Probabilistic Rateless Multiple Access for Machine-to-Machine Communication

Mahyar Shirvanimoghaddam, Member, IEEE, Yonghui Li, Senior Member, IEEE, Mischa Dohler, Fellow, IEEE, and Branka Vucetic, Fellow, IEEE

Abstract—Future machine to machine (M2M) communications need to support a massive number of devices communicating with each other with little or no human intervention. Random access techniques were originally proposed to enable M2M multiple access, but suffer from severe congestion and access delay in an M2M system with a large number of devices. In this paper, we propose a novel multiple access scheme for M2M communications based on the capacity-approaching analog fountain code to efficiently minimize the access delay and satisfy the delay requirement for each device. This is achieved by allowing M2M devices to transmit at the same time on the same channel in an optimal probabilistic manner based on their individual delay requirements. Simulation results show that the proposed scheme achieves a near optimal rate performance and at the same time guarantees the delay requirements of the devices. We further propose a simple random access strategy and characterize the required overhead. Simulation results show that the proposed approach significantly outperforms the existing random access schemes currently used in long term evolution advanced (LTE-A) standard in terms of the access delay.

Index Terms—Analog fountain codes, belief propagation, machine-to-machine communication, massive multiple access.

I. INTRODUCTION

Machine-to-Machine communications have emerged as a promising technology to enable trillions of multirole devices, namely machine-type communication (MTC) devices, to communicate with each other with little or no human intervention [1, 2]. It has many potential applications, such as intelligent transportation systems (ITS), healthcare monitoring, retail, banking, smart grids, home automation and so on. It is expected that in the next a few years, over 2 billion MTC devices will become directly attached to cellular networks to provide M2M communications [3]. Thus, there will be a massive number of MTC devices with no/low mobility [4] in each cell, which is significantly more than the number of users in current cellular networks. Moreover, M2M traffic involves a large number of short-lived sessions, attempting to deliver a small amount of data (few hundred bits) to the base station, which is quite different from those in human-to-human (H2H) communications. Such differences motivate researchers around the globe to optimize the current cellular networks to effectively enable M2M communications [5, 6].

Manuscript received March 13, 2015; accepted July 9, 2015. The material in this paper was presented in part at the 2014 IEEE International Symposium on Information Theory, Honolulu, HI.

M. Shirvanimoghaddam, Y. Li, and B. Vucetic are with the Center of Excellence in Telecommunications, School of Electrical and Information Engineering, The University of Sydney, Sydney, NSW 2006, Australia (e-mail: {mahyar.shirvanimoghaddam; yonghui.li; branka.vucetic}@sydney.edu.au).

M. Dohler is with King’s College London (e-mail: mischa.dohler@kcl.ac.uk).

In most existing wireless access networks, the first step in establishing an air interface connection, is to perform random access (RA) in a contention manner. Such a random access approach will not work effectively when the number of nodes is very large, due to frequent transmission collisions, which leads to network congestion, unexpected delays, packet loss, high energy consumption, and radio resource wastage [7]. In current RA approaches, when two or more devices select the same RA preamble, a collision will occur and the respective devices will not be scheduled for data transmission. To reduce the access delay in M2M communications, several schemes have been proposed, such as dynamic allocation [8, 9], slotted access, group-based [1, 10], pull-based, and access class barring [11, 12]. Although these approaches can reduce the access collisions to a certain degree, most of them still suffer from very high access delays in highly dense networks. The main idea behind these schemes is to delay the retransmission of the access request for a random/fixed amount of time, thus increasing the access probability within a relatively short time. This is however inefficient in M2M communications due to small short burst transmissions of devices which mainly do not require the whole RB for their transmissions. This means that for M2M communications with a very large number of devices there might not be enough RBs to be orthogonally allocated to the devices, which significantly increase the access delay even if the random access requests are delivered correctly.

Additionally, different MTC devices have diverse service requirements and traffic patterns. Generally, we can divide MTC traffic types into four different categories. The first type is the alarm traffic, which is completely random and its probability is very low; however, it has a very strict delay requirement. The second traffic type can be modeled by a random Poisson distribution with the parameters depending on the application [13]. The regular traffic, such as smart metering applications, is the third traffic type, and the last type is the streaming, like video surveillance applications. Current proposals for enabling M2M communications, did not consider the priorities among devices and different quality of service (QoS) requirements. These approaches have been generally designed for a fixed payload size and thus, cannot support M2M applications with different service requirements.

Recently, a systematic framework has been developed in [13, 14] to understand the fundamental limits of M2M communications in terms of power efficiency and throughput. Here, we consider a realistic model for M2M communications, which supports both regular and random traffics with different delay and service requirements. We develop a practical
transmission scheme for M2M communications and show that it can closely approach the fundamental throughput limit and satisfy the delay requirements of all devices. The main contributions of this paper are summarized next.

B. Contributions

1) Efficient Random Access Proposal: We consider the contention-based random access in the proposed probabilistic multiple access for M2M communications. In the contention process, each device transmits an RA preamble to the BS to establish a communication link with the BS. Unlike existing RA schemes, we group RA preambles into several sub-sets according to delay requirements of MTC devices. Then, a sub-set of RA preambles is assigned for those devices which have the same delay requirement. In the contention period, each device selects a specific RA preamble based on its delay requirement from the respective subset of RA preambles. By detecting the RA preamble of the device, the BS then knows which subset this RA preamble belongs to and thus knows its delay requirement. The BS detects the number of devices which have selected the same RA preamble and broadcasts this information to the devices. Based on this information, each device will select a random binary sequence with an appropriate length to share with the BS to construct the same code graphs in both devices and the BS. The devices which have selected the same RA preamble then transmit in the same RB, and their identities will be transmitted along with their payload data and will be later recognized by the BS after the decoding. In this way, each available RB will be effectively used by a number of devices and RA collisions is handled in an efficient manner to simultaneously support collided devices.

2) Efficient Maximum Throughput Transmission: Our second contribution in this work is to represent the multiple access process as a special kind of analog graph code. This enables us to draw on the recently proposed capacity approaching analog fountain codes [15] to design the optimal access protocol achieving the maximum common rate of a multiple access channel.

Fountain codes have been used to facilitate the multiple access transmissions over erasure channels [16–19], where collided packets are used in an effective way to increase the overall system throughput. In [20], it has been shown that the capacity of a Gaussian multiple access channel can be achieved by using spatially coupled sparse graph multi-user modulation with an iterative interference cancelation scheme at the receiver. However, this approach performs decoding and interference cancelation separately and its complexity increases significantly as the number of access users increases. To overcome this problem, we proposed a simple capacity-approaching analog fountain code (AFC) in [15] with linear complexity in both the encoder and decoder. Thanks to the interesting linear property of AFCs, the summation of AFC coded signals of multiple simultaneous transmitted users still forms an AFC code [21]. Thus the receiver can represent the multiple access transmission by a special kind of graph codes, and the BS can jointly decode them by using a standard belief propagation (BP) decoder. The proposed probabilistic multiple access AFC (MA-AFC) scheme enables the devices to transmit in a probabilistic manner in a way to form an equivalent AFC code at the BS, and approach the maximum common rate of the multiple access channel [21].

3) QoS aware Random Transmission: Different MTC devices have diverse traffic and service requirements. Existing random access schemes do not provide QoS guarantees in the random access transmissions. This precludes use in many practical M2M systems, where timing constraints are critical. How to alleviate access congestions in massive access transmissions while providing delay or QoS guarantees remains a significant hurdle in M2M access network design. In the proposed approach, the access probabilities of devices are optimized such that the delay requirement for each device is satisfied. We formulate an optimization problem based on the density evolution analysis of the AFC code to find the optimum code degree and access probability for each device to satisfy their delay requirements. We further show that the proposed approach can simultaneously satisfy the delay requirement for all devices.

Throughout the paper the following notations are used. We use a boldface small letter to denote a vector and the \(i^{th}\) entry of vector \(\mathbf{v}\) is denoted by \(v_i\). A boldface capital letter is used to denote a matrix and the \(i^{th}\) element in the \(j^{th}\) row of matrix \(\mathbf{G}\) is denoted by \(g_{i,j}\). The transpose of matrix \(\mathbf{G}\) is denoted by \(\mathbf{G}'\). A set is denoted by a curly capital letter and the cardinality of set \(S\) is represented by \(|S|\).

The rest of the paper is organized as follows. In Section II, we present the system model. The proposed probabilistic multiple access scheme for M2M communications is presented in Section III. Section IV presents the asymptotic analysis of the belief propagation decoding of the MA-AFC scheme based on the density evolution analysis. In Section V, an optimization problem is formulated to find the optimum code parameters. Simulation results are shown in Section VI, and finally Section VII concludes the paper.

II. SYSTEM MODEL

In this section, we first introduce the network and channel model in M2M communications. Then we discuss in more details the packet transmission and delay model for MTC devices.

A. Network Model and Random Access

We consider an M2M last mile access system, where \(N\) MTC devices are uniformly distributed in a single cell of radius \(r_0\) and communicate with a base station (BS) located at the origin. In this paper, we focus on M2M applications where MTC devices are deployed at fixed locations, such as smart meters, sensors or cameras in buildings, roads, bridges, etc. [22].

It is assumed that the channel between each device and the BS is a slow time-varying block fading channel, for which the channel remains constant within one transmission block but varies slowly from one block to the other. We consider a time division duplex (TDD)-based wireless access system, where the channel gain of the uplink is assumed to be the same as
that of the downlink [23]. With this assumption, each device can estimate the uplink channel gain from the pilot signal sent periodically over the downlink channel by the BS. Using the pilot signal, MTC devices also synchronize their timing to that of the BS. The BS, however, does not have the knowledge of any channel state information (CSI). This assumption is particularly relevant in M2M communications, where due to a large number of devices, it would be impractical for the BS to obtain CSI to every MTC device [24]. The channel between each device and the BS is modeled by path loss, large scale shadowing and small scale fading effects. The received power at the BS for the signal transmitted from device $U_i$ is given by:

$$P_{r,i} = P_{t,i} |h_i|^2,$$

where $P_{t,i}$ is the transmit power, $\gamma$ is the reference signal to noise ratio (SNR), defined as the average received SNR from a device transmitting at maximum power located at the cell edge, $h_i$ is the effective channel gain, defined as $|h_i|^2 = \chi_i r_i (\gamma_0 / \eta_i)^{\alpha}$, $r_i$ is the distance from $U_i$ to the BS, $\chi_i$ is a log-normal random variable representing large scale fading, $\gamma_0$ is the reference SNR, $\eta_i$ is small scale fading modeled by exponential random variable with parameter 1, and $\alpha$ is the propagation loss factor. We further assume that the devices perform uplink power control such that the average received SNR from all devices at the BS are the same [13] and equals to $\gamma_0$. Such an assumption has been considered for code division multiple access (CDMA) in [14].

Time is divided into slots of duration $\tau_s$ secs and total available bandwidth is $W$ Hz. Each time slot is referred to as a resource block (RB), where a fraction $\tau$, $0 < \tau < 1$, of the resource block is reserved for the random access procedure as shown in Fig. 1. Similar to [13], we assume that the number of active devices in each RB of duration $\tau_s$ is random and follows a Poisson process with rate $\lambda$. We also consider a contention-based random access strategy in M2M communications, where $N_s$ different random access preambles, which are orthogonal frequency division multiplexing (OFDM) symbols in the LTE-Advanced standard [7], are selected in a random access attempt. The BS also assigns an access probability to each device according to its delay requirement. The access probabilities will be later optimized in this paper to satisfy the delay requirement of all devices. The notation used in this paper is summarized in Table I for quick reference.

It is important to note that in many M2M applications, M2M and human-to-human (H2H) communication coexists. H2H users have high priority to obtain a connection to transmit their data to the BS. Most of research work has considered the contention-free random access for H2H users, that is the BS assigns a preamble and an RB to the H2H user in a timely manner [25]. In this work we only focus on M2M devices, where opportunistically contend for RBs through a contention-based random access, and H2H users are assumed to have access to the BS through the allocated RBs.

### B. Delay Model for MTC devices

In M2M communications, timing constraints for MTC devices typically range from 10 ms to several minutes due to the infrequent transmission features [1]. For instance, some M2M multimedia applications such as video surveillance systems with strict timing constraints have the delay requirement ranging from 10 ms to 40 ms. And smart meters usually transmit their data every several minutes or even every several hours. In Table II, we divide different M2M applications based on their delay requirements and traffic patterns, according to the 3rd Generation Partnership Project (3GPP) M2M standard [1]. More specifically, we divide the MTC devices into $N_d = 10$ groups based on their applications, where the maximum affordable delay in group $i$ is denoted by $t_i$.

As MTC devices with regular traffic have predictable activities and can be completely coordinated by the BS [22], we only focus on M2M devices with random activities with irregular traffic patterns in this paper. Each active MTC device $U_i$, $i \in \{1, 2, \ldots, N\}$, is assumed to have $k$ information symbols to transmit within a maximum delay of $t_j$, where $1 \leq j \leq N_d$. That is, these $k$ information symbols must be delivered to the BS within $t_j$ seconds after $U_i$ starts transmitting this message, i.e., $U_i$ belongs to the $j$th delay group. We also assume that the $i$th device has the packet arrival rate $\lambda_i$, that is in each RB, $\lambda_i$ packets are generated at the $i$th device, where each packet has to be delivered to the BS no later than $t_j$ seconds after its generation. Let us assume that $1/\lambda_i > t_j$. Therefore, if the previous packet has been delivered within maximum delay $t_j$, the new packet will be immediately sent by the device and there is no need to buffer it. On the other hand, if $1/\lambda_i < t_j$, we set the delay requirement of the device to $1/\lambda_i$ to avoid any
TABLE II
SEVERAL M2M APPLICATIONS AND THEIR DELAY CONSTRAINTS.

<table>
<thead>
<tr>
<th>Group</th>
<th>Delay Constraint</th>
<th>Example of Application</th>
<th>Traffic Type</th>
<th>Message Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>10 ms</td>
<td>Emergency Alarm [26]</td>
<td>Very Unlikely</td>
<td>Very small</td>
</tr>
<tr>
<td>2</td>
<td>20 ms</td>
<td>Intelligent transport system (ITS)</td>
<td>Regular/Irregular</td>
<td>Medium</td>
</tr>
<tr>
<td>3</td>
<td>40 ms</td>
<td>Streaming</td>
<td>Regular</td>
<td>Medium</td>
</tr>
<tr>
<td>4</td>
<td>100 ms</td>
<td>Control messages[26]</td>
<td>Regular</td>
<td>Very small</td>
</tr>
<tr>
<td>5</td>
<td>500 ms</td>
<td>eHealth</td>
<td>Regular</td>
<td>Medium</td>
</tr>
<tr>
<td>6</td>
<td>1 s</td>
<td>Monitoring the Devices [26]</td>
<td>Regular/Irregular</td>
<td>Medium</td>
</tr>
<tr>
<td>7</td>
<td>10 s</td>
<td>eHealth</td>
<td>Regular/Irregular</td>
<td>Medium</td>
</tr>
<tr>
<td>8</td>
<td>100 s</td>
<td>eHealth</td>
<td>Regular/Irregular</td>
<td>Medium</td>
</tr>
<tr>
<td>9</td>
<td>500 s</td>
<td>Smart meters</td>
<td>Regular</td>
<td>Medium</td>
</tr>
<tr>
<td>10</td>
<td>&gt;1000 s</td>
<td>Smart meters</td>
<td>Regular</td>
<td>Medium</td>
</tr>
</tbody>
</table>

Further delay due to queuing. It is also important to note that for devices with limited buffer size, the device can change the access probability according to the number of packets in its buffer. However, a packet in M2M communications has a small size (few bytes); therefore, we can assume that all the packets in the buffer of an MTC device can be transmitted in one RB [25]. Therefore in the rest of the paper, we only use the term delay requirement for each MTC device, and assume that the queuing delay has been taken into account when calculating the appropriate delay requirement of each device.

III. THE PROPOSED RATELESS PROBABILISTIC MULTIPLE ACCESS FOR M2M COMMUNICATIONS

In this section, we develop an efficient multiple access scheme for M2M communications based on analog fountain codes. For this aim, we first briefly introduce AFC codes. Then, we present the rateless probabilistic multiple access for M2M communications, referred to as multiple access analog fountain coding (MA-AFC).

A. Analog Fountain Codes

Analog fountain codes (AFC) were originally proposed in [15, 27] as an effective adaptive transmission scheme to approach the capacity of Gaussian channels. In AFC, the entire message of length \( k \) binary symbols is first modulated using binary phase shift keying (BPSK) modulation to obtain \( k \) modulated information symbols, \( b_i \in \{-1, 1\} \), where \( i \in \mathbb{Z}_k^1 \). Then, according to the predefined degree distribution function, \( d_c \) randomly selected modulated information symbols are linearly combined with real weight coefficients to generate one coded symbol. For simplicity, we assume that the code degree, \( d_c \), is fixed and weight coefficients are chosen from a finite weight set, \( W_s \) with \( D \) positive real members, as follows:

\[
W_s = \{w_i \in \mathbb{R}^+ | i \in \mathbb{Z}_D^1 \},
\]

where \( \mathbb{R}^+ \) is the set of positive real numbers. Let us assume that \( b \) is a BPSK modulated vector of dimension \( k \times 1 \), and \( G \) is the generator matrix of dimension \( m \times k \), then AFC coded symbols, \( c \), are generated as follows:

\[
c = Gb,
\]

where \( m \) is the number of coded symbols, only \( d_c \) elements of each row of matrix \( G \) are nonzero, and each nonzero element of \( G \) is randomly chosen from the weight set \( W_s \). To further enhance the performance, in [15] we proposed to use a high-rate precoder to encode the original data before applying the AFC code. To fully utilize the constellation plane, i.e., both in phase and quadrature phases, two consecutive coded symbols compose one signal as \( c_i + \sqrt{-1} c_{i+1} \). We also use the standard BP decoder to decode AFC codes. Further details of AFC encoding and decoding can be found in [15, 21].

B. Multiple Access AFC for M2M communications

The proposed probabilistic multiple access AFC (MA-AFC) scheme contains two steps. In the first step, namely contention phase, the random access requests are sent by the devices to the BS and a connection is established between the devices and the BS. In the second phase, namely data transmission phase, the actual payload data along with the device identity (ID) is transmitted by the devices to the BS.

1) The contention phase (CP): To apply AFC codes to M2M communication systems and satisfy delay requirements of all devices, we need to slightly modify the random access procedure in LTE-A [7]. The BS first partitions \( N_s \) RA preambles into \( N_d \leq N_s \) subsets, according to delay constraints of MTC devices, and each group of RA preambles are assigned to one delay group. The BS then broadcasts this information to all devices. Let \( S_i \) denote the set of RA preambles allocated to the \( i \)th delay group, then an MTC device with delay constraint \( d_i \) will randomly select an RA from \( S_i \) in the contention period. By detecting the RA preamble of the device, the BS will know which subset this RA preamble belongs to and thus know its delay requirement.

In the first step of CP, an MTC device with delay constraint \( d_i \) randomly selects an RA preamble from set \( S_i \) and transmits it to the BS. Let \( s \in S_i \) denote the selected RA preamble, then, it is possible that more than one devices with the same delay constraint \( d_i \) have chosen \( s \). As we assume that all MTC devices perform power control, such that the received signals from different devices will have the same received power, \( \gamma_0 \), the received power at the BS for the RA preamble \( s \) will be proportional to the number of devices which have selected RA preamble \( s \). Let \( \mu(s) \) denote the set of MTC devices, which have selected the same RA preamble \( s \). Since we assume that the RA preambles are orthogonal, the received signal corresponding to RA preamble \( s \) is given by:

\[
y = |\mu(s)|\sqrt{\gamma_0} s + z,
\]

where \( z \) is the additive white Gaussian noise (AWGN) vector. The average received signal power at the BS can then be calculated as \( |\mu(s)|^2\gamma_0 \). Since the BS knows \( \gamma_0 \), it can easily estimate \( |\mu(s)|^2 \).

In the second step of CP, the BS transmits a random access response (RAR) message including the number of devices which have selected each RA preamble, the allocated RB for each detected RA preamble, and timing information and optimized access probabilities. In the third step of CP, each device first calculates the length of a random seed based on
Let $p_{j}$ denote the access probability assigned to $U_j$. It is important to note that devices which have selected the same RA preamble will be assigned with the same access probability. In the data transmission phase, in each time instant $\ell$, device $U_j$ generates a binary random number, $I_{j,\ell}$, which is one with probability $p_{j}$, i.e., $p(I_{j,\ell} = 1) = 1 - p(I_{j,\ell} = 0) = p_{j}$. If this random number is one, then $U_j$ generates an AFC coded symbol, $u^{(j)}_{\ell}$ and sends it to the BS. The BS also generates the same binary random number as the one in $U_j$ as they have shared the same random seed, thus it already knows in which time instants device $U_j$ is sending a coded symbol. Using this random seed, both $U_j$ and the BS can construct the same AFC code structure, so the BS is able to perform the decoding on the received coded symbols.

For simplicity, we assume that devices use the same code degree $d_c$ and the same weight set $\mathcal{W}$ to generate AFC coded symbols. Let $\mathcal{U}$ denote the set of devices which are currently active and transmitting AFC coded symbols to the BS based on their access probabilities in the same RB. Then, the received signal at the BS in time instant $\ell$ is given by

$$y_{\ell} = \sum_{j \in \mathcal{U}} I_{j,\ell} h_j u^{(j)}_{\ell} + z_{\ell},$$

where $z_{\ell}$ is the AWGN with variance $\sigma^2$. As can be seen in (7), the received signal at the BS is the noisy version of the sum of coded symbols of devices in $\mathcal{U}$. Since coded symbols of each device have been generated by a linear combination of modulated information symbols, the received signal at the BS can be realized as a coded symbol of an analog fountain code, where modulated information symbols are from devices in $\mathcal{U}$. By using (3), (7) can be rewritten as follows:

$$y_{\ell} = \sum_{j \in \mathcal{U}} h_j I_{j,\ell} \sum_{r=1}^{k} g_{j,r}^{(j)} I_{\ell,r} + z_{\ell},$$

where $g_{j,r}^{(j)}$ is the $r^{th}$ modulated information symbol of $U_j$ and $g_{j,r}^{(j)}$ is the respective weight coefficient. Fig. 3 and Fig. 4 show the original and equivalent bipartite graphs of the AFC codes at the BS, respectively, when there are only two devices in $S$, $d_c = 2$, $I_{1,1} = I_{2,1} = 1$, and $I_{1,2} = I_{2,2} = 1$. As shown in these figures, the equivalent MA-AFC bipartite graph can be realized as a bipartite graph of an equivalent AFC code where the information symbols are from both users and weight coefficients are multiplied by the respective channel gains.

Let us define $X_{j,r}^{(\ell)} \triangleq I_{j,\ell} g_{j,r}^{(j)}$. Then its probability distribution is given by:

$$p(X_{j,r}^{(\ell)} = s) = \begin{cases} 1 - \frac{p_{j} d_{c}}{2K} & \text{if } s = 0 \\ \frac{p_{j} d_{c}}{2K} & \text{if } s \in \mathcal{W}, \end{cases}$$

and its mean and variance can be calculated as follows:

$$m_j = \mathbb{E}[X_{j,r}^{(\ell)}] = \sum_{s \in \mathcal{W}} \frac{p_{j} d_{c}}{2K} (s - s) = 0,$$

$$\sigma_{j}^{2} = \mathbb{E}[X_{j,r}^{(\ell)}^{2}] = \sum_{s \in \mathcal{W}} \frac{p_{j} d_{c}}{2K} s^{2} = \frac{1}{K} p_{j} d_{c} \sigma_{w}^{2},$$

where $\sigma_{w}^{2} = \frac{1}{K} \sum_{d_{c}=1}^{D} w_{d_{c}}^{2}$. Since in M2M communications the number of devices are very large and $X_{j,r}^{(\ell)}$’s are independent random variables, according to the Central Limit Theorem $\sum_{j \in \mathcal{U}} h_{j} \sum_{\ell=1}^{K} X_{j,\ell}$ will approximately follow a zero mean Gaussian distribution with the following variance:

$$\sigma^{2} = d_{c} \sigma_{w}^{2} \sum_{j \in \mathcal{U}} p_{j} |h_{j}|^{2}.$$
B. Asymptotic Performance Analysis of Multiple Access AFC based on the BP Decoding

Let us refer to information symbols of device $U_i$ as Type-$X_i$ variable nodes in the bipartite graph of MA-AFC codes for $i \in \mathbb{Z}_N^N$. Each received coded symbol at the BS may connect to modulated information symbols of various sets of devices. A simple way to represent coded symbols at the BS is to divide them into different types according to their connections to different sets of devices [29, 30]. For simplicity, we represent each non-empty set of devices by a vector $v$ of dimension $N$, where its $i^{th}$ entry, $v_i$, is 1 if $U_i$ belongs to the set; otherwise, it is zero. Here, we refer to coded symbols which are connected to the modulated information symbols of set $v$ as Type-$v$ check nodes. It is easy to show that the probability that a coded symbol is of Type-$v$, $q_v$, is given by:

$$q_v = \prod_{i=1}^{N} p_i^{v_i} (1 - p_i)^{1 - v_i},$$

and there exist $2^N - 1$ various types of check nodes, since the total number of nonempty subsets of devices is $2^N - 1$. In the BP decoder, messages, i.e., log likelihood ratios (LLRs), are passed between variable and check nodes in an iterative manner. The following lemma then gives the mean and variance of the messages between different types of variable and check nodes in each iteration of the BP decoder in an asymptotic case, when the number of variable and check nodes go to infinity.

Lemma 1: Let us consider an M2M system consisting of $N$ MTC devices wanting to simultaneously transmit their messages to a common BS. Each device $U_i$ is assigned with an access probability $p_i$, and code degree $d_i$. Let $t_i^{(v)}$ denote the message passed from a Type-$X_i$ variable node to a Type-$v$ check node in the $i^{th}$ iteration of the BP decoding algorithm at the BS. Then $t_i^{(v)}$ can be approximated by a Gaussian distribution with mean $m^{(v)}$ and variance $2m^{(v)}$, where $m^{(v)}$ can be calculated as follows:

$$m^{(v)} = \frac{e}{N} \log_2 (1 + N \gamma_0) \quad \text{bps/Hz},$$

where $d_v = md_v/k$, $\sigma_2^2 = \frac{1}{D} \sum_{i=1}^{D} w_i^2$, and $\Omega(x) = \frac{\gamma_0}{2x} \int_{-\infty}^{\infty} (1 - \tanh(x - y\sqrt{2})) \exp \left(-y^2 / 2\right) dy$. The proof of this lemma is provided in Appendix A. The BER for the $i^{th}$ device after $i$ iterations of the BP decoder, denoted by $p_{e,i}$, can then be calculated as follows:

$$p_{e,i}^{(t)} = Q \left( \sqrt{m^{(t)}} / 2 \right),$$

where $Q(x) = \frac{1}{\sqrt{\pi}} \int_{x}^{\infty} \exp \left(-z^2 / 2\right) dz$.

C. How much overhead is required in the contention phase?

Let $L_{cp,1}$ denote the length of the RA preamble. In the second phase of CP, the BS sends information about the

Fig. 4. Equivalent MA-AFC code graph at the BS.

IV. ANALYSIS OF THE RATELESS MULTIPLE ACCESS FOR M2M COMMUNICATIONS

In this section, we analyze the performance of the proposed MA-AFC scheme by using a density evolution [28] based approach. These results will be used as a basis to optimize the proposed rateless multiple access for M2M communications.

A. Optimum Coordinated Multiple Access

Let us first investigate the optimum coordinated multiple access, which will be used as a baseline for the proposed probabilistic multiple access AFC scheme. As shown in [13], the maximum common rate that can be achieved by $N$ devices transmitting in a typical RB with ordered effective channel gains $\{h_1\}$ and reference SNR $\gamma$ is given by:

$$R_c = \min_{j \in \mathbb{Z}_N^N} \frac{1}{j} \log_2 \left( 1 + \gamma \sum_{\ell=1}^{j} |h_\ell|^2 \right) \quad \text{bps/Hz},$$

which is simply obtained from the MAC capacity region, when the transmission rate of all the devices are the same. This can be used as an upper bound on the average system common rate

$$R_c = \frac{1}{N} \log_2 (1 + N \gamma_0) \quad \text{bps/Hz}.$$  

Therefore, the average system throughput increases with the number of devices with a logarithmic slope. The maximum payload size $L_{opt}$ which can be successfully transmitted by each device in a resource block of duration $\tau_s$ and bandwidth $W$ is then given by:

$$L_{opt} = \frac{W \tau_s}{N} \log_2 (1 + N \gamma_0) \quad \text{bits},$$

which is obtained without considering the overhead for the contention phase. In fact, (14) provides an upper bound on the maximum payload size which can be delivered from each active device at the BS.

$\text{Fig. 3. Original AFC code graph at the BS.}$

$\text{Fig. 4. Equivalent MA-AFC code graph at the BS.}$
number of devices which have selected a particular RA preamble. The average number of devices which have selected the same RA preamble is \(N/N_s\), where \(N\) is the number of active devices and \(N_s\) is the number of RA preambles. This information can be broadcasted by the BS by using at least \(L_{cp,2} = N_s \lceil \log_2 (N/N_s) \rceil\) information bits. In the third phase of CP, the users will select a random seed of length \(L(s)\) which is calculated by (6), for a given value of \(\epsilon\) and a given RA preamble \(s\). Thus, at most \(L_{cp,3} = N_s L(s)\) information bits have to be sent to the BS in the third phase of CP. Therefore, the total overhead can be calculated as follows:

\[
L_{ov} = L_{cp,1} + L_{cp,2} + L_{cp,3}
\]

\[
= L_{cp,1} + N_s \left\lceil \log_2 (N/N_s) \right\rceil + N_s \left\lceil \log_2 \left(1 + \frac{1}{1 - (1 - \epsilon)(N/N_s)^{-1}}\right) \right\rceil \text{ bits.} \tag{19}
\]

As the BS allocates RBs to devices and devices with the same RA preamble will transmit at the same RB, each device then knows at which RB it should transmit its message based on its delay requirement. In this case, the amount of overhead can be further reduced to

\[
L_{ov} = L_{cp,1} + \left\lceil \log_2 (N/N_s) \right\rceil + \left\lceil \log_2 \left(1 + \frac{1}{1 - (1 - \epsilon)(N/N_s)^{-1}}\right) \right\rceil \text{ bits.} \tag{20}
\]

Let \(k\) denote the message length of each device and \(R_c\) denote the achievable common rate for a given arrival rate \(\lambda\) using the proposed probabilistic multiple access approach. Then, this message can be successfully transmitted to the BS within a resource block of duration \(\tau_s\) and bandwidth \(W\) if \(k + L_{ov} \leq \tau_s W R_c\). In other words, the maximum message size (in bits) which can be supported by the proposed approach in an RB can be approximated by:

\[
k_{\text{max}} = \max\{\tau_s W R_c - L_{ov}, 0\}. \tag{21}
\]

V. Optimizing Access Probabilities With Delay Guarantees

The access probability and code degree for each device will be optimized in a way that delay constraints of all devices are satisfied. Here, we minimize the average degree of the received signal at the BS to further reduce the complexity of the decoder. Therefore, the BS should perform the following optimization process:

\[
\text{minimize } \sum_{i=1}^{N} p_i d_i
\]

subject to (a). \(Q \left(\sqrt{m_i} \right) < \delta\), (b). \(1 \leq d_i \leq D_s\) and (c). \(0 < p_i \leq 1\), for \(i \in \mathbb{Z}_N^+\), where \(\delta > 0\) is the target bit error rate, \(m_i(\infty) = \lim_{t \to \infty} m_i(t)\), and

\[
m_i(t) = \frac{\gamma |h_i|^2 \sigma_d^2 d_i t_i}{k} \sum_{v:v_{ji}=1}^{2N} 1 + \sum_{j=1}^{N} |h_j|^2 d_j \sigma_d^2 \Omega \left(m_j^{(t-1)}\right).
\]

It is important to note that \(\Omega(x)\) is close to zero for a large \(x\), so we can approximate (16) as \(m_i(t) \approx 2 \gamma |h_i|^2 \sigma_d^2 d_i t_i p_i\), and accordingly we have:

\[
m_i(t) = \frac{|h_i|^2 d_i p_i}{h_j^2 d_j p_j}, \tag{22}
\]

where we assume that \(\sigma_d^2\) is constant which is valid when the number of devices is very large. Let \(P_{c,i}\) denote the bit error rate of device \(U_i\) at the BS, then according to (18) we have:

\[
P_{c,i} \approx Q \left(\sqrt{\left(Q \left(1 \left(P_{c,i}\right)^{\frac{1}{2}}\right)^2 \frac{|h_i|^2 d_i p_i}{|h_j|^2 d_j p_j}\right)\right). \tag{23}
\]

More specifically, when the devices perform power control to have the same received SNR at the BS and use the same code degree \(d_c\), (23) can be further simplified as follows:

\[
P_{c,i} \approx Q \left(\sqrt{\left(Q \left(1 \left(P_{c,i}\right)^{\frac{1}{2}}\right)^2 \frac{p_i}{p_j}\right)\right). \tag{24}
\]

which clearly shows that the device which has a higher access probability is decoded sooner than those with lower access probabilities. Fig. 5 shows the BER for device \(U_2\) versus the BER of device \(U_1\) for the case that both devices have complete access to the channel and \(d_i = d_2\).

VI. Simulation Results

In this section, we investigate the performance of the proposed MA-AFC scheme in M2M communication systems. We first compare our proposed MA-AFC approach with the conventional massive RA technique in the LTE-A standard [31]. As shown in Fig. 6 we assume that time is divided into time frames of length 10 msec and the total system bandwidth is 20 MHz [22]. Each frame is further divided
into 10 subframes of length 1 msec. A subframe consists of 100 RBs, each with bandwidth $W = 0.2$ MHz. We also assume a total number of $N_s = 60$ preambles and for simplicity we assume several configurations of RAs, which includes 1 RA attempt per two time frames, 1 RA access attempt per time frame, and 2 RA attempts per time frame. Each successful device will then be allocated with 16 RBs for the data transmission in next subframes. We assume that the packet length of devices is small (few bytes), thus a device can successfully transmit its packet within one time frame.

The conventional random access approach, which is currently used in the LTE-A standard uses backoff approach [7]. In this approach, when a random access attempt is unsuccessful, the device will wait for a random time before transmitting the new access request. More specifically, before the $s^{th}$ random access request, the device will wait for time $t_{ra}^{(s)}$, where $t_{ra}^{(s)}$ is randomly drawn from range $[0, X_s - 1]$. The device can send a maximum number of $N_{ra}$ RA request using this backoff approach. More specifically, we assume maximum number of RA attempts $N_{ra} = 10$ and backoff indexes $X_s \in \{10, 20, 30, 40, 60, 80, 120, 160, 240, 320, 480, 960\}$ [22].

For the comparison, we use two metrics, blocking probability and average access delay. The blocking probability is defined as the ratio of the MTC devices failing to access the BS due to exceeding their maximum allowed random access attempts to the total number of MTC devices. The average access delay is defined as the average time elapsed from the time instant when a device sends the first access request until that it succeeds. Fig. 7 shows the blocking probability for both the proposed and conventional approaches. As can be seen in this figure, the blocking probability for the conventional approach is very high especially when the number of devices is large. However, devices in the proposed scheme can access to the BS with a very small blocking probability, due to the fact that the devices can transmit simultaneously and they do not need to wait until they select separate preambles.

Fig. 8 shows the average access delay versus the number of devices. As can be seen in this figure, the proposed approach outperforms the conventional approach and have roughly 70% less access delay. In Fig. 8, we considered two scenarios for MA-AFC. The access probability for all devices in Scenario 1 and Scenario 2 is set to $N_s/N$ and $10N_s/N$, respectively. It is important to note that the proposed approach in Scenario 2 has a lower access delay due to the fact that more devices can simultaneously transmit but it has higher complexity at the BS due to larger degrees of equivalent MA-AFC code at the BS.

We then simulate our proposed approach within one RB, and compare the results with the fundamental throughput limits in [13]. For this aim, we assume that the reference SNR is $\gamma = 0$ dB. This is equivalent to the case that a device transmits with 10 dBm (10mW) power over 1MHz bandwidth, a noise power spectral density of -174dBm/Hz, a receiver noise figure of 5dB, a spectral density of -174dBm/Hz, a receiver noise figure of 5dB, and a 3.76 pathloss exponent, a 128dB pathloss intercept at 1000m, and a cell radius of 1360m [13]. We also assume that $\tau_s = 1$ sec.

Let us first investigate the maximum common rate which can be achieved by the proposed probabilistic MA-AFC approach in the contention-free case. Fig. 9 shows the maximum common rate versus the arrival rate for the case that the devices perform the power control to have the same received SNR of 0 dB at the BS. We consider three cases for the MA-AFC approach with different access probabilities. As can be seen in this figure, the achievable rate of the proposed approach is very close to the optimal coordinated multiple access [13]. It is important to note that in the proposed approach we designed the access probabilities such that the number of devices which
Arrival rate ($\lambda$)

Payload size (bits)

Fig. 9. Maximum achievable common rate versus the arrival rate for the case that the devices perform power control to have the same received SNR of 0 dB at the BS.

Fig. 10. Maximum supported payload size versus the arrival rate. Devices perform power control to have the same received SNR of 0 dB at the BS. The devices use the same code degree of 8 and the same weight set $W_s$.

In Fig. 10, we show the maximum payload size which can be supported by the proposed approach in a resource block of duration $\tau_s = 1$ sec and bandwidth $W = 10$ kHz for different arrival rates. An upper bound on the maximum payload size which has been obtained from (14) is also shown in this figure. As can be seen in this figure, as the arrival rate increases, the payload size which can be supported decreases. This is because that the common rate is a decreasing function of the arrival rate (Fig. 9). Moreover, as the arrival rate increases, the overhead calculated in (20) increases, which reduces the supported payload size. Also, by increasing the access probability, a higher common rate can be achieved, which leads to higher supported payload sizes.

Fig. 11. BER versus the ratio of the number of coded symbols and the total number of information symbols. $m$ is the number of coded symbols, $N$ is the number of devices, and $k$ is the length of the message for each device.

Fig. 12. The average bit error rate for different groups of devices classified by delay requirements. The total number of devices is $N = 1000$, each group has 100 devices; the devices use the same code degree $d_c = 8$, access probabilities are calculated from (25) and $\gamma = 0$ dB. Dashed lines show the delay requirement for different groups of devices.

We now consider the case that devices have different delay requirements. We assume that the payload size of each device is 1 kb and 100 MTC devices belongs to each specific delay group, as shown in Table II. By using the optimization approach in Section IV, we can derive the optimized access probability for each group. Fig. 12 shows the average BER versus the delay for devices in each group of MTC devices specified in Table II, when all the devices has the same code degree $d_c = 8$. As can be seen in this figure, the delay requirement for devices in various groups have been simultaneously satisfied.
VII. CONCLUSIONS

In this paper, we proposed a novel rateless multiple access scheme to support a large number of devices in M2M communications. This is achieved by representing the multiple access process by a capacity approaching analog fountain code to minimize the access delay while achieving a higher throughput. In the proposed scheme, devices are allowed to transmit at the same time and the same channel, thus forming an equivalent analog fountain code at the base station (BS). The BS then performs joint decoding to recover all devices’ messages. We further analyzed the proposed multiple access analog fountain code by using the density evolution technique and formulated an optimization problem to find the optimum access probability for each device to satisfy the delay requirements of MTC devices. Simulation results showed that the proposed scheme closely approaches the optimal common-access analog fountain code by using the density evolution (BS). The BS then performs joint decoding to recover all

\[ \sum_{t' \in N(X_i) \setminus \ell} g_{t', X_i} \] can be approximated by \( \sigma^2 \hat{M}_{d,t}/k \), where \( m \) is the number of coded symbols. As a Type-\( X_i \) variable node is connected to a Type-\( \nu \) check node with probability \( q_{\nu} \), then the average total LLR value for a Type-\( X_i \) can be found as follow:

\[ m^{(t)}_{i} = \gamma_{X_i} \sigma^2_{\nu} \sum_{\nu \in N_i} \frac{2}{1 + \sigma^2_{\nu} q_{\nu}} \]

where \( \mathcal{N}(X_i) \) the set of all check nodes connected to a Type-\( X_i \) variable node except check node \( y_t \). As we assume that the number of variable and check nodes go to infinity, according to the law of large number, for large code degree \( d \),

\[ \sum_{t' \in N(X_i) \setminus \ell} g_{t', X_i} \] can be approximated by \( \sigma^2 \hat{M}_{d,t}/k \), where \( m \) is the number of coded symbols. As a Type-\( X_i \) variable node is connected to a Type-\( \nu \) check node with probability \( q_{\nu} \), then the average total LLR value for a Type-\( X_i \) can be found as follow:

\[ m^{(t)}_{i} = \gamma_{X_i} \sigma^2_{\nu} \sum_{\nu \in N_i} \frac{2}{1 + \sigma^2_{\nu} q_{\nu}} \]

where \( \mathcal{N}(X_i) \) the set of all check nodes connected to a Type-\( X_i \) variable node except check node \( y_t \). As we assume that the number of variable and check nodes go to infinity, according to the law of large number, for large code degree \( d \),

\[ \sum_{t' \in N(X_i) \setminus \ell} g_{t', X_i} \] can be approximated by \( \sigma^2 \hat{M}_{d,t}/k \), where \( m \) is the number of coded symbols. As a Type-\( X_i \) variable node is connected to a Type-\( \nu \) check node with probability \( q_{\nu} \), then the average total LLR value for a Type-\( X_i \) can be found as follow:

\[ m^{(t)}_{i} = \gamma_{X_i} \sigma^2_{\nu} \sum_{\nu \in N_i} \frac{2}{1 + \sigma^2_{\nu} q_{\nu}} \]

where \( \mathcal{N}(X_i) \) the set of all check nodes connected to a Type-\( X_i \) variable node except check node \( y_t \). As we assume that the number of variable and check nodes go to infinity, according to the law of large number, for large code degree \( d \),

\[ \sum_{t' \in N(X_i) \setminus \ell} g_{t', X_i} \] can be approximated by \( \sigma^2 \hat{M}_{d,t}/k \), where \( m \) is the number of coded symbols. As a Type-\( X_i \) variable node is connected to a Type-\( \nu \) check node with probability \( q_{\nu} \), then the average total LLR value for a Type-\( X_i \) can be found as follow:

\[ m^{(t)}_{i} = \gamma_{X_i} \sigma^2_{\nu} \sum_{\nu \in N_i} \frac{2}{1 + \sigma^2_{\nu} q_{\nu}} \]

where \( \mathcal{N}(X_i) \) the set of all check nodes connected to a Type-\( X_i \) variable node except check node \( y_t \). As we assume that the number of variable and check nodes go to infinity, according to the law of large number, for large code degree \( d \),

\[ \sum_{t' \in N(X_i) \setminus \ell} g_{t', X_i} \] can be approximated by \( \sigma^2 \hat{M}_{d,t}/k \), where \( m \) is the number of coded symbols. As a Type-\( X_i \) variable node is connected to a Type-\( \nu \) check node with probability \( q_{\nu} \), then the average total LLR value for a Type-\( X_i \) can be found as follow:

\[ m^{(t)}_{i} = \gamma_{X_i} \sigma^2_{\nu} \sum_{\nu \in N_i} \frac{2}{1 + \sigma^2_{\nu} q_{\nu}} \]

where \( \mathcal{N}(X_i) \) the set of all check nodes connected to a Type-\( X_i \) variable node except check node \( y_t \). As we assume that the number of variable and check nodes go to infinity, according to the law of large number, for large code degree \( d \),

\[ \sum_{t' \in N(X_i) \setminus \ell} g_{t', X_i} \] can be approximated by \( \sigma^2 \hat{M}_{d,t}/k \), where \( m \) is the number of coded symbols. As a Type-\( X_i \) variable node is connected to a Type-\( \nu \) check node with probability \( q_{\nu} \), then the average total LLR value for a Type-\( X_i \) can be found as follow:

\[ m^{(t)}_{i} = \gamma_{X_i} \sigma^2_{\nu} \sum_{\nu \in N_i} \frac{2}{1 + \sigma^2_{\nu} q_{\nu}} \]
Mahyar Shirvanimoghaddam received the B. Sc. degree with 1st Class Honours from University of Tehran, Iran, in September 2008, the M. Sc. Degree with 1st Class Honours from Sharif University of Technology, Iran, in October 2010, and the Ph.D. degree from The University of Sydney, Australia, in January 2015. He is currently a Postdoctoral Research Fellow at the School of Electrical Engineering and Computer Science, The University of Newcastle, Australia. His general research interests include channel coding techniques, cooperative communications, compressed sensing, machine-to-machine communications, and wireless sensor networks. Mahyar Shirvanimoghaddam is currently serving as the Reviewer of several international prestigious journals, e.g., IEEE Transactions on Communications, IEEE Transactions on Wireless Communications, IEEE Communication Letters, IEEE Transactions on Vehicular Technology, IEEE Wireless Communications Letters, IEEE Signal Processing Letters, and several international flagship conferences.

Yonghui Li (M04-SM09) received his Ph.D degree in November 2002 from Beijing University of Aeronautics and Astronautics. From 1999 to 2003, he was affiliated with Linkair Communication Inc, where he held a position of project manager with responsibility for the design of physical layer solutions for the LAS-CDMA system. Since 2003, he has been with the Centre of Excellence in Telecommunications, the University of Sydney, Australia. He is now an Associate Professor in School of Electrical and Information Engineering, University of Sydney. He is the recipient of the Australian Queen Elizabeth II Fellowship in 2008 and the Australian Future Fellowship in 2012. His current research interests are in the area of wireless communications, with a particular focus on MIMO, cooperative communications, coding techniques and wireless sensor networks. He holds a number of patents granted and pending in these fields. He is an executive editor for European Transactions on Telecommunications (ETT). He received the best paper awards from IEEE International Conference on Communications (ICC) 2014 and IEEE Wireless Days Conferences (WD) 2014.

Misha Dohler (S99M03SM07F14) is full Professor in Wireless Communications at Kings College London, Head of the Centre for Telecommunications Research, co-founder and member of the Board of Directors of the smart city pioneer Worldsensing, Fellow and Distinguished Lecturer of the IEEE, and Editor-in-Chief of the Transactions on Emerging Telecommunications Technologies and the Transactions on the Internet of Things.

He is a frequent keynote, panel and tutorial speaker, and has received numerous awards. He has pioneered several research fields, contributed to numerous wireless broadband, IoT/M2M and cyber security standards, holds a dozen patents, organized and chaired numerous conferences, has more than 200 publications, and authored several books. He has a citation h-index of 39 (top 1%)

He acts as policy, technology and entrepreneurship adviser, examples being Richard Bransons Carbon War Room, House of Parliament UK, UK Ministry BIS, EPSRC ICT Strategy Advisory Team, European Commission, Tech London Advocate, ISO Smart City working group, and various start-ups. He is also an entrepreneur, angel investor, passionate pianist and fluent in 6 languages. He has talked at TEDx. He had coverage by national and international TV & radio; and his contributions have featured on BBC News and the Wall Street Journal.

Branka Vucetic (M’83 SM’90 F’98) received the B.S.E.E., M.S.E.E., and Ph.D. degrees in 1972, 1978, and 1982, respectively, in electrical engineering, from The University of Belgrade, Belgrade, Yugoslavia. During her career, she has held various research and academic positions in Yugoslavia, Australia, and the UK. Since 1986, she has been with the Sydney University School of Electrical and Information Engineering in Sydney, Australia. She is currently the Director of the Centre of Excellence in Telecommunications at Sydney University. Her research interests include wireless communications, digital communication theory, coding, and multi-user detection.

In the past decade, she has been working on a number of industry sponsored projects in wireless communications and mobile Internet. She has taught a wide range of undergraduate, postgraduate, and continuing education courses worldwide. Prof. Vucetic has co-authored four books and more than two hundred papers in telecommunications journals and conference proceedings.