Temporal and Efficient Analysis of Services Availability

Johannes Klick, Stephan Lau
Matthias Wählisch, Volker Roth

{johannes.klick,stephan.lau,m.waehlisch,volker.roth}@fu-berlin.de

Freie Universität Berlin

Measuring Deployment of Internet Services

Objective

Identify hosts that provide a specific Internet service

Common Scanning Approaches

- IANA /0 – 4.3 billion IPv4 addresses
- IANA allocated – 3.7 billion IPv4 addresses
- BGP announced prefixes – 2.8 billion IPv4 addresses
- IP hitlists

Is this really a good idea?

Problems with Dumb Scanning

- Hitrates are often below two percent
- Abuse reports
- Rate limiting on routers
- Load on intrusion detection systems
- IP Blacklisting
- +++

We should scan less!

Proposed Solution: TASS

Topology Aware Scanning Strategy (TASS) in a nutshell:

1. Perfom a full IPv4 scan once
2. Select prefixes with a certain coverage of responsive hosts
3. Scan only the selected prefixes for a given time period

Result: Reduce scan traffic by 25-90 % and miss only 1-10% service responses

TASS in Detail:

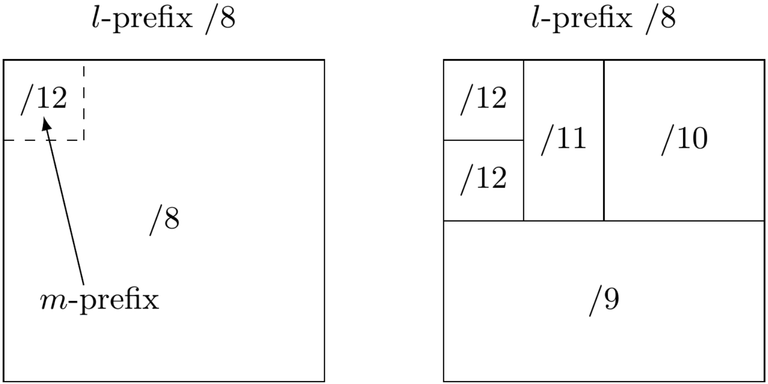
1. At time t_0, perform a full scan and output all responsive addresses. Let N be their number. Count the number of responsive addresses c_i in each responsive prefix i. The sum of all c_i is N.
2. Calculate the density $\rho_i = c_i / 2^{32-\text{prefix length}}$ of all responsive prefixes and their relative host coverage $\phi_i = c_i / N$ of responsive addresses.
3. Sort the prefixes in the descending order of density. Relabel prefixes so that $i < j \Leftrightarrow \rho_i > \rho_j$.
4. Find the smallest k so that $\sum_{i=1}^{k} \phi_i > \phi$.
5. Scan prefixes $1, \ldots, k$ repeatedly until time $t_0 + \Delta_t$, then start over at step 1.

Deriving Prefixes I

CAIDA Routeviews Prefix-to-AS database

1. Prefixes are not complementary
2. Less specific prefixes (l-prefixes) contain more specific prefixes (m-prefixes)
3. A single IP address can have multiple prefixes

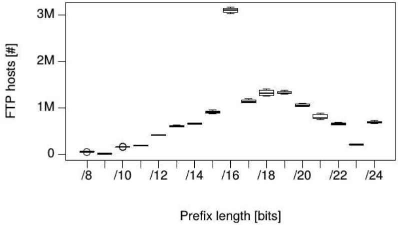
Deriving Prefixes II



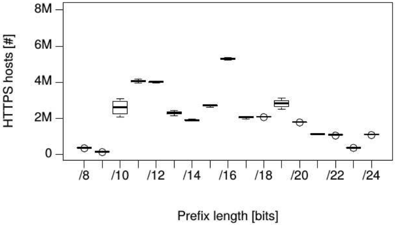
(a) Announced prefixes. **(b) Resulting m-prefixes.**

- The less specific *l*-prefix /8 contains the more specific *m*-prefix /12.
- The *l*-prefix is then decomposed into the more specific one and the remaining smaller prefixes

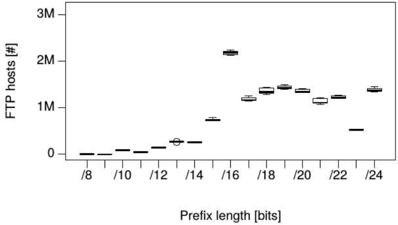
Host Stability vs. Prefix Length



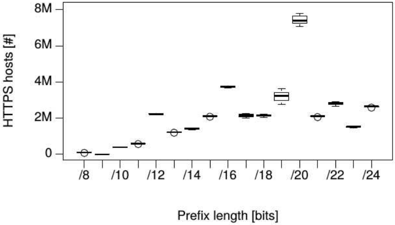
(a) FTP for less specific prefixes.



(b) HTTPS for less specific prefixes.



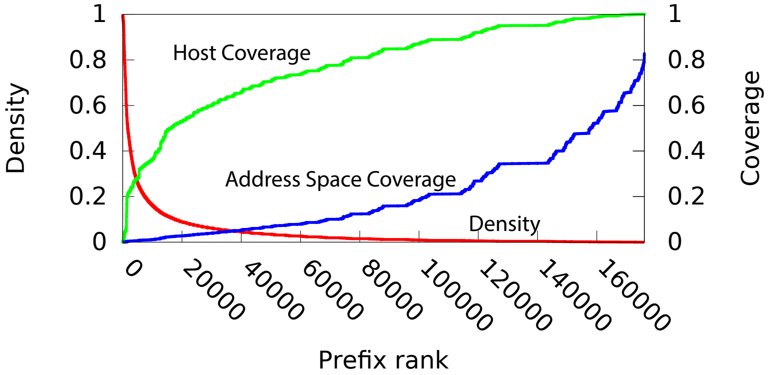
(c) FTP for more specific prefixes.



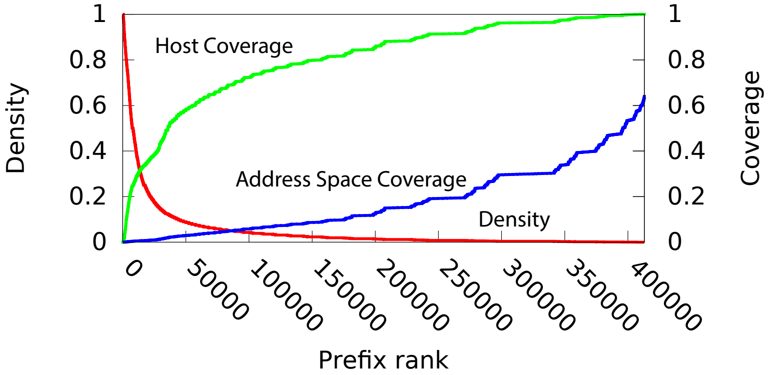
(d) HTTPS for more specific prefixes.

Host distribution over prefix lengths based on seven different measurements from 09/2015 to 03/2016. Datasource: censys.io.

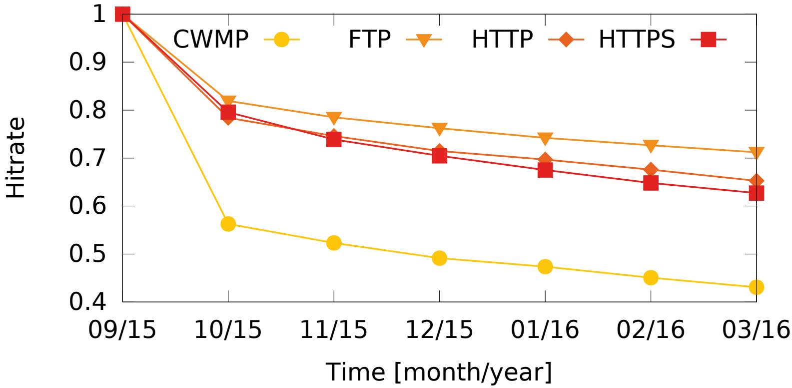
HTTPS (Less Specific Prefixes)



HTTPS (More Specific Prefixes)



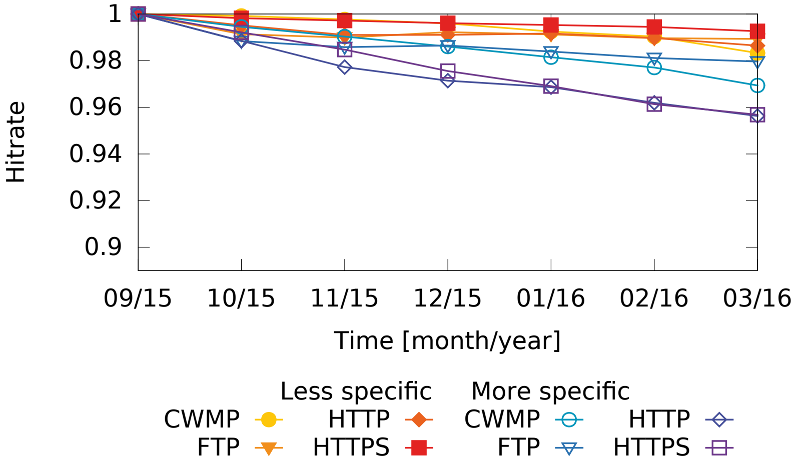
Accuracy over Time: IPv4 Hitlists



Hitrate of a IPv4 hitlist scan compared to IPv4 full scans.

Datasource: 4.1 TB from censy.io.

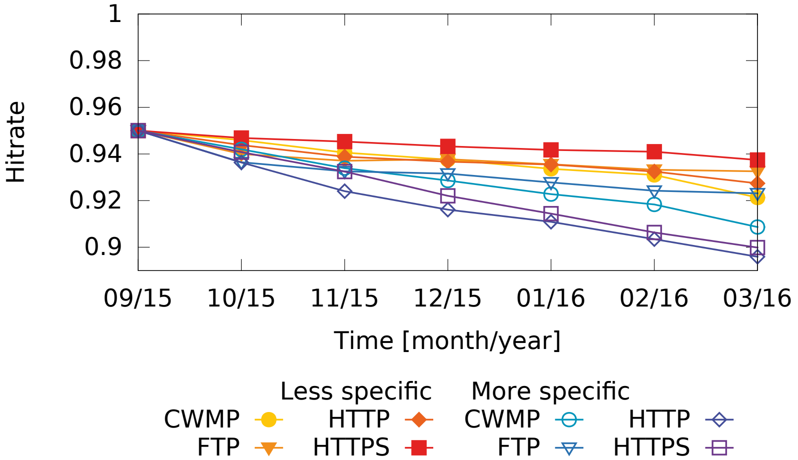
Accuracy over Time: TASS (Host Coverage 100%)



Hitrate of a TASS scan compared to IPv4 full scans.

Datasource: 4.1 TB from censy.io.

Accuracy over Time: TASS (Host Coverage 95%)



Hitrate of a TASS scan compared to IPv4 full scans.

Datasource: 4.1 TB from censy.io.

Future Work

- Detailed analysis of the skipped hosts
- Better understanding of service stability per AS type
- Analyses of longer time periods and more protocols
- IPv6 scans

Thanks!

More details:

„Towards Better Internet Citizenship: Reducing the Footprint of Internet-wide Scans by Topology Aware Prefix Selection“

`http://arxiv.org/pdf/1605.05856.pdf`

Towards Better Internet Citizenship: Reducing the Footprint of Internet-wide Scans by Topology Aware Prefix Selection

Johannes Klick
Freie Universität Berlin
johannes.klick@fu-berlin.de

Stephan Lau
Freie Universität Berlin
stephan.lau@fu-berlin.de

Matthias Wählisch
Freie Universität Berlin
m.waehlisch@fu-berlin.de

Volker Roth
Freie Universität Berlin
volker.roth@fu-berlin.de

arXiv:1605.05856v1 [cs.NI] 19 May 2016

ABSTRACT

Internet service discovery is an emerging topic to study the deployment of protocols. Towards this end, our community periodically scans the entire advertised IPv4 address space. In this paper, we question this principle. Being good Internet citizens means that we should limit scan traffic to what is necessary. We conducted a study of scan data, which shows that several prefixes do not accommodate any host of interest and the network topology is fairly stable. We argue that this allows us to collect representative data by scanning less. In our paper, we explore the idea to scan all prefixes once and then identify prefixes of interest for future scanning.

Based on our analysis of the *censys.io* data set (4.1 TB data encompassing 28 full IPv4 scans within 6 months) we found that we can reduce scan traffic between 25-90% and miss only 1-10% of the hosts, depending on desired trade-offs and protocols.

1. INTRODUCTION

Fast Internet-wide scanning is growing in popularity among researchers. At the time of writing, researchers regularly scan the Internet for vulnerable SSL certificates [5, 12], SSH public keys [10], and for the banners of plain text protocols such as SMTP, HTTP, FTP and Telnet [4]. The majority of researchers scan at least 2.8 billion addresses advertised in the IPv4 address space [5, 7, 10, 14, 17]. Hitrates, the fraction of probed addresses from which a response is received, are very often under two percent [6]. This means that most scan traffic is overhead. Most of these scans are done periodically for trend analyses, which exacerbates the amount of unnecessary scan traffic. For example, the ongoing Internet-wide research project *censys.io* [4, 6] probes the IANA allocated address space for 19 protocols on a continuous basis. This results in 72.2 billion generated IP-packets each week, or more than 3.754 trillion IP packets per year. Whereas scanning the IPv4 address space is feasible this is not any more the case for IPv6. When IPv6 becomes more popular, we need scanning strategies that limit scans to portions of the address space that are in use.

In this paper, we present the *Topology Aware Scanning Strategy (TASS)*, a new IP prefix based and topology aware scanning strategy for periodic scanning. *TASS* enables researchers to collect responses from 90-99% of the available hosts for six months by scanning only 10-75% of the announced IPv4 address space in each scan cycle (protocol dependent). *TASS* is seeded with the results of a full advertised IPv4 address scan for a given protocol and time period. The prefixes for all responses will be selected for periodic scans of the given protocol.

Periodic scanning of only selected prefixes reduces scan traffic significantly while hitting most of the hosts of interest. For instance, our analysis reveals that responsive prefixes obtained from a full FTP scan cover 98% of all FTP hosts 6 months later, at the cost of scanning only 57.4% of the advertised addresses. The scanning overhead can be optimized further by omitting prefixes with a low *density*. Here, *density* denotes the fraction of hosts per address space size. For example, if we limit prefix selection to a 95% coverage of the responsive addresses then we can still find 92.3% of the FTP hosts after six months while scanning only 20.6% of the announced address space.

For our evaluation of *TASS* we use 4.1 TB of data derived from 28 full IPv4 scans obtained from *censys.io* [4]. For common protocols we show that, following an initial scan of the full IPv4 address space, the hitrate for responsive prefixes decreases by about 0.3 percent per month compared to what a full scan would find. Consequently, periodical *TASS* scans are 1.25 to 10 times more efficient for a period of at least 6 months if researchers accept a single-digit percentage reduction in host coverage.

1