Bootstrapping Peer-to-Peer Networks[1]

Chris GauthierDickey Christian Grothoff

{chrisg,grothoff}@cs.du.edu

http://crisp.cs.du.edu/

Colorado Research Institute for Security and Privacy

[1] Funded in part by the National Science Foundation under Grant No. DUE–0416969. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect those of the National Science Foundation.

Overview

- What is P2P bootstrapping
- Existing solutions
- Using DNS to improve brute-force
- Experimental results

P2P bootstrapping

Three main definitions of “bootstrapping”:

- Starting a new P2P network (freshly designed protocol)
- Once a P2P network is running, any new peer that joins must be integrated into the network
- Before a new peer can be integrated into an existing network, the new peer must somehow obtain contact information to at least one node in the existing P2P network

P2P bootstrapping

For this work:

- Bootstrapping is the process that a new peer who intends to join a P2P network uses to discover contact information for another peer in the existing network.
- This discovery process may also be used to heal a partitioned network, but that is not the dominant use-case.

Our goal: complete decentralization

- **Pure** P2P networks are P2P networks that do not rely on **any** centralized services

⇒ All nodes are equal – no prime targets for adversaries

⇒ Ideally, we need to be able to decentralize all operations

⇒ Need decentralized bootstrapping!

Existing Solutions

- Public hostlist server
- Contact list shipped with software

Problems with Existing Solutions

- Public hostlist server:
  - Attacker can target hostlist server
  - Server operation maybe costly
  - Easy way for attacker to learn quickly about participants in the network
  - How does the end-user learn about changes in hostlist server addresses?
- Contact list shipped with software

Problems with Existing Solutions

- Public hostlist server
- Contact list shipped with software:
  - List might become outdated quickly
  - Easy way for attacker to learn quickly about participants in the network

Existing Solutions

- Public hostlist server
- Contact list shipped with software
- Brute force

Problems with Existing Solutions

- Public hostlist server
- Contact list shipped with software
- Brute force:
  - Expensive: $\frac{2^{32}}{N}$ operations where N is size of the network
  - N at the order of $2^{16} - 2^{20}$ for popular P2P networks

  $\Rightarrow$ Feasible, but not great

Bias

P2P users have cultural and linguistic bias:

- Content shared differs between regions
- Software user interface maybe available only in certain languages
- User groups form social networks providing regional support

This bias will be reflected in the distribution of peers in the IP address space!

Key Idea

Modify the brute-force scan-the-world approach to improve its performance;

bias it towards the most promising IP addresses based on the skewed user distributions observed in the real-world.

Approach

1. Partition the IP address space into regions and organizations (using reverse DNS lookups)
2. Given extensive lists of IP addresses of peers, determine which regions or organizations are most likely to use the network
3. Distribute summary of distribution information with the P2P software
4. Bootstrapping peers use distribution data to bias global IP scan

Too Much or Too Little Bias?

1. Peers could focus global scan only on most promising organization

⇒ High success rate initially

2. Most promising organization's network is likely small

⇒ Few peers in most active organization overburdened with bootstrap-requests

⇒ Organization may abandon network, resulting in lower success rate

⇒ Decentralization not really perfect

Too Much or Too Little Bias?

Goal: Need to select appropriate point between high cost of an unbiased scan-the-world approach and a strongly-biased scan-the-best approach!

Perfect Bias

1. Define minimum acceptable expected number of probes to bootstrap based on performance requirements (i.e., do not probe more than 1,000 IP addresses)

2. Do not scan organizations with a lower probability.

3. Scan other organizations proportional to network size and probability of success

⇒ Minimum performance requirements met.

⇒ All peers (except for those in low-probability organizations) have equal chance of being used for bootstrapping

Requirements and Assumptions

- Most users of the P2P network use the same port
- Trying to connect to the particular port is acceptable network use, even if the target machine does not participate in the P2P network itself
- Bootstrapping does not need to be instantaneous

Experimental Setup

- Tested three different P2P networks:
- IP address was assigned to an organization if it was in a contiguous range of addresses with first and last IP address sharing the same SOA (with initial ranges being determined using traditional IP address classes)
- Used simple TCP handshake to validate that initial point of contact was found

Network Size

P2P Network	Unique IPs	Port
Gnutella (8/2007)	377,246	6346
eDonkey (10/2007)	80,728	411
DirectConnect (10/2007)	175,139	4662

DNS Networks by SOA

Network Size (# IPs)	# SOAs
2^0 to 2^8 IPs	60,921
2^8 to 2^{16} IPs	14,577
2^{16} to 2^{24} IPs	1,296
2^{24} to 2^{32} IPs	22
Total	76,816

Bias (for Gnutella)

Organization (SOA)	# IPs	# Peers
ns.pc-network.ro	254	15 (5.91%)
ns1.netplanet.ro	254	12 (4.72%)
ns.rdstm.ro	11,244	517 (4.60%)
	...	
ns-a.bbtec.net	10,829,308	4 (0.00%)
rev1.kornet.net	10,857,115	1 (0.00%)
Total	2^{32}	3,741,099 (0.09%)

Performance of Bootstrapping

P2P Network	Gnutella	E2DK	DC
Random global scan	2425 ± 3089	1875 ± 1780	3117 ± 3080
Biased, TLD only	833 ± 897	18 ± 43	1252 ± 1874
Biased, domainname	1150 ± 1181	74 ± 86	623 ± 1599
Biased, subdomain	849 ± 820	56 ± 71	1786 ± 2545
Biased, FQN	817 ± 856	51 ± 92	1397 ± 2320
Recent hostlist	245 ± 245	7039 ± 7185	217 ± 211

Impact of Age (for Gnutella)

Year	Hostlist	Biased, TLD only
2004	1487 ± 1305	1257 ± 1333
2005	1124 ± 1138	1659 ± 1651
2006	546 ± 506	983 ± 1139
2007	246 ± 245	833 ± 897

Advantages

- Distribution information likely ages better than IP lists
- Distribution information is less problematic with respect to privacy requirements than IP lists or hostlist servers
- A biased global scan is fully decentralized
- Global scans can help heal fragmented networks

Questions

?

Copyright