Computer Communications 59 (2015) 12–23

Contents lists available at ScienceDirect

Computer Communications

journal homepage: www.elsevier.com/locate/comcom

Guaranteed capacity bounds in intermittently-connected networks: A resource-aware, holistic evaluation

Gabriel Sandulescu *, Aimaschana Niruntasukrat, Chalermpol Charnsripinyo

National Electronics and Computer Technology Center, Klong 1, Pathumthani, Thailand

ARTICLE INFO

Article history:
Received 27 November 2013
Received in revised form 2 October 2014
Accepted 20 December 2014
Available online 31 December 2014

Keywords:
Intermittently-connected networks
Disruption-tolerant networks
Capacity
Admission control
Quality of service

ABSTRACT

Capacity is difficult to estimate in sparse, intermittently-connected networks since timely feedback and reliable signalling is usually unavailable. This paper proposes a resource-aware, holistic framework for estimating and guaranteeing achievable capacity between pairs of nodes in such networks. The proposed framework builds on elements gathered autonomously by nodes, therefore results emerge from actual network properties: mobility, routing and resource distribution. Another important element is how data are quantized into messages at source and then injected in a single burst. Achievable capacity is formulated as a linear programming problem in periodic scenarios, then the model is extended using probabilistic bounds in random scenarios. The paper also discusses solutions below mathematical optimality, which are less complex and therefore more suitable for real-life implementation. Extensive simulations have been performed using real traces as well as synthetic mobility, different store-carry-forward protocols, homogeneous and heterogeneous resource distribution. Simulations indicate that this framework can be used to avoid resource exhaustion and network congestion in heterogeneous environments where no information about the system is known in advance. It may contribute to building admission control schemes and ensuring service guarantees.

1. Introduction

The well-established mobile communication paradigm relies today to a great extent on infrastructure-based solutions, such as the cellular network. However, the surging growth in global mobile data traffic and the anticipation of a frequency spectrum crunch [1] is pushing the research community to seek for alternatives. The possible usage of unlicensed frequency spectrum, a smaller radio range potentially leading to lower energy requirements, the desire to benefit from node mobility and patterns of social behaviour and the relative ubiquity of ad hoc connections are as many compelling incentives for transmitting data over opportunistic, intermittently-connected networks [2]. Since an end-to-end path is typically not available in these networks, communication is made possible only through intermediary nodes acting as data custodians, delivering messages towards the destination using store-carry-forward protocols [3].

As mobile nodes usually have only limited resources available, some nodes may become congested during overload periods and unusable right away, causing even more disconnections and further decreasing network performance. In order to tackle this problem, some recent works have proposed congestion control protocols acting either as adaptive forwarding schemes or adaptive replication schemes [4,5]. However, in many opportunistic scenarios (post-disaster, open urban communications, accident monitoring) nodes are freely joining and leaving the network and there is no effective way to ensure that the entire population of nodes use only a specific, uniform set of protocols. For these cases, admission control based on capacity evaluation may be used as a practical and powerful tool to support service guarantees.

In fully connected wireless networks feedback is quick and capacity can be interactively inquired or even brokered between different flows. The same strategy does not work in intermittently-connected networks because feedback is particularly long and snapshot information may lose pertinence before even reaching consumption points. However, in order to avoid network congestion and node resource exhaustion, achievable capacity needs to be estimated prior to injecting large amounts of data into the network.

Due to the proliferation of user-generated content, such as high resolution photos or videos created by users on their mobile devices, files to be sent can be large and their injection points will be usually concentrated in specific locations [6]. Sending large files

* Corresponding author.
E-mail address: gabriel.sandulescu@gmail.com (G. Sandulescu).

http://dx.doi.org/10.1016/j.comcom.2014.12.008
0140-3664/

over short time spans leads to bursty data transmissions that can pose scheduling and resource allocation problems [3]. In this paper, we are interested in cases where large data files are quantized (i.e., input data are divided into blocks called messages) then injected in one burst from fixed geographical locations.

This paper tries to answer the following question: how can end-to-end capacity be predicted in a network lacking end-to-end connectivity? The achievable capacity between an ordered pair of nodes is defined as the maximum number of bits that can be delivered from source to destination over a period not exceeding the latency target. This value depends on mobility, on the store-carry-forward protocol used but also on the distribution of resources in the system and on how data are quantized into messages at source. Injecting a larger amount of data than achievable capacity will drive some nodes into resource exhaustion and possibly lead to partial data delivery. Informally, we would like to know the maximum amount of data that can be injected at the source node and delivered integrally to destination in a given time such that no custodian node will be resource-exhausted. Additionally, we want to find out if a certain level of capacity can be guaranteed, also for the cases when mobility and workload are varying.

Network performance in sparse, intermittently-connected networks has been studied extensively either analytically, considering simplified hypotheses [7–10] or exploring the optimal routing metrics [11–13]. However, few works have focused specifically on capacity, such as Nicolò and Giaccone [14] who have proposed an offline framework, therefore covering only periodic scenarios. Conversely, the current paper attempts to evaluate capacity combining an online framework and a holistic approach[1] to node resources, while also covering scenarios involving random mobility. In another work [15], Small and Haas propose an analytical method for calculating throughput capacity in intermittently-connected networks but its usage is limited to the epidemic protocol and a single type of resource (bandwidth) while also considering that resources are homogeneously distributed in the network. In our work, end-to-end capacity is estimated based on actual node mobility, natural workload and the store-carry-forward protocol already deployed in the network. Our work also attempts to capture the effect of the bursty injection of relatively large input messages, since its effects on node resource exhaustion are usually severe and particularly difficult to fix in networks lacking timely feedback.

The main contribution of this paper is an evaluation framework for calculating the achievable capacity between two nodes in a finite-resource, intermittently-connected network. The novelty of this paper lies in the following points: first, the paper attempts to characterize node resources from a holistic perspective, which makes their influence on message delivery more complete and easier to evaluate. Second, because classical network signalling is impractical in networks lacking end-to-end connectivity, the proposed framework relies upon normal messages (also called bundles) to collect information about available resources encountered on their way and to build the delivery graph at destination. Third, considering different hypotheses on how messages are quantized, the paper proposes three methods for computing the end-to-end capacity from the delivery graph when the mobility pattern is known in advance (or periodic, e.g. public transportation, deep-space communications). Fourth, achievable capacity in random scenarios is approached by probabilistic bounds, such that upper and lower bounds are derived by sampling historical values. Finally, capacity guarantees are formulated for intermittently-connected networks in random scenarios given a particular confidence level and a number of samples.

[1] Since the proposed strategy is based on a global analysis of information available about three resource types (energy, buffer space and bandwidth) we have called our approach holistic.

We would like to note that a preliminary version of this study appeared in a conference version in [16]. As compared to the preliminary version, the current paper has been extended in the following directions:

(A) An analytical module has been devised in such a way as to guarantee that the proposed capacity bounds are not exceeded. This module is based on a distribution-free confidence interval for p-quantiles. By applying it, the current paper formulates service guarantees in intermittently-connected networks.
(B) Simulations have been performed not only against synthetic mobility but also against more realistic, traces-based mobility. In the current version we use traces available from a public database, the so-called San Francisco Cabs [17] traces.
(C) For the random mobility case this study now discusses both upper and lower bound on achievable capacity. Lower bound denotes capacity that can be achieved under lenient conditions, such as equi-sized messages and original protocol queue management. Upper bound denotes capacity that can be achieved only under optimal message quantization and optimal queue management. Only the lower bound was discussed in [16].
(D) In support of our capacity accuracy claims, we need to compare 2 delay datasets and state whether or not they may come from the same distribution. In the current paper this is done analytically, using Kolmogorov–Smirnov tests for two datasets, while [16] only offered weaker evidence based on the position of quartiles in support of this claim.

The rest of the paper is organized as follows. Section 2 reviews published works addressing related topics. Section 3 presents the network model considered in this paper. Section 4 provides a detailed explanation of the evaluation framework. Results obtained are presented in Section 5, while Section 6 concludes the paper.

2. Related work

Several works [7–10] have tried to assess performance limits in sparse networks by exploring the issue analytically. However, for reasons of mathematical tractability, their underlying model usually assumes a simple and known store-carry-forward protocol, a known and uniform mobility model, while they usually ignore several resource limitations, such as energy, buffer or bandwidth constraints. While these models are theoretically interesting, their estimations are difficult to use in real settings, except for the most simple ones, due to the over-simplification of the model and the heterogeneity of realistic settings.

On the other hand, there are works that have relied on simulation for their calculations [18–20]. They assume that protocol parameters can be discovered through simulation and then applied to real networks. While this approach appears to be more realistic than the former, the heterogeneity and dynamics of real-life scenarios usually exceed what can reasonably be studied by simulation. In practice, one of the sources of heterogeneity that is the most difficult to track is resource distribution over the node population [21].

The lack of a steady end-to-end path makes the traditional congestion control policy based on feedback impossible to apply in intermittently-connected networks. Specific schemes were proposed for these networks such as [4] which reacts by decreasing the load on the congested regions by offloading the traffic to uncongested ones or [5] which adaptively limits the replication a node performs during each encounter. Both schemes take into account only buffer space, ignoring energy or connection duration (or bandwidth) constraints. It appears however that proactive