Discussed papers weeks 4+5+6

TDTS21, vt 2022

Niklas Carlsson, *Linköping University*

LINKÖPING UNIVERSITY

How can we measure the Internet?

- Active measurements
  - Probes: Traceroute, ping, packet trains
  - Application simulation
- Passive measurement
  - Logs (WWW)
  - Monitors, sniffers

- What is measured?
  - Everything …

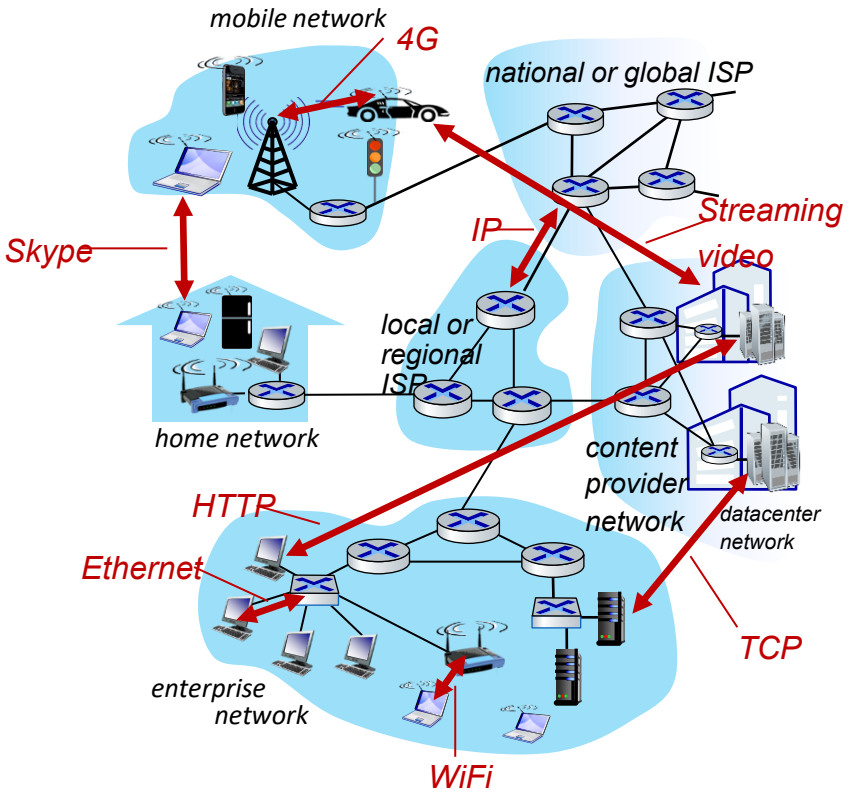
"you can't manage what you can't measure"

– quote claimed by many/several

- Where to measure? …

The Internet: a “nuts and bolts” view

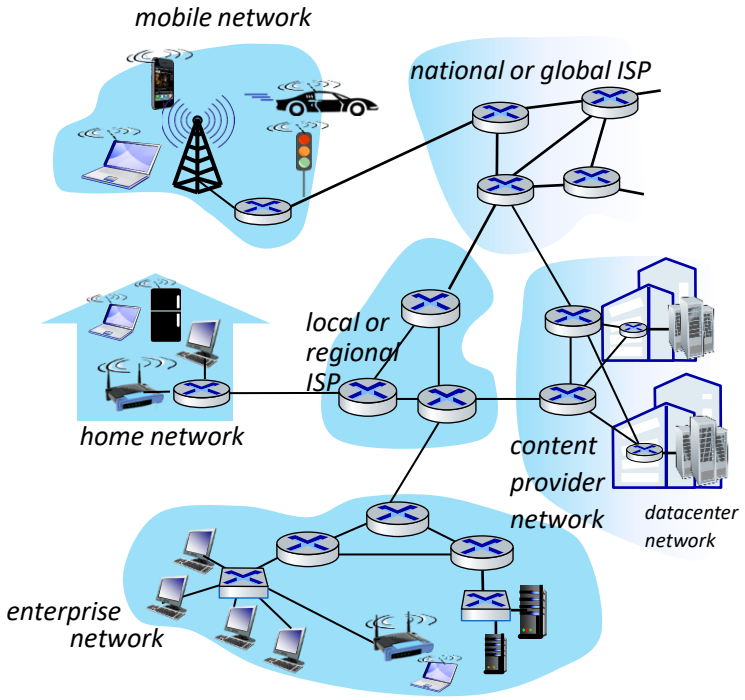
- *Internet:* “network of networks”
  - Interconnected ISPs
- *protocols are everywhere*
  - *control sending, receiving of messages*
  - *e.g., HTTP (Web), streaming video, Skype, TCP, IP, WiFi, 4G, Ethernet*
- *Internet standards*
  - *RFC: Request for Comments*
  - *IETF: Internet Engineering Task Force*



Slide from intro in in TDDE35/TDTS11/TDTS04/TDTS06 …

Internet structure: a “network of networks”

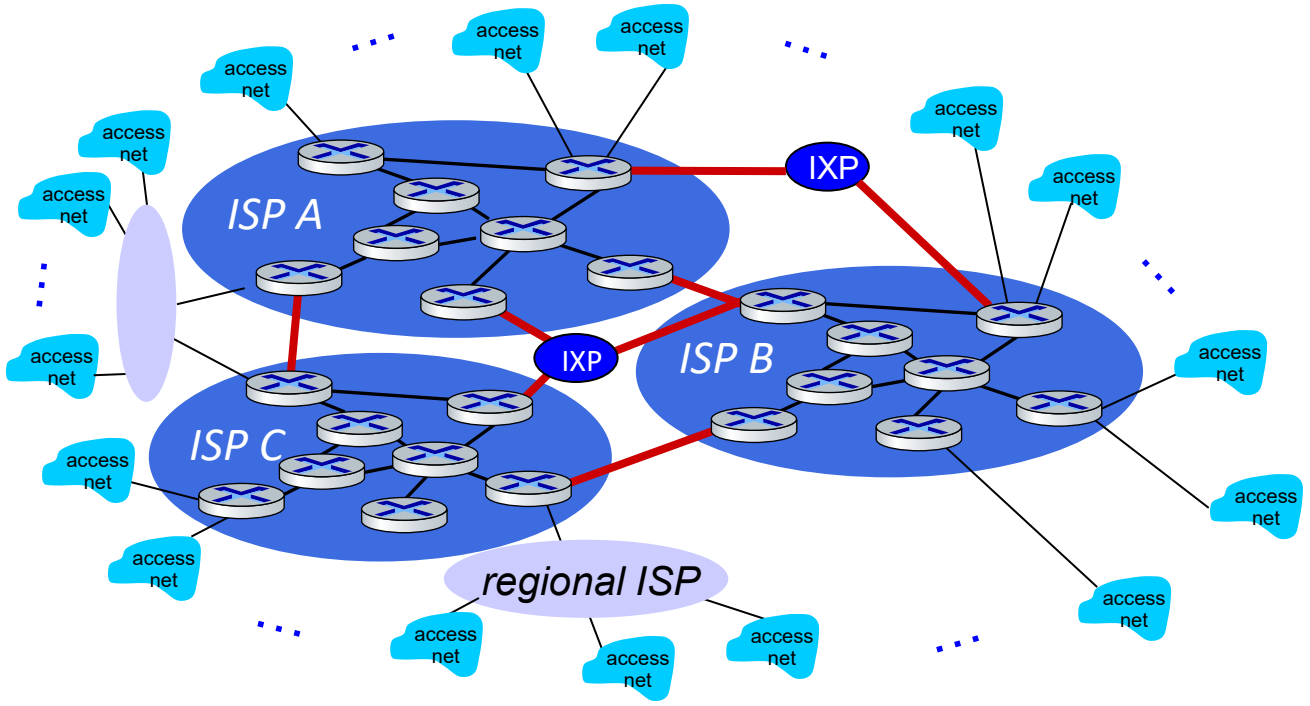
- hosts connect to Internet via access Internet Service Providers (ISPs)
- access ISPs in turn must be interconnected
  - so that *any* two hosts *(anywhere!)* can send packets to each other
- resulting network of networks is very complex
  - evolution driven by economics, national policies



Slide from intro in in TDDE35/TDTS11/TDTS04/TDTS06 …

Internet structure: a “network of networks”

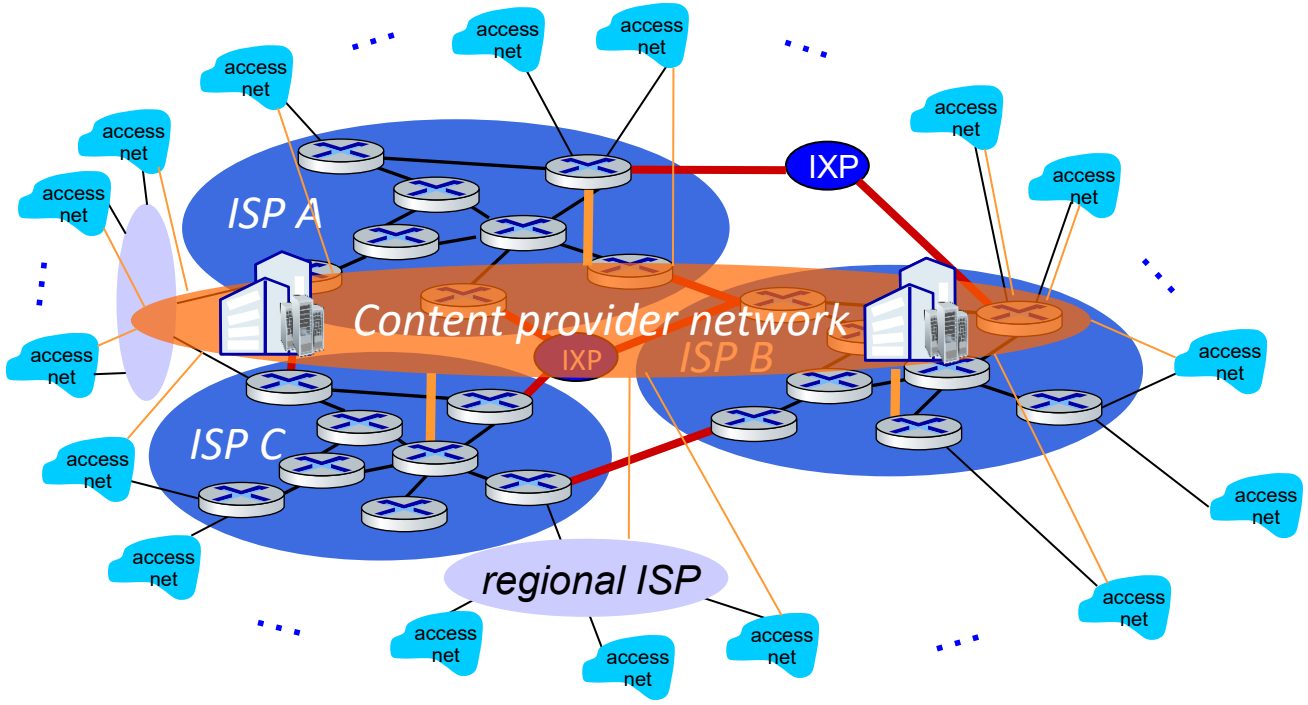
… and regional networks may arise to connect access nets to ISPs



Slide from intro in in TDDE35/TDTS11/TDTS04/TDTS06 …

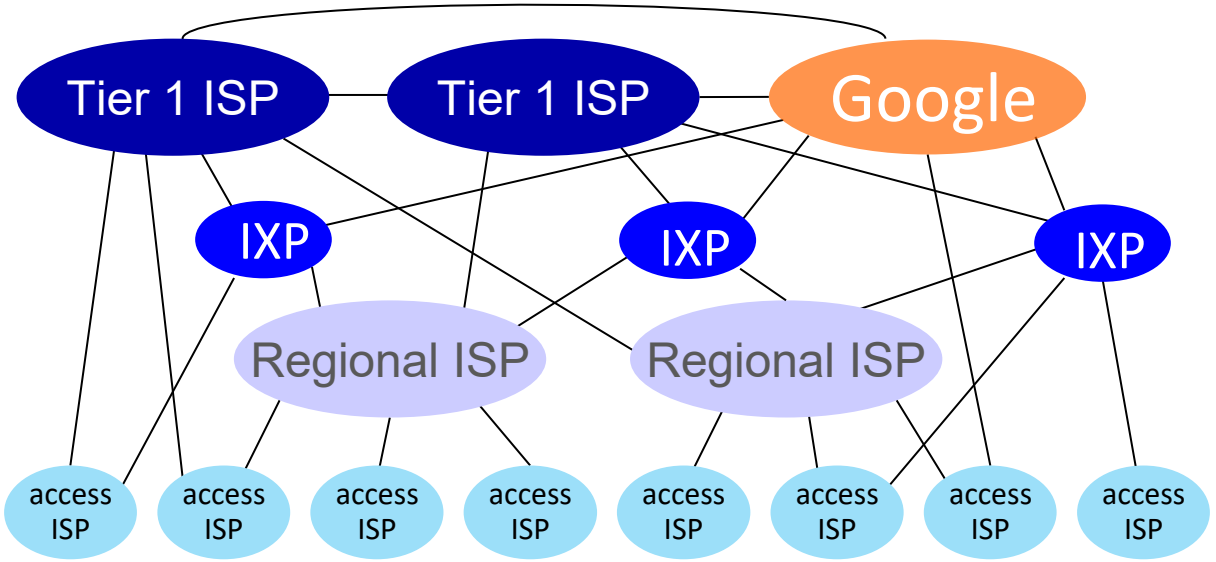
Internet structure: a “network of networks”

… and content provider networks (e.g., Google, Microsoft, Akamai) may run their own network, to bring services, content close to end users



Slide from intro in in TDDE35/TDTS11/TDTS04/TDTS06 …

Internet structure: a "network of networks"



At "center": small # of well-connected large networks

- "tier-1" commercial ISPs (e.g., Level 3, Sprint, AT&T, NTT), national & international coverage
- content provider networks (e.g., Google, Facebook): private network that connects its data centers to Internet, often bypassing tier-1, regional ISPs

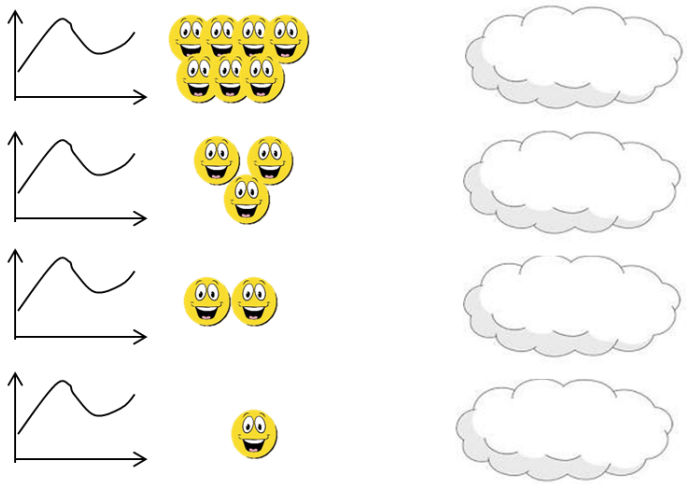
When should we measure the Internet?

- Diurnal and weekly traffic cycles
- Time scales depend on “what”, “how”, and “why”?
- Passive measurement are typically continuous
  - Can generate **huge** datasets
  - Log access problems
  - Privacy concerns
- Active measurements are typically discrete
  - Important characteristics can be missed
  - Probes can be filtered and/or detected

Publishing Internet Measurement Studies

- All major networking + security conferences & journals accept measurement papers
  - ACM SIGCOMM, IEEE INFOCOM, ACM SIGMETRICS
  - ACM CCS, NDSS, IEEE S&P, Usenix Security
  - IEEE/ACM ToN, IEEE TPDS
- Dedicated meetings
  - **ACM Internet Measurement Conf. (IMC)**
  - Passive & Active Measurements Conf. (PAM)
  - Traffic Measurement and Analysis Conf. (TMA)

Slide based on Carlsson & Eager (2014, 2018, 2022) …



Google™

amazon web services™

Microsoft

Shahin

Why this paper?

What you liked most?

Main **contributions**?

Category and **Context**; e.g

- Type of measurements?
- Type of data?
- Type of analysis?
- Type of contribution?

Correctness and **Clarity**

- If not, why not?

Understanding the Latency Benefits of Multi-Cloud Webservice Deployments

Zhe Wu and Harsha V. Madhyastha
University of California, Riverside
{zwu005,harsha}@cs.ucr.edu

ABSTRACT

To minimize user-perceived latencies, webservices are often deployed across multiple geographically distributed data centers. The premise of our work is that webservices deployed across multiple cloud infrastructure services can serve users from more data centers than that possible when using a single cloud service, and hence, offer lower latencies to users.

In this paper, we conduct a comprehensive measurement study to understand the potential latency benefits of deploying webservices across three popular cloud infrastructure services—Amazon EC2, Google Compute Engine (GCE), and Microsoft Azure. We estimate that, as compared to deployments on one of these cloud services, users in up to half the IP address prefixes can have their RTTs reduced by over 20% when a webservice is deployed across the three cloud services. When we dig deeper to understand these latency benefits, we make three significant observations. First, when webservices shift from single-cloud to multi-cloud deployments, a significant fraction of prefixes will see latency benefits simply by being served from a different data center in the same location. This is because routing inefficiencies that exist between a prefix and a nearby data center in one cloud service are absent on the path from the prefix to a nearby data center in a different cloud service. Second, despite the latency improvements that a large fraction of prefixes will perceive, users in several locations (e.g., Argentina and Israel) will continue to incur RTTs greater than 100ms even when webservices span three large-scale cloud services (EC2, GCE, and Azure). Finally, we see that harnessing the latency benefits offered by multi-cloud deployments is likely to be challenging in practice; our measurements show that the data center which offers the lowest latency to a prefix often fluctuates between different cloud services, thus necessitating replication of data.

Categories and Subject Descriptors

C.4 [**Performance of Systems**]: Measurement techniques; C.2.4 [**Communication Networks**]: Distributed Systems—*Distributed applications*; C.2.5 [**Communication Networks**]: Local and Wide-Area Networks—*Internet*

Keywords

Cloud services, Webservices, Latency

1 Introduction

The task of deploying a low-latency webservice has been simplified by Internet routing between any of those data centers and nearby users. For example, since Microsoft Azure has no data centers in South America, a webservice deployed on Azure has to serve users in Brazil from a data center in Texas, thus incurring RTTs of over 200ms. On the other hand, in the case of Amazon EC2, though users in Greece have a relatively nearby data center in Ireland, they will experience RTTs over 90ms due to circuitous routing to that data center [13].

Therefore, we argue that a webservice deployment that *spans multiple cloud services* can offer lower latencies to its clients, than that possible when using a single cloud service. This is because a multi-cloud webservice deployment has a larger set of data centers to choose from when serving its users. As a result, a webservice can take advantage of the fact that 1) one cloud service may have a data center in a particular region while another may not, or 2) even if multiple cloud services have data centers in a region, the data center in one of those services may have lower latencies to users in that region due to less circuitous routing. Going back to our example above, users in Greece and Brazil will have their latencies reduced to less than 60ms and 25ms, respectively, when a webservice is deployed across both EC2 and Azure.

Webservice deployments spanning multiple cloud services will however come at the expense of greater implementation and management complexity, and depending on the webservice's workload and data replication policy, also potentially incur higher costs than single-cloud deployments; though, note that deploying a webservice is challenging even within a single cloud service if users are served from multiple data centers. While estimating the cost associated with a webservice is best left to the webservice's administrator, in this paper, we focus on understanding the latency benefits that multi-cloud webservice deployments can potentially offer. Our goal is to enable webservice administrators to decide whether the cost and complexity of distributing their applications across multiple cloud services is worth the commensurate latency benefits.

Towards this end, we continually measure latencies for 5 weeks from 265 PlanetLab sites to three popular cloud services—Amazon EC2, Google Compute Engine (GCE), and Microsoft Azure. We use these latency measurements to estimate latencies from data centers in the three cloud services to roughly 90K IP prefixes. We find that, when a webservice is deployed across the three cloud services, users in 20–50% of prefixes will see at least a 20% reduction in RTT as compared to single-cloud deployments. When we dig deeper to understand the sources of these latency benefits, we make the following three observations.

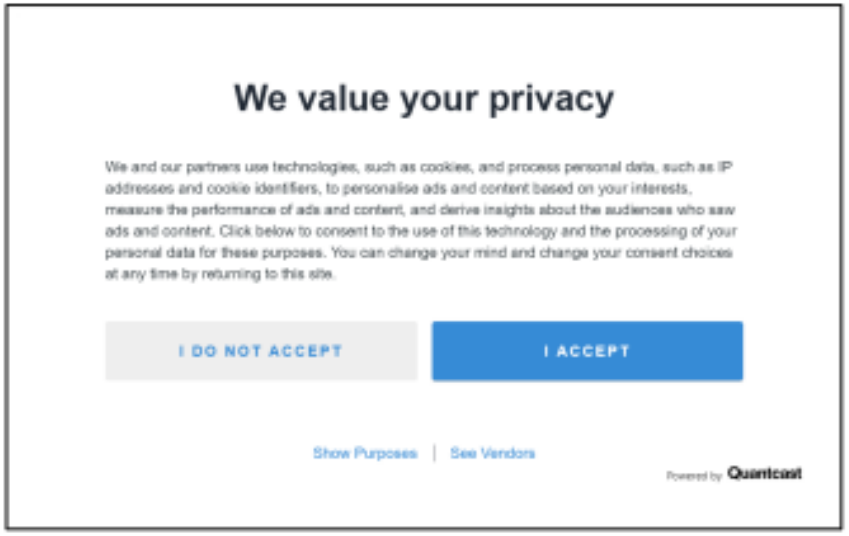


Figure A.1: Default version of Quantcast's consent dialog. The dialog is shown as a modal popup with a dark-gray background covering the rest of the page.

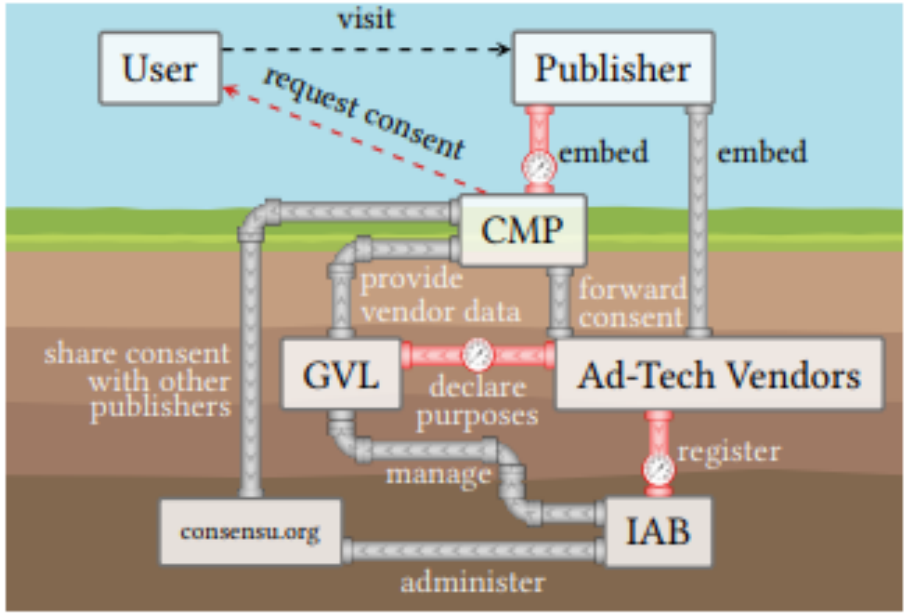


Figure 2: Surfacing the web's new compliance engine: Publishers embed CMPs, which display consent prompts to users, forward consent decisions to ad-tech vendors and also share it globally across websites. In the background, the IAB orchestrates this through its Transparency and Consent Framework (TCF).

Martin

Why this paper?

What you liked most?

Main **contributions**?

Category and **Context**; e.g.

- Type of measurements?
- Type of data?
- Type of analysis?
- Type of contribution?

Correctness and **Clarity**

- If not, why not?

Measuring the Emergence of Consent Management on the Web

Maximilian Hils
University of Innsbruck
maximilian.hils@uibk.ac.at

Daniel W. Woods
University of Innsbruck
daniel.woods@uibk.ac.at

Rainer Böhme
University of Innsbruck
rainer.boehme@uibk.ac.at

ABSTRACT

Privacy laws like the General Data Protection Regulation (GDPR) and the California Consumer Privacy Act (CCPA) have pushed internet firms processing personal data to obtain user consent. Uncertainty around sanctions for non-compliance led many websites to embed a Consent Management Provider (CMP), which collects users' consent and shares it with third-party vendors and other websites. Our paper maps the formation of this ecosystem using longitudinal measurements. Primary and secondary data sources are used to measure each actor within the ecosystem. Using 161 million browser crawls, we estimate that CMP adoption doubled from June 2018 to June 2019 and then doubled again until June 2020. Sampling 4.2 million unique domains, we observe that CMP adoption is most prevalent among moderately popular websites (Tranco top 50-10k) but a long tail exists. Using APIs from the ad-tech industry, we quantify the purposes and lawful bases used to justify processing personal data. A controlled experiment on a public website provides novel insights into how the time-to-complete of two leading CMPs' consent dialogues varies with the preferences expressed, showing how privacy aware users incur a significant time cost.

CCS CONCEPTS

• **Networks** → **Network measurement**; • **Information systems** → **Online advertising**; • **Security and privacy** → *Privacy protections*; Usability in security and privacy.

KEYWORDS

GDPR, CCPA, consent, privacy, web measurement

ACM Reference Format:
Maximilian Hils, Daniel W. Woods, and Rainer Böhme. 2020. Measuring the Emergence of Consent Management on the Web. In *ACM Internet Measurement Conference (IMC '20), October 27–29, 2020, Virtual Event, USA.* ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3419394.3423647

1 INTRODUCTION

Vendors harvesting personal data prefer operating beyond the user's attention as evidenced by the use of secret tracking technologies [1, 29, 38]. This was tolerated by websites who rely on advertising revenues [51]. Sanctions associated with recent privacy laws threaten this state of affairs. In the EU, the General Data Protection Regulation (GDPR) requires firms processing personal data to establish a legal basis, such as by obtaining user consent. In the US, the California Consumer Privacy Act (CCPA) requires websites to collect the consent of minors and also to allow users to opt-out of the sale of their personal data. To comply with both laws, an infrastructure of consent must be designed so that users can consent to the privacy practices of websites and Ad-tech vendors.

In the past, each website offered a unique privacy policy and dialogue. This diversity overwhelmed users who could not commit hundreds of hours to reading each privacy policy [6, 36] nor navigate novel interface designs without making errors [2]. Privacy advocates argued that users should set preferences in the browser to avoid such problems [9, 27, 34], whereas Ad-tech companies lobbied against standardized privacy. However, the new imperative to obtain consent creates problems for Ad-tech vendors who must manage and document heterogeneous forms of consent collected across multiple websites.

Consent management providers (CMPs) emerged in the last three years to standardize the collection of online consent. These intermediaries define legal terms and conditions, present these to users via an embedded consent dialogue, store the resulting signal, and share it with third-parties. In essence, CMPs have created a consent ecosystem involving users, websites, and third-party vendors. For example, one CMP allows websites to collect consent for a 'Global Vendor List' with a membership fee of 1200 €, which was termed the commodification of consent [60].

The rise of CMPs represents a new stage in how privacy preferences are communicated, with previous stages including cookies settings in browsers [37] or custom cookie banners on websites [53]. This paper offers a longitudinal study of the formation of a consent ecosystem orchestrated by CMPs. We introduce the notion of a consent flow—from users through consent dialogues to a website and then onto third-parties—and make measurements at each interface. This complements post-GDPR related work relying on snapshots of relatively small samples of domains, which is shown in Figure 1.

Our insights include:

- Using 161 million browser crawls, we measure CMP adoption

Table 1: Global density of cloud provider endpoints, and their backbone network infrastructure.

	Datacenters per continent						Backbone
	EU	NA	SA	AS	AF	OC	N/W
Amazon EC2 (AMZN)	6	6	1	6	1	1	Private
Google (GCP)	6	10	1	8	-	1	Private
Microsoft (MSFT)	14	10	1	15	2	4	Private
Digital Ocean (DO)	4	6	-	1	-	-	Semi
Alibaba (BABA)	2	2	-	16	-	1	Semi
Vultr (VLTR)	4	9	-	1	-	1	Public
Linode (LIN)	2	5	-	3	-	1	Public
Amazon Lightsail (LTSL)	4	4	-	4	-	1	Private
Oracle (ORCL)	4	4	1	7	-	2	Private
IBM (IBM)	6	6	-	1	-	-	Semi
Total	**52**	**62**	**4**	**62**	**3**	**12**	

Max

Why this paper?

What you liked most?

Main **contributions**?

Category and **Context**; e.g.,

- Type of measurements?
- Type of data?
- Type of analysis?
- Type of contribution?

Correctness and **Clarity**

- If not, why not?

Cloudy with a Chance of Short RTTs

Analyzing Cloud Connectivity in the Internet

The Khang Dang[♮†] Nitinder Mohan[♮†] Lorenzo Corneo[♯] Aleksandr Zavodovski[♯]
Jörg Ott[♮] Jussi Kangasharju[♭]

[♮]Technical University of Munich [♯]Uppsala University [♭]University of Helsinki
[†]Equal contribution

ABSTRACT

Cloud computing has seen continuous growth over the last decade. The recent rise in popularity of next-generation applications brings forth the question: "Can current cloud infrastructure support the low latency requirements of such apps?" Specifically, the interplay of wireless last-mile and investments of cloud operators in setting up direct peering agreements with ISPs globally to current cloud reachability and latency has remained largely unexplored.

This paper investigates the state of end-user to cloud connectivity over wireless media through extensive measurements over six months. We leverage 115,000 wireless probes on the Speedchecker platform and 195 cloud regions from 9 well-established cloud providers. We evaluate the suitability of current cloud infrastructure to meet the needs of emerging applications and highlight various hindering pressure points. We also compare our results to a previous study over RIPE Atlas. Our key findings are: (i) the most impact on latency comes from the geographical distance to the datacenter; (ii) the choice of a measurement platform can significantly influence the results; (iii) wireless last-mile access contributes significantly to the overall latency, almost surpassing the impact of the geographical distance in many cases. We also observe that cloud providers with their own private network backbone and direct peering agreements with serving ISPs offer noticeable improvements in latency, especially in its consistency over longer distances.

CCS CONCEPTS

• **Networks** → **Public Internet**; **Network measurement**.

KEYWORDS

Cloud connectivity, Last-mile latency, Peering, Edge computing

ACM Reference Format:
The Khang Dang, Nitinder Mohan, Lorenzo Corneo, Aleksandr Zavodovski, Jörg Ott and Jussi Kangasharju. 2021. Cloudy with a Chance of Short RTTs: Analyzing Cloud Connectivity in the Internet. In *ACM Internet Measurement Conference (IMC '21), November 2–4, 2021, Virtual Event, USA.* ACM, New

1 INTRODUCTION

Cloud computing has become the core enabler for an ever-increasing growth of networked services on the Internet over the past decade [21]. Cloud providers have made significant investments to expand their global footprint, not just by deploying datacenters in new locations [9] but also installing private backbones interconnecting vast geographical regions [8, 29, 90], deploying Point-of-Presence (PoPs) at Intenet eXchange Points (IXPs) [2] and colocation facilities [47] closer to their customers [67]. Due to these advancements in the backbone, the cloud infrastructure was able to handle the sudden rise in user traffic as the majority population moved to work-from-home model around the globe in 2020 [31].

Beyond improving cloud computing infrastructure, interest has recently grown in "edge computing", a paradigm deploying compute servers closer to the users and outside the managed cloud infrastructure, e.g., on ISP premises [28] or in city-owned buildings [53]. The trend of edge computing is primarily driven by a widespread *belief* that the current cloud infrastructure is too sparsely deployed to support the latency requirements of next-generation mission-critical applications [23], such as AR/VR [56], autonomous vehicles [49], etc. However, the cloud infrastructure has improved dramatically since the inception of edge computing in 2009 [72]. Along with advances in the backbone, cloud hypergiants have also invested heavily in installing new datacenters in previously under-provisioned locations [76]. Furthermore, many new small-to-medium-sized cloud providers, such as Vultr, Linode, DigitalOcean, etc., have entered the market and focus their services on specific geographical regions.

However, the growth in the cloud ecosystem has remained largely unnoticed by researchers. This can be primarily attributed to a shortage of impartial studies that investigate the state of cloud reachability and the factors that impact it globally. Few previous works in this space are either out-of-date since they do not capture the recent expansion of cloud infrastructure [48], cover only a limited set of cloud providers [8], or use vantage points that do not consider users in home environments using wireless connectivity [22]. In this paper, we plug this gap in research by providing a

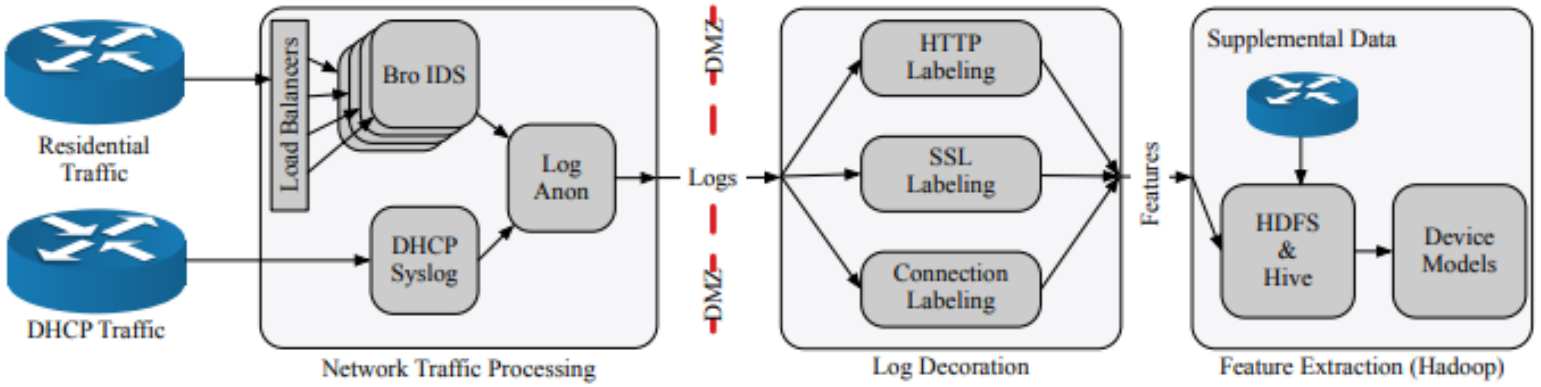


Figure 1: System architecture overview. Network traffic is first processed into logs and its addresses anonymized. The next stage replays the network traffic logs to extract further information and label each connection with (also anonymized) MAC address information. The decorated logs are then stored in Hive where they are labeled with security incidents, security practice features, and behavioral features. Lastly, device models are created for analysis.

August

Why this paper?

What you liked most?

Main **contributions**?

Category and **Context**; e.g.,

- Type of measurements?
- Type of data?
- Type of analysis?
- Type of contribution?

Correctness and **Clarity**

- If not, why not?

Measuring Security Practices and How They Impact Security

Louis F. DeKoven
University of California, San Diego
ldekoven@cs.ucsd.edu

Audrey Randall
University of California, San Diego
aurandal@eng.ucsd.edu

Ariana Mirian
University of California, San Diego
amirian@cs.ucsd.edu

Gautam Akiwate
University of California, San Diego
gakiwate@cs.ucsd.edu

Ansel Blume
University of California, San Diego
ablume@ucsd.edu

Lawrence K. Saul
University of California, San Diego
saul@cs.ucsd.edu

Aaron Schulman
University of California, San Diego
schulman@cs.ucsd.edu

Geoffrey M. Voelker
University of California, San Diego
voelker@cs.ucsd.edu

Stefan Savage
University of California, San Diego
savage@cs.ucsd.edu

ABSTRACT

Security is a discipline that places significant expectations on lay users. Thus, there are a wide array of technologies and behaviors that we exhort end users to adopt and thereby reduce their security risk. However, the adoption of these "best practices" — ranging from the use of antivirus products to actively keeping software updated — is not well understood, nor is their practical impact on security risk well-established. This paper explores both of these issues via a large-scale empirical measurement study covering approximately 15,000 computers over six months. We use passive monitoring to infer and characterize the prevalence of various security practices in situ as well as a range of other potentially security-relevant behaviors. We then explore the extent to which differences in key security behaviors impact real-world outcomes (i.e., that a device shows clear evidence of having been compromised).

CCS CONCEPTS

• **Security and privacy** → *Intrusion detection systems*; • **Networks**;

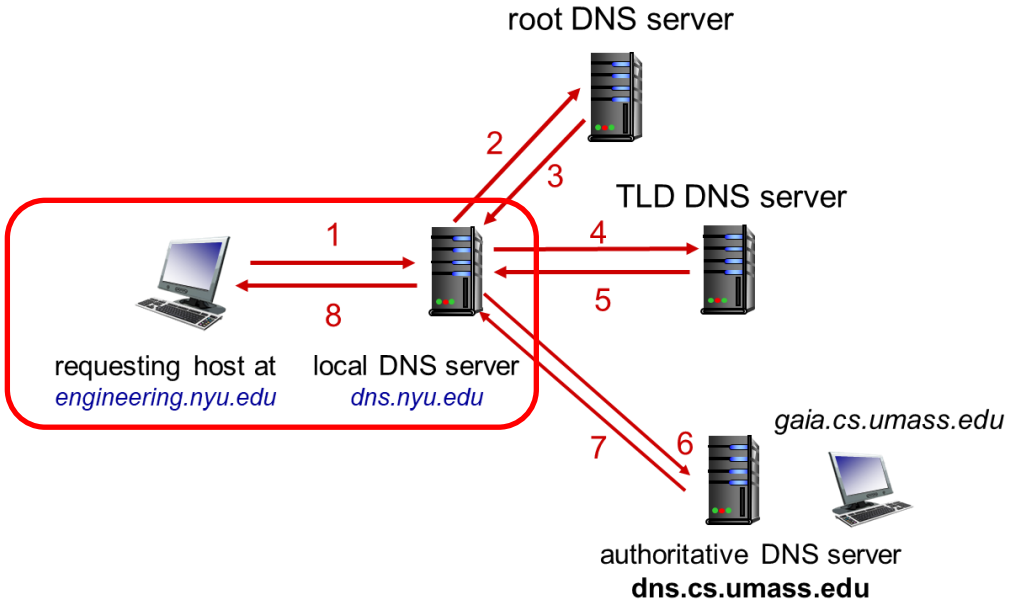
ACM Reference Format:
Louis F. DeKoven, Audrey Randall, Ariana Mirian, Gautam Akiwate, Ansel Blume, Lawrence K. Saul, Aaron Schulman, Geoffrey M. Voelker, and Stefan Savage. 2019. Measuring Security Practices and How They Impact Security. In *Internet Measurement Conference (IMC '19), October 21–23, 2019, Amsterdam, Netherlands.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3355369.3355571

1 INTRODUCTION

Ensuring effective computer security is widely understood to require a combination of both appropriate technological measures and prudent human behaviors; e.g., rapid installation of security updates to patch vulnerabilities or the use of password managers to ensure login credentials are distinct and random. Implicit in this status quo is the recognition that security is not an intrinsic property of today's systems, but is a byproduct of making appropriate choices — choices about what security products to employ, choices about how to manage system software, and choices about how to engage (or not) with third-party services on the Internet. Indeed, the codifying of good security choices, commonly referred to as security policy or "best practice", has been a part of our lives as long as security has been a concern.

However, establishing the value provided by these security practices is underexamined at best. First, we have limited empirical data about which security advice is adopted in practice. Users have a plethora of advice to choose from, highlighted by Reeder et al.'s recent study of expert security advice, whose title — "152 Simple Steps to Stay Safe Online" — underscores both the irony and the variability in such security lore [35]. Clearly few users are likely to follow all such dicta, but if user behavior is indeed key to security, it is important to know which practices are widely followed and which have only limited uptake.

A second, more subtle issue concerns the efficacy of security practices when followed: Do they work? Here the evidence is scant. Even practices widely agreed upon by Reeder's experts, such as keeping software patched, are not justified beyond a rhetorical argument. In fact, virtually all of the most established security best practices — including "use antivirus software", "use HTTPS/TLS", "update your software regularly", "use a password manager", and so on — have attained this status without empirical evidence quantifying their impact on security outcomes. Summarizing this state of affairs, Herley writes, "[Security] advice is complex and growing, but the benefit is largely speculative or moot", which he argues leads rational users to reject security advice [17].



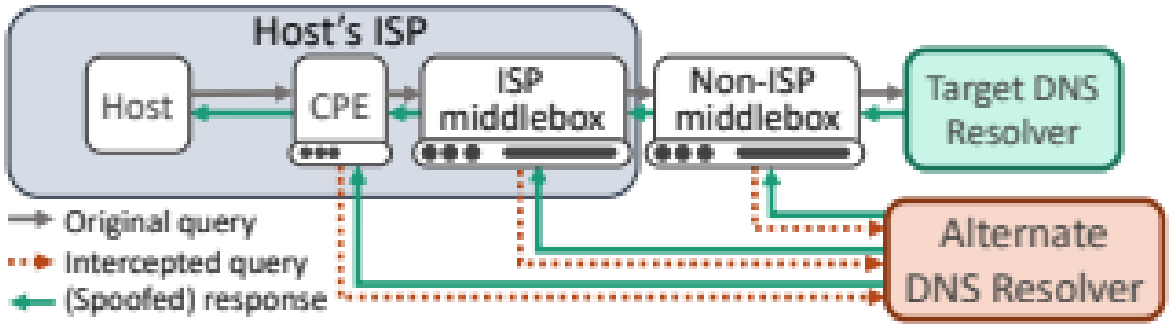


Figure 1: Locations where interception can occur.

Marco

Why this paper?

What you liked most?

Main **contributions**?

Category and **Context**; e.g.,

- Type of measurements?
- Type of data?
- Type of analysis?
- Type of contribution?

Correctness and **Clarity**

- If not, why not?

Home is Where the Hijacking is: Understanding DNS Interception by Residential Routers

Audrey Randall
UC San Diego
aurandal@eng.ucsd.edu

Enze Liu
UC San Diego
e7liu@eng.ucsd.edu

Ramakrishna Padmanabhan
CAIDA/UC San Diego
ramapad@caida.org

Gautam Akiwate
UC San Diego
gakiwate@cs.ucsd.edu

Geoffrey M. Voelker
UC San Diego
voelker@cs.ucsd.edu

Stefan Savage
UC San Diego
savage@cs.ucsd.edu

Aaron Schulman
UC San Diego
schulman@cs.ucsd.edu

ABSTRACT

DNS interception — when a user's DNS queries to a target resolver are intercepted en route and forwarded to a different resolver — is a phenomenon of concern to both researchers and Internet users because of its implications for security and privacy. While the prevalence of DNS interception has received some attention, less is known about *where* in the network interception takes place. We introduce methods to identify where DNS interception occurs and who the interceptors may be. We identify when interception is performed before the query exits the ISP, and even when it is performed by the Customer Premises Equipment (CPE) in the user's own home. We believe that these techniques are vital in the light of the ongoing debate concerning the value of privacy-enhancing DNS transport.

CCS CONCEPTS

• **Networks** → **Home networks**; **Network measurement**.

ACM Reference Format:
Audrey Randall, Enze Liu, Ramakrishna Padmanabhan, Gautam Akiwate, Geoffrey M. Voelker, Stefan Savage, and Aaron Schulman. 2021. Home is Where the Hijacking is: Understanding DNS Interception by Residential Routers. In *ACM Internet Measurement Conference (IMC '21), November 2–4, 2021, Virtual Event, USA.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3487552.3487817

1 INTRODUCTION

In principle, devices are free to direct their DNS queries to the recursive resolver of their choosing. Indeed, it is this freedom that has enabled the growth of public resolvers such as those offered by Google, Cloudflare, and others. However, a key underlying assumption is that DNS queries are faithfully forwarded as they are addressed. Unfortunately, this is not always so.

DNS queries sent by a user's device can be *intercepted* en route

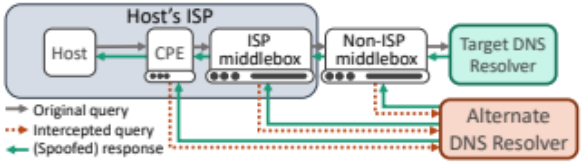


Figure 1: Locations where interception can occur.

spoofs responses so they appear to have been sent by the intended resolver. Transparent interception is difficult to detect because the alternate resolver does *not* have to modify the response. Even if the reason for the interception is benign — such as to prevent malware from evading DNS filtering — the interception of requests and misrepresentation of responses raise serious ethical concerns [14, 45] and can also interfere with the correct operation of protocols such as DNSSEC [14, 31].

While prior work has identified the broad prevalence of transparent interception [24, 31, 49], there are no established techniques for establishing *where* the interception is implemented. Indeed, there are a range of different points in the network where such interception might take place.

DNS redirection, another form of DNS manipulation, has also been found to occur in several parts of the network. DNS redirection occurs when a DNS resolver returns an altered response for specific queries and may occur with or without DNS interception. DNS redirection has been discovered *in Customer Premises Equipment (CPE)* to block resolution of specific domain names [17], *in ISPs* to replace NXDOMAIN responses with advertisements [30, 48] or enhance security and performance [44], and *outside of ISPs* to implement country-level censorship [4, 5, 16, 27]. Transparent interception has been far less extensively studied, although we are

Slide based on Bertmar et al. (WPES 2021) …

Slide based on Bertmar et al. (WPES 2021) …

Jennifer

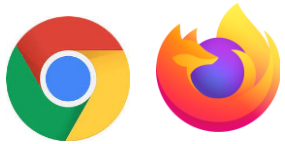
James

Persona

No blocker

Blocker

Location



Browser

Shiwei

Why this paper?

What you liked most?

Main **contributions**?

Category and **Context**; e.g.

- Type of measurements?
- Type of data?
- Type of analysis?
- Type of contribution?

Correctness and **Clarity**

- If not, why not?

On the Potential for Discrimination via Composition

Giridhari Venkatadri
Northeastern University

Alan Mislove
Northeastern University

ABSTRACT

The success of platforms such as Facebook and Google has been due in no small part to features that allow advertisers to target ads in a fine-grained manner. However, these features open up the potential for discriminatory advertising when advertisers include or exclude users of protected classes—either directly or indirectly—in a discriminatory fashion. Despite the fact that advertisers are able to *compose* various targeting features together, the existing mitigations to discriminatory targeting have focused only on individual features; there are concerns that such composition could result in targeting that is more discriminatory than the features individually.

In this paper, we first demonstrate how compositions of individual targeting features can yield discriminatory ad targeting even for Facebook's restricted targeting features for ads in special categories (meant to protect against discriminatory advertising). We then conduct the first study of the potential for discrimination that spans across three major advertising platforms (Facebook, Google, and LinkedIn), showing how the potential for discriminatory advertising is pervasive across these platforms. Our work further points to the need for more careful mitigations to address the issue of discriminatory ad targeting.

CCS CONCEPTS

• **Security and privacy** → **Social aspects of security and privacy**; • **Information systems** → **Online advertising**; **Social networks**.

KEYWORDS

Targeted advertising, Advertising platforms, Discriminatory ad targeting

ACM Reference Format:
Giridhari Venkatadri and Alan Mislove. 2020. On the Potential for Discrimination via Composition. In *ACM Internet Measurement Conference (IMC '20), October 27–29, 2020, Virtual Event, USA.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3419394.3423641

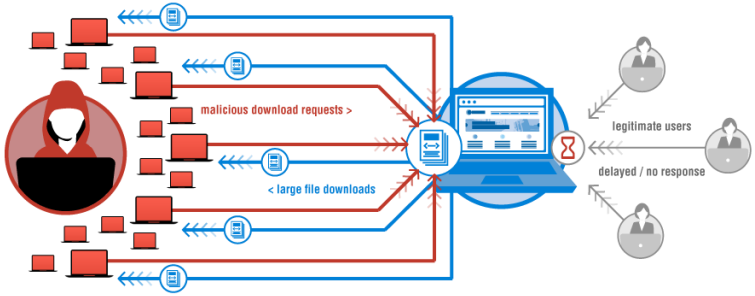
1 INTRODUCTION

Online advertising platforms such as Facebook, Google, and LinkedIn leverage their rich user databases to allow advertisers to target ads to particular users on their platforms. While the ability to selectively target relevant users is advantageous to advertisers—targeting raises concerns that advertisers could knowingly (or unknowingly) target users in order to selectively exclude users of certain sensitive populations (such as users of particular genders, ages, races, or other historically disadvantaged groups). Such discriminatory targeting, while concerning in and of itself, could also run afoul of law for advertisements related to housing, credit, and employment, where special legal protections exist [1–3].

This concern of discriminatory targeting was first raised in the context of Facebook's platform (the largest and most mature of these platforms), where it was shown that an advertiser could explicitly exclude users with certain "ethnic affinities" (such as African American) when targeting housing ads [16]. Subsequent research demonstrated that the problem was not limited to options that explicitly mentioned a protected class, and many additional options exist that are strongly correlated with protected classes [37]. In response to the uproar—including lawsuits from the National Fair Housing Alliance [14] and the U.S. Department of Housing and Urban Development (HUD) [15]—Facebook made a number of changes to its targeting options. These changes included deploying a restricted interface for housing, credit, and employment ads that has more limited targeting options [12].

Unfortunately, there are two key omissions in terms of understanding and protecting against discrimination in ad targeting. First, the previous discussion and proposed mitigations have focused on *individual* targeting options that happen to be correlated with a particular sensitive population; indeed, Facebook's above-mentioned restrictions to mitigate discriminatory advertising primarily focused on disabling access to many individual targeting options. However, these advertising platforms typically allow advertisers to *compose* multiple such options together in various ways. Thus, two (or more) targeting options that are individually only mildly correlated with a protected class (and therefore only mildly discriminatory), may end up being more significantly correlated when used in *conjunction* with each other. For example, the population interested in electrical engineering, or the population interested in sports cars, might each be somewhat skewed towards men; however, the population interested in electrical engineering *and* sports cars might be significantly more skewed. As a result, limiting individual targeting options alone may be insufficient to prevent discriminatory advertising. Indeed, if composing targeting options in general tends to yield more skew than individual targeting options, even an honest advertiser who uses multiple targeting options may end

IoT and DDoS attacks



THE WALL STREET JOURNAL.

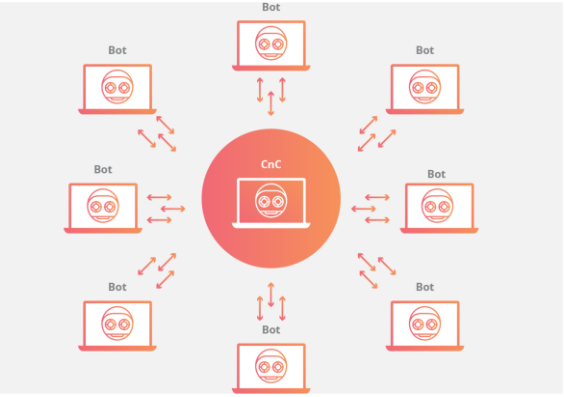
October 21, 2016

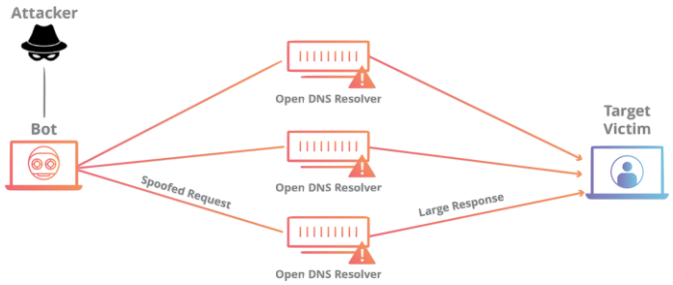
Cyberattack Knocks Out Access to Websites

Popular sites such as Twitter, Netflix and PayPal were unreachable for part of the day

twitter amazon web services PayPal NETFLIX SOUNDCLOUD

Spotify GitHub reddit New York Times





Control lots of machines

Amplification

Yinan

Why this paper?

What you liked most?

Main **contributions**?

Category and **Context**; e.g

- Type of measurements?
- Type of data?
- Type of analysis?
- Type of contribution?

Correctness and **Clarity**

- If not, why not?

A Haystack Full of Needles: Scalable Detection of IoT Devices in the Wild

Said Jawad Saidi
Max Planck Institute for Informatics

Anna Maria Mandalari
Imperial College London

Roman Kolcun
Imperial College London

Hamed Haddadi
Imperial College London

Daniel J. Dubois
Northeastern University

David Choffnes
Northeastern University

Georgios Smaragdakis
TU Berlin
Max Planck Institute for Informatics

Anja Feldmann
Max Planck Institute for Informatics/Saarland University

ABSTRACT

Consumer Internet of Things (IoT) devices are extremely popular, providing users with rich and diverse functionalities, from voice assistants to home appliances. These functionalities often come with significant privacy and security risks, with notable recent large-scale coordinated global attacks disrupting large service providers. Thus, an important first step to address these risks is to know *what* IoT devices are *where* in a network. While some limited solutions exist, a key question is whether device discovery can be done by Internet service providers that only see sampled flow statistics. In particular, it is challenging for an ISP to efficiently and effectively track and trace activity from IoT devices deployed by its millions of subscribers—all with sampled network data.

In this paper, we develop and evaluate a scalable methodology to accurately detect and monitor IoT devices at subscriber lines with limited, highly sampled data in-the-wild. Our findings indicate that millions of IoT devices are detectable and identifiable within hours, both at a major ISP as well as an IXP, using *passive*, sparsely *sampled* network flow headers. Our methodology is able to detect devices from more than 77% of the studied IoT manufacturers, including popular devices such as smart speakers. While our methodology is effective for providing network analytics, it also highlights significant privacy consequences.

CCS CONCEPTS

• **Security and privacy** → *Network security*; • **Networks** → *Network monitoring*; **Public Internet**; **Network measurement**;

KEYWORDS

Internet of Things, IoT detection, IoT secuirty and privacy, Internet Measurement

ACM Reference Format:
Said Jawad Saidi, Anna Maria Mandalari, Roman Kolcun, Hamed Haddadi, Daniel J. Dubois, David Choffnes, Georgios Smaragdakis, and Anja Feldmann. 2020. A Haystack Full of Needles: Scalable Detection of IoT Devices in the Wild. In *ACM Internet Measurement Conference (IMC '20), October 27–29, 2020, Virtual Event, USA.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3419394.3423650

1 INTRODUCTION

The number of IoT devices deployed within homes is increasing rapidly. It is estimated that at the end of 2019, more than 9.5 billion IoT devices were active, and the IoT population will increase to 20 billion by 2025 [1]. Such devices include virtual assistants, smart home control, cameras, and smart TVs. While users deploy some IoT devices explicitly, they are often unaware of the security threats and privacy consequences of using such devices [2]. Major Internet Service Providers (ISPs) are developing strategies for dealing with the large-scale coordinated attacks from these devices.

Existing solutions focus on instrumenting testbeds or home environments to collect and analyze full packet captures [3–5], local search for IoT anomalies [6, 7], active measurements [8, 9], or data from antivirus companies running scan campaigns from users homes [7]. In isolation, these data sources do not provide enough insights for preventing network-wide attacks from IoT devices [10]. Detecting IoT devices from an ISP can help to identify suspicious traffic and what devices are common among the subscriber lines generating that traffic.

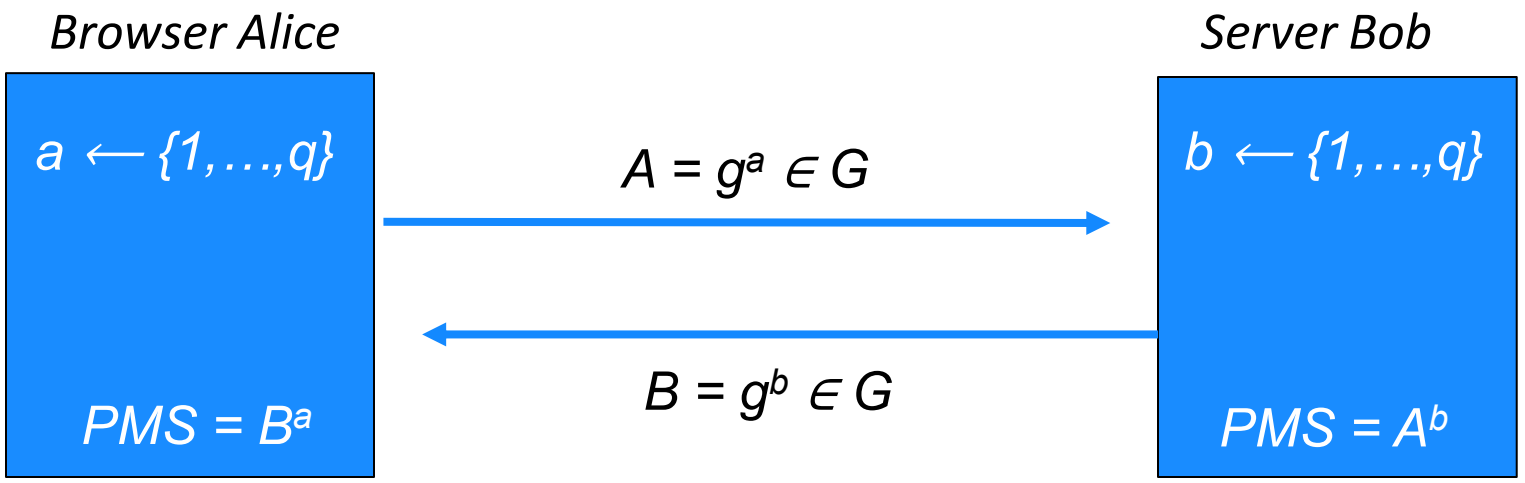
In this paper, we present a methodology for detecting home IoT devices in-the-wild at an ISP, and an Internet Exchange Point (IXP), by relying on passive, sampled network traces and active probing experiments. We build on the insight that IoT devices typically

Slide from TDDE53…

TLS overview: (1) DH key exchange

Anonymous key exchange secure against eavesdropping:

The Diffie-Hellman protocol in a group $G = \{1, g, g^2, g^3, \ldots, g^{q-1}\}$



$$PreMasterSecret = g^{ab} = (g^b)^a = B^a = (g^a)^b = A^b$$

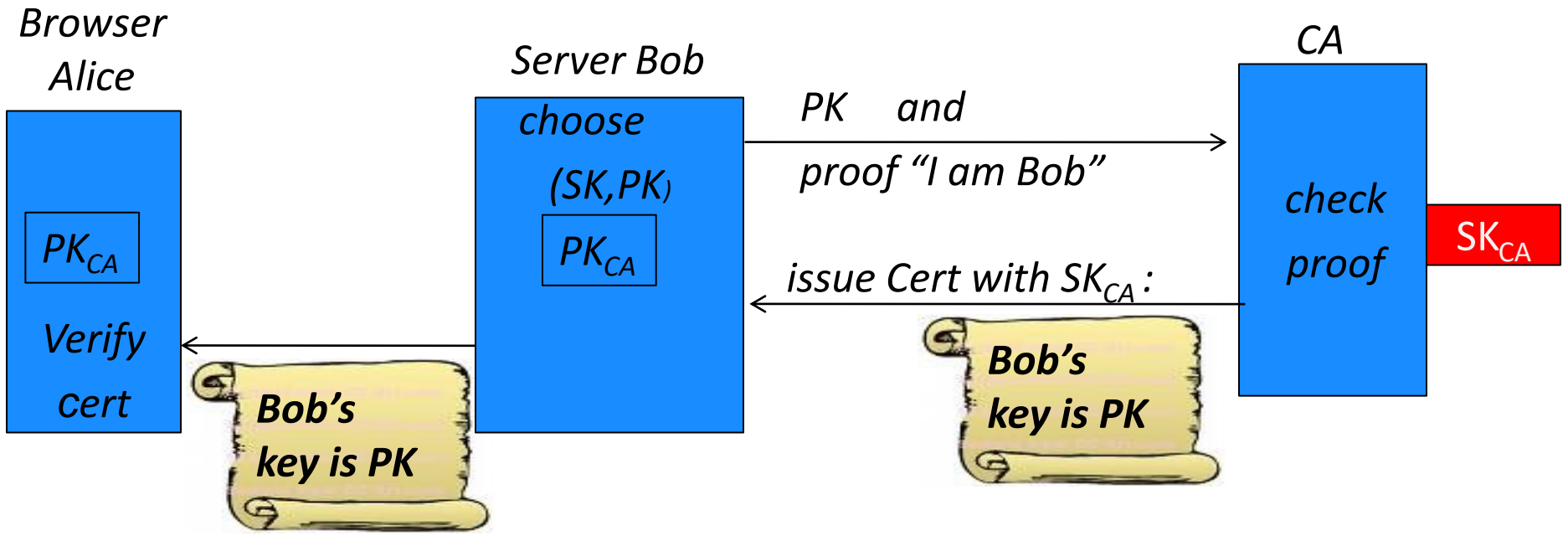
Used to establish a shared secret

Group G is publicly known

Slide from TDDE53…

TLS overview: (2) Certificates

How does Alice (browser) obtain PK_{Bob} ?



Bob uses Cert for an extended period (e.g. *one year)*

Ellen

Why this paper?

What you liked most?

Main **contributions**?

Category and **Context**; e.g.,

- Type of measurements?
- Type of data?
- Type of analysis?
- Type of contribution?

Correctness and **Clarity**

- If not, why not?

IoTLS: Understanding TLS Usage in Consumer IoT Devices

Muhammad Talha Paracha
Northeastern University

Daniel J. Dubois
Northeastern University

Narseo Vallina-Rodriguez
IMDEA Networks / ICSI / AppCensus Inc.

David Choffnes
Northeastern University

ABSTRACT

Consumer IoT devices are becoming increasingly popular, with most leveraging TLS to provide connection security. In this work, we study a large number of TLS-enabled consumer IoT devices to shed light on how effectively they use TLS, in terms of establishing secure connections and correctly validating certificates, and how observed behavior changes over time. To this end, we gather more than two years of TLS network traffic from IoT devices, conduct active probing to test for vulnerabilities, and develop a novel black-box technique for exploring the trusted root stores in IoT devices by exploiting a side-channel through TLS *Alert Messages*. We find a wide range of behaviors across devices, with some adopting best security practices but most being vulnerable in one or more of the following ways: use of old/insecure protocol versions and/or ciphersuites, lack of certificate validation, and poor maintenance of root stores. Specifically, we find that *at least* 8 IoT devices still include distrusted certificates in their root stores, 11/32 devices are vulnerable to TLS interception attacks, and that many devices fail to adopt modern protocol features over time. Our findings motivate the need for IoT manufacturers to audit, upgrade, and maintain their devices' TLS implementations in a consistent and uniform way that safeguards all of their network traffic.

CCS CONCEPTS

• **Security and privacy** → **Network security**; **Embedded systems security**; • **Networks** → *Network measurement*; **Network security**;

KEYWORDS

Internet of Things, IoT, Transport Layer Security, TLS, network security, embedded systems security, measurement techniques

ACM Reference Format:
Muhammad Talha Paracha, Daniel J. Dubois, Narseo Vallina-Rodriguez, and David Choffnes. 2021. IoTLS: Understanding TLS Usage in Consumer IoT Devices. In *ACM Internet Measurement Conference (IMC '21), November 2–4, 2021, Virtual Event, USA.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3487552.3487830

1 INTRODUCTION

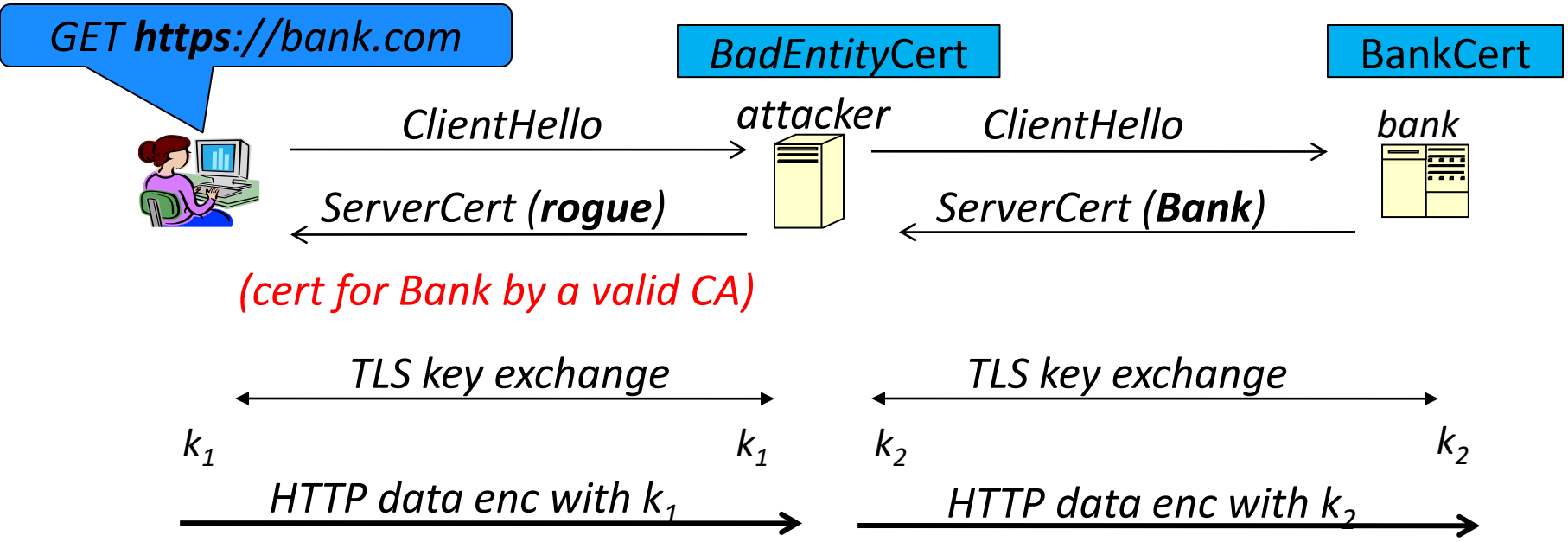
Consumer Internet-of-Things (IoT) devices such as voice assistants, smart TVs and video doorbells are popular, with their prevalence projected to be 75 billion by 2025 [14]. Most IoT devices rely on TLS, the de facto secure transport protocol, to provide confidentiality, integrity and authenticity of their network communications [26]. Numerous prior works have shown that TLS security properties can be compromised due to development errors (e.g., [31]), insecure configurations (e.g., [39]), and outdated clients (e.g., [20]). While TLS usage has been studied extensively in mobile applications and web browsers (e.g., [47], [49], [37]), there is little insight into its effectiveness in the IoT ecosystem (e.g., [26]).

More specifically, there exists a research gap in understanding whether TLS implementations in IoT devices: (*i*) establish connections using secure TLS versions and ciphersuites, (*ii*) correctly perform certificate validation while using a generally trusted set of root certificates, and (*iii*) adopt new features as the protocol evolves over time (e.g., modern ciphersuites). There are several challenges that prevent the use of existing methodologies to study these aspects of IoT devices. *First*, understanding TLS support on a significant number of IoT devices requires blackbox testing techniques; this is because source code is generally unavailable and firmware analysis is not scalable. *Second*, most IoT devices provide limited ways to trigger TLS traffic for measurement—the timing, destination, and contents of their communication are all dependent on device functionality and interactions. *Third*, existing vantage points offer limited opportunities to track device behavior over time (e.g., recent work considers only manufacturer-level device tracking using ISP/IXP data [53]).

In this work, we address these challenges to study a large number of TLS-enabled consumer IoT devices (with over 200 million units sold collectively). *First*, we shed light on the security of TLS implementations and configurations in these devices using existing and novel active measurement techniques that require only TLS traffic interception. *Second*, based on the insight that devices generate significant network traffic when powered on, we automate device reboots using smart plugs to trigger TLS activity for our experiments. And *third*, we analyze ≈2 years of network traffic

Slide from TDDE53 …

Man in the middle attack using rogue cert



Attacker proxies data between user and bank.
Sees all traffic and can modify data at will.

David

Why this paper?

What you liked most?

Main **contributions**?

Category and **Context**; e.g.,

- Type of measurements?
- Type of data?
- Type of analysis?
- Type of contribution?

Correctness and **Clarity**

- If not, why not?

Investigating Large Scale HTTPS Interception in Kazakhstan

Ram Sundara Raman
University of Michigan
ramaks@umich.edu

Leonid Evdokimov
Independent
leon@darkk.net.ru

Eric Wurstrow
University of Colorado Boulder
ewust@colorado.edu

J. Alex Halderman
University of Michigan
jhalderm@umich.edu

Roya Ensafi
University of Michigan
ensafi@umich.edu

ABSTRACT

Increased adoption of HTTPS has created a largely encrypted web, but these security gains are on a collision course with governments that desire visibility into and control over user communications. Last year, the government of Kazakhstan conducted an unprecedented large-scale HTTPS interception attack by forcing users to trust a custom root certificate. We were able to detect the interception and monitor its scale and evolution using measurements from in-country vantage points and remote measurement techniques. We find that the attack targeted connections to 37 unique domains, with a focus on social media and communication services, suggesting a surveillance motive, and that it affected a large fraction of connections passing through the country's largest ISP, Kazakhtelecom. Our continuous real-time measurements indicated that the interception system was shut down after being intermittently active for 21 days. Subsequently, supported by our findings, two major browsers (Mozilla Firefox and Google Chrome) completely blocked the use of Kazakhstan's custom root. However, the incident sets a dangerous precedent, not only for Kazakhstan but for other countries that may seek to circumvent encryption online.

CCS CONCEPTS

• **General and reference** → **Measurement**; • **Security and privacy** → **Security protocols**; **Web protocol security**; • **Social and professional topics** → **Governmental surveillance**; *Technology and censorship.*

KEYWORDS

HTTPS, Interception, Kazakhstan, MitM, Certificates

ACM Reference Format:
Ram Sundara Raman, Leonid Evdokimov, Eric Wurstrow, J. Alex Halderman

1 INTRODUCTION

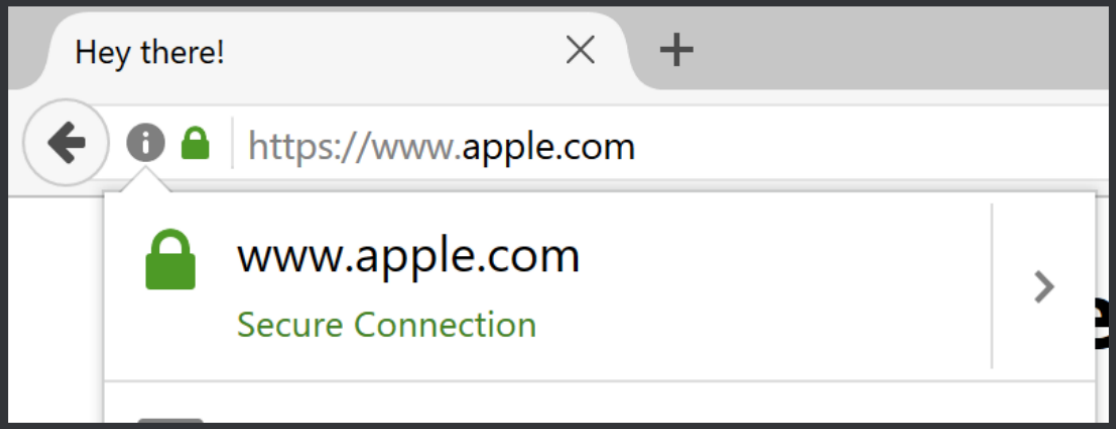
HTTPS protects billions of users: 74–95% of daily web traffic is now encrypted, providing much-needed privacy and security [1, 23]. At the same time, deep packet inspection technologies that inspect HTTPS connections have also advanced [29, 46, 50]. Although enterprise-level interception is common despite being fraught with security issues [17, 40], large-scale interception at the ISP or national level has been limited, even as increased adoption of HTTPS challenges mass surveillance and keyword-based censorship [5, 19].

Last year, in an unprecedented move, the Republic of Kazakhstan became the first country to deploy carrier-grade HTTPS interception on a national level. Starting on July 17, 2019,[1] Kazakhstan launched an HTTPS interception man-in-the-middle (MitM) attack, after instructing citizens to install a government-issued root certificate on all devices and in every browser for "security" purposes [8]. This interception, which the government described as a "pilot", covered large portions of the country's network and was active intermittently until being shut down on August 7, 2019.

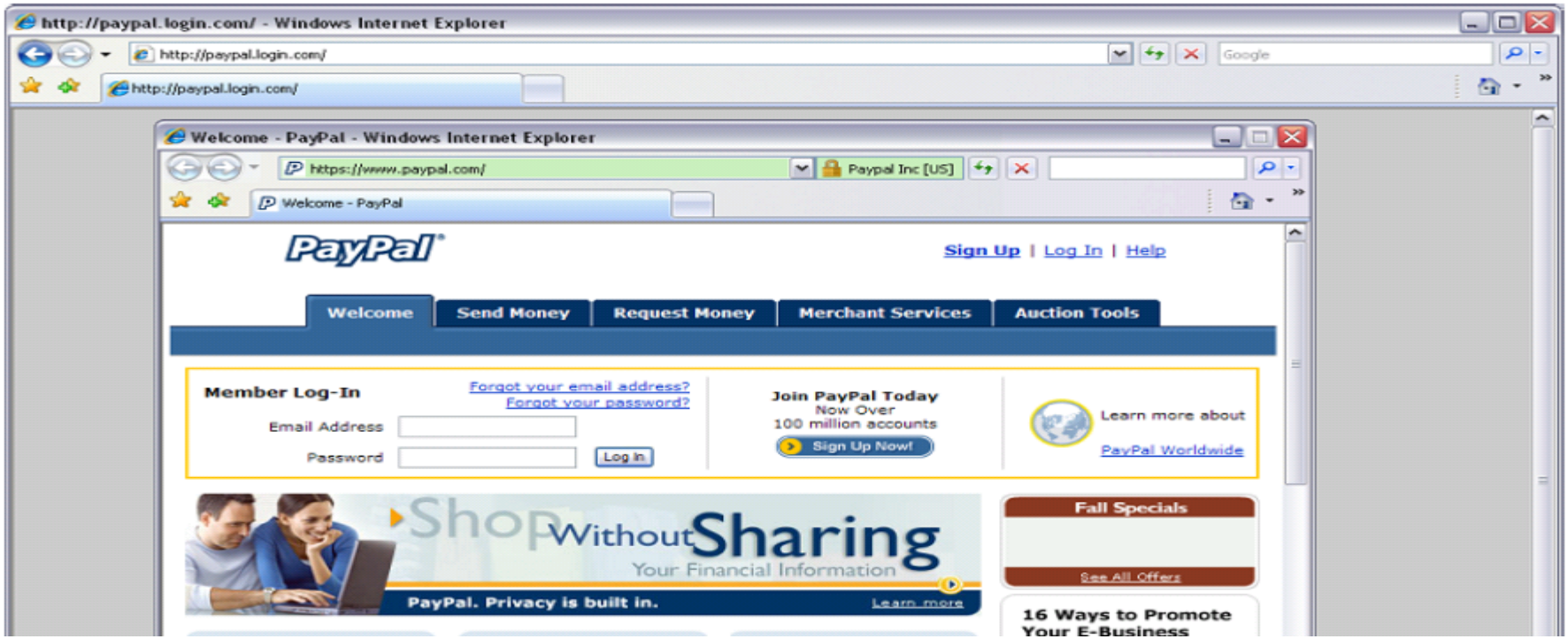
While the attack was going on, we worked to understand the interception technique, measure its scope, and identify its likely targets. We first detected the interception using data from Hyperquack, a recently introduced remote technique for detecting keyword-based network interference [50]. Beginning on July 20, Hyperquack's HTTPS measurements to some (but not all) of 82 available vantage points in Kazakhstan detected rogue untrusted certificates for popular destinations such as `google.com` and `facebook.com`. The certificates were issued by the Kazakh government's custom root CA, `Qaznet Trust Network`. We later confirmed these detections with direct measurements from local virtual private servers (VPSes) and 52 in-country RIPE Atlas nodes.

We determined that the interception system would trigger on TLS connections passing through certain network locations in Kazakhstan when a targeted domain was present in the TLS Server Name

Notice anything odd?



Examples by Dan Boneh



Niklas

Why this paper?

What you liked most?

Main **contributions**?

Category and **Context**; e.g.,

- Type of measurements?
- Type of data?
- Type of analysis?
- Type of contribution?

Correctness and **Clarity**

- If not, why not?

Are You Human? Resilience of Phishing Detection to Evasion Techniques Based on Human Verification

Sourena Maroofi
sourena.maroofi@
univ-grenoble-alpes.fr
Univ. Grenoble Alpes, CNRS,
Grenoble INP, LIG
France

Maciej Korczyński
maciej.korczynski@grenoble-inp.fr
Univ. Grenoble Alpes, CNRS,
Grenoble INP, LIG
France

Andrzej Duda
andrzej.duda@grenoble-inp.fr
Univ. Grenoble Alpes, CNRS,
Grenoble INP, LIG
France

ABSTRACT

Phishing is one of the most common cyberattacks these days. Attackers constantly look for new techniques to make their campaigns more lucrative by extending the lifespan of phishing pages. To achieve this goal, they leverage different anti-analysis (i.e., evasion) techniques to conceal the malicious content from anti-phishing bots and only reveal the payload to potential victims. In this paper, we study the resilience of anti-phishing entities to three advanced anti-analysis techniques based on human verification: Google re-CAPTCHA, alert box, and session-based evasion. We have designed a framework for performing our testing experiments, deployed 105 phishing websites, and provided each of them with one of the three evasion techniques. In the experiments, we report phishing URLs to major server-side anti-phishing entities (e.g., Google Safe Browsing, NetCraft, APWG) and monitor their occurrence in the blacklists. Our results show that Google Safe Browsing was the only engine that detected all the reported URLs protected by alert boxes. However, none of the anti-phishing engines could detect phishing URLs armed with Google re-CAPTCHA, making it so far the most effective protection solution of phishing content available to malicious actors. Our experiments show that all the major server-side anti-phishing bots only detected 8 out of 105 phishing websites protected by human verification systems. As a mitigation plan, we intend to disclose our findings to the impacted anti-phishing entities before phishers exploit human verification techniques on a massive scale.

CCS CONCEPTS

• **Security and privacy → Phishing**;

ACM Reference Format:
Sourena Maroofi, Maciej Korczyński, and Andrzej Duda. 2020. Are You

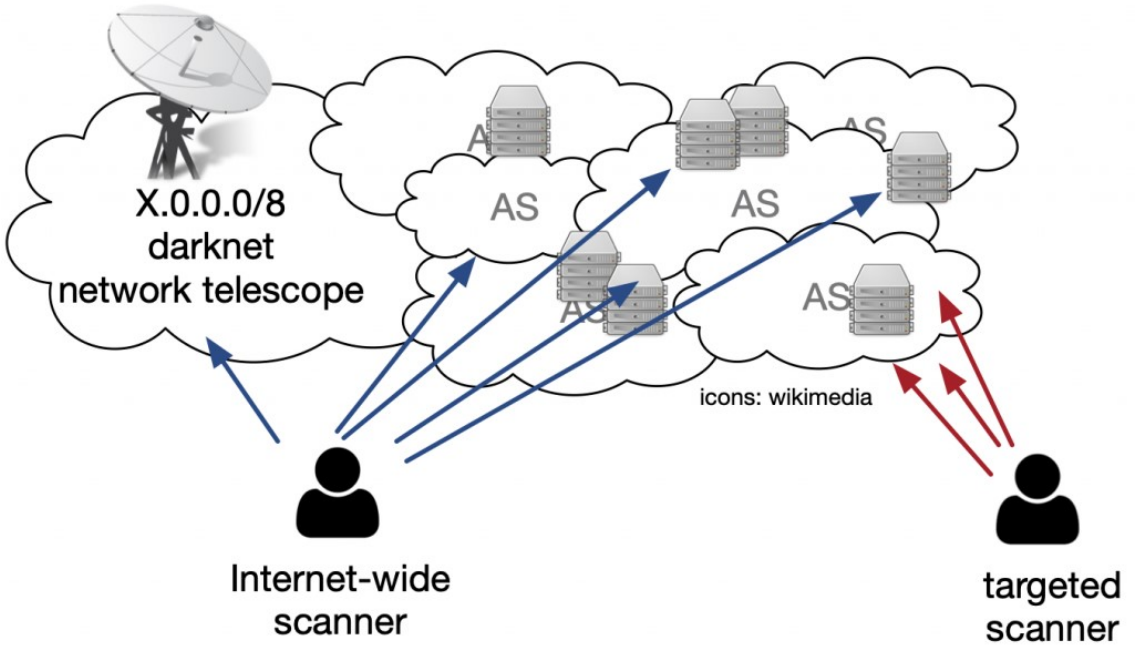
1 INTRODUCTION

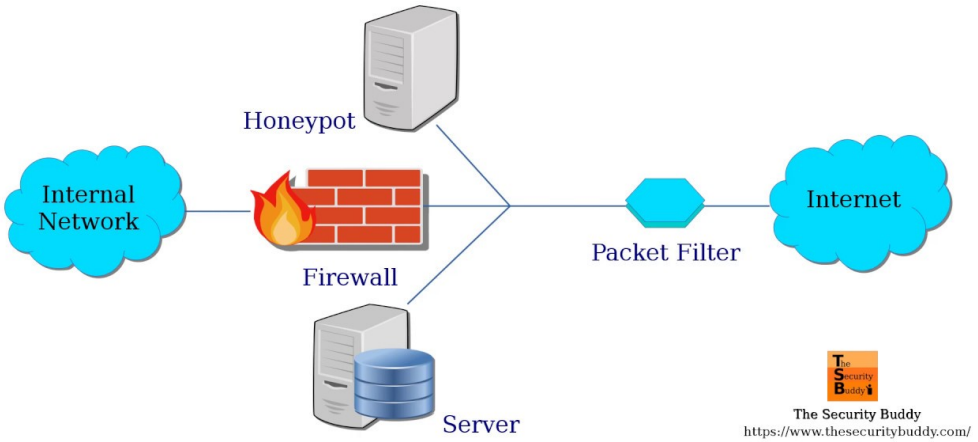
Phishing is a form of social engineering with the goal of collecting credentials of end-users usually achieved by either email spoofing [1] or directly luring victims to enter their sensitive information into a fake website that matches the look and feel of the legitimate one [2]. There have been much research [3–7] and industry efforts to combat phishing, just to mention the Anti-Phishing Working Group (APWG) [8], PhishTank [9], or recently formed the COVID-19 Cyber Threat Coalition [10]. Nevertheless, according to the latest report from IBM X-force, phishing is the number one initial infection vector among attackers [11].

As with any other cybercriminal activity, phishers and security organizations are constantly in battle. While phishers try to develop new or misuse the existing techniques such as URL shorteners [12] to make their attacks more effective, anti-phishing organizations try to adapt their methods to detect phishing attacks swiftly. One of the most effective techniques used by miscreants is the *anti-analysis* also known as *evasion* [13]. The term refers to a wide range of techniques used by attackers to prevent automatic threat analysis [14]. While these techniques are very popular among malware developers [15], phishers also begin to use them [16, 17]. Malicious actors leverage evasion techniques to tell anti-phishing bots and humans apart. If the end-user is human, they reveal the malicious payload while for anti-phishing bots, they deliver a benign page to evade detection.

Previous work analyzed the existing and well known evasion techniques such as web-cloaking [18], URL redirection [19], using URL shorteners [12], and code obfuscation [20]. These techniques can affect the detection time, yet all major anti-phishing systems can cope with them [21]. For example, to detect web-cloaking, we can send two requests to the server, one with a user-agent related to a known anti-phishing bot (e.g., *googlebot*) and the other one with a

https://blog.apnic.net/2019/12/11/whos-scanning-your-network/





Joakim

Why this paper?

What you liked most?

Main **contributions**?

Category and **Context**; e.g

- Type of measurements?
- Type of data?
- Type of analysis?
- Type of contribution?

Correctness and **Clarity**

- If not, why not?

Open for hire: attack trends and misconfiguration pitfalls of IoT devices

Shreyas Srinivasa
Aalborg University, Denmark

Jens Myrup Pedersen
Aalborg University, Denmark

Emmanouil Vasilomanolakis
Aalborg University, Denmark

ABSTRACT

Mirai and its variants have demonstrated the ease and devastating effects of exploiting vulnerable Internet of Things (IoT) devices. In many cases, the exploitation vector is not sophisticated; rather, adversaries exploit misconfigured devices (e.g. unauthenticated protocol settings or weak/default passwords). Our work aims at unveiling the state of IoT devices along with an exploration of the current attack landscape. In this paper, we perform an Internet-level IPv4 scan to unveil 1.8 million misconfigured IoT devices that may be exploited to perform large-scale attacks. These results are filtered to exclude a total of 8, 192 devices that we identify as honeypots during our scan. To study current attack trends, we deploy six state-of-art IoT honeypots for a period of 1 month. We gather a total of 200, 209 attacks and investigate how adversaries leverage misconfigured IoT devices. In particular, we study different attack types, including denial of service, multistage attacks and attacks from infected online hosts. Furthermore, we analyze data from a /8 network telescope covering a total of 81 billion requests towards IoT protocols (e.g. CoAP, UPnP). Combining knowledge from the aforementioned experiments, we identify 11, 118 IP addresses (that are part of the detected misconfigured IoT devices) that attacked our honeypot setup and the network telescope.

ACM Reference Format:
Shreyas Srinivasa, Jens Myrup Pedersen, and Emmanouil Vasilomanolakis. 2021. Open for hire: attack trends and misconfiguration pitfalls of IoT devices. In *ACM Internet Measurement Conference (IMC '21), November 2–4, 2021, Virtual Event, USA.* ACM, New York, NY, USA, 21 pages. https://doi.org/10.1145/3487552.3487833

1 INTRODUCTION

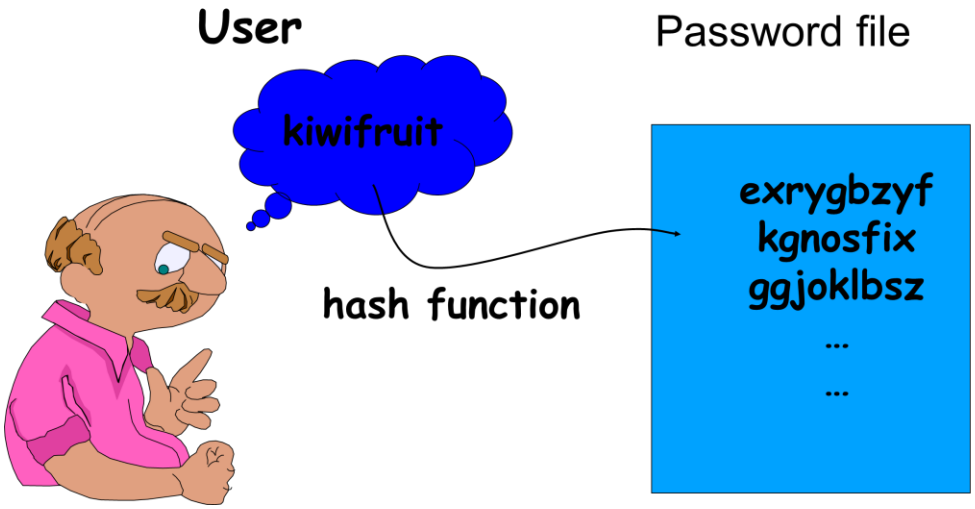
With the adoption of IoT, there is an increase of misconfigured devices on the Internet. Some are incorrectly configured or left with default configuration, thereby making them vulnerable [28]. Misconfigured IoT devices are exploited on a large scale by malware like Mirai that infect vulnerable devices with bots [44]. A device is considered to be *misconfigured* if its incorrect configuration leads facilitated through botnets. For instance, many variants of the Mirai botnet and newer IoT malware like GitPaste-12 [13], Kaiji [9], RHOMBUS [49] continue to look for vulnerable devices on the Internet [44]. Furthermore, recent research shows the possibilities of DoS attacks through messaging protocols like MQTT [87, 88] and CoAP [91].

According to the ENISA Threat Landscape Report 2020, malware attacks are the leading and emerging threats worldwide [16]. While it is known that botmasters look for vulnerable devices with misconfigured protocols of Telnet and SSH, research suggests that bot deployments are now possible with IoT-based protocols like MQTT, AMQP, and UPnP [4, 31, 51, 82]. With the increasing adoption of IoT in diverse sectors like Industry 4.0, healthcare, and critical infrastructure, we argue that this poses a significant threat.

Heretofore, there has been research on the underlying IoT vulnerabilities and proposing honeypots to analyze the threat actors for specific protocols [32, 46, 63, 99]. However, to the best of our knowledge, no work combines an active search for misconfigured devices with an analysis of the attack trends in IoT by deploying multiple honeypots and studying the traffic flow received on a network telescope. In this paper, we unveil the vulnerable aspects of misconfigured services on IoT devices and emphasize the importance of authentication and authorization in IoT protocols and devices.

Our contributions are summarized as follows:

- We perform Internet-wide scans on six protocols: Telnet, MQTT, CoAP, AMQP, XMPP, and UPnP. As a result, we unveil 1.8 million misconfigured IoT devices that can either be infected with bots or be leveraged for a (D)DoS amplification attack. In addition, we use open datasets to complement our findings. Furthermore, our scan takes into account the existence of honeypots. To deal with the lack of ground truth knowledge for deployed honeypots on the Internet, we analyze the response banners from our scan and the static banners returned by open-source honeypots. Hence, we filter out from the results 8, 192 systems that we classify as



h(password | salt)

- Based on analysis of 3.3 million leaked passwords (2014).

1. 123456	13. letmein
2. password	14. abc123
3. 12345	15. 111111
4. 12345678	16. mustang
5. qwerty	17. access
6. 123456789	18. shadow
7. 1234	19. master
8. baseball	20. michael
9. dragon	21. superman
10. football	22. 696969
11. 1234567	23. 123123
12. monkey	24. batman
	25. trustno1

Hampus

Why this paper?

What you liked most?

Main **contributions**?

Category and **Context**; e.g.,

- Type of measurements?
- Type of data?
- Type of analysis?
- Type of contribution?

Correctness and **Clarity**

- If not, why not?

Tripwire: Inferring Internet Site Compromise

Joe DeBlasio, Stefan Savage, Geoffrey M. Voelker and Alex C. Snoeren
UC San Diego
{jdeblasio,savage,voelker,snoeren}@cs.ucsd.edu

ABSTRACT

Password reuse has been long understood as a problem: credentials stolen from one site may be leveraged to gain access to another site for which they share a password. Indeed, it is broadly understood that attackers exploit this fact and routinely leverage credentials extracted from a site they have breached to access high-value accounts at other sites (e.g., email accounts). However, as a consequence of such acts, this same phenomena of password reuse attacks can be harnessed to indirectly infer site compromises—even those that would otherwise be unknown. In this paper we describe such a measurement technique, in which unique honey accounts are registered with individual third-party websites, and thus access to an email account provides indirect evidence of credentials theft at the corresponding website. We describe a prototype system, called Tripwire, that implements this technique using an automated Web account registration system combined with email account access data from a major email provider. In a pilot study monitoring more than 2,300 sites over a year, we have detected 19 site compromises, including what appears to be a plaintext password compromise at an Alexa top-500 site with more than 45 million active users.

CCS CONCEPTS

• **Security and privacy** → *Intrusion detection systems*; *Authentication*; *Web application security*; *Phishing*; Social network security and privacy; • **Social and professional topics** → **Computer crime**;

KEYWORDS

Password Reuse, Website Compromise, Cybercrime, Webmail

ACM Reference Format:
Joe DeBlasio, Stefan Savage, Geoffrey M. Voelker and Alex C. Snoeren. 2017. Tripwire: Inferring Internet Site Compromise. In *Proceedings of IMC '17, London, United Kingdom, November 1–3, 2017,* 14 pages.
https://doi.org/10.1145/3131365.3131391

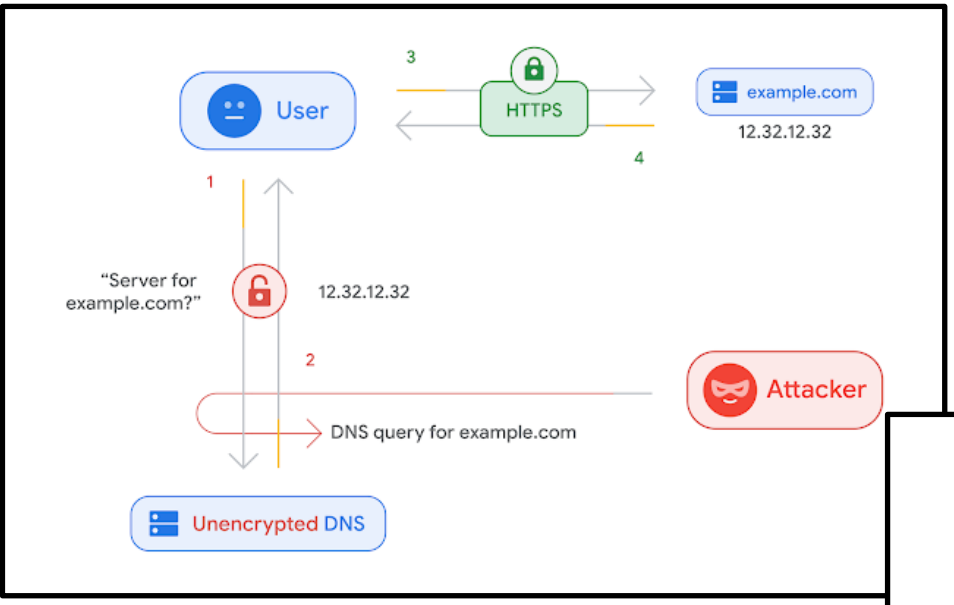
Thus, attackers seeking to compromise such accounts need only acquire the user's credentials.

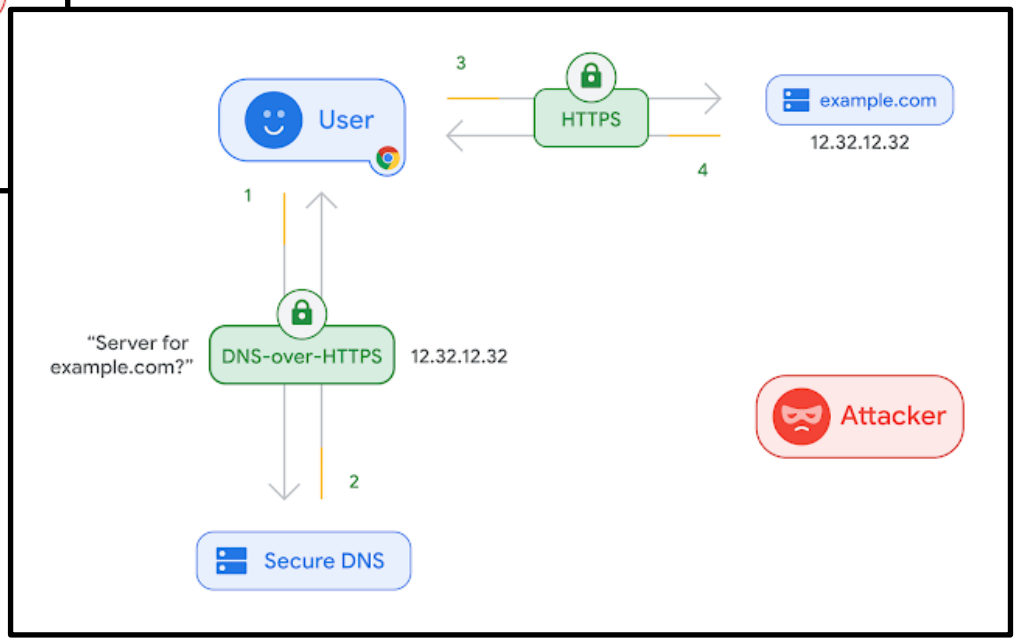
While there are a range of vectors by which account credentials can be compromised—including phishing, brute force and malware—perhaps the most pernicious arises from the confluence of data breaches and account reuse. In this scenario, an unrelated site is compromised such that all of its user accounts and passwords (or, more commonly, password hashes) are exposed. An attacker can then leverage this information to access any other accounts a user may have using the same credentials (a situation exacerbated by the widespread use of a user's email address as their standard user name) [3, 8, 14]. In one recent study, Das et al. estimated that over 40% of users reuse passwords [27] and our own anecdotal experience with stolen bulk account data suggests that up to 20% of stolen credentials may share a password with their primary email account.

Moreover, opportunities for such attacks abound, with reports of data breaches now commonplace: in just the past year, reports have surfaced of 117 million account credentials stolen from Linkedin [10] and 360 million from Myspace [4]. In 2014, Hold Security reported obtaining credentials for more than a billion users from breaches on several Internet services [39]. Indeed, the market for stolen credentials is thriving, with credentials being sold in bulk for under a penny a piece [41]. The value of these credentials lies in their ability to be used across sites, enabling account compromise at sites otherwise wholly unaffected by the unrelated original site's compromise.

The most sensitive and important credentials on offer are those associated with major email providers (e.g., Gmail, Live/Hotmail, Yahoo, etc.) because in modern usage it is these accounts that are the foundation for one's Internet footprint. In particular, online services commonly require an email address to register, to reconfirm accounts, to communicate key information and to reset or recover passwords. Thus, access to someone's email account can be sufficient to gain access to a broad array of other services as well. Indeed, while email accounts with such services are routinely compromised en masse, only one major email provider has had a public breach to date (Yahoo, in late 2014 [21]).

https://blog.chromium.org/2020/05/a-safer-and-more-private-browsing-DoH.html





Frans

Why this paper?

What you liked most?

Main **contributions**?

Category and **Context**; e.g.,

- Type of measurements?
- Type of data?
- Type of analysis?
- Type of contribution?

Correctness and **Clarity**

- If not, why not?

An End-to-End, Large-Scale Measurement of DNS-over-Encryption: How Far Have We Come?

Chaoyi Lu[1,2], Baojun Liu[3], Zhou Li[4], Shuang Hao[5], Haixin Duan[1,2,6], Mingming Zhang[1], Chunying Leng[1], Ying Liu[1], Zaifeng Zhang[7] and Jianping Wu[1]
[1]Institute for Network Sciences and Cyberspace, Tsinghua University
[2]Beijing National Research Center for Information Science and Technology (BNRist), Tsinghua University
[3]Department of Computer Science and Technology, Tsinghua University [4]University of California, Irvine
[5]University of Texas at Dallas [6]Qi An Xin Technology Research Institute [7]360 Netlab

ABSTRACT

DNS packets are designed to travel in unencrypted form through the Internet based on its initial standard. Recent discoveries show that real-world adversaries are actively exploiting this design vulnerability to compromise Internet users' security and privacy. To mitigate such threats, several protocols have been proposed to encrypt DNS queries between DNS clients and servers, which we jointly term as DNS-over-Encryption. While some proposals have been standardized and are gaining strong support from the industry, little has been done to understand their status from the view of global users.

This paper performs by far the first end-to-end and large-scale analysis on DNS-over-Encryption. By collecting data from Internet scanning, user-end measurement and passive monitoring logs, we have gained several unique insights. In general, the service quality of DNS-over-Encryption is satisfying, in terms of accessibility and latency. For DNS clients, DNS-over-Encryption queries are less likely to be disrupted by in-path interception compared to traditional DNS, and the extra overhead is tolerable. However, we also discover several issues regarding how the services are operated. As an example, we find 25% DNS-over-TLS service providers use invalid SSL certificates. Compared to traditional DNS, DNS-over-Encryption is used by far fewer users but we have witnessed a growing trend. As such, we believe the community should push broader adoption of DNS-over-Encryption and we also suggest the service providers carefully review their implementations.

CCS CONCEPTS

• **Networks** → **Application layer protocols**; **Network measurement**; *Naming and addressing*.

KEYWORDS

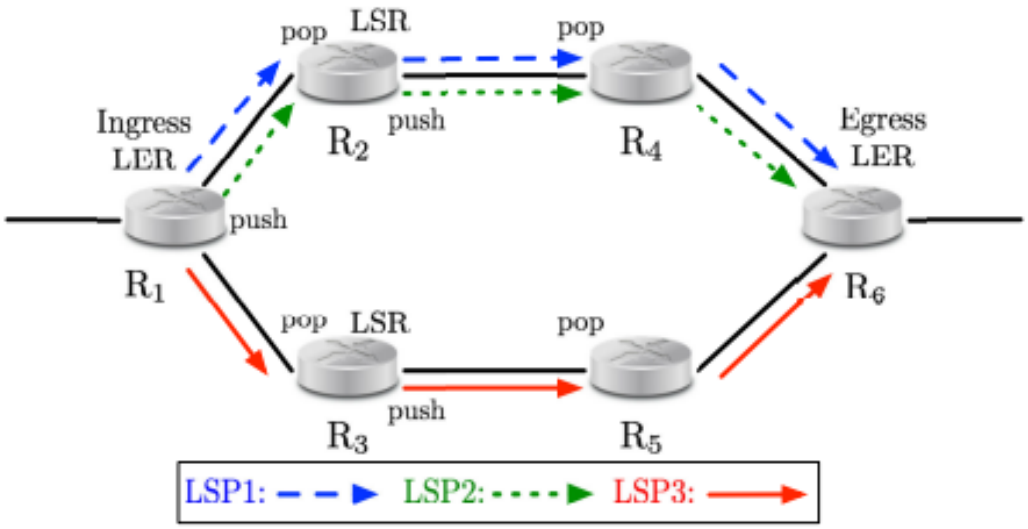
Domane Name System, DNS Privacy, DNS-over-TLS, DNS-over-HTTPS, DNS Measurement

ACM Reference Format:
Chaoyi Lu[1,2], Baojun Liu[3], Zhou Li[4], Shuang Hao[5], Haixin Duan[1,2,6], Mingming Zhang[1], Chunying Leng[1], Ying Liu[1], Zaifeng Zhang[7] and Jianping Wu[1]. 2019. An End-to-End, Large-Scale Measurement of DNS-over-Encryption: How Far Have We Come?. In *Internet Measurement Conference (IMC '19), October 21–23, 2019, Amsterdam, Netherlands.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3355369.3355580

1 INTRODUCTION

Domain Name System (DNS) is one of the fundamental building blocks of the Internet, mapping a user-friendly domain name to numerical IP addresses. According to its initial IETF standard, DNS packets are transmitted over UDP protocol *in clear-text*. Therefore, communication integrity and confidentiality are absent. Unfortunately, this design makes DNS communications vulnerable to attacks like eavesdropping and tampering [29]. In fact, real-world adversaries have been exploiting DNS to harm Internet users. As an example, released secret documents show that NSA has been covertly monitoring and hijacking DNS traffic, under the MoreCowBell [44] and QuantumDNS [12] projects. A recent study also shows that network middleboxes are actively intercepting DNS packets and rerouting them to alternative resolvers [60].

One of the mainstream approaches to mitigating such threat is to *encrypt DNS communications*. To this end, various techniques are proposed, including DNS-over-TLS (DoT), DNS-over-HTTPS (DoH), DNS-over-QUIC and DNSCrypt. In this paper, we jointly term them as DNS-over-Encryption (DoE). Although most of the protocols have only been established for a few years, some have


pop
LSR
pop
Ingress
LER
push
R2
push
R4
Egress
LER
push
R1
R6
pop
LSR
pop
push
R3
R5
LSP1:
LSP2:
LSP3:

Mohammad

Why this paper?

What you liked most?

Main **contributions**?

Category and **Context**; e.g.,

- Type of measurements?
- Type of data?
- Type of analysis?
- Type of contribution?

Correctness and **Clarity**

- If not, why not?

MPLS Under the Microscope: Revealing Actual Transit Path Diversity

Yves Vanaubel
Université de Liège
Belgium
yves.vanaubel@ulg.ac.be

Pascal Mérindol
Université de Strasbourg
France
merindol@unistra.fr

Jean-Jacques Pansiot
Université de Strasbourg
France
pansiot@unistra.fr

Benoit Donnet
Université de Liège
Belgium
benoit.donnet@ulg.ac.be

ABSTRACT

Traffic Engineering (TE) is one of the keys for improving packet forwarding in the Internet. It allows IP network operators to finely tune their forwarding paths according to various customer needs. One of the most popular tool available today for optimizing the use of networking resources is MPLS. On the one hand, operators may use MPLS and label distribution mechanisms such as RSVP-TE in conjunction with BGP to define multiple transit paths (for a given edge pair) verifying different constraints on their network. On the other hand, when operators simply enable LDP for distributing MPLS labels in order to improve the scalability of their network, another kind of path diversity may appear thanks to the ECMP feature of IGP routing.

In this paper, using an MPLS labels analysis, we demonstrate that it is possible to better understand the transit path diversity deployed within a given ISP. More specifically, we introduce the Label Pattern Recognition (LPR) algorithm, a method for analyzing `traceroute` data including MPLS information. LPR reveals the actual usage of MPLS according to the inferred label distribution protocol and is able to make the distinction between ECMP and TE multi-path forwarding. Based on an extensive and longitudinal `traceroute` dataset obtained from CAIDA, we apply LPR and find that each ISP behavior is really specific in regard to its MPLS usage. In particular, we are able to observe independently for each ISP the MPLS path diversity and usage, and its evolution over time. Globally speaking, the main outcomes of our study are that (i) the usage of MPLS has been increasing over the the last five years with

Categories and Subject Descriptors

C.2.1 [**Network Architecture and Design**]: Network topology

General Terms

Measurements

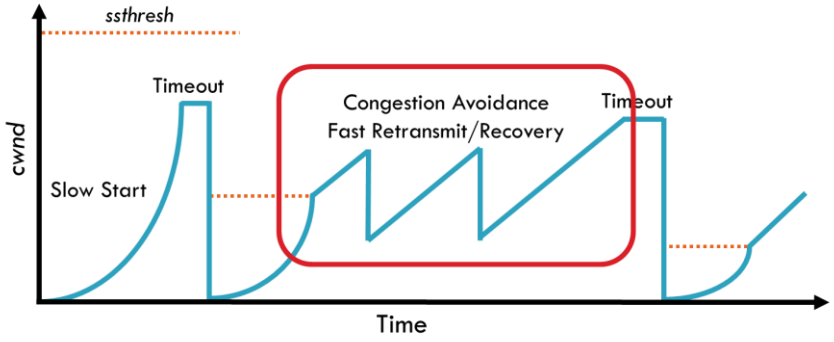
Keywords

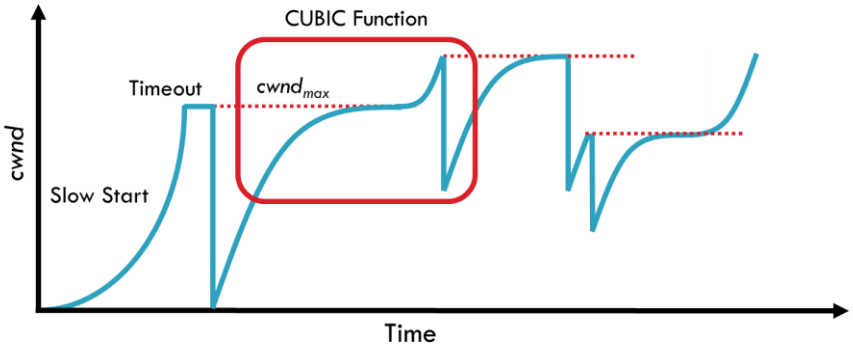
network discovery; MPLS; ECMP; multipath; LDP; RSVP-TE; traffic engineering; traceroute

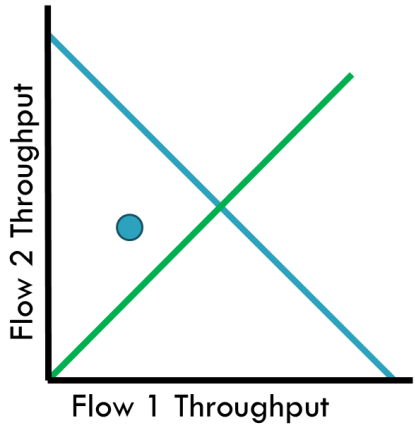
1. INTRODUCTION

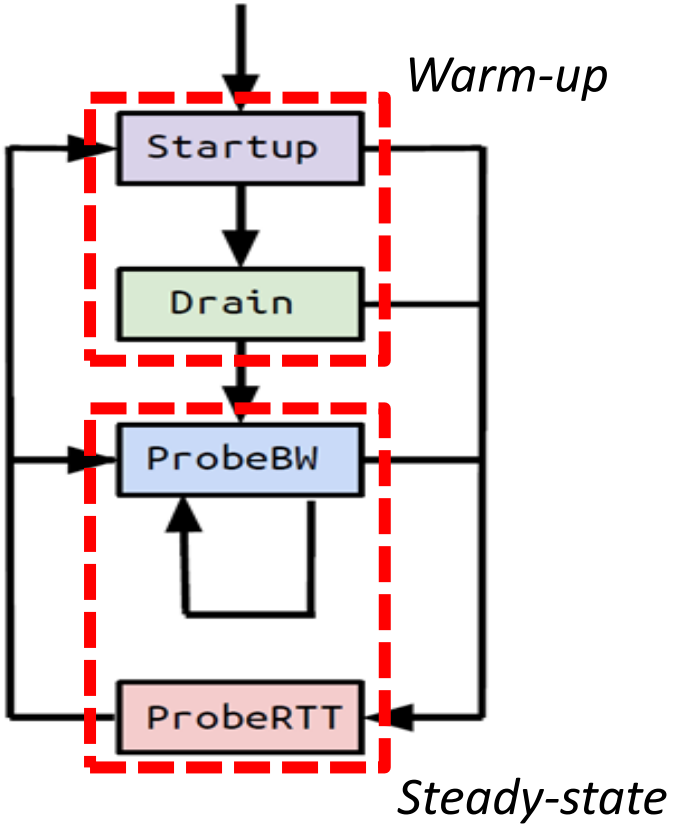
One of the cornerstones of the Internet is the way data is forwarded through routing paths. Typically, most of the IP flows are treated the same way whatever their specific Quality of Service (QoS) needs, their destination, or their origin. This absence of privileges and flow distinction is called *best effort routing* or *Internet neutrality*. Tools allowing operators to easily enable path diversity and, so, to perform *Traffic Engineering* (TE) are *Equal Cost MultiPath* (ECMP) load balancers [1] at the IP level and *Multiprotocol Label Switching* (MPLS [2]).

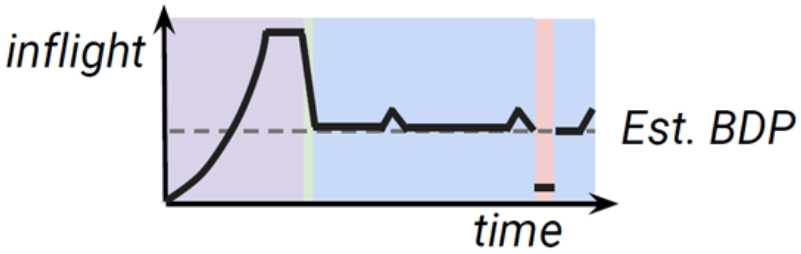
Historically, MPLS has been designed to reduce the time required to make forwarding decisions thanks to the insertion of *labels* before the IP header. Nowadays, it is commonly believed that MPLS is mainly used for providing additional virtual private networks (VPN) services [3] and TE capabilities [4, 5]. Recently, a few studies focused on MPLS,


ssthresh
Timeout
Congestion Avoidance
Fast Retransmit/Recovery
Timeout
cwnd
Slow Start
Time


CUBIC Function
Timeout
$cwnd_{max}$
cwnd
Slow Start
Time


Flow 2 Throughput
Flow 1 Throughput


Warm-up
Startup
Drain
ProbeBW
ProbeRTT
Steady-state


inflight
Est. BDP
time

Rodrigo

Why this paper?

What you liked most?

Main **contributions**?

Category and **Context**; e.g.,

- Type of measurements?
- Type of data?
- Type of analysis?
- Type of contribution?

Correctness and **Clarity**

- If not, why not?

Revisiting TCP Congestion Control Throughput Models & Fairness Properties At Scale

Adithya Abraham Philip, Ranysha Ware,
Rukshani Athapathu, Justine Sherry, Vyas Sekar
Carnegie Mellon University
United States of America

Abstract

Much of our understanding of congestion control algorithm (CCA) throughput and fairness is derived from models and measurements that (implicitly) assume congestion occurs in the last mile. That is, these studies evaluated CCAs in "small scale" edge settings at the scale of tens of flows and up to a few hundred Mbps bandwidths. However, recent measurements show that congestion can also occur at the core of the Internet on inter-provider links, where thousands of flows share high bandwidth links. Hence, a natural question is: Does our understanding of CCA throughput and fairness continue to hold at the scale found in the core of the Internet, with 1000s of flows and Gbps bandwidths?

Our preliminary experimental study finds that some expectations derived in the edge setting do not hold at scale. For example, using loss rate as a parameter to the Mathis model to estimate TCP NewReno throughput works well in edge settings, but does not provide accurate throughput estimates when thousands of flows compete at high bandwidths. In addition, BBR – which achieves good fairness at the edge when competing solely with other BBR flows – can become very unfair to other BBR flows at the scale of the core of the Internet. In this paper, we discuss these results and others, as well as key implications for future CCA analysis and evaluation.

CCS Concepts

• **Networks** → **Protocol testing and verification**; **Transport protocols**; **Network measurement**.

Keywords

congestion control, computer networks, fairness, throughput, tcp, bbr, reno, cubic

ACM Reference Format:
Adithya Abraham Philip, Ranysha Ware,, Rukshani Athapathu, Justine

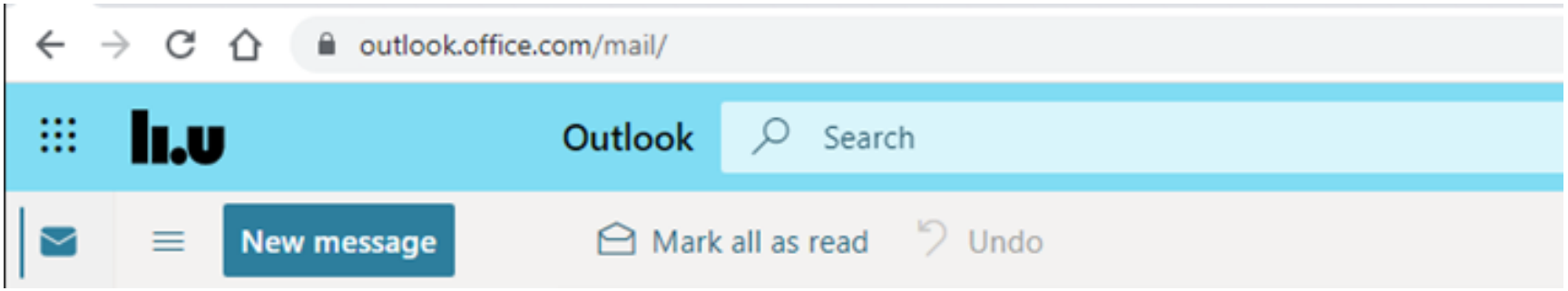
1 Introduction

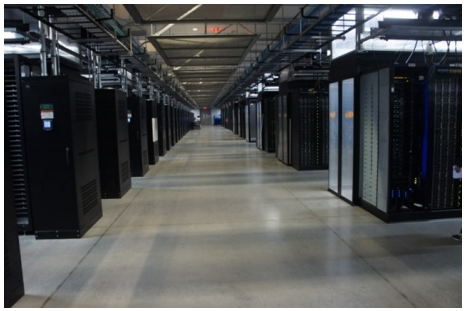
Congestion control algorithms (CCAs) are a fundamental building block of the modern Internet, and the networking community has been analyzing and designing CCAs for over three decades [12, 18, 26, 37, 39, 48]. Of specific interest to us is the wide-area setting. In this setting, *throughput* and *fairness* are important CCA properties, as they determine the effectiveness with which data can be transferred across the Internet, and the ability for multiple TCP flows to co-exist. To this end, many past efforts have used systematic models and experimental studies to understand these properties. For instance, the model by Mathis et al. [37] and Padhye et al. [41] predict the throughput of a NewReno flow as a function of packet loss and round-trip time (RTT), and the BBR model by Ware et al. [48] predicts the throughput of BBR when competing with other CCAs. Application developers can use such results to decide which CCA best suits the network conditions they experience or to debug performance issues.

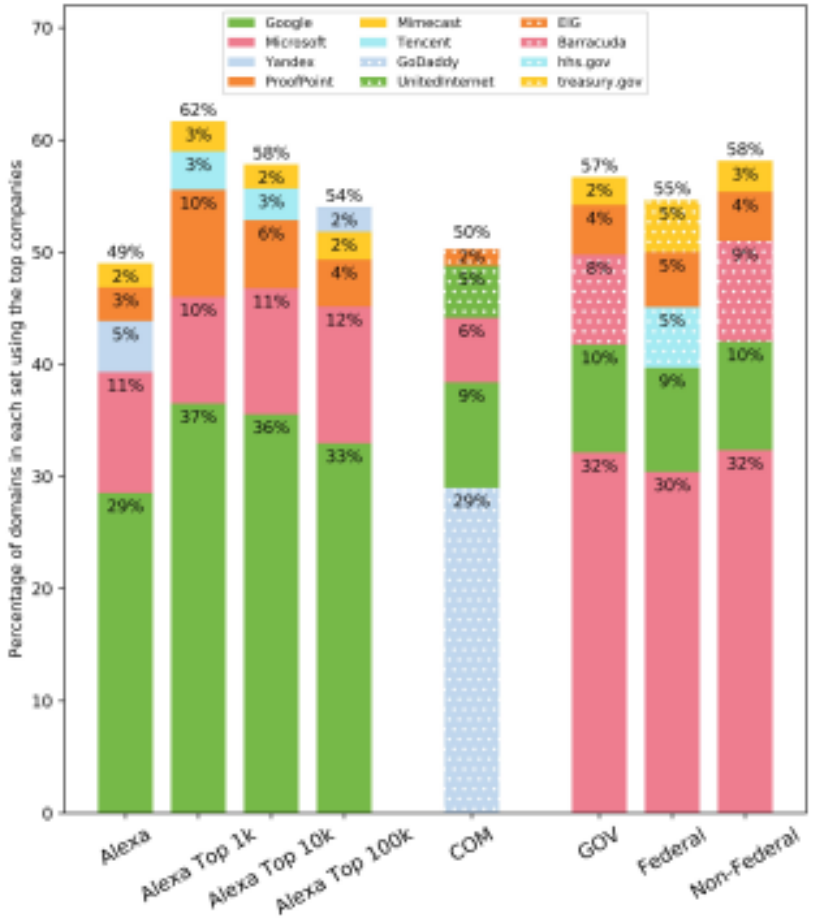
One implicit assumption in many of these results is that in the wide-area setting, congestion almost always occurs close to the *edge* or last mile of the network. To this end, many findings have been derived in contexts that emulate congestion occurring at the edge (e.g., residential link settings). Specifically, they consider a few or tens of flows competing for a shared bottleneck link with a capacity of a few tens or hundreds of Mbps [26, 37, 45, 52].

Many measurements suggest, however, that this assumption of congestion-at-the-edge may not always hold. Indeed, measurements both new [21] and old [11] show that congestion does occur at inter-domain links. These settings are characterized by higher flow counts and a larger network pipe [4, 13]. Indeed, past work on router buffer sizing has shown that CCA properties can change in such a setting [13].

This raises a natural question: do the known findings and models about TCP throughput [37] and fairness [20, 26, 28, 39] derived from edge-link settings, hold at the scale found in the core of the Internet?


outlook.office.com/mail/
Outlook
Search
New message
Mark all as read
Undo




Google
Microsoft
Yandex
ProofPoint
Mimecast
Tencent
GoDaddy
UnitedInternet
EIG
Barracuda
hhs.gov
treasury.gov
Percentage of domains in each set using the top companies
0
10
20
30
40
50
60
70
49%
62%
58%
54%
50%
57%
55%
58%
Alexa
Alexa Top 1k
Alexa Top 10k
Alexa Top 100k
COM
GOV
Federal
Non-Federal

Maximilian

Why this paper?

What you liked most?

Main **contributions**?

Category and **Context**; e.g.,

- Type of measurements?
- Type of data?
- Type of analysis?
- Type of contribution?

Correctness and **Clarity**

- If not, why not?

Who's Got Your Mail? Characterizing Mail Service Provider Usage

Enze Liu
UC San Diego
e7liu@eng.ucsd.edu

Gautam Akiwate
UC San Diego
gakiwate@cs.ucsd.edu

Mattijs Jonker
University of Twente
m.jonker@utwente.nl

Ariana Mirian
UC San Diego
amirian@eng.ucsd.edu

Stefan Savage
UC San Diego
savage@cs.ucsd.edu

Geoffrey M. Voelker
UC San Diego
voelker@cs.ucsd.edu

ABSTRACT

E-mail has long been a critical component of daily communication and the core medium for modern business correspondence. While traditionally e-mail service was provisioned and implemented independently by each Internet-connected organization, increasingly this function has been outsourced to third-party services. As with many pieces of key communications infrastructure, such centralization can bring both economies of scale and shared failure risk. In this paper, we investigate this issue empirically — providing a large-scale measurement and analysis of modern Internet e-mail service provisioning. We develop a reliable methodology to better map domains to mail service providers. We then use this approach to document the dominant and increasing role played by a handful of mail service providers and hosting companies over the past four years. Finally, we briefly explore the extent to which nationality (and hence legal jurisdiction) plays a role in such mail provisioning decisions.

CCS CONCEPTS

• **Information systems** → **World Wide Web**; • **World Wide Web** → **Internet communications tools**; • **Internet communications tools** → **E-mail**.

ACM Reference Format:
Enze Liu, Gautam Akiwate, Mattijs Jonker, Ariana Mirian, Stefan Savage, and Geoffrey M. Voelker. 2021. Who's Got Your Mail? Characterizing Mail Service Provider Usage. In *ACM Internet Measurement Conference (IMC '21), November 2–4, 2021, Virtual Event, USA.* ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3487552.3487820

1 INTRODUCTION

displacing the postal service for such matters over the previous two decades.

However, unlike the postal service (and many other forms of person-to-person communication) e-mail is not centrally administered, but is organized such that each Internet domain owner, by virtue of their DNS MX record, can make unique provisioning decisions about how and where they will accept e-mail delivery. Thus, organizations are free to provision separate e-mail services for each domain they own, to share service among domains they operate, or to outsource e-mail entirely to third-party providers. These choices, in turn, can have significant implications for the resilience, security, legal standing, performance and cost of e-mail service.

In particular, concerns have been raised in recent years about the general risks of increasing Internet service centralization and consolidation [5, 10, 17]. For example, centralization amplifies the impact of (even rare) service failures [4, 15, 25]. Similarly, a single data breach in a widely-used service can put thousands of customers' data at risk.[1] Finally, the legal jurisdiction in which a given service provider operates is implicitly imposed on the data managed by that provider. For instance, as a U.S. company, Google-managed data is subject to the Stored Communications Act, which provides data access to the government under warrant even if the data belongs to a foreign party not residing in the U.S..

Indeed, while historically e-mail was provisioned and implemented independently by each organization (*i.e.*, hosting a local mail server acting as a full-fledged Mail Transfer Agent), the rise of third-party enterprise mail service providers (notably Google and Microsoft) has challenged that assumption; indeed, there are compelling reasons to believe that that global e-mail service is also increasingly subject to a significant degree of centralization. How-

Christine

Why this paper?

What you liked most?

Main **contributions**?

Category and **Context**; e.g.,

- Type of measurements?
- Type of data?
- Type of analysis?
- Type of contribution?

Correctness and **Clarity**

- If not, why not?

Home is Where the Hijacking is: Understanding DNS Interception by Residential Routers

Audrey Randall
UC San Diego
aurandal@eng.ucsd.edu

Enze Liu
UC San Diego
e7liu@eng.ucsd.edu

Ramakrishna Padmanabhan
CAIDA/UC San Diego
ramapad@caida.org

Gautam Akiwate
UC San Diego
gakiwate@cs.ucsd.edu

Geoffrey M. Voelker
UC San Diego
voelker@cs.ucsd.edu

Stefan Savage
UC San Diego
savage@cs.ucsd.edu

Aaron Schulman
UC San Diego
schulman@cs.ucsd.edu

ABSTRACT

DNS interception — when a user's DNS queries to a target resolver are intercepted en route and forwarded to a different resolver — is a phenomenon of concern to both researchers and Internet users because of its implications for security and privacy. While the prevalence of DNS interception has received some attention, less is known about *where* in the network interception takes place. We introduce methods to identify where DNS interception occurs and who the interceptors may be. We identify when interception is performed before the query exits the ISP, and even when it is performed by the Customer Premises Equipment (CPE) in the user's own home. We believe that these techniques are vital in the light of the ongoing debate concerning the value of privacy-enhancing DNS transport.

CCS CONCEPTS

• **Networks** → **Home networks**; **Network measurement**.

ACM Reference Format:
Audrey Randall, Enze Liu, Ramakrishna Padmanabhan, Gautam Akiwate, Geoffrey M. Voelker, Stefan Savage, and Aaron Schulman. 2021. Home is Where the Hijacking is: Understanding DNS Interception by Residential Routers. In *ACM Internet Measurement Conference (IMC '21), November 2–4, 2021, Virtual Event, USA.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3487552.3487817

1 INTRODUCTION

In principle, devices are free to direct their DNS queries to the recursive resolver of their choosing. Indeed, it is this freedom that has enabled the growth of public resolvers such as those offered by Google, Cloudflare, and others. However, a key underlying assumption is that DNS queries are faithfully forwarded as they are addressed. Unfortunately, this is not always so.

DNS queries sent by a user's device can be *intercepted* en route

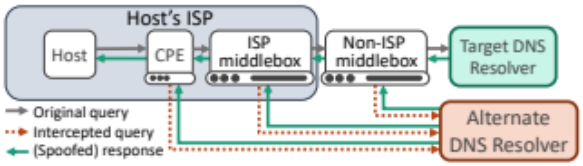


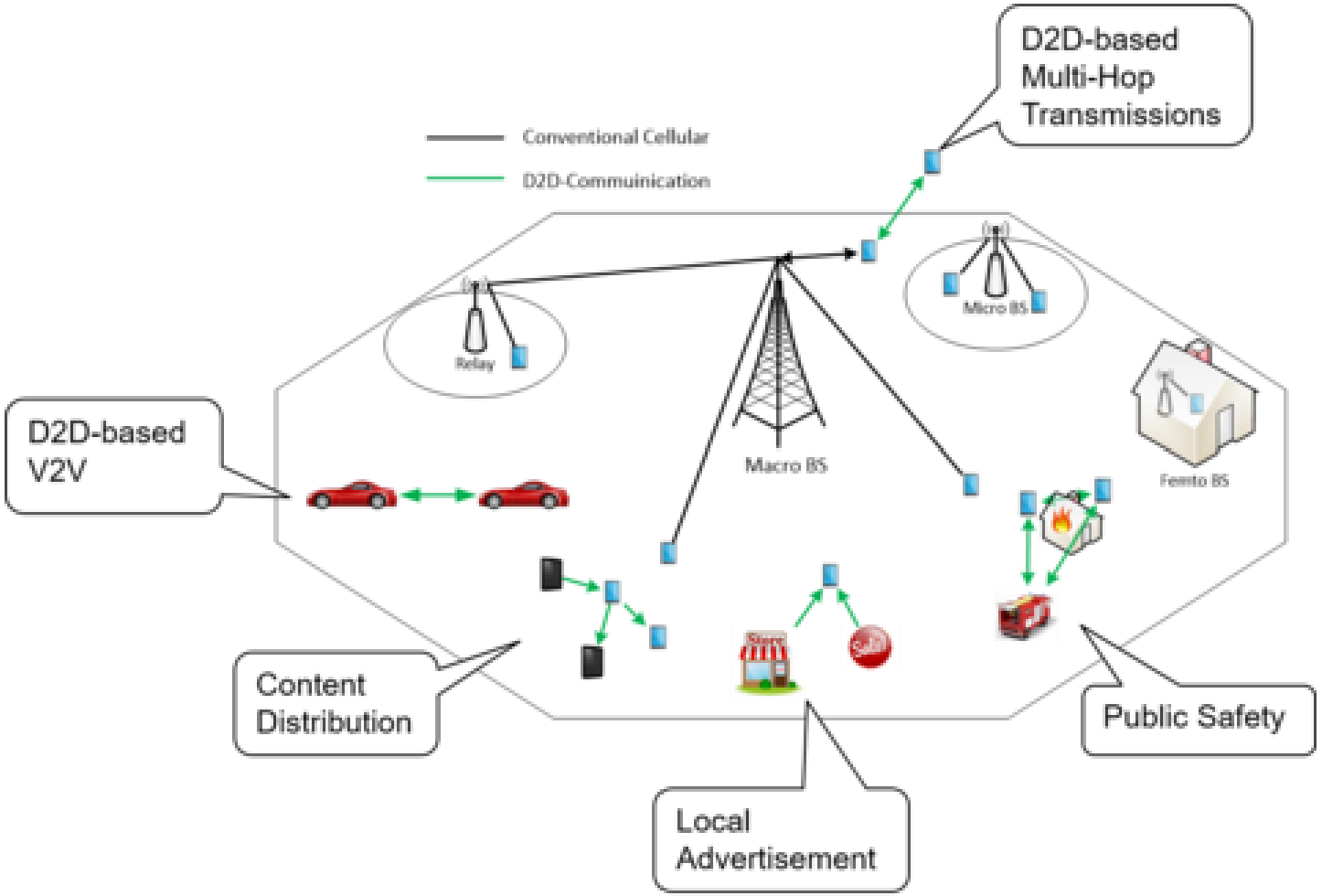
Figure 1: Locations where interception can occur.

spoofs responses so they appear to have been sent by the intended resolver. Transparent interception is difficult to detect because the alternate resolver does *not* have to modify the response. Even if the reason for the interception is benign — such as to prevent malware from evading DNS filtering — the interception of requests and misrepresentation of responses raise serious ethical concerns [14, 45] and can also interfere with the correct operation of protocols such as DNSSEC [14, 31].

While prior work has identified the broad prevalence of transparent interception [24, 31, 49], there are no established techniques for establishing *where* the interception is implemented. Indeed, there are a range of different points in the network where such interception might take place.

DNS redirection, another form of DNS manipulation, has also been found to occur in several parts of the network. DNS redirection occurs when a DNS resolver returns an altered response for specific queries and may occur with or without DNS interception. DNS redirection has been discovered *in Customer Premises Equipment (CPE)* to block resolution of specific domain names [17], *in ISPs* to replace NXDOMAIN responses with advertisements [30, 48] or enhance security and performance [44], and *outside of ISPs* to implement country-level censorship [4, 5, 16, 27]. Transparent interception has been far less extensively studied, although we are

https://www.ei.tum.de/lkn/research/dfg-d2d/



Armelle

Why this paper?

What you liked most?

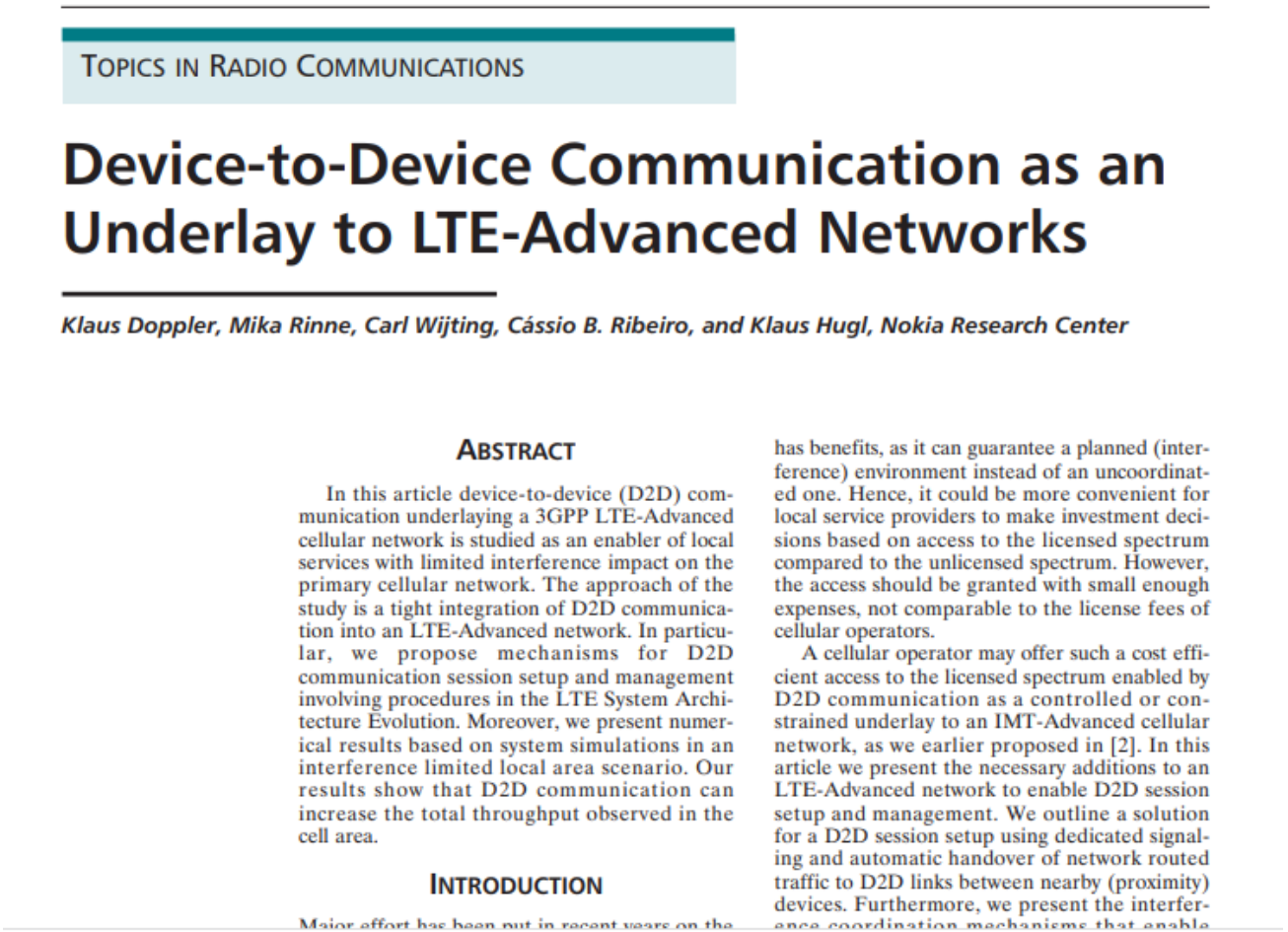
Main **contributions**?

Category and **Context**; e.g.,

- Type of measurements?
- Type of data?
- Type of analysis?
- Type of contribution?

Correctness and **Clarity**

- If not, why not?



TOPICS IN RADIO COMMUNICATIONS

Device-to-Device Communication as an Underlay to LTE-Advanced Networks

Klaus Doppler, Mika Rinne, Carl Wijting, Cássio B. Ribeiro, and Klaus Hugl, Nokia Research Center

ABSTRACT

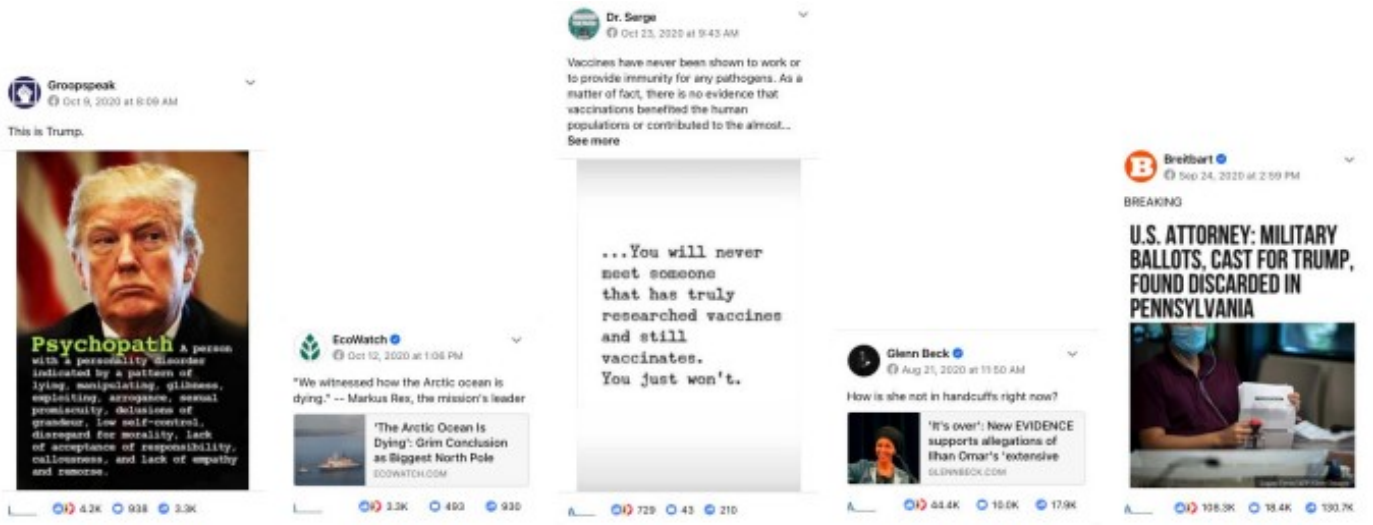
In this article device-to-device (D2D) communication underlaying a 3GPP LTE-Advanced cellular network is studied as an enabler of local services with limited interference impact on the primary cellular network. The approach of the study is a tight integration of D2D communication into an LTE-Advanced network. In particular, we propose mechanisms for D2D communication session setup and management involving procedures in the LTE System Architecture Evolution. Moreover, we present numerical results based on system simulations in an interference limited local area scenario. Our results show that D2D communication can increase the total throughput observed in the cell area.

INTRODUCTION

Major effort has been put in recent years on the

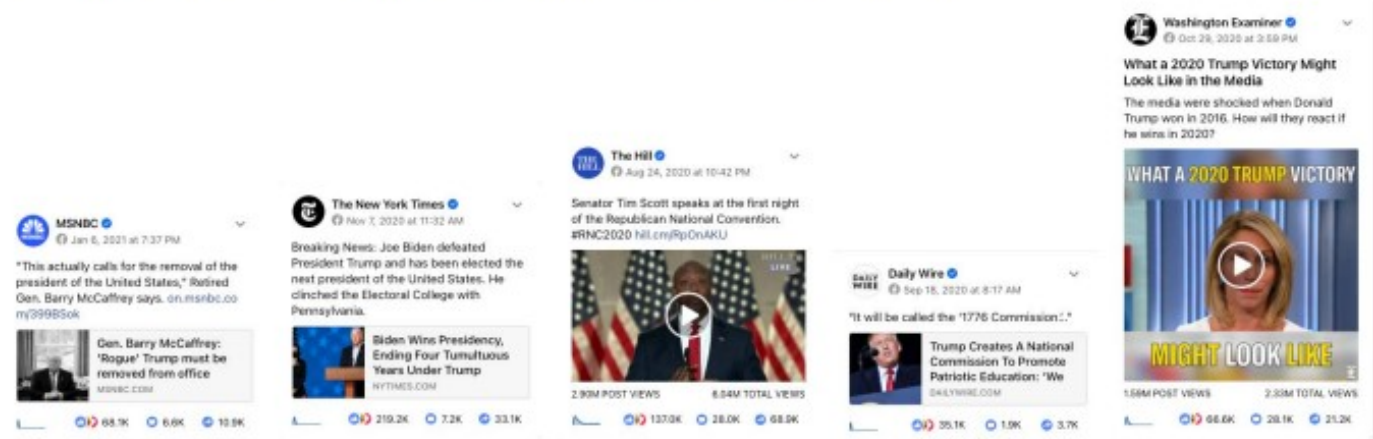
has benefits, as it can guarantee a planned (interference) environment instead of an uncoordinated one. Hence, it could be more convenient for local service providers to make investment decisions based on access to the licensed spectrum compared to the unlicensed spectrum. However, the access should be granted with small enough expenses, not comparable to the license fees of cellular operators.

A cellular operator may offer such a cost efficient access to the licensed spectrum enabled by D2D communication as a controlled or constrained underlay to an IMT-Advanced cellular network, as we earlier proposed in [2]. In this article we present the necessary additions to an LTE-Advanced network to enable D2D session setup and management. We outline a solution for a D2D session setup using dedicated signaling and automatic handover of network routed traffic to D2D links between nearby (proximity) devices. Furthermore, we present the interference coordination mechanisms that enable



(a) Groopspeak (b) EcoWatch (c) Dr. Serge (d) Glenn Beck (e) Breitbart

Figure 10: Misinformation: Sample Facebook posts ordered by the publisher's political leaning from far left (a) to far right (e). Publishers classified as misinformation based on NewsGuard and Media Bias/Fact Check data.



(a) MSNBC (b) The New York Times (c) The Hill (d) The Daily Wire (e) Washington Examiner

Figure 11: Non-Misinformation: Sample Facebook posts ordered by the publisher's political leaning, far left (a) to far right (e). Publishers classified as non-misinformation based on NewsGuard and Media Bias/Fact Check data.

Alireza

Why this paper?

What you liked most?

Main **contributions**?

Category and **Context**; e.g.,

- Type of measurements?
- Type of data?
- Type of analysis?
- Type of contribution?

Correctness and **Clarity**

- If not, why not?

Understanding Engagement with U.S. (Mis)Information News Sources on Facebook

Laura Edelson*
New York University
New York, NY, USA
lj992@nyu.edu

Minh-Kha Nguyen*
Université Grenoble Alpes
Grenoble, France

Ian Goldstein
New York University
New York, NY, USA

Oana Goga
Université Grenoble Alpes, CNRS, Inria, Grenoble INP, LIG
Grenoble, France

Damon McCoy
New York University
New York, NY, USA

Tobias Lauinger
New York University
New York, NY, USA

ABSTRACT

Facebook has become an important platform for news publishers to promote their work and engage with their readers. Some news pages on Facebook have a reputation for consistently low factualness in their reporting, and there is concern that Facebook allows their misinformation to reach large audiences. To date, there is remarkably little empirical data about how often users "like," comment and share content from news pages on Facebook, how user engagement compares between sources that have a reputation for misinformation and those that do not, and how the political leaning of the source impacts the equation. In this work, we propose a methodology to generate a list of news publishers' official Facebook pages annotated with their partisanship and (mis)information status based on third-party evaluations, and collect engagement data for the 7.5 M posts that 2,551 U.S. news publishers made on their pages during the 2020 U.S. presidential election. We propose three metrics to study engagement (1) across the Facebook news ecosystem, (2) between (mis)information providers and their audiences, and (3) with individual pieces of content from (mis)information providers. Our results show that misinformation news sources receive widespread engagement on Facebook, accounting for 68.1 % of all engagement with far-right news providers, followed by 37.7 % on the far left. Individual posts from misinformation news providers receive consistently higher median engagement than non-misinformation in every partisanship group. While most prevalent on the far right, misinformation appears to be an issue across the political spectrum.

CCS CONCEPTS

• **Security and privacy → Social aspects of security and pri-**

KEYWORDS

Facebook, news, misinformation, engagement, measurement.

ACM Reference Format:
Laura Edelson, Minh-Kha Nguyen, Ian Goldstein, Oana Goga, Damon McCoy, and Tobias Lauinger. 2021. Understanding Engagement with U.S. (Mis)Information News Sources on Facebook. In *ACM Internet Measurement Conference (IMC '21), November 2–4, 2021, Virtual Event*. ACM, New York, NY, USA, 20 pages. https://doi.org/10.1145/3487552.3487859

1 INTRODUCTION

After the 2016 U.S. presidential election, there was broad public concern [3, 14] about the impact that online misinformation might have had on public confidence in the fairness of the American electoral system. In response to scrutiny from lawmakers and their users, Facebook announced several initiatives [22, 23] aimed at reducing misinformation on its platforms.

To date, there is little public data about how widespread the misinformation problem on Facebook is. Prior work about misinformation on digital platforms has focused either on mechanisms of spread [13, 18, 21], or on absolute measurements of fake news [15, 19]. With few notable exceptions [34], research has not widely studied the interplay of misinformation and partisanship on digital platforms.

In this work, we aim to shed light on user engagement within the news ecosystem on Facebook. To the best of our knowledge, we are the first to characterize engagement based on the political leaning and factualness of news sources. We explore engagement with (mis)information news from three different perspectives:

(1) What share of overall engagement with U.S. news sources is

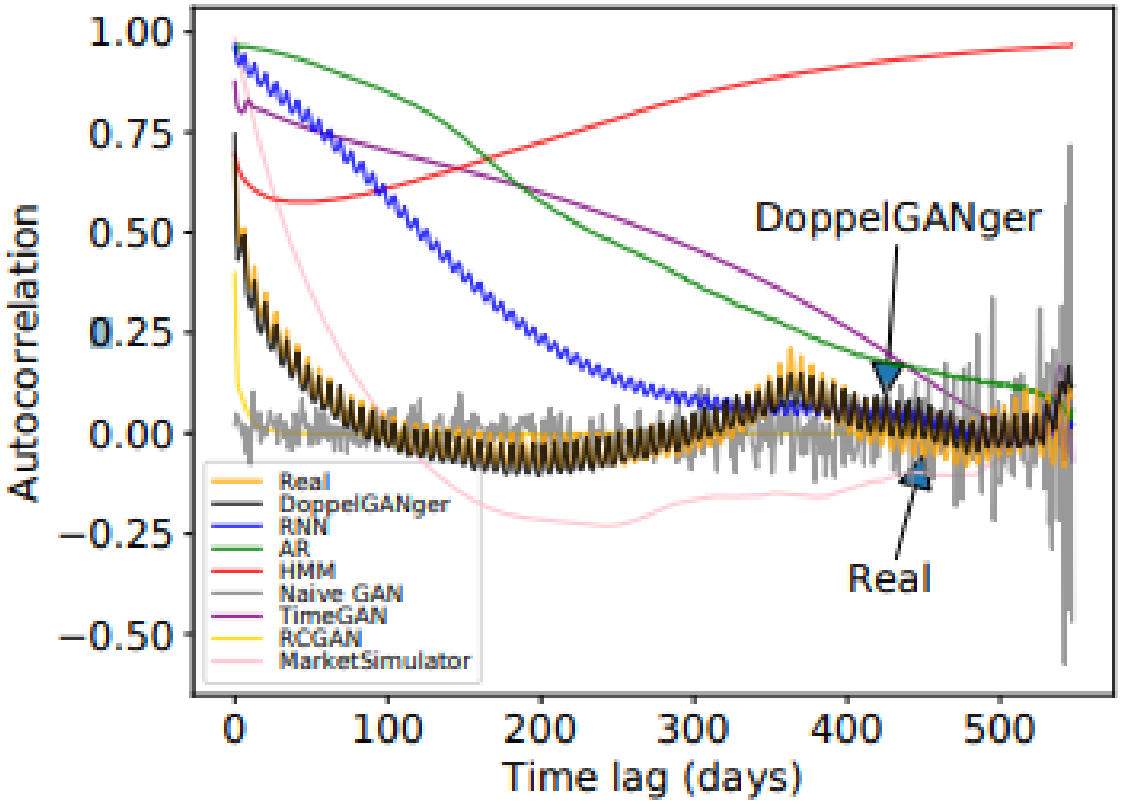


Figure 1: Autocorrelation of daily page views for Wikipedia Web Traffic dataset.

Minh-Ha

Why this paper?

What you liked most?

Main **contributions**?

Category and **Context**; e.g.,

- Type of measurements?
- Type of data?
- Type of analysis?
- Type of contribution?

Correctness and **Clarity**

- If not, why not?

Using GANs for Sharing Networked Time Series Data: Challenges, Initial Promise, and Open Questions

Zinan Lin
Carnegie Mellon University
Pittsburgh, PA
zinanl@andrew.cmu.edu

Alankar Jain
Carnegie Mellon University
Pittsburgh, PA
alankarjain91@gmail.com

Chen Wang
IBM
New York, NY
Chen.Wang1@ibm.com

Giulia Fanti
Carnegie Mellon University
Pittsburgh, PA
gfanti@andrew.cmu.edu

Vyas Sekar
Carnegie Mellon University
Pittsburgh, PA
vsekar@andrew.cmu.edu

ABSTRACT

Limited data access is a longstanding barrier to data-driven research and development in the networked systems community. In this work, we explore if and how generative adversarial networks (GANs) can be used to incentivize data sharing by enabling a generic framework for sharing synthetic datasets with minimal expert knowledge. As a specific target, our focus in this paper is on time series datasets with metadata (e.g., packet loss rate measurements with corresponding ISPs). We identify key challenges of existing GAN approaches for such workloads with respect to fidelity (e.g., long-term dependencies, complex multidimensional relationships, mode collapse) and privacy (i.e., existing guarantees are poorly understood and can sacrifice fidelity). To improve fidelity, we design a custom workflow called DoppelGANger (DG) and demonstrate that across diverse real-world datasets (e.g., bandwidth measurements, cluster requests, web sessions) and use cases (e.g., structural characterization, predictive modeling, algorithm comparison), DG achieves up to 43% better fidelity than baseline models. Although we do not resolve the privacy problem in this work, we identify fundamental challenges with both classical notions of privacy and recent advances to improve the privacy properties of GANs, and suggest a potential roadmap for addressing these challenges. By shedding light on the promise and challenges, we hope our work can rekindle the conversation on workflows for data sharing.

CCS CONCEPTS

• **Networks** → **Network simulations**; • **Computing method-**

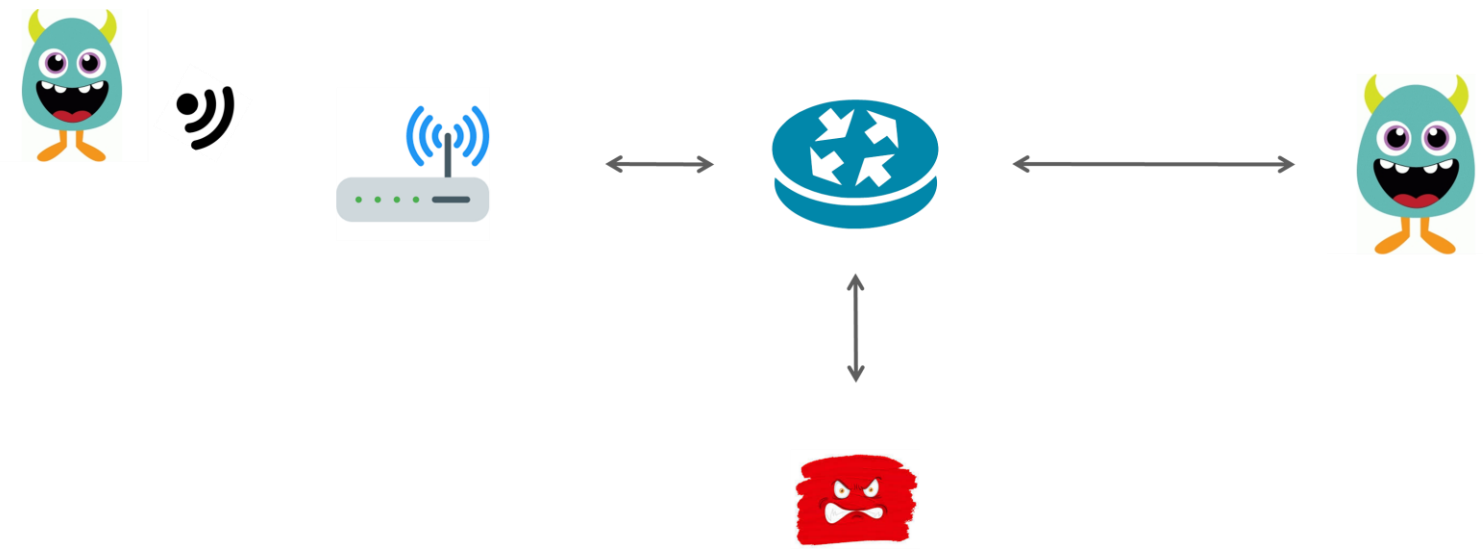
ACM Reference Format:
Zinan Lin, Alankar Jain, Chen Wang, Giulia Fanti, and Vyas Sekar. 2020. Using GANs for Sharing Networked Time Series Data: Challenges, Initial Promise, and Open Questions. In *ACM Internet Measurement Conference (IMC '20), October 27–29, 2020, Virtual Event, USA.* ACM, New York, NY, USA, 20 pages. https://doi.org/10.1145/3419394.3423643

1 INTRODUCTION

Data-driven techniques [51] are central to networking and systems research (e.g., [11, 16, 23, 48, 60, 60, 71, 74, 83, 104]). This approach allows network operators and system designers to explore design choices driven by empirical needs, and enable new data-driven management decisions. However, in practice, the benefits of data-driven research are restricted to those who possess data. Even when collaborating stakeholders have plenty to gain (e.g., an ISP may need workload-specific optimizations from an equipment vendor), they are reluctant to share datasets for fear of revealing business secrets and/or violating user privacy. Notable exceptions aside (e.g., [2, 77]), the issue of data access continues to be a substantial concern in the networking and systems communities.

One alternative is for data holders to create and share *synthetic* datasets modeled from real traces. There have been specific successes in our community where experts identify the key factors of specific traces that impact downstream applications and create generative models using statistical toolkits [5, 29, 31, 43, 64, 68, 78–81, 97, 98, 109, 110, 115]. Unfortunately, this approach requires significant human expertise and does not easily *generalize* across workloads and use cases.

The overarching question for our work is: Can we create high-



Suleman

Why this paper?

What you liked most?

Main **contributions**?

Category and **Context**; e.g.,

- Type of measurements?
- Type of data?
- Type of analysis?
- Type of contribution?

Correctness and **Clarity**

- If not, why not?

Detection, Classification, and Analysis of Inter-Domain Traffic with Spoofed Source IP Addresses

Franziska Lichtblau
TU Berlin
franziska@inet.tu-berlin.de

Florian Streibelt
TU Berlin
florian@inet.tu-berlin.de

Thorben Krüger
TU Berlin
thorben@inet.tu-berlin.de

Philipp Richter
TU Berlin
prichter@inet.tu-berlin.de

Anja Feldmann
TU Berlin
anja@inet.tu-berlin.de

ABSTRACT

IP traffic with forged source addresses (i.e., spoofed traffic) enables a series of threats ranging from the impersonation of remote hosts to massive denial-of-service attacks. Consequently, IP address spoofing received considerable attention with efforts to either suppress spoofing, to mitigate its consequences, or to actively measure the ability to spoof in individual networks. However, as of today, we still lack a comprehensive understanding both of the prevalence and the characteristics of spoofed traffic "in the wild" as well as of the networks that inject spoofed traffic into the Internet.

In this paper, we propose and evaluate a method to passively detect spoofed packets in traffic exchanged between networks in the inter-domain Internet. Our detection mechanism identifies both source IP addresses that should never be visible in the inter-domain Internet (i.e., unrouted and bogon sources) as well as source addresses that should not be sourced by individual networks, as inferred from BGP routing information. We apply our method to classify the traffic exchanged between more than 700 networks at a large European IXP. We find that the majority of connected networks do not, or not consistently, filter their outgoing traffic. Filtering strategies and contributions of spoofed traffic vary heavily across networks of different types and sizes. Finally, we study qualitative characteristics of spoofed traffic, regarding both application popularity as well as structural properties of addresses. Combining our observations, we identify and study dominant attack patterns.

CCS CONCEPTS

• **Networks** → **Network measurement**; **Network security**;

KEYWORDS

ACM Reference Format:
Franziska Lichtblau, Florian Streibelt, Thorben Krüger, Philipp Richter, and Anja Feldmann. 2017. Detection, Classification, and Analysis of Inter-Domain Traffic with Spoofed Source IP Addresses. In *Proceedings of ACM Internet Measurements Conference (IMC'17).* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3131365.3131367

1 INTRODUCTION

The Internet Protocol (IP) provides a unified and simple abstraction for communication over the Internet. It identifies hosts by their IP addresses, allowing for data exchanges across heterogeneous networks. While the simplicity of the Internet Protocol has proven immensely powerful it comes with inherent limitations, such as the lack of packet-level authenticity. Routers perform only a lookup for the destination address of incoming packets, the authenticity of source IP addresses of packets is not validated on the path between sender and receiver.

The resulting ability to forge the source IP address of a packet (i.e., *spoofing*) enables a series of cybersecurity threats, ranging from the impersonation of remote hosts to massive denial-of-service Attacks, causing major disruptions of Internet services [48]. In response, the IETF developed best practices for ingress traffic filtering at autonomous system (AS) borders [23]. The spoofing problem also received considerable attention from the research community with systems and architectures that have the potential to either limit or prevent spoofing in the Internet (e.g., [6, 24, 32]). However, these mitigation approaches have not succeeded in eliminating spoofing in production environments: Attacks involving spoofed source IP addresses remain widespread [17, 37].

The measurement community has been very successful in detecting the *ability* to spoof in individual networks using active mea-