Computer Networks 53 (2009) 2321–2339

Contents lists available at ScienceDirect

Computer Networks

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

T-Man: Gossip-based fast overlay topology construction ☆

Márk Jelasity [a,*], Alberto Montresor [b], Ozalp Babaoglu [c]

[a] Research Group on AI, University of Szeged and HAS, P.O. Box 652, H-6701 Szeged, Hungary
[b] Dipartmento di Ingegneria e Scienza dell'Informazione, University of Trento, via Sommarive 14, I-38050 Povo (TN), Italy
[c] Dipartimento di Scienze dell'Informazione, University of Bologna, mura Anteo Zamboni 7, I-40126 Bologna, Italy

ARTICLE INFO

Article history:
Available online 1 April 2009

Keywords:
Gossip-based protocols
Overlay networks
Bootstrapping
Self-organizing middleware

ABSTRACT

Large-scale overlay networks have become crucial ingredients of fully-decentralized applications and peer-to-peer systems. Depending on the task at hand, overlay networks are organized into different topologies, such as rings, trees, semantic and geographic proximity networks. We argue that the central role overlay networks play in decentralized application development requires a more systematic study and effort towards understanding the possibilities and limits of overlay network construction in its generality. Our contribution in this paper is a gossip protocol called T-Man that can build a wide range of overlay networks from scratch, relying only on minimal assumptions. The protocol is fast, robust, and very simple. It is also highly configurable as the desired topology itself is a parameter in the form of a ranking method that orders nodes according to preference for a base node to select them as neighbors. The paper presents extensive empirical analysis of the protocol along with theoretical analysis of certain aspects of its behavior. We also describe a practical application of T-Man for building Chord distributed hash table overlays efficiently from scratch.

1. Introduction

Overlay networks have emerged as perhaps the single-most important abstraction when implementing a wide range of functions in large, fully-decentralized systems. The overlay network needs to be designed appropriately to support the application at hand efficiently. For example, application-level multicast might need carefully controlled random networks or trees, depending on the multicast approach [1,2]. Similarly, decentralized search applications benefit from special overlay network structures such as random or scale-free graphs[3,4], super-peer networks [5], networks that are organized based on proximity and/or capacity of the nodes [6,7], or distributed hash tables (DHT-s), for example, [8,9].

In current work, protocol designers typically assume that a given network exists for a long period of time, and only a relatively small proportion of nodes join or leave concurrently. Furthermore, applications either rely on their own idiosyncratic procedures for implementing join and repair of the overlay network or they simply let the network evolve in an emergent manner based on external factors such as user behavior.

We believe that there is room and need for interesting research contributions on at least two fronts. The first concerns the question whether a single framework can be used to develop flexible and configurable protocols without sacrificing simplicity and performance to tackle the plethora of overlay networks that have been proposed. The second front concerns scenarios in overlay construction that are often overlooked, such as massive joins and leaves, as well as quick and efficient bootstrapping of a desired overlay from scratch or some initial state. Current approaches either fail

☆ This work was completed while the authors were with the University of Bologna, Italy

* Corresponding author. Tel.: +36 62 544127.
E-mail addresses: jelasity@inf.u-szeged.hu (M. Jelasity), montreso@dit.unitn.it (A. Montresor), babaoglu@cs.unibo.it (O. Babaoglu).
URL: http://www.inf.u-szeged.hu/~jelasity/ (M. Jelasity).

1389-1286/$ - see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.comnet.2009.03.013

or are prohibitively expensive in such scenarios. Combining results on these two fronts would enable several interesting possibilities. These include: (i) overlay network creation *on demand*, (ii) deployment of temporary and adaptive decentralized applications with custom overlay topologies that are designed on-the-fly, (iii) federation or splitting of different existing architectures [10].

In this paper we address both questions and present an algorithm called T-Man for creating a large class of overlay networks from scratch. The algorithm is highly configurable: the network to be created is defined compactly by a *ranking method*. The ranking method formalizes the following idea: when shown a set of nodes, we assume each node in the network is able to decide which ones it likes from the set more and which ones it likes less (we will later use this ability of nodes to help them have neighbors they like as much as possible). In other words, each node can order any set of nodes. Formally speaking, the ranking method is able to order any set of nodes given a so called *base node*. By defining an appropriate ranking method, we will be able to build a wide variety of topologies, including sorted rings, trees, toruses, clustering and proximity networks, and even full-blown DHT networks, such as the Chord ring with fingers. T-Man relies only on an underlying peer sampling service [11] that creates an initial overlay network with random links as the starting point.

The algorithm is gossip-based: all nodes periodically communicate with a randomly-selected neighbor and exchange (bounded) neighborhood information in order to improve the quality of their own neighbor set. This approach, while requiring no more messages than the heartbeats already present in proactive repair protocols, is simple, and achieves fast and robust convergence as we demonstrate.

In this paper we limit our study to the overlay construction problem. Using T-Man for overlay maintenance is also possible [12] with performance and cost that are not dramatically different from existing periodic repair protocols currently used in most overlay networks. The originality and attractiveness of T-Man as a maintenance protocol lies in its generality and configurability. The main contribution of this paper is to show that a single, generic gossip-based algorithm can *create* many different overlay networks *from scratch* quickly and efficiently.

1.1. Related work

Related work in bootstrapping include the algorithm of Voulgaris and van Steen [13] who propose a method to jump-start Pastry [9]. This protocol is specifically tailored to Pastry and its message complexity is significantly higher than that of T-Man. More recently, the bootstrapping problem has been addressed in other specific overlays [14–16]. These algorithms, although reasonably efficient, are specific to their target overlay networks.

An approach closer to T-Man is Vicinity, described in [17]. Although Vicinity was inspired by the earliest version of T-Man, it does contain notable original components related to overlay maintenance, such as churn management, and other techniques to boost performance.

Finally, we mention related work that use gossip-based probabilistic and lightweight algorithms. We note that these algorithms are targeted neither at efficient bootstrapping, nor at generic topology management. Massoulié and Kermarrec [18] propose a protocol to evolve a topology that reflects proximity. More recent protocols applying similar principles include [19] and [20]. Repair protocols used extensively in many DHT overlays also belong to this category (e.g., [8,21,22]).

1.2. Contribution

Our contribution with respect to related work is threefold. First, we introduce a lightweight probabilistic protocol that can construct a wide range of overlay networks based on a compact and intuitive representation: the ranking method. The protocol has a small number of parameters, and relies on minimal assumptions, such as nodes being able to obtain a random sample from the network (the peer sampling service). The protocol is an improved and simplified version of earlier variants presented at various workshops [12,23,10]. Second, we develop novel insights for the tradeoffs of parameter settings based on an analogy between T-Man and epidemic broadcasts. We describe the dynamics of the protocol considering it as an epidemic broadcast, restricted by certain factors defined by the parameters and properties of the ranking method (that is, the properties of the desired overlay network). We also analyze storage complexity. Third, we present novel algorithmic techniques for initiating and terminating the protocol execution. We describe how to construct the Chord overlay as a practical application of T-Man. We present extensive simulation results that support the efficiency and reliability of T-Man.

1.3. Road map

Sections 2 and 3 present the system model and the overlay construction problem. Section 4 describes the T-Man protocol. In Section 5 we present theoretical and experimental results to characterize key properties of the protocol and to give guidelines on parameter settings. Section 6 presents practical extensions to the protocol related to bootstrapping and termination, and extensive experimental results are also given to examine the behavior of the protocol in different failure scenarios. Section 7 presents a practical application: the creation of the Chord overlay network [8]. Section 8 concludes the paper.

2. System model

We consider a set of nodes connected through a routed network. Each node has an address that is necessary and sufficient for sending it a message. Furthermore, all nodes have a *profile* containing any additional information about the node that is relevant for the definition of an overlay network. Node ID, geographical location, available resources, etc. are all examples of profile information. The address and the profile together form the *node descriptor*. At times, we will use "node descriptor" and "node" interchangeably if this does not cause confusion.