. 2727–2732.

[12] X. Chen and D. Guo, "Many-access channels: The Gaussian case with random user activities," in *IEEE Inter. Symp. Inf. Theory (ISIT)*, Jun. 2014.

[13] T.-Y. Chen, X. Chen, and D. Guo, "Many-broadcast channels: Definition and capacity in the degraded case," in *IEEE Inter. Symp. Inf. Theory (ISIT)*, Jun. 2014.

[14] L. Zheng and D. N. C. Tse, "Communication on the grassmann manifold: A geometric approach to the noncoherent multiple-antenna channel," *IEEE Trans. Inf. Theory*, vol. 48, no. 2, pp. 359–383, Feb. 2002.

[15] A. Lozano, R. W. Heath, and J. G. Andrews, "Fundamental limits of cooperation," *IEEE Trans. Inf. Theory*, vol. 59, no. 9, pp. 5213–5226, Sep. 2013.

[16] M. J. Wainwright, "Information-theoretic limits on sparsity recovery in the high-dimensional and noisy setting," *IEEE Trans. Inf. Theory*, vol. 55, no. 12, pp. 5728–5741, Dec. 2009.

[17] Y. Wu and S. Verdú, "Optimal phase transitions in compressed sensing," *IEEE Trans. Inf. Theory*, vol. 58, no. 10, pp. 6241–6263, Oct. 2012.

[18] M. Akcakaya and V. Tarokh, "Shannon-theoretic limits on noisy compressive sampling," *IEEE Trans. Inf. Theory*, vol. 56, no. 1, pp. 492–504, Jan. 2010.

[19] D. L. Donoho, A. Javanmard, and A. Montanari, "Information-theoretically optimal compressed sensing via spatial coupling and approximate message passing," *IEEE Trans. Inf. Theory*, vol. 59, no. 11, pp. 7434–7464, Nov. 2013.

[20] D. L. Donoho, "Compressed sensing," *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1289–1306, Apr. 2006.

[21] E. Candes, J. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Communications on Pure and Applied Mathematics*, vol. 59, no. 8, pp. 1207–1223, Aug. 2006.

[22] M. Bayati and A. Montanari, "The LASSO risk for Gaussian matrices," *IEEE Trans. Inf. Theory*, vol. 58, no. 4, pp. 1997–2017, Apr. 2012.

[23] M. J. Wainwright, "Sharp thresholds for high-dimensional and noisy sparsity recovery using l_1-constrained quadratic programming (LASSO)," *IEEE Trans. Inf. Theory*, vol. 55, no. 5, pp. 2183–2202, May 2009.

[24] D. Needell and J. A. Tropp, "CoSaMP: Iterative signal recovery from incomplete and inaccurate samples," *Appl. Comput. Harmon. Anal.*, vol. 26, pp. 301–321, 2008.

[25] J. A. Tropp and S. J. Wright, "Computational methods for sparse solution of linear inverse problems," *Proc. IEEE*, vol. 98, no. 6, pp. 948–958, Jun. 2010.

[26] D. L. Donoho, A. Maleki, and A. Montanari, "Message-passing algorithms for compressed sensing," *Proceedings of National Academy of Sciences*, vol. 106, no. 45, pp. 18 914–18 918, Nov. 2009.

[27] ——, "The noise-sensitivity phase transition in compressed sensing," *IEEE Trans. Inf. Theory*, vol. 57, no. 10, pp. 6920–6941, Oct. 2011.

[28] D. Baron, S. Sarvotham, and R. G. Baraniuk, "Bayesian compressive sensing via belief propagation," *IEEE Trans. Signal Process.*, vol. 58, no. 1, pp. 269–280, Jan. 2010.

[29] S. Rangan, "Generalized approximate message passing for estimation with random linear mixing," in *IEEE Inter. Symp. Inf. Theory (ISIT)*, Jul. 2011, pp. 2168–2172.

[30] M. Bayati and A. Montanari, "The dynamics of message passing on dense graphs, with applications to compressed sensing," *IEEE Trans. Inf. Theory*, vol. 57, no. 2, pp. 764–785, Feb. 2011.

[31] A. Montanari, "Graphical models concepts in compressed sensing," in *Compressed Sensing*, Y. Eldar and G. Kutyniok, Eds. Cambridge University Press, 2012, ch. 9, pp. 394–438.

[32] Z. Chen and W. Yu, "Massive device activity detection by approximate message passing," in *IEEE Inter. Conf. Acoustic, Speech, Signal Process. (ICASSP)*, Mar. 2017.

[33] L. Liu and W. Yu, "Massive device connectivity with massive MIMO," 2017, preprint.