Fast and Accurate Cardinality Estimation in Cellular-Based Wireless Communications

Mohammad G. Khoshkholgh*, Victor C. M. Leung*, Kang G. Shin**
* Department of Electrical and Computer Engineering, the University of British Columbia
** Department of Electrical and Computer Science, the University of Michigan

***Abstract*—Cardinality estimation is of fundamental importance in various wireless communication applications. The performance of any adaptive medium access control in the uplink channel is in fact affected by its accuracy. This paper provides a fresh look at this fundamental problem, and proposes a novel scheme for fast and accurate cardinality estimation. This scheme utilizes the experienced outage probability in detecting IDs of the devices in the uplink. It is observed that for practically relevant values of Signal-to-Interference-plus-Noise Ratio (SINR) threshold, e.g., $\beta \leq 10$ dB, and spreading sequence length, e.g., $N_c \geq 7$, the base station is able to estimate the true cardinality with vanishing error in only one shot, even if the network is so populated, e.g., the number of active devices $N \geq 100$. Otherwise, one may apply the proposed method in couple of consecutive shots such that in each of which portion of the cardinality is estimated.**

I. Introduction

Estimating the number of operating users is the prerequisite task of any successful adaptive scheduler in wireless communications. Generally, in the literature it is simply assumed that the scheduler perfectly knows the cardinality in advance [1]. In practice, a number of steps should be taken for this goal. For example, authors in [2], [3] explained methods for estimating cardinality based on scrutinizing the history of the scheduled traffic. In reality, on the other hand, the dynamic of the system imposes uncertainties on the availability and furthermore the accuracy of such knowledge, particularly noting the emergence of new wireless communication applications that their traffic is in nature event-driven and accurately unpredictable [4]. Forth generation Long Term Evolution (LTE) cellular communications are expected to handle machine-type communications (M2M communications) and sensory traffic. Note that M2M traffic is mainly event-driven, and generally hard to predict [4]. In these systems it is predicted that in each time instance hundreds of machines, sensors, and devices will be turned on and seeking association to the central base stations (BS), which, apart from the system design and configuration, demands cautious cardinality estimation. Without effective cardinality estimation the medium access control mechanisms are not successful to evenly handle the traffic demands.

In this paper the main focus is on the cardinality estimation in wireless communication systems. For this goal, let firstly borrow the following constructive cardinality problem, also known as German tank problem, [5]:

Problem 1: *The tanks of a country's army are numbered 1 to N. In a war this country loses n random tanks to the enemy, who discovers that the captured tanks are numbered. Assume X_1, X_2, …, X_n are the numbers of the captured tanks. What is the estimation of N according to the captured tanks?*

In [5] a solution suggested via evaluation of $E \max_i X_i$, where E is the expectation operator. Undoubtedly, in the specific case of wireless communications we encounter similar situations. Particularly, if we interpret *tanks* as the users (machines, sensors, or devices) asking for the connectivity to the BS by sending out service requests in the uplink channel, *enemy* is the BS who needs to inspect the received requests for estimating the cardinality. A new cardinality estimation problem in this context is then formulated as:

Problem 2: *Assume N users with ID numbers X_1, X_2, …, X_N are asking for service via sending out requests, simultaneously, to the access point in the designated time slot. How the access point could estimate N from processing the detected users IDs?*

In this paper the main focus is on solving this problem. This problem is fundamentally important since the performance of almost any proposed medium access control (MAC) protocol is explicitly affected by miss-evaluation of the cardinality. For instance a common choice of access probability broadcasted by the BS at the start of a frame is $1/\hat{N}$ where $\hat{N}$ is the estimation of the population available at the BS [6]. Obviously, if $|\hat{N} - N| \gg 0$ the network may collapse. In essence, a wrong medium access probability can lead to overflow of packets in queues because of high collision incidents (when $1/\hat{N} \gg 1/N$) and/or frequent deferral of the transmission (when $1/\hat{N} \ll 1/N$), and consequently instability of the whole system.

In this paper we propose a novel method inspired by the Problem 1's solution for the cardinality estimation in a generic wireless communication model. Obvious applications of the proposed method, however, could be in M2M communications [4], wireless sensor networks [7], and uplink channels in cellular communications. Note that, the usage of the solution of the Problem 1 for solving Problem 2 is not trivial due mainly to some inherent differences between two problems. In contrast to the Problem 1, here random behaviors such as AWGN, channel fading, multi-user interference, and collision incidents among spreading sequences are affecting the ID detection stage that in turn making the problem much more involved.

We first derive an expression for the probability of ID's de-

978-1-4799-8406-0/15/$31.00 ©2015 IEEE

tections. This result incorporates the effects of many pertinent factors including the length of the spreading sequence, the probability that spreading sequences conflicts, the Signal-to-Interference-plus-Noise Ratio (SINR) threshold, and transmission power of the devices. The developed algorithm then uses this detection probability for estimating the cardinality. Our numerical study confirms that for practically relevant values of SINR threshold, e.g., $\beta \leq 10$ dB, and spreading sequence length, e.g., $N_c \geq 7$, the BS is able to estimate the true cardinality with vanishing error in only one shot, even if the network is so populated, e.g., the number of active devices $N \geq 100$. For the cases that the network is heavily loaded and/or the spreading sequence is short or the SINR threshold is large, we further suggest that to apply the proposed method in couple of consecutive shots such that in each of which portion of the cardinality is estimated.

Cardinality estimation is vastly studied in the literature. In [7] authors adopted Good-Turing method for network size estimation in sensor networks with ALOHA access protocol. However, effect of the mentioned random behaviors are overlooked. In [6] a new mechanism is devised for estimating the network traffic assuming ALOHA protocol with collision model. Here in this paper we consider capture model [1] that is more realistic in wireless communications. Sequential pooling with mobile access point is also proposed in [8] for wireless sensor networks focusing on reducing the number of transmissions. In RF-ID systems the reader's main task is to estimate the cardinality of the network comprising of perhaps ten thousand tags. The proposed algorithms for RF-ID systems may not be applicable for the scenarios this paper is interested on as these algorithms generally require hundreds to thousands iterations before the final convergence [9]. Besides, these algorithms are not designed to incorporate fading fluctuations and multi-user interference, and only can deal with simplified probabilistic errors [10].

The rest of this paper is organized as follows. In Section II we present the system model. Then in Section III we mathematically analyze the cardinality estimation. We further evaluate the accuracy of the proposed method in numerical evaluations. Finally Section IV concludes the paper.

II. System Model

Consider a wireless system including a BS and N active users randomly distributed in the coverage area. This system model is generic enough to cover many wireless communication scenarios including sensor networks and M2M communications. For instance, in M2M networks machines independently turn on in the case of an event (surveillance and health care applications) or periodically (periodic reporting in smart grid communications) for having connection to the application server via the BS [4]. There is M number of spreading codes available to the users with length N_c where

$$M = 2^{N_c} - 1. \tag{1}$$

With the urge of service request each user randomly peaks one of the stored codes, and transmits its ID to the access point. We assume users are aware of the window allocated to the service request via periodically broadcasted information by the BS at the start of each communication session. Note that the spreading code sequences are fixed. A direct consequence of this generic assumption is that there is non-zero probabilities that couple of users air their IDs with the similar code sequence, which will result in collision among the sequences at the BS. Since, we have ruled out any priori information or estimation of the network size it may not be generally possible to choose N_c beforehand. For example in [2] it is shown that by selecting the code length as

$$N_c \approx \log\left(1 + \frac{N-1}{p_C}\right), \tag{2}$$

the collision probability $p_C \ll 1$ is guaranteed. However, such a design procedure may not be robust in reality as the BS does not have N in advance.

We assume the cell is circular with radius R and the BS is located at the origin. Users are randomly distributed through the following pdf

$$f_{D_i}(x) = \frac{2}{R^2}x, \tag{3}$$

in which $D_i \in [0, R]$ is the distance between user i and the BS.

III. Cardinality Estimator

Assume in a given time instance N users send out the service requests. Each user i selects randomly a spreading code c_i and transmits its ID to the BS with transmission power P. Here for simplicity we assume the synchronized system. More investigation needs to study the asynchronous model. Moreover, we rule out the adaptive power allocation in our analysis. Assuming the users have degrees of knowledge about the channel state information including fading gain and path-loss attenuation, approaches like truncated power control can be adopted [11]. The BS then tries to detect the ID's. Considering matched filter (MF) model, the performance of detection then is a function of the Signal-to-Noise-Ratio (SINR), which is defined as

$$\mathrm{SINR}_i = \frac{H_i D_i^{-\alpha}}{\eta + \frac{1}{M}\sum\limits_{j\in\mathcal{U}_i} H_j D_j^{-\alpha} + \sum\limits_{j\in\mathcal{C}_i} H_j D_j^{-\alpha}}, \tag{4}$$

where η is defined as

$$\eta = \mathrm{SNR}^{-1} = \frac{\sigma^2}{P}, \tag{5}$$

considering σ^2 the AWGN noise. In this model we have used the following notations and assumptions: $\{H_i\}$ are i.i.d. exponential random variables which are constant during the entire frame duration including cardinality estimation procedure; α is the path-loss exponent and has values in the interval $[2, 6]$; $\{D_i^{-\alpha}\}$ are distance-dependent path-loss attenuations noticing that D_i is the distance between user i and the BS which is random and is drawn according to the pdf (3); $\mathcal{C}_i$ represents a set of the users who also use the sequence code c_i so are with collision with the user i's spreading sequence; $\mathcal{U}_i$ represents a set containing users who have sequences different than c_i though might be in conflict with each other that in not the concern in the definition of SINR of user i. Note that MF can reduce the effect of the interference from users belong $\mathcal{U}_i$ by the factor $1/M$.

In the formulation of SINR there are three independent random variables including users locations, channel fading, and sequence selection. The following gives the probability of the reception as a function of the number of users N. For the time being let assume N is specified.

Theorem 1: Assume N users attempt for the channel. The reception probability of user i, $\psi_i(N)$, is given by

$$\psi_i = \frac{2p_U}{R^2}\int\limits_0^R xe^{-\eta\beta x^\alpha}\left[1 - {}_2F_1(1,\delta,1+\delta;\frac{-Mx^{-\alpha}}{\beta R^\alpha})\right]^{N-1} dx$$

$$+\frac{2p_C}{R^2}\sum_{n=1}^{N-1}\binom{N-1}{n}\left(\frac{1}{M}\right)^n\left(1-\frac{1}{M}\right)^{N-n-1}\times$$

$$\int\limits_0^R xe^{-\eta\beta x^\alpha}\left[1 - {}_2F_1(1,\delta,1+\delta;\frac{-Mx^{-\alpha}}{\beta R^\alpha})\right]^{N-n-1}\times$$

$$\left[1 - {}_2F_1(1,\delta,1+\delta;-\frac{x^{-\alpha}}{\beta R^\alpha})\right]^n dx, \qquad (6)$$

where

$$p_U = Pr\{i \in U\} = \frac{1}{M}\left(1-\frac{1}{M}\right)^{N-1}, \qquad (7)$$

$$p_C = Pr\{i \in C\} = 1-\left(1-\frac{1}{M}\right)^{N-1}. \qquad (8)$$

Proof: See the Appendix.

Note that $\psi_i(N) = \psi(N)\ \forall i$. According to Theorem 1 the BS detects a user's ID with probability $\psi(N)$. Having some of these ID's detected now we concoct a procedure inspired from [5] for estimating N as the following. Denote the maximum detected ID by Y as

$$Y = \max_i X_i. \qquad (9)$$

The probability mass function of Y can be described as

$$Pr\{Y=k\} = \psi_k(N)\prod_{j=k+1}^{N}(1-\psi_j(N)),\ \ \forall k = 0,1,\ldots,N$$
$$= \psi(N)(1-\psi(N))^{N-k}. \qquad (10)$$

Note that it is straightforward to check that $\sum\limits_{k=0}^{N} Pr\{Y=k\} = 1$ as follows:

$$\sum_{k=0}^{N} Pr\{Y=k\} = (1-\psi(N))^N + \sum_{k=1}^{N}\psi(N)(1-\psi(N))^{N-k}$$
$$= (1-\psi(N))^N + \psi(N)(1-\psi(N))^{N-1}$$
$$\times\frac{(1-\psi(N))^{-N}-1}{\psi(N)}(1-\psi(N)) = 1. \qquad (11)$$

Now we evaluate the expected value of Y as a function of N:

$$\begin{aligned} EY\ &= \sum_{k=1}^{N} k\psi(N)(1-\psi(N))^{N-k} \\ &= \psi(N)(1-\psi(N))^{N+1}\sum_{k=1}^{N}\frac{d}{d\psi(N)}(1-\psi(N))^{-k} \\ &= \psi(N)(1-\psi(N))^{N+1}\frac{d}{d\psi(N)}\frac{(1-\psi(N))^{-N}-1}{\psi(N)} \\ &= \frac{1-\psi(N)}{\psi(N)}\left[(1-\psi(N))^N-1\right]+N. \qquad (12)\end{aligned}$$

Now we assume EY is in fact the maximum detected ID, which the BS has the knowledge of its value. Then the BS can solve the above equation for N to obtain an estimate of it denoted by $\hat{N}$.

The performance of the proposed cardinality estimator is presented in Fig. 1 and Fig. 2. For the simulation we have considered a single cell with radius 1 Km and scattered N users, which N is of course unknown to the BS, randomly in the coverage area. Maximum detected ID is assigned to EY. After that we have solved (12) to estimate $\hat{N}$. Fig. 1 illustrates the error for different values of N_c assuming that $\beta = 4$ dB. As it is seen error is a decreasing function of N_c. For a system with long enough spreading sequence $N_c \geq 8$ the error is almost zero even for a very highly crowded network $N \geq 100$. By increasing N, the chance of collision among the spreading sequences will increase that reduce the effectiveness of the proposed method.

As it is seen for the case of $N_c = 6$ the effectiveness of the proposed method is reduced when N is increased. In such a case, there is a good chance that some users's IDs conflicts, assuming that the corresponding SINRs are sufficiently large that the IDs are detectable at the BS, which ensues the recognizance of only one user. Besides, due to high imposed intra-cell interference the chance that the BS correctly detects IDs are unfortunately very low. Augmenting N_c to 7 can, on the other hand, dramatically enhances the accuracy of the estimator comparing to the case of $N_c = 6$. This is because the collision probability of the spreading sequences is somehow reduces exponentially. Besides, the detrimental impact of the intra-cell interference is significantly reduced.

What Fig. 2 indicates is the error for different values of β assuming that $N_c = 9$. As it is seen by increasing β, $N - \hat{N}$ is also increased that is due to lowering the probability of ID detection. Recall that to detect an ID the corresponding SINR has to exceed the SINR threshold β. When the network is populated, the level of experienced intra-cell interference is generally higher, which makes it harder to detect the IDs if the SINR threshold is large.

Note that the BS may assign a number of sequential slots, say T, for cardinality estimation if N_c is low enough or β is high enough. For such scenarios, each user selects each slot with probability $1/T$ and sends out its service request. The BS applies the concocted procedure on each time slot to estimate

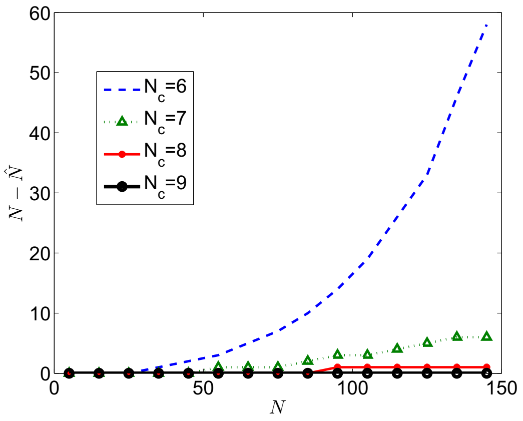

Fig. 1. Cardinality estimator error versus N for different values of N_c.

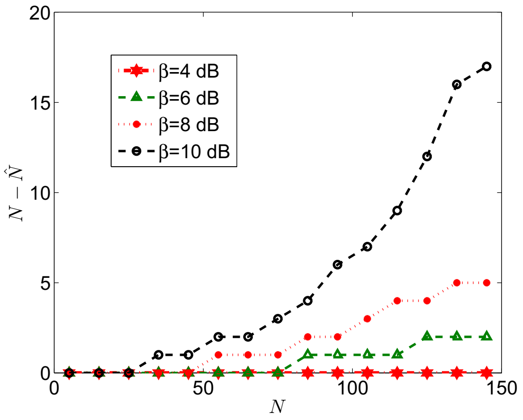

Fig. 2. Cardinality estimator error versus N for different values of β when $N_c = 9$.

the cardinality of the users using that time slot. Then the BS only needs to sum up the estimated users across the time slots.

IV. Conclusion

In this paper we have focused on cardinality estimation in wireless communications. Our formulation considered the impact of channel fading, random distribution of the users, multi-user interference, and collision among the spreading sequences. We then devised an approach to accurately estimate the number of active users. The numerical results indicated that for Signal to Interference plus Noise Ratio (SINR), $\beta \leq 10$ dB, and spreading sequence length, $N_c \geq 7$, the access point was able to estimate the true cardinality with vanishing error even for very populated systems $N \geq 100$ in one shot.

Appendix

Let p_U denote the probability that a randomly selected code is unique. p_C is the collision probability among the codes. According to the results of [2] we already know that

$$p_U = Pr\{i \in U\} = \frac{1}{M}\left(1 - \frac{1}{M}\right)^{N-1} \tag{13}$$

$$p_C = Pr\{i \in C\} = 1 - \left(1 - \frac{1}{M}\right)^{N-1}. \tag{14}$$

One may then evaluate the reception probability of the i-th user's request as

$$\psi_i = p_U Pr\left\{\mathrm{SINR}_i \geq \beta | i \in \mathcal{U}\right\} + p_C Pr\left\{\mathrm{SINR}_i \geq \beta | i \in \mathcal{C}\right\}. \tag{15}$$

Consider the event $\{\mathrm{SINR}_i \geq \beta | i \in U\}$. Its associated probability can be written as

$$\begin{aligned}
&Pr\left\{\mathrm{SINR}_i \geq \beta | i \in U\right\} \\
&\overset{(a)}{=} E_{D_1,\ldots,D_N} e^{-\mathrm{NSR}\beta D_i^\alpha} E_{\{H_j\}_{j\neq i}} e^{-\beta D_i^\alpha \frac{1}{M} \sum\limits_{j\neq i} H_j D_j^{-\alpha}} \\
&\overset{(b)}{=} E_{D_1,\ldots,D_N} e^{-\mathrm{NSR}\beta D_i^\alpha} \prod_{j\neq i} \frac{1}{1 + \frac{1}{M}\beta D_i^\alpha D_j^{-\alpha}} \\
&= E_{D_i} e^{-\mathrm{NSR}\beta D_i^\alpha} E_{\{D_j\}_{j\neq i}} \prod_{j\neq i} \frac{1}{1 + \frac{1}{M}\beta D_i^\alpha D_j^{-\alpha}} \\
&\overset{(c)}{=} E_{D_i} e^{-\mathrm{NSR}\beta D_i^\alpha} \prod_{j\neq i} E_{D_j} \frac{1}{1 + \frac{1}{M}\beta D_i^\alpha D_j^{-\alpha}} \\
&\overset{(d)}{=} E_{D_i} e^{-\mathrm{NSR}\beta D_i^\alpha} \left[\frac{2}{R^2}\int\limits_0^R \frac{x}{1 + \frac{1}{M}\beta D_i^\alpha x^{-\alpha}} dx\right]^{N-1} \\
&\overset{(e)}{=} \frac{2}{R^2}\int\limits_0^R x e^{-\mathrm{NSR}\beta x^\alpha} \left[1 - {}_2F_1(1, \delta, 1+\delta; -M\frac{x^{-\alpha}}{\beta R^\alpha})\right]^{N-1} dx,
\end{aligned} \tag{16}$$

where we have used the following steps: (a) H_i is exponentially distributed random variable. Furthermore, channel fluctuations and distance-dependence path-loss fluctuations are independent. (b) for $i \neq j$ H_i and H_j are independent exponentially random variables. (c) $\{D_i\}$ are independent random variables. (d) noticing (3). (e) ${}_2F_1(.)$ is Gauss hypergeometric function.

Now consider the second term of (15). We may proceed as the following:

$$\begin{aligned}
&Pr\left\{\mathrm{SINR}_i \geq \beta | i \in C\right\} \\
&= E_{D_1,\ldots,D_N} e^{-\mathrm{NSR}\beta D_i^\alpha} E \prod_{j\in\mathcal{U}_i} \frac{1}{1 + \frac{1}{M}\beta D_i^\alpha D_j^{-\alpha}} \\
&\times \prod_{j\in\mathcal{C}_i} \frac{1}{1 + \beta D_i^\alpha D_j^{-\alpha}},
\end{aligned} \tag{17}$$

where $\mathcal{U}_i$ is the set of users with different spreading sequence and $\mathcal{C}_i$ is the set of the users with the identical sequence to c_i. Let $|.|$ stand for cardinality, (17) is reduced to

$$\begin{aligned}
&E_{D_i} e^{-\mathrm{NSR}\beta D_i^\alpha} E \left[\frac{2}{R^2}\int\limits_0^R \frac{x}{1 + \frac{1}{M}\beta D_i^\alpha x^{-\alpha}} dx\right]^{|\mathcal{U}_i|} \\
&\times \left[\frac{2}{R^2}\int\limits_0^R \frac{x}{1 + \beta D_i^\alpha x^{-\alpha}} dx\right]^{|\mathcal{C}_i|},
\end{aligned}$$

$$= E_{D_i} e^{-\mathrm{NSR}\beta D_i^{\alpha}} \sum_{n=1}^{N-1} \binom{N-1}{n} \left(\frac{1}{M}\right)^{n} \left(1-\frac{1}{M}\right)^{N-n-1}$$

$$\left[\frac{2}{R^2}\int\limits_0^R \frac{x}{1+\frac{1}{M}\beta D_i^{\alpha} x^{-\alpha}}dx\right]^{N-n-1}$$

$$\times \left[\frac{2}{R^2}\int\limits_0^R \frac{x}{1+\beta D_i^{\alpha} x^{-\alpha}}dx\right]^{n}$$

$$= \frac{2}{R^2} \sum_{n=1}^{N-1} \binom{N-1}{n} \left(\frac{1}{M}\right)^{n} \left(1-\frac{1}{M}\right)^{N-n-1}$$

$$\int\limits_0^R x e^{-\mathrm{NSR}\beta x^{\alpha}} \left[1-\ {}_2F_1(1,\delta,1+\delta;-M\frac{x^{-\alpha}}{\beta R^{\alpha}})\right]^{N-n-1}$$

$$\left[1-\ {}_2F_1(1,\delta,1+\delta;-\frac{x^{-\alpha}}{\beta R^{\alpha}})\right]^{n} dx. \qquad (18)$$

Back substitution of (16) and (18) in (15), the final result is obtained.

References

[1] S. Adireddy and L. Tong, "Exploiting decentralized channel state information for random access," *IEEE Transactions on Information Theory*, vol. 51, no. 2, pp. 537–561, Feb. 2005.

[2] H. S. Dhillon, H. C. Huang, H. Viswanathan, and R. A. Valenzuela, "Power-efficient system design for cellular-based machine-to-machine communications," *IEEE Transactions on Wireless Communications*, vol. 12, no. 11, pp. 5740–5753, Nov. 2013.

[3] R. Talak and N. B. Mehta, "Optimal timer-based best node selection for wireless systems with unknown number of nodes," *IEEE Transactions on Communications*, vol. 61, no. 11, pp. 4475–4485, Nov. 2013.

[4] M. Z. Shafiq, L. Ji, A. X. Liu, J. Pang, and J. Wang, "Large-scale measurement and characterization of cellular machine-to-machine traffic," *IEEE/ACM Transactions on Networking*, vol. 21, no. 6, pp. 1960–1973, Dec. 2014.

[5] S. Ghahramani, *Fundamentals of Probability with Random Processes*, 3rd ed. Pearson Perentice Hall, 2005.

[6] H. Wu, C. Zhu, R. J. La, X. Liu, and Y. Zhang, "FASA: Accelerated S-ALOHA using access history for event-driven M2M communications," *IEEE/ACM Transactions on Networking*, vol. 21, no. 6, pp. 1904–1917, Dec. 2013.

[7] C. Budianu, S. B.-David, and L. Tong, "Estimation of the number of operating sensors in large-scale sensor networks with mobile access," *IEEE Transactions on Signal Processing*, vol. 54, no. 5, pp. 1703–1715, May 2006.

[8] A. Leshem and L. Tong, "Estimating sensor population via probabilistic sequential polling," *IEEE Signal Processing Letters*, vol. 12, no. 5, pp. 395–398, May 2005.

[9] T. Li, S. S. Wu, S. Chen, and M. C. K. Yang, "Generalized energy-efficient algorithms for the RFID estimation problem," *IEEE/ACM Transactions on Networking*, vol. 20, no. 6, pp. 1978–1990, Dec. 2012.

[10] W. Luo, S. Chen, Y. Qiao, and T. Li, "Missing-tag detection and energytime tradeoff in large-scale RFID systemswith unreliable channels," *IEEE/ACM Transactions on Networking*, vol. 22, no. 4, pp. 1079–1091, Aug. 2014.

[11] A. J. Goldsmith, *Wireless Communications*. Cambridge University Press, 2005.