Quick Discovery of Mobile Devices in the Many-User Regime - Carrier Sensing or Simultaneous Detection?

Altug Karakurt*, Atilla Eryilmaz*, C. Emre Koksal*
*Department of Electrical and Computer Engineering, The Ohio State University
{karakurt.1, eryilmaz.2, koksal.2}@osu.edu

Abstract—**We consider the problem of detecting the active wireless stations among a very large population. This problem is highly relevant in applications involving passive and active RFID tags and dense IoT settings. The state of the art mainly utilizes interference avoiding (e.g., CSMA-based) approaches with the objective of identifying one station at a time. We first derive basic limits of the achievable delay with interference avoiding paradigm. Then, we consider the setting in which each station is assigned a signature sequence, picked at random from a specific alphabet and active stations transmit their signatures simultaneously upon activation. The challenge at the detector is to detect all active stations from the combined signature signal with low probability of misdetection and false positives. We show that, such an interference embracing approach can substantially reduce the detection delay, at an arbitrarily low probability of both types of detection errors, as the number of stations scale. We show that, under a randomized activation model the collision embracing detection scheme achieves $\Theta\big(\frac{\log^2(n)}{\log(\log(n))}\big)$ delay while the expected delay of existing CSMA schemes are $\Omega\big(\log^2(n)\big)$ for a population of n stations. Finally, we discuss large-scale implementation issues such as the design of low-complexity detection schemes and present numerical investigations.**

I. Introduction

Multi-station detection is commonly used in various large scale tracking tasks such as sensor networks, supply chain management and stock control. In this problem setting, there is a large population of stations, a subset of which is present in the range of a detector at the time and the goal is to detect this subset with low probability of error. One of the key challenges of multi-station detection is the multiple access aspect of the network due to the unique challenges of the problem. Many applications involve stations that have very limited energy and computational capabilities, such as mobile IoT nodes that are powered with very limited batteries. Some applications even contain passive stations that use backscatter techniques to communicate, such as passive RFID tags. In order to be applicable to such cases, multi-station detection methods address the potentially limited computational capacity and energy of the stations, as well as ensuring robustness to the varying station population in detection range, scalability and the required quick detection rates.

The existing methods used in practice are dominantly collision avoidance techniques. One approach is to have the detector broadcast messages, based on which the stations decide whether they would transmit at the time or not. The broadcast messages are usually based on a tree search algorithm to find hierarchical subsets that contain active stations [1]. Another line of work employs random access methods inspired by slotted Aloha [2], [3], where each station randomly picks a time-slot within a time-frame of a pre-determined size. The detector controls the frame-size based on the observed successes and failures on previous time-frames.

There are also studies that investigate methods for simultaneous station detection, that are proposed as extensions to collision avoidance methods. One of these approaches is to use multi-antenna decoders as in [4], where multiple linear combinations of transmitted signals are received at the antennas and detection is done by solving these linear equations. Another line of research [5] employs signal processing inspired techniques to cluster the complex-valued samples of the received mixture. Later work [6], [7], [8] improve this method by taking additional information into account, such as considering the temporal information instead of only using projection of the samples on the plane. These studies share a common caveat inherent in the clustering technique; as the station population increases, the number of clusters grow exponentially and they become indistinguishable.

In [9], authors design binary codes and propose a scheme to decode multiple stations at once over a binary-OR channel. These described simultaneous detection methods work as part of collision avoidance techniques aim to decode the collided packets when an undesired collision occurs. Due to this use-case, these methods expect only a few packets collide.

As an alternative perspective to the established collision avoidance schemes, we describe a new coding scheme that embraces collisions. A major goal of this approach is to ensure scalability to be feasible for applications that involve large populations of stations. We propose to have all active stations transmit at once and detect them simultaneously from the received signal mixture. The work in [10] also embraces collisions to detect stations simultaneously. Despite its practical appeal this work omits theoretical guarantees and due to its use of clustering techniques, suffers from the aforementioned lack of scalability.

In our problem formulation, we propose statistical models that capture the statistics of stations' activation. By considering these, we are able to connect the detection problem with other fields like compressed sensing. This enables numerous algorithms to be employed as potential solutions to the large-scale detection problem.

The contributions of this paper can be summarized as

- We present basic limits of achievable delay by collision avoidance approaches in multi-station detection.

- We propose a communication scheme for the simultaneous station detection and derive bounds for its error probability and delay performance under the proposed activation models. We show that the proposed approach outperforms collision avoidance methods. More specifically, we show that under the independent activation model the delay of proposed simultaneous detection scheme is $\Theta\big(\frac{\log^2(n)}{\log(\log(n))}\big)$, while collision avoidance approaches yield $\Omega\big(\log^2(n)\big)$ delay scaling.
- We discuss practical issues related to implementing our scheme and describe a computationally feasible decoding approach and present numerical investigations.

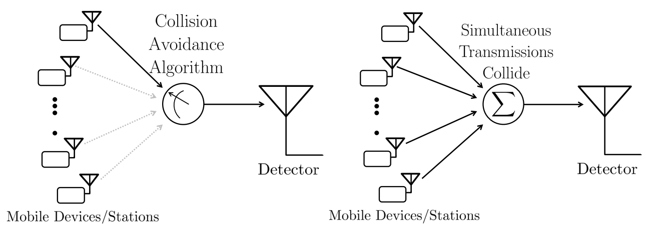
II. System Model and Problem Description

Let there be a total of n stations in the system. We call a station $active$ (state 1) if it is within the reading range of the detector and $passive$ (state 0) otherwise. We consider two cases that model the statistics of the distribution of states of stations. We employ these models later in our analyses and numerical investigations.

- *Sparse Activation Model:* There exists an upper bound $k_{max} \in \mathbb{Z}_+$ on the number of active stations at a time. This bound is independent of n.
- *Independent Activation Model:* The states of stations are i.i.d. Bernoulli random variables with some $p \in [0, 1]$, where p is allowed to be a function of n.

The goal of multi-station detection problem is to detect the set of active stations at a time with as small delay as possible while achieving low detection error probability. Here, we refer to the time it takes for a scheme to attempt detecting the set of active tags in the system as the $delay$ of the scheme. It is measured in terms of the number of transmitted symbols. We report this performance metric in the evaluations we present in the following sections.

For detection of stations, users interact with the central station using a variety of protocols. These protocols can be classified as collision embracing or collision avoiding. These paradigms are presented in Figure 1. As mentioned in the previous section, the literature on multiple access of multi-station detection dominantly relies on collision avoidance approaches that aim to coordinate present stations to transmit their signatures one at a time. Next, we describe the activation process in our collision embracing scheme.



(a) Collision Avoidance Scheme

(b) Collision Embracing Scheme

Figure 1: Perspectives of Collision Avoidance and Collision Embracing Schemes

We associate unique real-valued sequences, called *signature sequences* to each station to distinguish them from each other. We propose to have all stations transmit their signature sequences simultaneously and detect the set of present stations from the received mixture of these signals. The aforementioned collision avoidance approaches yield iterative solutions to detect stations one at a time, while our approach views this task as a one-shot problem.

Let m denote the length of the signature sequences of stations and these sequences be stored in a codebook $A_{m\times n}$ of real symbols satisfying the power constraint $\frac{1}{m}\sum_{j=1}^{m} a_{ij}^2 \leq P,\ \forall i \in \{1, 2, ..., n\}$ for a given power constraint $P \in \mathbb{R}$. In our analyses and numerical investigations we use random Gaussian codebooks, entries of which are independent and identically distributed (i.i.d.) as $a_{ij} \sim \mathcal{N}(0, P)$. We use $\text{SNR} = \frac{P}{N_0/2}$ to denote the signal-to-noise ratio per transmitted symbol.

Let the states of stations be denoted by the state-vector $\mathbf{s} \in \{0, 1\}^n$ and s_i denote the state of the i^{th} station. An active station transmits its signature sequence to show its presence, which for station i would be $\{A_{ij}\}_{j=1}^{m}$. Letting $\mathbf{y}$ denote the m-dimensional vector received at the detector, we obtain the following model

$$\mathbf{y} = A\mathbf{s} + \mathbf{w}, \tag{1}$$

where $\mathbf{w} = \{w_i\}_{i=1}^{m}$ is additive white Gaussian noise with $w_i \sim \mathcal{N}(0, N_0/2)$. Using $\mathbf{y}$ and the known codebook A, the detector infers the set of active stations.

Let $\hat{\mathbf{s}}$ denote the recovered state-vector from $\mathbf{y}$ by the detector. Since the simultaneous detection of stations is a one-shot process, we consider the event $\{\mathbf{s} \neq \hat{\mathbf{s}}\}$ as a detection error. On the other hand, in the case of an erroneous detection a collision avoidance scheme allows the detector to ask for a re-transmission. Since the proposed scheme does not allow any re-transmissions, we aim to achieve the following.

Problem Statement: Our goal is to design a codebook A that achieves low detection delays and yields a probability of detection error that vanishes with increasing number of stations, n.

Since the multi station detection problem is commonly used in large scale applications, the detection method is required to be applicable to large populations of stations and achieve low delay scaling to still be feasible. In our analyses, we guarantee low probability of detection error for large n, which is usually the case in practice. We provide theoretical guarantees on the achieved delay and error probabilities by the proposed scheme in the next sections.

In addition to the mentioned related literature, the information theoretic *many-user information* paradigm presented in [11] is similar to ours under certain conditions. This paper aims to characterize the capacity of Gaussian channel. What distinguishes this work is that they allow the number of transmitters to grow with the block-size, instead of assuming it to be constant. The independent activation model we propose becomes in line with this perspective when p is a function of n. In Section IV, we consider this case in our analysis and compare the generality of our results with the related investigations in [11].

III. Basic Limits of Simultaneous Detection

We present theoretical guarantees for the simultaneous detection scheme when maximum likelihood (ML) decoding is used by the detector. Let $\mathcal{E}$ denote the event of an erroneous detection, that is the event of having a station whose state is detected incorrectly. and let T_{SD} denote the delay of the simultaneous detection scheme.

We first present a bound on probability of detection error for finitely many stations, that is independent of the station activation model. Then, using this result we analyze $\mathrm{E}[T_{SD}]$ under both of the proposed activation models.

Lemma 1. *The probability of detection error for the proposed simultaneous detection scheme using maximum likelihood decoding satisfies*

$$\mathrm{P}(\mathcal{E}) \leq \sum_{k=1}^{n} \binom{n}{k} \left(1+\frac{k}{4}\mathrm{SNR}\right)^{-\frac{m}{2}}. \qquad (2)$$

Proof. We proceed by following the random coding argument (as in [12], Chapter 5). Let $x(\mathbf{s}) = A\mathbf{s}$ denote the encoding function that maps the state-vector $\mathbf{s}$ to its corresponding codeword, where the codebook $A = \{a_{ij}\}_{m\times n}$ consists of i.i.d. Gaussian random variables $a_{ij} \sim \mathcal{N}(0, P)$. Let

$$P_2(\mathbf{s}_i, \mathbf{s}_j) = \mathrm{Q}\left(\frac{||x(\mathbf{s}_i) - x(\mathbf{s}_j)||}{\sqrt{2N_0}}\right),$$

which is the probability that an ML-detector confuses the state-vectors $\mathbf{s}_i$ and $\mathbf{s}_j$ with each other under an AWGN channel [12]. For a given codebook and a particular state-vector $\mathbf{s}_i$, using union bound and the fact that there are 2^n possible state-vectors we obtain the following.

$$\begin{aligned}\mathrm{P}(\mathcal{E}) &= \sum_{i=1}^{2^n} \mathrm{P}(\mathbf{s}_i)\mathrm{P}(\mathcal{E}|\mathbf{s}_i) \\ &\leq \sum_{i=1}^{2^n} \mathrm{P}(\mathbf{s}_i) \sum_{j=1, j\neq i}^{2^n} P_2(\mathbf{s}_i, \mathbf{s}_j).\end{aligned}$$

Let $N_{\mathbf{v}}$ denote the ℓ_1-norm of a vector $\mathbf{v}$. Then, $N_{|\mathbf{s}_i - \mathbf{s}_j|}$ denotes the number of differing indices between the two state-vectors $\mathbf{s}_i$ and $\mathbf{s}_j$.

$$\sum_{j=1, j\neq i}^{2^n} P_2(\mathbf{s}_i, \mathbf{s}_j) \leq \sum_{k=1}^{n} \binom{n}{k} \mathrm{P}(N_{|\mathbf{s}_i - \mathbf{s}_j|} = k). \qquad (3)$$

Let $\hat{\mathbf{s}}$ to denote the decoded state-vector. By the construction of our codebook, we have $|x(\mathbf{s}) - x(\hat{\mathbf{s}})| = |x(\mathbf{s} - \hat{\mathbf{s}})|$. Using this, we define a new random variable $D_{N_{|\mathbf{s}-\hat{\mathbf{s}}|}}$ as

$$D_{N_{|\mathbf{s}-\hat{\mathbf{s}}|}} = ||x(\mathbf{s}) - x(\hat{\mathbf{s}})|| = \sqrt{\sum_{i=1}^{m} \left(\sum_{j=1}^{N_{|\mathbf{s}-\hat{\mathbf{s}}|}} a_{ij}\right)^2}.$$

We rewrite (3) using this definition and the fact that $\mathrm{Q}(x) < e^{-\frac{x^2}{2}}$.

$$\begin{aligned}\sum_{j=1, j\neq i}^{2^n} P_2(\mathbf{s}_i, \mathbf{s}_j) &\leq \sum_{k=1}^{n} \binom{n}{k} \mathrm{Q}\left(\frac{D_k}{\sqrt{2N_0}}\right) \\ &< \sum_{k=1}^{n} \binom{n}{k} \mathrm{E}\left[e^{-\frac{D_k^2}{4N_0}}\right].\end{aligned}$$

Note that $(\sum_{i=1}^{k} a_{ij}) \sim \mathcal{N}(0, kP)$ and $(\frac{D_k^2}{4N_0}) \sim \frac{kP}{4N_0}\chi_m^2$, where χ_m^2 denotes the chi-squared distribution with m-degrees of freedom. Hence, this expectation is the evaluation of the moment generating function of χ_m^2 distribution at $t = -\frac{kP}{4N_0}$.

$$\mathrm{E}\left[e^{-\frac{D_k^2}{4N_0}}\right] = \left(1 + \frac{kP}{2N_0}\right)^{-\frac{m}{2}}.$$

Plugging this expression in (3) we obtain the following bound.

$$\begin{aligned}\mathrm{P}(\mathcal{E}) &\leq \left(\sum_{i=1}^{2^n} \mathrm{P}(s_i)\right) \left(\sum_{k=1}^{n} \binom{n}{k} \left(1 + \frac{kP}{2N_0}\right)^{-\frac{m}{2}}\right) \\ &= \sum_{k=1}^{n} \binom{n}{k} \left(1 + \frac{kP}{2N_0}\right)^{-\frac{m}{2}} \qquad \square\end{aligned}$$

This result is independent of the station activation model since prior probabilities of sets of stations being active decouple from our bound for probability of error and disappear by adding up to one.

We aim to choose m large enough to make sure that the probability of detection error vanishes as $n \to \infty$, while achieving low delay. Using Lemma 1, we first prove a sufficient condition on m towards this goal and use it to provide theoretical guarantees for the delay of our scheme under both of the proposed activation models.

As long as we can guarantee that probability of detection error vanishes with increasing n, the delay performance of our scheme converges to m as $n \to \infty$ since the probability of correct detection of all stations approaches one. Hence, the detection task would be completed in one trial with probability converging one.

1) Sparse Activation Model: In this model, the maximum number of active stations is bounded by an integer k_{max} that is independent of n. Note that we do not have any restrictions on the distribution of station population as long as its support is bounded by $[0, k_{max}]$.

Theorem 1. *Let simultaneous detection scheme be used with ML-decoding. Suppose that the number of active stations at a time is bounded by some $k_{max} \in \mathbb{Z}_+$. If $m(n) = \omega(\log(n))$ then for any SNR > 0, $\lim_{n\to\infty} \mathrm{P}(\mathcal{E}) = 0$.*

Proof. See Appendix A.

Corollary 1. *For large enough n, the probability of correct detection for simultaneous detection scheme with ML-decoding gets arbitrarily close to 1. Therefore, for any SNR > 0 a delay of*

$$T_{SD} = \frac{4k_{max}\log(n) - 2\log(2k_{max}!)}{\log(1 + k_{max}\mathrm{SNR})}$$

is achievable with a probability of error that vanishes with increasing n.

Proof. This result follows from the last step of the proof of Theorem 1. $\square$

In the proof of these result, we exploit the boundedness of the number of active stations to obtain a tight bound, which is not an unrealistic expectation for many practical applications. For example, sparsity can be guaranteed in inventory related applications involving RFID tags since activation of tags require them to be in the limited reading range of the tag reader. Hence, no matter how large scale of a tag population the problem has and how densely they are placed, the range of tag reader might limit the maximum number of active stations at a time. Another domain that involves such sparsity is massive Internet of Things (IoT) applications, where despite the possibly large population of devices they are likely to be located much more sparsely compared to an inventory application. Using such structural properties of application at hand, one can bound the maximum possible number of the active stations in the range of the detector.

2) Independent Activation Model: In this model, the states of stations are i.i.d. Bernoulli random variables with parameter p, which is allowed to scale with n. For our analysis, we set $p(n) = \frac{f(n)}{n}$, where $f(n)$ is a monotonically increasing sub-linear function of n. This choice is due to the proof technique we are using, where under the restriction of $p(n) = \omega(1/n)$ we bound the probability of the population of active stations deviating from the expected population.

Theorem 2. *Let simultaneous detection scheme be used with ML-decoding. Suppose that the states of stations are i.i.d. Bernoulli random variables with $p(n) = \frac{f(n)}{n}$, where $f(n)$ is a sub-linear monotonically increasing function of n. If*

$$m(n) = \omega\left(\frac{2f(n)\log(n) - \log\left(\lceil 2f(n)\rceil !\right)}{\log(f(n))}\right)$$

then

$$\lim_{n\to\infty} \mathrm{P}(\mathcal{E}) = 0.$$

Proof. See Appendix B.

Corollary 2. *For large enough n, the probability of correct detection for simultaneous detection scheme with ML-decoding gets arbitrarily close to 1. Therefore, for $p(n) = \frac{f(n)}{n}$ and any SNR > 0 a delay of*

$$T_{SD} = \frac{4f(n)\log(n) - 2\log(\lceil 2f(n) - 1\rceil !)}{\log\left(1 + \frac{f(n)}{2}\mathrm{SNR}\right)}$$

is achievable with a probability of error that vanishes with increasing n.

Proof. This result follows from the last step of the proof of Theorem 2. □

This activation model is more challenging than the sparse activation model. As discussed previously, many applications can be modeled using the sparse activation model due to problem specific properties like the possibly limited reading range of detectors or sparse placement of stations. However, this setting is still of interest as it demonstrates the robustness of the scheme to the variation in active number of stations. Moreover, the availability of such theoretical guarantees might expand the possible applications of multi-station detection and make new use-cases feasible. Note that, under this model the expected number of active stations is unbounded and despite this challenge, the proposed simultaneous scheme achieves a desirable delay performance.

Corollary 3. *As a special case of the previous result, for $p(n) = \frac{c\log(n)}{n}$ for some $c \in \mathbb{R}_+$ and any SNR > 0 a delay of*

$$T_{SD} = \frac{4c\log^2(n) - 2\log\left(\lceil 2c\log(n) - 1\rceil !\right)}{\log\left(1 + \frac{c\log(n)}{2}\mathrm{SNR}\right)}$$

is achievable with a probability of error that vanishes with increasing n.

As mentioned in Section I, part of the work presented in [11] is similar to our scheme under the independent activation model when p scales with n. In that study, *minimum user identification cost* is reported as a measure, which is analogous to m in our formulation. It is defined as the minimum signature length for detection of the set of active transmitters, which guarantees a vanishing error probability as $n \to \infty$. Using our notation, the main result for this part of this study (Theorem 2 in [11]) requires the following hypothesis for $p(n) = \frac{f(n)}{n}$.

$$\lim_{n\to\infty} ne^{-\delta f(n)} = 0 \text{ for all } \delta > 0,$$

which is not satisfied by our benchmark setting of $f(n) = c\log(n)$, since for $\delta = 1/c$ this limit yields 1. This condition is equivalent to requiring $f(n) = \omega(\log(n))$. Regardless of the result being used, we must have $f(n) = o(n)$ to ensure $p(n) \in [0, 1]\ \ \forall n$. Therefore, while our result is applicable for any $f(n)$ that is $o(n)$, the similar theorem in [11] requires $f(n)$ to be $\omega(\log(n))$ and $o(n)$.

IV. Basic Delay Limits of Collision Avoidance Schemes

In this section, we present the fundamental limits of the delay performance of collision avoidance schemes. We first derive a bound for a genie-aided scheme assuming perfect coordination. Despite being unachievable, the scaling of this optimistic result is outperformed by the proposed simultaneous detection scheme. We then use dynamic programming to present a fundamental limit for any CSMA scheme that detects stations one at a time. Finally, we turn to a RFID tag detection method used in practice, called QueryTree [1], as a special case of our previous analyses and derive a bound for it.

A. Genie-Aided Collision Avoidance Scheme with Perfect Coordination

Regardless of how they schedule their transmissions, all collision avoidance schemes try to coordinate the stations to maximize the ratio of time-slots at which there is exactly one transmission. Hence, the optimal situation for a collision avoidance scheme is to achieve perfect coordination and have exactly one station transmit at each time-slot, without any idle slots. We now analyze this optimistic case by considering a scheme, where we assume that there is a genie that assigns each station a time-slot to achieve perfect coordination without any overhead.

We denote the delay of this scheme by T_{Genie}. We employ the independent activation model with $p(n) = \frac{c\log(n)}{n}$ and assume a power constraint of P and an AWGN channel with a noise variance of $N_0/2$.

Lemma 2. *Suppose that the states of stations are i.i.d. Bernoulli random variables with $p(n) = \frac{c\log(n)}{n}$, for some $c \in \mathbb{R}_+$. The delay performance of genie-aided scheme with perfect coordination satisfies*

$$\mathrm{E}[T_{Genie}] \geq \frac{2c\log^2(n)}{\log(2)\log\left(1+\mathrm{SNR}\right)} \tag{4}$$

Proof. Under the independent activation model with $p(n) = \frac{c\log(n)}{n}$, the expected number of active stations at a time is $\log(n)$. Therefore, even if there is exactly one transmitter at a time at all time-slots without any errors due to channel impairments, the expected number of time-slots needed to have each station transmit their signature is at least $\log(n)$. Since there are n stations in the system, the binary signature length of each station must be at least $\log_2(n)$.

By Theorem 9.1.1 of [13], the capacity of a Gaussian channel with power constraint P and noise variance $N_0/2$ is the following.

$$C = \frac{1}{2}\log\left(1+\mathrm{SNR}\right) \text{ bits/real symbol} \tag{5}$$

Therefore, if the stations use capacity achieving channel coding, a binary string of length $\log_2(n)$ would be transmitted by $\frac{\log_2(n)}{C}$ symbols. □

B. Query Tree Protocol

As proposed in [1], QueryTree protocol converts the multiple access of RFID tag identification to a tree-search problem. The tag reader has a hierarchical tree of the set of tags in the system and traverses it in a depth-first manner. At the beginning of each time-frame, the reader broadcasts the prefix of the current node of the tree and the tags whose signatures start with the prefix respond. If the reader observes collision, it keeps traversing down the branch. Otherwise records the response (or no response) and trims the sub-tree below the current node and backtracks to the next unresolved node in the tree.

Similar to the previous result, we let the expected number of active tags be $c\log(n)$ and let $m = \log(n)$ bits, which is the lower bound on m due to the population of n tags in the system. By Equation (1) of [1], for $n > 54$ the expected number of queries sent by the reader for identifying $c\log(n)$ tags, denoted by q, is bounded as $\mathrm{E}[q] \geq 2.881c\log(n) - 1$. We assume that the time-frames are of a fixed length of $\log_2(n)$, which is the minimum due to the signature length. Let N_b denote the total number of bits exchanged for identifying $c\log(n)$ tags. Since we are omitting the broadcasted queries from the reader in N_b, we obtain the following bound.

$$\mathrm{E}[N_b] \geq (2.881c\log(n) - 1)\log_2(n)$$

By (5), using a capacity achieving channel coding scheme, N_b bits can be transmitted in $\frac{N_b}{C}$ symbols. Hence, the expected delay of QueryTree, denoted by T_{QT}, satisfies

$$\mathrm{E}[T_{QT}] \geq \frac{5.762c\log^2(n) - \log(n)}{\log(2)\log\left(1+\mathrm{SNR}\right)}. \tag{6}$$

C. Fundamental Delay Limit of CSMA Schemes

Consider a CSMA scheme, where the number of active stations, denoted by k, are known at the detector and the frame-size is chosen optimally to minimize the detection delay. Since any other CSMA scheme without centralized coordination would achieve a delay at least as much as this scheme, the delay obtained by this optimal algorithm serves as a fundamental limit on CSMA approaches.

We formulate this scheme using dynamic programming. We use the expected delay of the scheme as the objective function $J(k,q)$ with the control parameter q, which denotes the choice of the number of time-slots. At the end of each frame, the size of the next frame is chosen optimally based on the number of undetected active stations. We use $q^*(k)$ and $J^*(k)$ to denote the optimal frame-size and corresponding minimum expected delay.

$$J^*(k) = \min_{q\geq 1}\left\{q + \sum_{l\leq k}\mathrm{P}(k\to l|q)J^*(l)\right\},$$

where $\mathrm{P}(k\to l|q)$ denotes the probability of detecting $(k-l)$ stations and reducing the number of remaining stations to l. In that event, the scheme would choose q accordingly to incur the expected delay of $J^*(l)$. Adding the size of the current frame and the expected delay of detecting the remaining active stations, we obtain the equation above. One caveat of this formulation is that the channel is implicitly assumed to be error-free and the only source of error is collision of multiple packets.

The main challenge of this problem is to compute $\mathrm{P}(k\to l|q)$ terms. We use the fact that the decision of a station on whether to use a particular time-slot or not is conditionally independent from its decisions for the rest of time-slots. Let $X(k,q)$ be the random variable that denotes the number of successful decoding in q time-slots and in the presence of k stations. This value can be decomposed as the following, where $p_i(k,q)$ is the probability of exactly i stations choosing a fixed time-slot.

$$p_i(k,q) = \binom{k}{i}\left(\frac{1}{q}\right)^i\left(1-\frac{1}{q}\right)^{(k-i)}$$

$$X(k,q) = \begin{cases} X(k,q-1), & \text{w.p. } p_0(k,q) \\ X(k-1,q-1), & \text{w.p. } p_1(k,q) \\ X(k-i,q-1), & \text{w.p. } p_i(k,q),\ \ k\geq i\geq 2 \end{cases}$$

Since $\mathrm{P}(k\to l|q) = \mathrm{P}(X(k,q) = (k-l))$, we can use the random variable X as a surrogate to compute the probabilities iteratively.

$$\begin{aligned}\mathrm{P}(X(k,q)=i) &= p_0(k,q)\mathrm{P}(X(k,q-1)=i) \\ &\quad + p_1(k,q)\mathrm{P}(X(k-1,q-1)=i-1) \\ &\quad + \sum_{j=2}^{k} p_j(k,q)\mathrm{P}(X(k-j,q-1)=i)\end{aligned}$$

We numerically computed the $(J^*(k), q^*(k))$ for $k \in \{1,...,300\}$ and obtained the optimal frame-size to be $q^*(k) = k$. Moreover, the optimal expected delay turned out to be an approximately linear curve with its slope rapidly converging to e as shown in Figure 2. Based on these

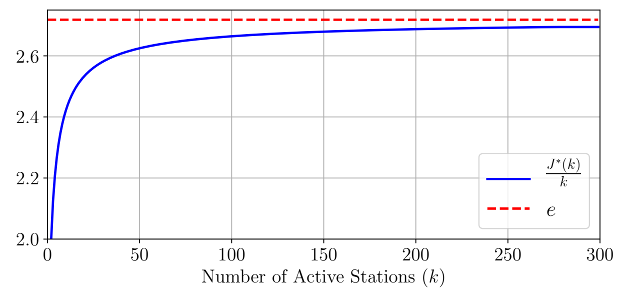


Figure 2: The slope of $J^*(k)$ converges to e with growing number of active stations.

observations, we present the following conjecture and use that expression in the following discussions.

Conjecture 1. *The expected delay of the CSMA scheme that chooses the frame-sizes optimally satisfies the following, where k denotes the number of active stations.*

$$\lim_{k \to \infty} |J^*(k) - ek| = 0 \tag{7}$$

Since we are interested in the scaling of this expression and the convergence is observed to occur rapidly, we approximate the expected delay cost of this scheme as $J^*(k) \approx ek$. Finally, we plug in $\mathrm{E}[k] = \log(n)$ and as in QueryTree analysis, we lower bound the duration of each time-slot with $\log(n)$. Letting T^*_{CSMA} denote the delay of this optimal scheme, we obtain a symbol-level expected delay as the following with the use of a capacity achieving coding.

$$\mathrm{E}[T^*_{CSMA}] \approx \frac{e \log^2(n)}{\log(2)\log(1+\text{SNR})}. \tag{8}$$

D. Discussion

Due to the common use of multi-station detection in large-scale applications, the scaling of delay performance with the station population plays an important role in the evaluation of multi-access schemes. Observe that the detection times of existing collision avoidance schemes scale with $\Omega(\log^2(n))$, while the proposed scheme achieves $\Theta\left(\frac{2c\log^2(n)-\log(\lceil 2f(n)\rceil!)}{\log(\log(n))}\right)$ with probability approaching 1 as $n \to \infty$.

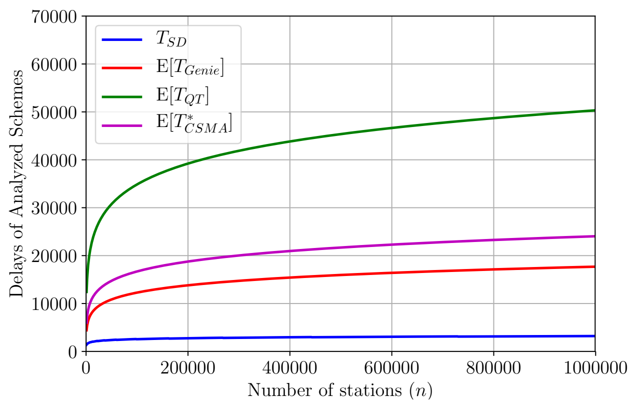


Figure 3: Presented expressions for delay performances for SNR=-15 dB

To demonstrate the difference between these scalings, we plot these expressions in Figure 3 for SNR $= -15$ dB. The gap between the delay of the proposed scheme and the derived fundamental limit of CSMA schemes increases significantly for large n. This suggests potentially significant improvements for large scale applications, which are common in many contexts that employ multi-station detection.

V. Implementation of Simultaneous Detection & Numerical Results

Despite its value for our theoretical work, ML-decoding is not feasible for practical implementations due to its computational complexity growing exponentially with n. Hence, we direct our attention to more applicable decoding methods.

Due to various other challenges in the decoding process, we assume that the number of active stations, denoted by k, is known at the decoder and omit the estimation of this parameter. Estimating the number of active stations has been studied extensively in the literature, especially in the context of RFID tag detection. For high SNR levels, even simpler methods like power detection can be used to accurately estimate k. We assume that the decoder employs such a method k is known.

A. Near-far Problem

In this paper we assumed that the transmissions are over an AWGN channel. This model can also be used to approximate problems concerning fading channels, where the gains of channels between stations and the detector are close to each other. In many applications, this assumption is not very unrealistic as long as the stations have a line-of-sight with the detector and are separated from it by approximately the same distance. The following discussion on the practical decoding methods assumes this structure. It is important to note that if this assumption can not be satisfied in an application, one might suffer from the near-far problem.

More specifically, due to factors like significant differences between the distances of stations from the detector and high variation among channel gains in the presence of fading channels, signals received from some of the stations might have much more power than the others. If not addressed, the unbalanced powers of signals might cause problems in the detection of k and especially complicate the detection of stations whose signals were received with relatively smaller power. In that case, additional steps like channel estimation might be introduced to the decoding methodology.

B. Multi-User Decoding Methods

As a feasible alternative to ML-decoding, the decoding methods proposed for multi-user decoding methods can be used with our scheme. One of the established applications of this framework is the multiple-input multiple-output (MIMO) decoding problems, which we use as an example to describe how multi-user detection can be adapted to be used with simultaneous detection scheme. One can interpret each stream in a MIMO channel as the transmission of a station, which replaces the channel gain coefficients in the original problem setting with the signature sequences of stations. Using this analogy, various MIMO decoding methods become feasible candidates for replacing ML-decoding to use with our scheme. Although some not straight-forward, decoding methods like minimum mean squared error decoding, decorrelator, matched filter [14], semidefinite relaxation and sphere decoding [15] can be adapted to be used in place of ML-decoding with simultaneous detection scheme. Interpreting channel gain as the codewords of same power brings the implicit assumption of having approximately the same channel gains for all stations. We omit the details of particular decoding methods since they fall out of the scope of this paper.

C. Decoding with Orthogonal Matching Pursuit

Another alternative is to consider compressed sensing methods for decoding under the sparse activation model. In this model, the state-vector $\mathbf{s}$ is a sparse vector and the decoding process aims to recover it from the vector $A\mathbf{s}$ with additive noise, which falls in the framework of compressed sensing and methods like orthogonal matching pursuit (OMP) [16] can be used to detect the set of active stations. Similar issues to near-far problem might appear with this method, too, since factors like channel fading are not considered in many of the compressed sensing problem formulations.

OMP is an influential iterative compressed sensing algorithm that greedily detects the set of non-zero indices of a sparse vector $\mathbf{s}$ from a noisy observation such as our model $\mathbf{y} = A\mathbf{s} + \mathbf{w}$ under the sparse activation model. The sparse signal to be decoded in our case is the state vector $\mathbf{s}$. OMP keeps a residue parameter $\mathbf{r}_t$, which is initialized as $\mathbf{r}_0 = \mathbf{y}$ and has a stopping threshold τ.

Let $A(S)$ denote the submatrix of A obtained by only keeping the set of columns of A in the index set S. At each iteration, the algorithm searches for the index of $\mathbf{s}$ that maximizes its projection onto $\mathbf{r}_t$, then subtracts this projection from the residue. The algorithm keeps executing these steps until $||r||_2 \leq \tau$ is achieved.

Since we assume that k is known and the choice of τ is not straight-forward, we use a variant of this algorithm, which we call OMP-k. Its only difference from the original OMP algorithm is the stopping rule, where instead of using a threshold we run the algorithm for exactly k iterations. The pseudo-code of this method is presented in Algorithm 1, where I denotes the identity vector and $< \mathbf{r}_t, \mathbf{x}_t >$ is the dot product of the two vectors.

Algorithm 1 OMP-k

Require: The active number of stations k
  $\mathbf{r}_1 = \mathbf{y}$ and $S = \emptyset$
  for $t = 1, 2, .., k$ **do**
    $x_t = \arg\max_{x_t \in \mathbf{x}} |< r_t, x_t >|$
    $S = S \cup \{x_t\}$
    $P_t = A(S)(A(S)^T A(S))^{-1} A(S)^T$
    $r_{t+1} = (I - P_t)y$
  end for
  return S

D. Numerical Investigations

1) Simulation Setting: We conducted numerical investigations under the sparse activation model, using OMP-k decoding for $k_{max} = 20$. In our simulations, we use the implementation of the algorithm in the scikit-learn [17] scientific computing software library. In order to capture the behavior of our policy in various settings we report results for a set of different SNR levels and span a large range of n with our simulations. As the performance metric for our investigations, we report the accuracy of the station detection, where an attempt is considered successful if the states of all stations are correctly detected. All reported results are obtained using random Gaussian codebooks and by averaging the results of 50 independent simulations.

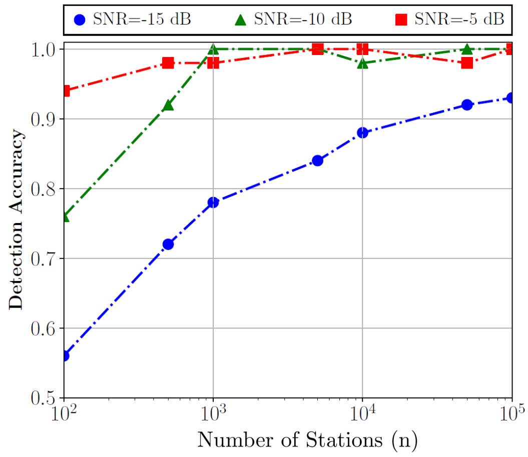


Figure 4: The accuracy of OMP-k decoding for m chosen as the lower bound in Corollary 1 under the sparse activation model

2) Numerical Results: We set m to be equal to the bound presented in Corollary 1, which is the minimum block-length that ensures vanishing probability of error. With this choice, we simulated the detection process for growing station population and various SNR levels. The results of this simulation is presented in Figure 4.

Note that despite the presented delay guarantees for the simultaneous detection scheme were asymptotic and with ML-decoding, accuracies very close to 1 are achieved with minimum choice of m for relatively small n values using OMP. This result shows the potential of the simultaneous detection scheme for practical applications. Note that one can obtain even higher accuracies for such small station populations by setting m larger than the lower bound and slightly sacrificing from the delay.

VI. Discussion & Conclusion

In this paper we introduced a multiple access scheme for the simultaneous detection of mobile stations. Unlike the vast majority of existing approaches, we embrace collisions rather than trying to avoid them. We presented both finite-regime and asymptotic theoretical guarantees for this scheme concerning its delay performance and error probability. Another contribution of this work is the analysis of collision avoidance schemes and derivation of fundamental limits on their delay performances. Using these results, we showed that the proposed simultaneous scheme yields more desirable scaling of detection delay compared to the any delay achievable by a collision avoidance scheme. Finally, we discussed the practical challenges and presented lower complexity variants of our decoding method and presented numerical investigations showing the applicability and high performance of the proposed scheme.

REFERENCES

[1] Ching Law, Kayi Lee, and Kai-Yeung Siu. Efficient memoryless protocol for tag identification. In *Proc. of the 4th Int. Workshop on Discrete Algorithms and Methods for Mobile Computing and Communications*, DIALM '00, pages 75–84, New York, NY, USA, 2000. ACM.

[2] Harald Vogt. *Efficient Object Identification with Passive RFID Tags*, pages 98–113. Springer Berlin Heidelberg, Berlin, Heidelberg, 2002.

[3] EPCglobal. EPC radio-frequency identity protocols class-1 generation-2 UHF RFID protocol for communications at 860 MHz - 960 MHz version 2.0.1, 2015.

[4] Gianluca Carroccia and Gaia Maselli. Inducing collisions for fast rfid tag identification. *IEEE Communications Letters*, 19(10):1838–1841, 2015.

[5] Christoph Angerer, Robert Langwieser, and Markus Rupp. Rfid reader receivers for physical layer collision recovery. *IEEE Transactions on Communications*, 58(12):3526–3537, December 2010.

[6] Jelena Kaitovic and Markus Rupp. Improved physical layer collision recovery receivers for rfid readers. In *2014 IEEE Int. Conf. on RFID (IEEE RFID)*, pages 103–109. IEEE, 2014.

[7] Pan Hu, Pengyu Zhang, and Deepak Ganesan. Laissez-faire: Fully asymmetric backscatter communication. In *ACM SIGCOMM Computer Communication Review*, volume 45, pages 255–267. ACM, 2015.

[8] Jiajue Ou, Mo Li, and Yuanqing Zheng. Come and be served: Parallel decoding for cots rfid tags. In *Proc. of the 21st Annual Int. Conf. on Mobile Computing and Networking*, pages 500–511. ACM, 2015.

[9] Linghe Kong, Liang He, Yu Gu, Min-You Wu, and Tian He. A parallel identification protocol for rfid systems. In *INFOCOM, 2014 Proceedings IEEE*, pages 154–162. IEEE, 2014.

[10] Jue Wang, Haitham Hassanieh, Dina Katabi, and Piotr Indyk. Efficient and reliable low-power backscatter networks. In *Proc. of the ACM SIGCOMM 2012 Conf. on Applications, Technologies, Architectures, and Protocols for Computer Communication*, pages 61–72. ACM, 2012.

[11] Xu Chen, Tsung-Yi Chen, and Dongning Guo. Capacity of gaussian many-access channels. *IEEE Trans. on Information Theory*, 63(6):3516–3539, June 2017.

[12] John M. Wozencraft and Irwin M. Jacobs. *Principles of Communication Engineering*. John Wiley & Sons Inc, 1966.

[13] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley Series in Telecommunications and Signal Processing. Wiley-Interscience, 2nd ed edition, 2006.

[14] David Tse and Pramod Viswanath. *Fundamentals of wireless communication*. Cambridge university press, 2005.

[15] Joakim Jalden. *Maximum likelihood detection for the linear MIMO channel*. PhD thesis, 2004.

[16] Yagyensh Chandra Pati, Ramin Rezaiifar, and Perinkulam Sambamurthy Krishnaprasad. Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition. In *Twenty-Seventh Asilomar Conf. on Signals, Systems and Computers*, pages 40–44. IEEE, 1993.

[17] Fabian Pedregosa et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research (JMLR)*, 12:2825–2830, 2011.

APPENDIX A
PROOF OF THEOREM 1

Due to the k_{max} bound, we have $N_{|\mathbf{s}^{(i)}-\mathbf{s}^{(j)}|} \leq 2k_{max}$. Therefore, the probability of making errors with more than $2k_{max}$ mismatches is zero under this scheme. By Lemma 1 and by $\frac{n^k}{k!} \geq \binom{n}{k}$, we obtain the following.

$$\mathrm{P}(\mathcal{E}) \leq \sum_{k=1}^{2k_{max}} \frac{n^k}{k!}\left(1+\frac{kP}{2N_0}\right)^{-\frac{m}{2}}$$

Since this is a finite sum, it suffices to show that each summand of the bound vanishes to show that $\mathrm{P}(\mathcal{E}) \to 0$. Hence, following is an equivalent condition.

$$\lim_{n\to\infty}\left\{\frac{n^k}{k!}\left(1+\frac{kP}{2N_0}\right)^{-\frac{m}{2}}\right\} = 0,\ \forall k \leq 2k_{max} \tag{9}$$

By taking the logarithm of both sides of (9), we obtain the following.

$$\lim_{n\to\infty}\left\{k\log(n) - \log(k!) - \frac{m}{2}\log\left(1+\frac{kP}{2N_0}\right)\right\} = -\infty$$

The following condition on the choice of m as a function of n, ensures the stated condition above is satisfied.

$$m(n) \geq 2\frac{k\log(n)-\log(k!)}{\log\left(1+\frac{kP}{2N_0}\right)} + g(n),$$

where $g(n)$ is any monotonically increasing non-negative function of n. If this condition is satisfied for $k = 2k_{max}$, then it holds for any other feasible k. Finally, setting $k = 2k_{max}$ and reorganizing terms yield the desired result. $\square$.

APPENDIX B
PROOF OF THEOREM 2

Let $\epsilon > 0$. We partition the event $\mathcal{E}$ over the number of active stations, denoted by $N_{\mathbf{s}}$.

$$\begin{aligned}\mathrm{P}(\mathcal{E}) &= \mathrm{P}(\mathcal{E}|N_{\mathbf{s}} \geq np(1+\epsilon))\mathrm{P}(N_{\mathbf{s}} \geq np(1+\epsilon)) \\ &+ \mathrm{P}(\mathcal{E}|N_{\mathbf{s}} < np(1+\epsilon))\mathrm{P}(N_{\mathbf{s}} < np(1+\epsilon))\end{aligned}$$

Bounding the first and fourth terms of the right hand-side by 1, we end up with the following bound. Then, we show that both terms on the bound vanish as $n\to\infty$.

$$\mathrm{P}(\mathcal{E}) \leq \mathrm{P}(N_{\mathbf{s}} \geq np(1+\epsilon)) + \mathrm{P}(\mathcal{E}|N_{\mathbf{s}} < np(1+\epsilon))$$

Since s_i are i.i.d. Bernoulli random variables with parameter $p = \frac{f(n)}{n}$, we can use Chernoff-Hoeffding bound.

$$\mathrm{P}(N_{\mathbf{s}} \geq np(1+\epsilon)) \leq e^{-\frac{\epsilon^2 np}{3}} = e^{-\frac{\epsilon^2 f(n)}{3}}$$

Therefore, $\lim_{n\to\infty}\mathrm{P}(N_{\mathbf{s}} \geq np(1+\epsilon)) = 0$.

To bound $\mathrm{P}(\mathcal{E}|N_{\mathbf{s}} < np(1+\epsilon))$, we use (2) with the fact that $\frac{n^k}{k!} \geq \binom{n}{k}$.

$$\begin{aligned}\mathrm{P}(\mathcal{E}|N_{\mathbf{s}} < np(1+\epsilon)) &\leq \sum_{k=1}^{2\lfloor f(n)(1+\epsilon)\rfloor} \frac{n^k}{k!}\left(1+\frac{kP}{2N_0}\right)^{-\frac{m}{2}} \\ &\leq 2f(n)(1+\epsilon)\max_k\left\{\frac{n^k}{k!}\left(1+\frac{kP}{2N_0}\right)^{-\frac{m}{2}}\right\}\end{aligned}$$

Instead of finding the maximizing k value for this bound, we choose m such that this bound vanishes for all k. Similar to (9), we require the following.

$$\lim_{n\to\infty}\left\{2f(n)(1+\epsilon)\frac{n^k}{k!}\left(1+\frac{kP}{2N_0}\right)^{-\frac{m}{2}}\right\} = 0,\ \forall k$$

Since this expression holds for arbitrarily small $\epsilon > 0$, we can drop the $(1+\epsilon)$ coefficient from our expression by choosing it small enough. Taking the logarithm of both sides yields the following condition.

$$m(n) \geq 2\frac{\log(2f(n)) + k\log(n) - \log(k!)}{\log\left(1+\frac{k}{4}\mathrm{SNR}\right)} + g(n)$$

where $g(n)$ is a monotonically increasing nonnegative function of n. Setting $k = \lceil 2f(n)\rceil$ and reorganizing terms complete the proof. $\square$