

On the Capacity of Slotted Aloha With Ancillary Channels

Andrea Munari, *Member, IEEE*, Gianluigi Liva, *Senior Member, IEEE*, and Matteo Berioli, *Member, IEEE*

Abstract—In this letter, we focus on a legacy system operated with slotted-Aloha and complemented by a redundancy channel where nodes transmit replicas of their packets. The number of replicas follows a probability distribution, and successive interference cancelation is applied across channels. Leaning on the theory of codes on graphs and algebraic tools, we prove that the system can provide arbitrarily small error rate up to a certain load, beyond which packet losses have to be undergone with finite probability. Tight upper bounds on capacity are derived for both regions, characterizing the achievable performance as a function of the deployed ancillary resources. Simulation results for moderate MAC frame length are also provided.

Index Terms—Aloha, random access, density evolution.

I. INTRODUCTION

ALBEIT simple, random access schemes still play a paramount role in wireless communications, enabling data delivery when no coordination is available. A major boost to the attainable performance has recently been triggered by letting users transmit replicas of their packets within a frame, and by applying at the receiver successive interference cancelation (SIC) techniques to resolve collisions. Borrowing tools from the theory of codes on graphs, such an approach can be effectively modeled resorting to bi-partite graphs [1]. In this setting, [2] showed that channel loads close to 1 [pkt/slot] can be served with vanishing error probability, while [3] strengthened the result proving that the throughput efficiency can be made arbitrarily close to 1 for asymptotically large frames. Despite such encouraging results, a vast fraction of the systems deployed nowadays still resort to simpler random access policies. From this standpoint, we focus on a *legacy* channel operated with plain slotted Aloha (SA) which is accessed (potentially saturated) by a number of *basic* nodes, and tackle the problem of upgrading the system to support a larger user population. To this aim, we complement the channel with a set of additional resources, where nodes send some replicas of their packets following a probability distribution.¹ SIC is performed across

channels, facing the intrinsic inefficiency of having only a fraction of the slots available to smartly place redundancy as opposed to [2]. In this context, we show that the system can indeed provide vanishingly small error probability for all the injected traffic up to a certain load, deriving a tradeoff between capacity and repetition rate. Moreover, we provide an upper bound to the performance of the scheme when operated beyond such load threshold, characterizing the maximum throughput that can be aimed for given the offered redundancy. Via density evolution, we show analytically that the derived capacity bounds are tight for asymptotically-large MAC frames. Moreover, simulation results for moderate-size frames are provided and discussed. To obtain some of the key results, we follow an algebraic approach based on [4], [5], which creates a parallel between the reception pattern seen over a set of slots and a system of linear equations over a finite field. The tool proves to be very powerful in identifying accurate bounds for the considered class of schemes.

II. SYSTEM MODEL

We consider a set of M users sending information in the form of data packets (or bursts) to a common receiver by means of random access procedures. Time is divided in frames, each composed of N_a slots of duration T_s , set to allow the transmission of one burst by the slot-synchronous users. We refer to these resources as the *primary* or *legacy* channel C_a , which is accessed following SA. Thus, every node randomly selects one of the N_a slots to send a burst over each frame, and we define the offered load as $G = M/N_a$ [pkt/slot].

The legacy channel is complemented by an additional set of orthogonal resources, dubbed *secondary* or *redundancy* channel C_b . For each frame populating C_a , N_b slots of duration T_s are allocated to the secondary channel. The ratio $\alpha = N_b/N_a$ is introduced to quantify the amount of redundancy made available to the system. Terminals access C_b by randomly distributing over the N_b slots replicas of the packet sent over the primary channel in the corresponding frame. The number of replicas ℓ is drawn independently at every frame by each node from a common probability mass function Λ , so that with probability Λ_ℓ , $\ell \in \{1, \dots, L\}$ one burst is sent over C_a , while $\ell - 1$ copies are delivered over C_b . Furthermore, every packet contains a pointer to the slot number in C_a or C_b of all its twin replicas. An ideal channel connects terminals to the common receiver, while collisions are regarded as destructive.² A packet is then correctly decoded as soon as no other bursts were transmitted over the same slot, whereas the presence of two or more overlapping packets prevents decoding any of them. At the receiver side, SIC is applied. Accordingly, whenever a burst is successfully retrieved, the interference contribution of its replicas is subtracted from the signal received over the

Manuscript received August 20, 2014; revised January 9, 2015; accepted January 30, 2015. Date of publication February 9, 2015; date of current version April 8, 2015. The associate editor coordinating the review of this paper and approving it for publication was H. Luo.

A. Munari is with the Institute of Networked Systems, Rheinisch-Westfälische Technische Hochschule (RWTH) University, 52076 Aachen, Germany (e-mail: andrea.munari@inets.rwth-aachen.de).

G. Liva is with the Institute of Communications and Navigation, German Aerospace Center (DLR), 82234 Wessling, Germany (e-mail: gianluigi.liva@dlr.de).

M. Berioli is with Zodiac Inflight Innovations, Triagnosys GmbH, 82234 Wessling, Germany (e-mail: matteo.berioli@triagnosys.com).

Digital Object Identifier 10.1109/LCOMM.2015.2401558

¹A relevant example is the Automatic Identification System (AIS) maritime standard, which enforces vessels to periodically distribute beacons in VHF band for collection at dedicated satellites. Due to the large satellite footprint, AIS traffic can be modeled as a heavily congested SA channel, supported by an ancillary frequency band recently allocated by regulatory boards for ships to transmit replicas of their packets. The capacity of such composed systems is yet to be understood.

²The ideal channel assumption holds for both legacy and ancillary channel, since as a first approximation losses are mostly due to collisions. Extensions of the presented approach may be foreseen on the footsteps of [6].

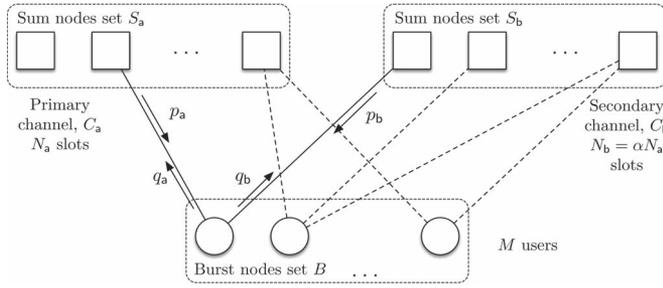


Fig. 1. Tripartite graph representing the proposed access scheme.

corresponding slots, indicated by the piggybacked pointers. This may allow decoding of a packet that was previously experiencing a collision, and the procedure continues iteratively until either the whole frames are cleaned or only slots with collisions remain. All the results derived next rely on perfect SIC (the assumption is justified by the analysis provided in [2]). Moreover, as in [2] we focus on an asymptotic setting where the number of slots allocated to frames and the population M tend to infinity following the proportions specified by the ratios G and α . This allows deriving upper bounds on the load that can be sustained with arbitrarily high reliability, characterizing system capacity. As to performance, we study: i) the probability P_L that an information unit cannot be retrieved at the end of the SIC process, i.e., that none of its replicas could be decoded; ii) the aggregate throughput \mathcal{T} , i.e., the average number of retrieved packets per available slot. By definition of G , and by the fact that $N_a(1 + \alpha)$ slots can be accessed for transmission, $\mathcal{T} = G(1 - P_L)/(1 + \alpha)$ follows.

The system can be conveniently represented resorting to the tripartite graph $\mathcal{G} = (B, S_a, S_b, E)$ of Fig. 1, where each of the M users is mapped to a *burst node* (BN) in the set B . Slots in the primary channel populate the set of *sum nodes* (SNs) in S_a , while redundancy slots composing C_b are mapped to *sum nodes* in S_b . An edge $e \in E$ connects a burst node b to a sum node s if and only if the terminal associated to b transmits one of its replicas in slot s . For the considered scheme, the degree of a BN follows the probability distribution Λ , where one of the ℓ outgoing edges flows to S_a , while the remaining ones connect to S_b . The introduced model allows to capture the behavior of the iterative SIC process borrowing tools from the theory of density evolution for multi-edge type low-density parity-check codes over the binary erasure channel [1]. To this aim, let us start iteration i considering slots in C_a , and focus on a BN b of degree k . Let e be the edge connecting b with $s \in S_a$, and define as $q_{a,(i)}$ the probability that e is still unknown at the beginning of step i . This event occurs when, at the end of iteration $i-1$, none of the $k-1$ bursts transmitted by the terminal over C_b was revealed. Denoting as $p_{b,(i-1)}$ the probability that one of the $k-1$ edges connecting b to S_b is not removed at iteration $i-1$, we get: $q_{a,(i)} = p_{b,(i-1)}^{k-1}$. Similarly, let us define as $1 - q_{a,(i)}$ the probability that e is removed after inspection of s . This is verified as soon as all the other edges connected to s were removed as outcome of iteration $i-1$, so that the only remaining burst can be decoded. If s has degree d , $1 - p_{a,(i)} = (1 - q_{a,(i)})^{d-1}$.

Focusing on C_b , consider a SN $r \in S_b$ of degree d , and let now e indicate an edge connecting the degree- k BN b to r . We denote as $q_{b,(i)}$ the probability that e is not removed after the inspection of s . This condition is met when neither the edge con-

necting b to s at iteration i nor any of the other $k-2$ edges from b to S_b at iteration $i-1$ were removed. Indicating as $p_{b,(i-1)}$ the probability of one of the latter events, we obtain $q_{b,(i)} = q_{a,(i)} p_{b,(i-1)}^{k-2}$. Finally, $p_{b,(i)}$ can be evaluated following the same argument used for $q_{a,(i)}$, leading to $1 - p_{b,(i)} = (1 - q_{b,(i)})^{d-1}$.

The presented relations express the probability for an edge to be removed from the graph given the degree of the node (either SN or BN) it is connected to. The conditioning can be removed in the asymptotic setting by averaging over such edge distributions [1], [2], computing the evolution of the erasure probabilities at iteration i as:

$$\begin{aligned} q_{a,(i)} &= \sum_{\ell} \Lambda_{\ell} p_{b,(i-1)}^{\ell-1} & q_{b,(i)} &= p_{a,(i)} \sum_{\ell} \lambda_{\ell} p_{b,(i-1)}^{\ell-2} \\ p_{a,(i)} &= 1 - e^{-Gq_{a,(i)}} & p_{b,(i)} &= 1 - e^{-\frac{G(\bar{\Lambda}-1)}{\alpha} \cdot q_{b,(i)}} \end{aligned} \quad (1)$$

where $\lambda_{\ell} = \Lambda_{\ell}(\ell - 1)/(\bar{\Lambda} - 1)$, and $\bar{\Lambda} = \sum_{\ell} \ell \Lambda_{\ell}$ is the average number of bursts transmitted by a node over the available resources. Equations in (1) can be iterated to study the probability that a burst is retrieved at the end of the SIC process, setting initial conditions to $p_{a,(0)} = 1 - e^{-G}$, $q_{a,(0)} = 1$, $p_{b,(0)} = 1 - e^{-G(\bar{\Lambda}-1)/\alpha}$, $q_{b,(0)} = 1$. This, in turn, allows to derive the sought average packet loss rate for the scheme under consideration, as an information unit is lost when none of its replicas could be retrieved:

$$P_L = p_a \sum_{\ell} \Lambda_{\ell} p_b^{\ell-1}$$

where $\{p_a, p_b\}$ are the convergence values of iterations (1).³

III. CAPACITY BOUNDS

The framework of Section II highlights the key role played by the degree distribution Λ on system performance. From this standpoint, the presented scheme intrinsically differs from other approaches (e.g., [2], [5]), in that only a fraction of the slots can be used to distribute replicas. We thus start by characterizing this limitation and evaluating the associated tradeoffs. To this aim, we introduce the *system capacity with vanishingly small error probability* \mathcal{C}_0 :

$$\mathcal{C}_0(\alpha, R) = \sup_{\{\Lambda, \bar{\Lambda}=1/R\}} \{\mathcal{T}(\Lambda) | \alpha, P_L \rightarrow 0\}$$

which assesses the maximum throughput the system can deliver while keeping $P_L \rightarrow 0$, as a function of the redundancy α and of the average repetition rate $R = 1/\bar{\Lambda}$. Within this framework, the following key results hold:

Theorem 1: For any $\alpha > 0$, $\mathcal{C}_0(\alpha, R) \leq G^*/(1 + \alpha)$, where G^* is the unique solution in $(0, 1)$ of

$$R = \left[1 + \frac{\alpha}{G} \ln \left(\frac{\alpha}{\alpha - G + (1 - e^{-G})} \right) \right]^{-1}.$$

Proof: By definition of \mathcal{C}_0 , we are interested in operating the channels such that data can be successfully retrieved out of each frame. Let us consider an alternative model for the system under consideration, where each of the information units generated by the M users is mapped onto an element of a

³In the asymptotic setting, the neighborhood of a node in the graph is a tree with probability close to 1. This allows neglecting dependencies in the collision patterns (i.e., cycles in the graph representation), and leads (1) to hold. The impact of cycles on finite length frames will be discussed later.

Galois field $\text{GF}(2^l)$ of suitable cardinality, so that all replicas of a unit correspond to the same symbol. In turn, the superposition of packets over a slot is mapped at the receiver to the sum in $\text{GF}(2^l)$ of the elements corresponding to the transmitted bursts. When considering jointly C_a and C_b , then, a system of k linear equations in M unknowns is built at the receiver and, by the Rouché-Capelli theorem, a solution can be found only if $k \geq M$. On the other hand, a slot contributes with an equation if and only if at least one node transmitted over it. For C_a this happens with probability $1 - e^{-M/N_a}$, while for the redundancy channel the condition is met with probability $1 - \exp(-M \times (\bar{\Lambda} - 1)/N_b)$ since on average $\bar{\Lambda} - 1$ replicas are sent by each user over C_b . Adding up the average number of non-empty slots over the two channels, and recalling the definition of R , a necessary condition for retrieving the packets is obtained as

$$N_a(1 - e^{-G}) + \alpha N_a \left(1 - e^{-\frac{G(1-R)}{\alpha R}}\right) \geq M. \quad (2)$$

The statement follows by simple manipulations of (2), and by the definition of $C_0(\alpha, R)$. ■

Corollary 1: For any $\alpha > 0$, let G be a load at which the system can operate with vanishingly small error probability. Then, $G < G_{\max}$, where G_{\max} is the unique real solution of:

$$\alpha = G - (1 - e^{-G}). \quad (3)$$

Proof: Dividing both members of (2) by αN_a , the statement follows immediately as necessary condition for the inequality to hold. ■

The theorem offers interesting insights on the capacity region for the system by stating an upper bound to C_0 , shown in Fig. 2 for the exemplary case $\alpha = 1$. The curve reports a limit to the maximum throughput that can be targeted for a given R or, equivalently, the maximum normalized traffic $G/(1 + \alpha)$ that can be sustained over the set of available resources ensuring vanishing P_L . Following this interpretation, a bound to the capacity when the rate is made arbitrarily low is given by Corollary 1 as $C_0(\alpha, R) \leq G_{\max}/(1 + \alpha)$. For the case under consideration, $G_{\max}/(1 + \alpha) \simeq 0.92$, whereas for a RA system where the complete channel is available to distribute burst replicas, a normalized load arbitrarily close to 1 can be approached for sufficiently low rates [3]. Eq. (3), thus, effectively captures the lower efficiency induced by the sub-optimal use of the primary channel. Moreover, the corollary offers a powerful tool in terms of system design, as it allows to determine the minimum amount of redundancy α that needs to be allocated (i.e., the cost to be paid) in order to reliably sustain a target load over the legacy system.⁴ The tightness of the bound can be verified analytically in the asymptotic setting resorting to the framework of Section II. To this aim, the behavior of the system given Λ has been evaluated by iterating the recursions in (1), while the degree distribution has been optimized resorting to differential evolution [7], targeting the maximum load providing vanishing P_L with degree $L \leq 15$. The results, reported by red-marked points in Fig. 2, show how the characterization of C_0 in Theorem 1 is indeed very accurate. Let us also recall that the presented scheme resorts to a simple form of coding, where

⁴As per (3), an exponential growth in $G_{\max}(\alpha)$ is obtained when $\alpha < 1$, later subsiding to a linear trend. Thus, values of α up to 1 offer a strong payoff, whereas solutions allocating more redundancy become appealing to support high loads in congested primary channels.

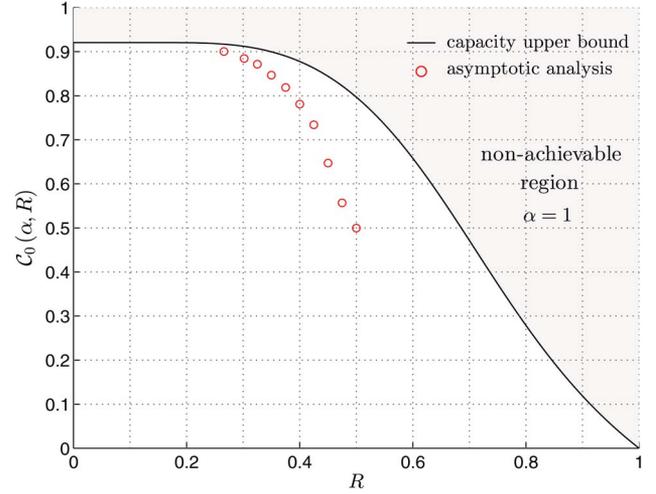


Fig. 2. Upper bound to the system capacity with vanishing error probability vs rate. Red points report analytical performance in the asymptotic setting.

a terminal sends equal burst replicas. Then, the system cannot be operated at rate higher than 1/2, with the low efficiency of repetition coding becoming pronounced when R is increased. Conversely, the algebraic approach used for Theorem 1 does not make any specific assumption on the type of redundancy sent over C_b . Thus, the statement applies to other forms of linear block coding as well (e.g., [2]), extending the capacity bound to any R .

Following Corollary 1, two operating regions can be identified. In fact, while for $G < G_{\max}$ distributions for which the throughput approaches the offered load can be found, when $G \geq G_{\max}$ it is intrinsically not possible to offer a vanishingly small error probability to all the users. Although typically neglected in the literature, such an *overload region* becomes relevant in practical scenarios, as additional resources are likely to be added exactly when the primary channel is congested. From this standpoint, let us remove the constraints on R to introduce the *capacity of the system at load G* as

$$\mathcal{C}(\alpha, G) = \sup_R \{\mathcal{T}|\alpha, G\}$$

and characterize the maximum attainable throughput when no reliability condition is imposed. Then, we have

Theorem 2: For any $\alpha > 0$, $\mathcal{C}(\alpha, G)$ is upper-bounded as:

$$\begin{cases} \mathcal{C}(\alpha, G) \leq \frac{G}{1+\alpha} & \text{for } G < G_{\max} \\ \mathcal{C}(\alpha, G) \leq \frac{G\Lambda_1^* e^{-G\Lambda_1^*} + G(1-\Lambda_1^*)}{1+\alpha} & \text{for } G \geq G_{\max} \end{cases}$$

where Λ_1^* is the unique solution in $[0,1)$ of

$$\alpha = G(1 - \Lambda_1) - e^{-G\Lambda_1} \left(1 - e^{-G(1-\Lambda_1)}\right).$$

Proof: For $G < G_{\max}$ the proposition stems directly from Theorem 1, as up to the load threshold all the traffic can be retrieved with arbitrarily small P_L . Let us focus instead on a specific G in the overload region, and consider a generic distribution Λ . Removing the assumption of vanishing P_L , we allow a fraction Λ_1 of the nodes to transmit only over C_a . Accordingly, an upper bound to the throughput can be derived assuming that all the information units sent by the fraction $1 - \Lambda_1$ of nodes that access the redundancy channel are correctly retrieved. In this case, the burst sent by each of them

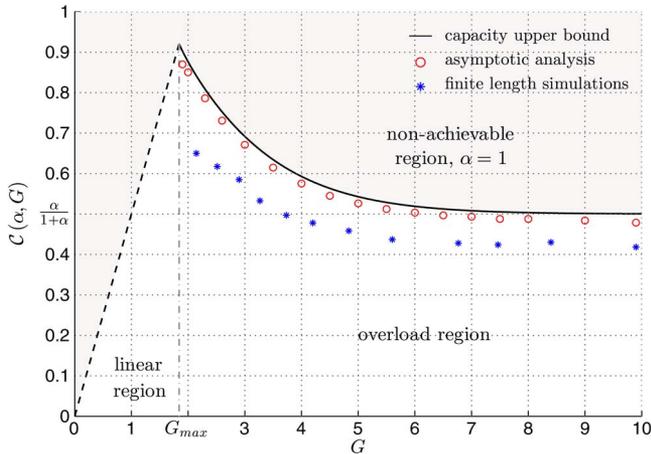


Fig. 3. Upper bound to the capacity of the system vs channel load.

over the legacy channel can be removed through SIC, reducing the load experienced in C_a to $G\Lambda_1$. Thus, C_b allows to collect at most $M(1 - \Lambda_1)$ data units, while the remaining $M\Lambda_1$ ones experience the loss rate of a SA channel. It follows that $\mathcal{T} \leq [G\Lambda_1 e^{-G\Lambda_1} + G(1 - \Lambda_1)] / (1 + \alpha)$. Observing that the bounding function is monotonically decreasing in Λ_1 , a limit to the system capacity is obtained by finding the minimum value Λ_1^* for which all the traffic on C_b can be decoded with vanishing P_L for the load G of interest. To achieve this, we apply once more the algebraic approach followed for the proof of Theorem 1. In this case, though, we aim at retrieving just the $M(1 - \Lambda_1)$ unknowns that are also sent over C_b . Conversely, packets transmitted by the $M\Lambda_1$ nodes that only access C_a represent symbols that do not contribute to the resolution of the system of equations. Therefore, when considering C_a , only non-empty slots that contain solely packets transmitted by the $M(1 - \Lambda_1)$ terminals of interest are to be considered. On average, their number evaluates to $N_a e^{-G\Lambda_1} (1 - e^{-G(1-\Lambda_1)})$. On C_b , instead, all non-empty slot contribute, for an average number $\alpha N_a (1 - e^{-G(1-\Lambda_1)(1-R)/\alpha R})$ of equations. By the Rouché-Capelli theorem, a necessary condition for the inequality to hold is derived as $\alpha > G(1 - \Lambda_1) - e^{-G\Lambda_1} (1 - e^{-G(1-\Lambda_1)})$. The sought value of Λ_1^* for any G follows directly. ■

The upper bound on system capacity is reported in Fig. 3 for $\alpha = 1$, highlighting the linear and overload regions. The tightness of the result is again verified analytically in the asymptotic setting (red markers), optimizing for each G the degree distribution Λ through differential evolution to maximize the system throughput. As expected, capacity decreases more than linearly when the load is raised, eventually approaching $\alpha / (1 + \alpha)$. Indeed, for very large G , the primary channel gets congested, bringing almost no value to \mathcal{T} and not supporting the SIC procedures on the secondary channel. The optimal working condition for the system is thus to consider C_b alone, which can approach a throughput of one packet per each of the N_b slots. By normalization, $\mathcal{T} \rightarrow N_b / (N_a + N_b) = \alpha / (1 + \alpha)$. The characterization of the overload region is particularly useful for system design: not only does it set the performance that can be targeted given a certain α , but also, for any desired operating point G on the legacy channel, it quantifies the minimum amount of redundancy that has to be provided to approach such limit. To evaluate the behavior of the

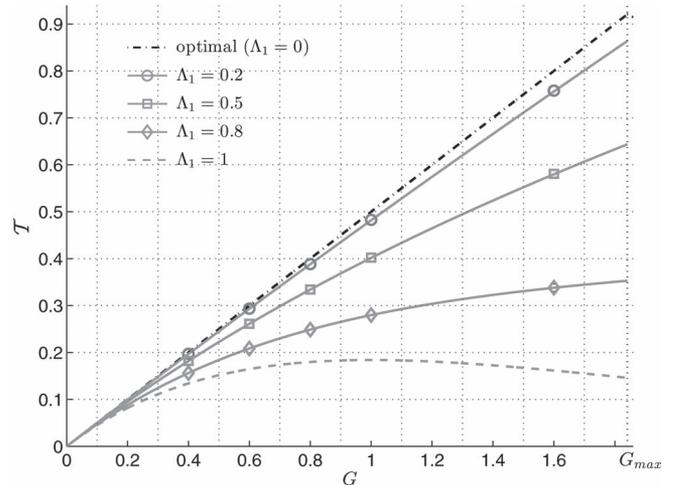


Fig. 4. Maximum achievable throughput in the linear load region when only a fraction $1 - \Lambda_1$ of nodes is sending redundancy. Lines report capacity bounds, while markers have been obtained via iterations of (1). $\alpha = 1$.

scheme for finite-length frames, simulations were performed (blue markers in Fig. 3). Here, the system was operated over 1000-slot frames using the optimized distributions obtained via the asymptotic analysis. As discussed, in the finite-length regime SIC is hindered by the presence of cycles and correlated collision patterns. Nonetheless, the analytical trends are confirmed in realistic configurations as well, validating the presented design tools.

Finally, we remark that the inequality derived in Theorem 2, $\mathcal{T} \leq [G\Lambda_1 e^{-G\Lambda_1} + G(1 - \Lambda_1)] / (1 + \alpha)$, can also be applied for $G < G_{max}$. In this case, operating the system with $\Lambda_1 > 0$ is suboptimal, as some nodes will not be guaranteed a vanishing P_L , but may be of practical relevance. In fact, if we think of the legacy channel as operated per an established SA standard, Λ_1 may represent the portion of *basic* terminals which cannot access the new resources, complemented by a fraction $1 - \Lambda_1$ of smarter, while standard-compliant, nodes. Fig. 4 reports the bound on \mathcal{T} when $G < G_{max}$, for $\alpha = 1$ and different values of Λ_1 . In the case under study, already for Λ_1 equal to 0.8 and 0.5, the peak throughput over the legacy scheme increases by a factor 2 and 3.5, respectively. Such results prompt the use of ancillary channels as a promising approach to increase the capacity of SA.

REFERENCES

- [1] T. Richardson and R. Urbanke, *Modern Coding Theory*. Cambridge, U.K.: Cambridge Univ. Press, 2008.
- [2] E. Paolini, G. Liva, and M. Chiani, "Coded Slotted ALOHA: A graph-based method for uncoordinated multiple access," *IEEE Trans. Inf. Theory*, submitted for publication. [Online]. Available: <http://arxiv.org/abs/1401.1626>
- [3] K. Narayanan and H. Pfister, "Iterative collision resolution for slotted ALOHA: An optimal uncoordinated transmission policy," in *Proc. ISTC*, 2012, pp. 136–139.
- [4] R. Urbanke and B. Rimoldi, "Coding for the F -adder channel: Two applications of reed-solomon codes," in *Proc. IEEE ISIT*, 1973, p. 85.
- [5] C. Stefanovic and P. Popovski, "ALOHA random access that operates as a rateless code," *IEEE Trans. Commun.*, vol. 61, no. 11, pp. 4653–4662, Nov. 2013.
- [6] C. Stefanovic and P. Popovski, "Coded slotted ALOHA with varying packet loss rate across users," in *Proc. IEEE GlobSIP*, 2013, pp. 787–790.
- [7] R. Storn and K. Price, "Differential evolution—A simple and efficient adaptive scheme for global optimization over continuous spaces," *J. Global Optim.*, vol. 11, no. 4, pp. 341–399, Dec. 1995.