Characterizing Two-tier Topology in Modern File-Sharing Applications

Author's name Ommitted for Anonymity

ABSTRACT

Modern unstructured Peer-to-Peer (P2P) file-sharing applications are becoming increasing popular and use a two-tier architecture to improve their scalability. Despite their impact on the Internet, little is known about the behavior of these applications, in particular the characteristics of their unstructured overlay topology. Deriving such characterizations requires capturing accurate snapshots of the overlay topology which is inherently challenging due to its dynamic nature.

This paper presents the first detailed characterization of two-tier overlay topologies in a popular unstructured P2P application, namely Gnutella. We describe fundamental challenges in capturing accurate snapshots, present a methodology for quantifying the accuracy of snapshots, and demonstrate that inaccurate snapshots can lead to erroneous conclusions–such as a power-law degree distribution. We developed a set of measurement techniques into a parallel P2P crawler, called Cruiser, to efficiently and accurately capture snapshots of large scale two-tier overlay topologies. We have more than 15,000 snapshots, captured during the past six months by Cruiser. We characterize the graph-related properties of individual snapshots and the dynamics of the overlay topology, and investigate their underlying causes and implications. This study provides essential insights into the behavior of overlay topologies which are necessary to improve the design and evaluation of file-sharing applications.

Keywords

Peer-to-peer, File sharing, Overlay topology, Gnutella

1. INTRODUCTION

During recent years, the Internet has witnessed the explosive growth in popularity of various Peer-to-Peer (P2P) applications. In particular, today's P2P file-sharing applications (*e.g.*, FastTrack, eDonkey, Gnutella) are extremely popular with millions of clients [] which undoubtedly have an impact on Internet traffic [1]. In file-sharing P2P applications, participating peers often form an overlay which is used by individual peers to search for desired resources (*e.g.*, files) among other participating peers. There is neither any control, nor any constraint or coordination in overlay formation by participating peers in these applications. The overlay topology is formed by peers joining (or leaving) the network at any arbitrary time, based on some loosely defined (and possibly different) set of rules, and may change their connections to the network in response to departing neighbors. In a nutshell, participating peers collectively form an "unstructured" and dynamically changing overlay as peers join and leave the system. A new generation of popular file-sharing applications have adopted a *two-tier* architecture to improve their scalability. In this architecture, a subset of peers, called *ultrapeer*, form a top-level overlay whereas other participating peers, called *leaf peers*, are connected to the system through multiple ultrapeers.

Design and simulation-based evaluation of (both structured and unstructured) file sharing applications have received lots of attention during recent years. However, little is known about characteristics of popular P2P file sharing applications over the Internet, in particular, after introduction of the two-tier architecture by these applications. Deriving such characterizations provides

[1]it is really difficult to accurately quantify the overall contribution of these applications in Internet traffic because they choose different port numbers

unique insights into the behavior of these decentralized systems in a real setting (*i.e.*, realistic group size, degree of heterogeneity, workload, network and peer dynamics) which is very hard (if not impossible) to obtain through simulation or modeling. Deriving such characterizations is essential to examine any potential performance bottleneck, limitations or design anomalies in P2P systems in practice, as well as their impact on the Internet. For example, such characterization enables us to build an overlay topology generator that is necessary for meaningful evaluation of P2P applications. There are three key aspects of unstructured P2P file sharing applications that should be characterized: *(i)* query workload [**?**, **?**], *(ii)* file distribution (or replication) [**?**, **?**] and overlay network topology [**?**]. While these three issues are equally important, to our surprise, there has not been any recent study on characterization of two-tier overlay topology in file sharing applications. We are aware of only two earlier studies on this issue. Ripeanu et al. [**?**] mapped the Gnutella network, and Saroiu et al. [**?**] briefly examined the resiliency of the Gnutella overlay topology in the face of attack. [XXX, any other study on other p2p systems] These studies are outdated (almost three years old), have only conducted a limited analysis on the old Gnutella protocol (before introduction of the two-tier architecture) that was significantly smaller than today's P2P applications.

Accurate characterization of overlay topology for a large scale P2P application is difficult. A common approach in these studies is to examine properties of snapshots of the overlay that are captured by a crawler. However, capturing accurate snapshots of these systems is hard for two reasons: *(i)* the dynamics nature of peer participation (*i.e.*, churn) and their pair-wise connection, and *(ii)* a large number of peers are not directly reachable. Furthermore, accuracy of a captured snapshots can not be verified since there is no reference snapshot for comparison[XXX]. Previous studies either deployed slow crawlers which inevitably lead to distorted (*i.e.*, stretched) snapshots of the overlay [**?**], or partially crawled the overlay [**?**] which is likely to capture biased (and non-representative) snapshots. [XXX, we may add that char of small and homogeneous systems such as Akamai does not represent a true p2p apps][these are outdated anyway] More importantly, to our knowledge, none of the previous measurement-based studies have examined the accuracy of their captured snapshots. In the absence of any reliable characterization of overlay topology, researchers use ad-hoc overlay for evaluation of their protocol.

In this paper, we present the first detailed characterization of the two-tier overlay topology in Gnutella based on recent measurements. This study makes two important contributions: **Measurement Methodology:** We describe fundamental challenges in capturing accurate snapshots, present a methodology for quantifying the accuracy of snapshots, and demonstrate that inaccurate snapshots can lead to erroneous conclusions– such as a power-law degree distribution. We developed a set of measurement techniques into a parallel P2P crawler, called *Cruiser*, to efficiently and accurately capture snapshots of large scale two-tier overlay topologies. Cruiser effectively leverages the two-tier architecture of the overlay along with the new handshaking mechanism in Gnutella to capture a complete snapshot of the Gnutella overlay in a few minutes. This implies that crawling speed by Cruiser is order of magnitude faster than any previous P2P crawler, and thus its captured snapshots are significantly more accurate. Having more accurate snapshots allows us to examine the dynamics of the overlay much shorter timescales which was not feasible in previous studies.

Characterization of Two-Tier Overlay: We have captured more than 15,000 snapshots of the Gnutella network during the past six months by Cruiser. We leverage this dataset to characterize the graph-related properties of individual snapshots and the dynamics of the overlay topology, and investigate their underlying causes and implications. To the extent possible, we conduct our analysis in a generic (*i.e.*, Gnutella-independent) fashion to ensure the applicability to other P2P systems. This study provides essential insights into the behavior of overlay topologies which are necessary to improve the design and evaluation of file-sharing applications. Some of our findings can be summarized as follows[to be revised]:

- The overall node degree does not exhibit a power-law distribution, differing from previous studies [**?**, **?**, **?**].
- A non-negligible portion of ultrapeers cannot accept incoming connections.
- The size of the Gnutella network has dramatically grown over the past couple of years. Despite this increase, the diameter of the topology remains low. More importantly the distribution of pair-wise path lengths has become more homogeneous with lower mean value. These desired properties have been maintained by the introduction of semi-structure to the topology, and increasing the degree of peers in the top-level overlay.
- The overall topology has become denser and exhibits clear small-world properties.
- Despite variations in the total number of peers with time of day, a large number of peers are available at any time, and the semi-structure remains balanced. (*i.e.*, the ratio between leaf to ultrapeers remains relatively constant).

1.1 Why Examining Gnutella?

Most popular P2P file-sharing applications, including eDonkey, FastTrack, and Gnutella, have adopted the two-tier overlay topology but they have different popularity. We need to answer the question that "why Gnutella is a good candidate for such characterization?". Gnutella is an open protocol with several mature implementations. First, Several evidences from different sources show that Gnutella network has been rapidly growing during recent months. Our direct measurement shows that the size of Gnutella snapshots (average number of participating peer at any point of time) has doubled over past 6 months, from 400K to 800K peers per snapshot, and more than XXX unique IP addresses. [XXX, graph for growth in size] Examination of Internet2 weekly logs revealed that Gnutella was ranked second or third in terms of its contribution in observed traffic among file-sharing application last year. However, it currently contributes significantly more than any other P2P file sharing applications [2]. Finally, unofficial information about population of users for different P2P applications [] shows that Gnutella has about one third of users of most popular P2P applications.

Second, Gnutella is the most popular P2P file-sharing with open source protocol. This eliminates (or at least significantly reduces) any "incompatibility" error in our measurement that could easily occur in other P2P applications that have been reverse-engineered, namely FastTrack and eDonkey. Furthermore, Gnutella provides a new handshaking feature for crwalers that significantly improves efficiency and accuracy of collected information by crawlers.

The rest of this paper is organized as follows. XXX¡ TBD] In Section **??**, we present our measurement methodology, and describe how we capture and postprocess snapshots of the Gnutella network, and examine their accuracy. A detailed characterization of the Gnutella overlay is presented in Section **??**. Section **??** presents a summary of related work. Finally, Section **??** concludes the paper and presents our future plans.

mention that impact of p2p applications on the network is hard o quantify since there is no reliable approach to capture their associated traffic

[2]These statistics should be view with a fairly big grain of salt because they often use port number to detect associated traffic to a particular P2P application. This approach is not accurate since P2P applications can use different port numbers.