Traffic Shadowing: A Global Investigation of Internet Traffic Observation and User Data Reutilization

Yunpeng Xing, Chaoyi Lu, Baojun Liu, Member, IEEE, Ruixuan Li, and Haixin Duan, Member, IEEE

***Abstract*—We conduct a large-scale measurement of an insufficiently explored form of traffic manipulation—traffic shadowing. It refers to the covert process by which users' Internet traffic is silently observed, stored, and later exploited to generate unsolicited requests. A comprehensive analysis of it aids in diagnosing faulty devices, identifying user data leaks, and evaluating the related security and privacy risks. Previous studies on traffic shadowing suffer from limited probing destinations, no measurements from residential networks, or insufficient analysis. In this paper, we select a large number of IPv4 open servers as destination IPs to conduct a large-scale global traffic shadowing measurement (Phase I) from data center networks, and conduct a China-wide measurement in residential networks using a local web advertising service provider (Phase II). We find traffic decoys sent to 3.7×10^3 open DNS resolvers, 1.4×10^3 open HTTP servers and 1.9×10^3 open HTTPS servers are affected. More than 90% of unsolicited requests appear more than 1 hour after initial decoys. One decoy may trigger multiple unsolicited requests. We investigate the location of user data flows and find that DNS traffic is primarily exiting from open resolvers, while HTTP/HTTPS traffic can be sniffed along the path or at the destination. Between traffic observers and unsolicited visitors, there is a prevalent flow of user data. Our findings encourage the technical community to adopt proactive measures to address this phenomenon, such as promoting the deployment of encryption protocols and "oblivious" solutions.**

***Index Terms*—Traffic Shadowing, Network Monitoring, Security Probing**

I. Introduction

Internet traffic has been exposed to various forms of manipulation, such as Internet censorship [67, 13], session interception [40, 12], and middlebox interference [19, 28]. This paper explores a more passive and covert form of traffic manipulation that has been less studied: *traffic shadowing*. It describes the process in which on-path observers sniff and capture packets during transmission, causing them to reappear as unsolicited requests later, even when no legitimate clients are waiting for responses. For example, an APNIC blog [35] reported that a quarter of DNS queries are repeated, sometimes hours or even days after their initial origination.

The underlying causes of traffic shadowing can be either benign or malicious. Benign motives include security companies gathering threat intelligence or intrusion detection systems proactively scanning domains, while malicious drivers encompass attackers collecting exploitable information or individuals harvesting sensitive user privacy data. Existing research fails to adequately uncover the root causes behind this phenomenon. Moreover, even when stemming from benign intentions, the execution process may compromise user security and privacy due to insufficient regulatory constraints. Therefore, conducting comprehensive measurements of traffic shadowing is essential to assess its impact on user security and privacy.

As a thorough analysis of traffic shadowing, there are three major questions to be answered. The first question is: **what is the landscape of traffic shadowing?** This includes the protocols for affected traffic and unsolicited requests, as well as the affected proportion. Then, considering the traffic shadowing process, we can identify three components: *Internet users*, *traffic observers*, and *unsolicited visitors*. *Users* are victims. *Traffic observers* sniff user traffic and may transfer it to those seeking to utilize the data. *Unsolicited visitors* are the actual exploiters of user data, and they could either be the traffic observers themselves or organizations collaborating with them. This raises the second question: **what are the characteristics of traffic observers and unsolicited visitors?** Furthermore, the observation of user data and the flow of this data from traffic observers to unsolicited visitors—both of which constitute unsolicited manipulations of users' privacy—are critical factors in the privacy risks associated with traffic shadowing. Therefore, there is the last and even the most important question: **How does data flow within the traffic shadowing ecosystem?**

Research gap. Previous studies have limitations that prevent them from adequately addressing the three questions characterizing traffic shadowing. Xing et al. [64] perform a global measurement leveraging VPN-based vantage points, revealing that three key protocols—DNS, HTTP, and TLS—are susceptible, particularly DNS queries to public resolvers and HTTP requests to high-ranking websites. Their methodology relies on commercial data center VPNs to deploy traffic decoys, while overlooking residential networks. Given that traffic shadowing is often motivated by data harvesting, its behavior in residential networks warrants closer scrutiny. Although datacenter VPN nodes are distributed across many countries, they are predominantly concentrated within data centers established by VPN providers in each nation. As a result, their network diversity

Manuscript received 26 August 2025; revised 28 February 2026; accepted 28 May 2026; approved by IEEE/ACM TRANSACTIONS ON NETWORKING Associate Editor Sara Alouf. Date of publication 00 Month 0000; date of current version 00 Month 0000. This work is supported by the National Key Research and Development Program of China (No. 2023YFB3105600) and the National Natural Science Foundation of China (62102218). An earlier version of this article was presented at Internet Measurement Conference 2024 (IMC 2024) [DOI: 10.1145/3646547.3689023]. (Corresponding authors: Chaoyi Lu and Baojun Liu.)

Yunpeng Xing, Baojun Liu, Ruixuan Li, and Haixin Duan are with the Institute for Network Sciences and Cyberspace, BNRist, Tsinghua University, Beijing 100084, China (e-mail: xyp23@mails.tsinghua.edu.cn; lbj@tsinghua.edu.cn; lrx.goat@gmail.com; duanhx@tsinghua.edu.cn).

Chaoyi Lu is with Zhongguancun Laboratory, Beijing 100190, China (e-mail: lucy@zgclab.edu.cn).

Digital Object Identifier 00.0000/TNET.0000.0000000

is significantly lower than that of residential network vantage points (distributed in many actual users' home networks). In [64], DNS probing traffic is sent only to 36 DNS servers, and their HTTP/HTTPS probing traffic is sent only to the top 1k tranco sites, which are mainly distributed across large data center networks. The limited scope of destination addresses restricts the diversity of their results. Their study does not address how data flows within the traffic shadowing ecosystem, which is essential for assessing the associated privacy risks and developing effective countermeasures. The APNIC Blog [35] analyzes traffic shadowing solely on the DNS protocol, lacking analysis of traffic observers, unsolicited visitors, and data flow patterns. The technical analysis on FireEye [44] focuses on a single software vendor.

Our study. We reuse the decoy-based method proposed in the previous work, make two improvements to the measurement, and conduct a deeper analysis. We spread traffic decoys (DNS, HTTP, TLS), each containing a unique domain name, to trigger traffic shadowing. The domain name is controlled by us and resolved to our http/https honeypot so that we can receive unsolicited requests on the DNS authoritative server and on the honeypot, identifying traffic shadowing against our decoys. We use a traceroute-based method to locate traffic observation to a specific hop on the path or at the destination. Our measurement consists of two phases: Phase I is a global-scale measurement conducted in data center networks, and Phase II is a China-wide measurement in residential networks. In Phase I, to improve path diversity, we send traffic decoys from 4 controlled VPSs to 1.6×10^5 open servers on the Internet, without using commercial VPN services (unlike [64]). In Phase II, we use a web advertising service provider in China [3], considering that [64] found traffic shadowing to be prevalent in China.

Results. After a six-month measurement (Oct. 12, 2024 to Apr. 12, 2025), we find traffic decoys sent to 3.7×10^3 open DNS resolvers, 1.4×10^3 open HTTP servers and 1.9×10^3 open HTTPS servers are affected. HTTP/TLS traffic is observed whether by on-path sniffers (31% and 46%) or by open servers at the decoy traffic destination (69% and 54%). 99% of the observed DNS traffic is observed at the destination (the open resolvers). Traffic observers share user data with unsolicited visitors belonging to different organizations with them. The unsolicited visitors query the domain name in decoys and then send unsolicited http/https[1] requests to our honeypots. Traffic decoys do not only trigger unsolicited requests once when sent; instead, they are stored for a long period (even 6 months) and repeatedly exploited. More than 90% unsolicited requests appear one hour later than initial decoys. A significant proportion of source IPs of unsolicited requests are listed in the Spamhaus IP blocklist [2], indicating that they may be involved in other Internet scanning activities. We find indicative results that may imply user data transfer. Traffic decoys are sniffed in the networks of 130 organizations that have on-path devices, yet the associated unsolicited requests originate from another set of 26 organizations. The organizations hosting on-path devices are mainly ISPs and data centers, while the organizations from which the unsolicited requests originate are primarily carriers, cloud service providers, and security companies. In the Chinese residential network, the proportion of traffic decoys triggering unsolicited requests is at least one order of magnitude higher than in data center networks, while the behavioral characteristics are generally consistent with those in data center networks.

[1]To make it easier to distinguish, we use lowercase letters for the protocols of unsolicited requests and uppercase letters for the protocols of the traffic decoy.

Contributions of this work include:

- We improve the approach of the previous work [64]: We send decoy to 1.5×10^5 destination IPs to improve path diversity, rather than adopting commercial VPN services to introduce numerous source IPs for the same purpose. We develop a method to deploy traffic decoys through advertising services, thereby introducing 7.3×10^4 residential network vantage points.
- Compared with the previous work, we conduct a more in-depth analysis and obtain more findings, including the distribution and working characteristics of sniffing devices, the characteristics of unsolicited visitors' devices, and the patterns of user data flow among different organizations.

II. Background and Related Work

Traditional types of traffic manipulation: Proactive and obvious. Traffic manipulation, a long-existing phenomenon, has been studied by many researchers. Governments and network operators perform *Internet censorship* to prevent users from accessing prohibited content. Previous studies focus on the strategies adopted by censors [67, 65], and the strategies for censorship circumvention [13, 7, 26, 62]. Researchers conduct global measurements to reveal the scale and characteristics of Internet censorship[11, 45, 47, 50]. Some measurements focus on specific countries, examining their particular Internet censorship systems[5, 30, 63, 66, 68]. Another type of traffic shadowing is *session interception*, which may be used to enhance security and performance. Researchers analyze the characteristics of such behavior, the affected networks, and the underlying motives. Previous studies, for example, show that DNS interception primarily occurs in residential networks [52], where ISPs use it to reduce traffic leaving their network, thus minimizing settlement fees paid to other ISPs [40]. Some studies focus on HTTP interception devices and find that they may engage in activities such as page tampering or ad injection [12, 59, 15, 72]. Traffic manipulation also includes field tampering at the TCP and IP layers [19, 28, 56], with studies on this often being proposed under the terminology of *middlebox* [16, 17]. Middleboxes may be deployed for security or performance reasons, but they violate the end-to-end principle, potentially affecting network availability or introducing additional security issues[42, 27, 33, 60].

Traffic shadowing: a passive and covert format of traffic manipulation. A key distinction between traffic shadowing and other types of traffic manipulation is that the traffic sent by users is not interfered with and can still reach the server, and the server's response is returned to the user as expected. Therefore, traffic shadowing is difficult for users to perceive and challenging for researchers to measure. To detect traffic shadowing, we cannot simply check the tampered fields in

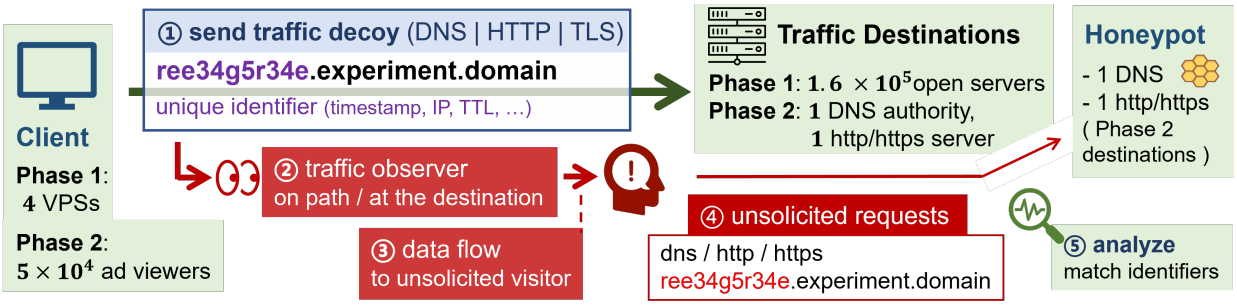

Fig. 1: Overview of methodology

the packets received by the server, as is done in other traffic manipulation detection methods. Due to the covert nature of traffic observation and the uncertainty of unsolicited request timing, prior work [35, 64] sends traffic decoys and performs asynchronous matching between server logs and the decoys to detect traffic shadowing. A blog of APNIC [35] uses an online ad service to send DNS queries containing random strings to detect shadow traffic. They find that one quarter of the incoming DNS queries on the authority servers are shadow traffic. They find that half of the shadow traffic comes one day after the initial probing traffic. Xing et al. [64] extend the measured protocol to DNS, HTTP, and TLS. They find that the unsolicited requests triggered may appear hours or even days after the initial decoys, and may use the same or a different protocol than the initial decoys. They find that most unsolicited requests are for benign purposes such as security scanning. A technical analysis [44] shows that a specific security software (FireEye) sniffs users' traffic and actively probes domain names to detect malicious websites.

III. Approach

We use the decoy-based method proposed by a previous work [64] to detect traffic shadowing, and make 2 improvements: (1) We select 1.6×10^5 servers as destinations for decoy traffic to ensure sufficient path coverage, rather than relying on commercial VPN nodes to provide source IPs while using only a small number of destination IPs; (2) We integrate vantage points in residential networks. Next, we will give an overview of the approach, introduce our improvements (summarized in Table II), and discuss the limitations.

A. A Decoy-based Method

A distinctive characteristic of traffic shadowing is unsolicited, reappearing requests with no original client waiting for them. So we spread traffic decoy and compare incoming traffic and decoys we sent to detect traffic shadowing, shown in Figure 1. **Step 1.** We send traffic decoys to some destinations, each embedded with a unique identifier. The specific settings for the traffic sources and destinations depend on the experimental phase: Phase I (III-B) and Phase II (III-C). **Step 2.** These traffic decoys are observed during network transmission or after reaching the destination server. **Step 3.** User data flows from traffic observers to unsolicited visitors. **Step 4.** Unsolicited visitors are prompted by the domain name in our decoy to send unsolicited requests to our honeypots. **Step 5.**

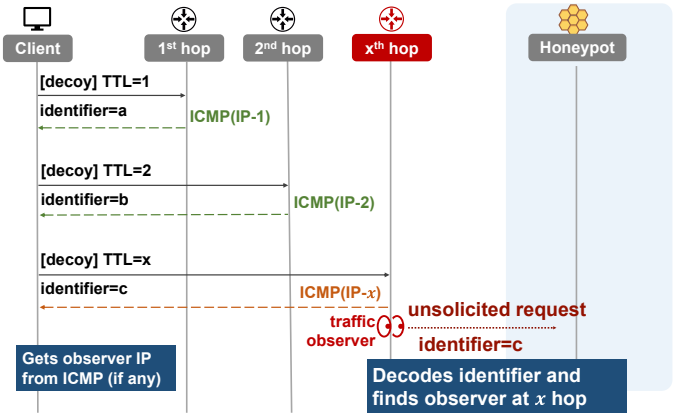

Fig. 2: Traceroute-based method

We compare the requests received by the honeypots with the traffic decoys sent by the client to detect traffic shadowing.

Generate traffic decoys. The traffic decoys we send include three protocols critical to the Internet: DNS, HTTP, and TLS (port 443) [2]. Each protocol's request packet contains a domain name field: qname in DNS queries, hostname in HTTP requests, and SNI in TLS (v1.2) client hello messages. We use TLS 1.2 instead of TLS 1.3 for several reasons: The ECH security extension in TLS 1.3 encrypts the SNI, reducing the likelihood of user privacy leaks. Although TLS 1.3 offers improved security and performance, some older servers have not yet deployed TLS 1.3 [32], and in some countries [1], TLS 1.3 traffic is blocked.

We embed unique identifiers within the domain names. The format of the domain name is:

$$< identifier > .experiment.domain$$

*.*experiment.domain* is the (anonymized) wildcard domain name of our HTTP/HTTPS server and is purchased exclusively for the experiment. According to our design, unsolicited visitors will be directed to our honeypot.

Detect traffic shadowing through analyzing our honeypots' logs. Our honeypot includes the DNS authoritative server and a web honeypot (i.e., an HTTP/HTTPS server built by Nginx), which can log all DNS queries and HTTP/HTTPS requests targeting the experimental domain. The identifier in

[2]In a preliminary experiment, we also tested several other protocols that transmit credentials in plaintext—including FTP, TELNET, and SMTP—but observed no instances of attackers sniffing and reusing user credentials.

the domain name of the requests reveals the initial decoy observed. Unsolicited requests of traffic shadowing may be mixed with the normal logs triggered by our own requests, and the requests caused by other types of traffic manipulation (e.g., DNS interception [40] and transparent HTTP proxy [12]) or some normal mechanisms (e.g., reverse proxy server that forwards HTTP requests based on the hostname field). Considering traffic shadowing's stealthiness and the time required for data flowing from the traffic observer to the unsolicited visitor, unsolicited requests triggered by traffic shadowing usually have a longer delay. So, we set a time threshold to filter these incoming requests. Only requests with a time delay large enough compared to the initial decoy are treated as 'shadowing traffic'. There are nine possible protocol combinations of the initial decoy (DNS, HTTP, TLS) and the unsolicited request (dns, http, https), represented as: [3]

$$\{DNS|HTTP|TLS\} - \{dns|http|https\}$$

Localize traffic observers. To determine the specific hop where the traffic observer is located, we use a traceroute-based method, which is also used by some other measurement studies on traffic manipulation [44, 51], shown in Figure 2. On a path, we send 64 decoy packets with TTL values ranging from 1 to 64, each with a unique identifier. Only decoy traffic with a sufficiently large TTL can reach the traffic observer, potentially triggering unsolicited requests. So, if, along a given path, unsolicited requests are not triggered until the decoy's initial TTL reaches x, then we conclude that the observers are x hops away from the VP. If the device hosting the traffic observer sends an ICMP TTL Exceeded message to us, we can determine the traffic observer's IP address as the source IP of the IP-ICMP packet. Among the three types of traffic decoys we spread, both HTTP and TLS are based on TCP and require a connection to be established. For traceroute, we perform the first two normal TCP handshakes and set the TTL of the third handshake to 1, which carries the payload.

B. Phase I: A Global Measurement from Data Center Networks

Path diversity in spreading traffic decoys (the green arrow from *client* to *traffic destinations* in Figure 1) is essential for accurate and comprehensive traffic shadowing measurements. There are two main approaches in prior works to increase path diversity: sending probe packets from a large number of source addresses [45, 64] or sending probe packets from a few clients to a large number of destination servers [29]. To avoid excessive experimental overhead, there is a trade-off between these two approaches. The previous study [64] adopted the first method. However, its measurement results indicate that measurement results' variations due to different source IPs are significantly smaller than those caused by changing destination IPs. In addition, other prior work [40] also indicates that traffic manipulators may selectively manipulate data destined for specific IP addresses. Therefore, in this paper, we increase the number of destination addresses significantly while using only a small number of source IPs to obtain the maximum measurement results with minimal measurement overhead.

[3]They are the nine types of traffic shadowing we detect. We use uppercase to represent the decoy protocol and lowercase to represent the protocol of unsolicited requests.

TABLE I: VP Diversity

	Phase I			Phase II$^+$
	DNS	HTTP	TLS	
IP	61369	50865	50011	73682
AS	20308	20376	20643	104
CC*	223/**71**	229/**60**	226/**58**	1
Province*	1702	1562	1504	31/**29**

$^+$ The four types of decoys we spread in Phase II are sent together and share the same vantage points.
* The bold numbers represent only countryCodes or provinces with more than 100 IP addresses.

Our method: use a large number of open servers on the Internet as traffic destinations. For cost reasons, we do not set up our own servers as traffic decoy destinations; instead, we leverage publicly accessible servers on the Internet, specifically those with open ports (UDP/53 for DNS, TCP/80 for HTTP, TCP/443 for TLS). Servers with UDP port 53 open are generally open resolvers [39]. Servers with TCP ports 80/443 open are generally web servers or reverse proxy servers. **1.** We exclude the private IPv4 address ranges from the IPv4 address space and then sample 1% of the remaining addresses. **2.** We scan the ports (UDP/53, TCP/80, TCP/443) of the sampled IP addresses, and we retain only those with open ports. **3.** After one month, we perform an additional port scan on these IP addresses to select those with consistently open ports. **4.** To avoid straining a single AS, we preserve at most 5 IP addresses per AS per the protocol of decoy. The final number of destination addresses we select is shown in Table I. The selected destination addresses are distributed relatively evenly across countries, rather than concentrated in certain countries, as detailed in Appendix A.

We send the three types of traffic decoys (DNS, HTTP, TLS) from four VPSs (located in US, UK, AU, ZA) to the destinations mentioned above. With the root privileges of these VPSs, we send IP packets with custom TTL for localization (Figure 2) and embed our decoy in the payload. In the decoy packets, the destination address at the IP layer is the IP address of the open server, but the embedded domain name in the application layer (qname for DNS, host name for HTTP, and server name indication for TLS) is resolved to our own HTTP/HTTPS honeypot, so that the domain name will direct the unsolicited traffic to our own honeypot.

C. Phase II: A China-wide Measurement from Residential Networks

Residential networks are generally more valuable to traffic observers than data center networks, because they provide access to actual user data, while data center networks typically offer only aggregated server data. Therefore, traffic shadowing is more likely to occur in residential networks, and it is valuable to detect it there. Previous work [64] indicates that most traffic observers are in China, so we intend to deploy vantage points within residential networks in China for further measurement. Previous works have various methods to obtain vantage points in residential networks, including inviting vol-

TABLE II: Compared with the previous work, we increase the number of destination IPs and reduce the number of source IPs in probing data center networks, and integrate residential networks in addition.

	IMC24 [64]	This work
Data Center Network Path Diversity		
DNS	4364 × 36 * [16101]	4 × 61369 [89070]
HTTP	4364 × 2325 [28953]	4 × 50865 [199858]
TLS	4364 × 2325 [24298]	4 × 50011 [203311]
Residential Network Path Diversity		
DNS	0	73682 × 1
HTTP	0	73682 × 1
TLS	0	73682 × 1
Analysis		
landscape	◐	●
traffic observer	◐	●
unsolicited visitor	◐	●
data flow	○	●

* Source IP addresses × Destination IP addresses
[...] number of the nodes in the traceroute paths
○ No analysis. ◐ Limited analysis. ● Analyzed.

unteers to run specific software [25], executing measurement scripts in the background of popular software clients [40], utilizing web advertising services [35], and so on. We need to select a method that adheres to ethical standards and ensures adequate vantage points.

Our method: use web advertising service to integrate vantage points from residential networks. We subscribe to a web advertising service provider in China [3]. It allows us to submit JavaScript scripts, which are embedded as web page advertisements on many partner websites. The scripts are automatically executed when users browse these sites, sending our traffic decoys. In the JavaScript script, we do not have root privileges, so we can only use high-level APIs to access HTTP or HTTPS URLs. As a result, we can not send custom IP packets to arbitrary IP addresses just as we did with VPS in Phase I. So, we cannot perform traceroute to localize traffic observers, and we cannot set the IP-layer destination addresses to different IP addresses than those resolved by the domain name in the application layer. Our traffic decoys are all directly sent to our own honeypots (DNS is sent to local recursive resolvers, and HTTP/HTTPS is sent from the client to our web honeypot).

The limitation of the advertising service's capabilities also means we cannot send HTTP or TLS decoys individually: when we access an HTTP or HTTPS URL, the browser automatically performs a DNS query and then makes the web request. To address the challenge of distinguishing among decoy protocols, we design four specialized decoy types.

- **DNS(CName)** The authoritative server resolves a regular domain query to a CName domain, and resolves the CName to a regular IP. The HTTP/HTTPS requests sent by the browser will contain the original domain name. The CName is set as our DNS(CName) decoy.
- **DNS(NoAnswer)** If the authoritative server rejects the DNS query (responding with Servfail or NXDomain), the JavaScript on the browser will not make subsequent HTTP(S) requests. This decoy helps determine whether unsolicited visitors continue to access the user's data, even when they know there is no valid response.
- **DNS+HTTP** Accessing a regular HTTP URL will trigger both a DNS query and an HTTP request, which is our DNS+HTTP decoy. By comparing the unsolicited requests triggered by this decoy with those triggered by DNS(CName) and DNS(NoAnswer) decoys, the effect of the HTTP decoy can be highlighted.
- **DNS+HTTPS** Similar with DNS+HTTP decoy, we access an regular HTTPS URL.

Please note that the protocol combination of traffic shadowing in this phase is 12 (i.e. 4 decoys × 3 unsolicited requests) instead of 9 in Phase I.

D. Limitation

We can't detect traffic shadowing when unsolicited requests fall outside our honeypots (not dns or http/https), or when protocols beyond our decoys are observed. If some unsolicited requests have a time delay shorter than our threshold, they will also be dropped by our method to fully exclude duplicate requests caused by normal mechanisms or other types of traffic manipulation. The design of our domain name, particularly the generation of random subdomain identifiers, may result in some false positives, as security products may treat certain domain names as automatically generated as part of attacks. When analyzing the IP addresses of unsolicited visitors, we use Spamhaus, a widely recognized IP blacklist, to assess their maliciousness. However, this approach may be affected by false positives generated by the IP blacklist.

Phase I limitation. The discrepancy between the domain names in the traffic decoys and the destination IP addresses may lead to the omission of certain shadow traffic, as some shadow traffic might be sent directly to the destination IP rather than to the domain name (i.e., the honeypot server). The source IP addresses of decoys in Phase I are only four, causing a limitation to path diversity. But we have a large number of traffic destinations, and the previous work [64] on traffic shadowing has shown that the difference caused by varying source IPs is much smaller than that caused by varying destination IPs. Our method to localize traffic observers is disturbed by noise (e.g., no ICMP response). In the measurement results section, we will provide the proportion of paths that can accurately localize traffic observers.

Phase II limitation. We select only residential vantage points in China and do not use residential vantage points from other countries. This is because we find in the results of the previous work [64] that most traffic observers are located in China. In addition, due to the unique nature of the Chinese Internet, many previous studies have also focused their measurements on China [40, 70]. Because we can only use high-level JavaScript APIs to access HTTP or HTTPS URLs, we can't use the traceroute-based method to localize traffic observers. We acknowledge that limitations in this aspect may constrain our research on traffic shadowing, but it must be emphasized that Phase II serves as a complement to Phase I rather than a standalone effort. Through our market research, we found that obtaining residential network (instead of data center) VPNs in China is challenging. Our measurements of traffic shadowing conducted via Chinese residential

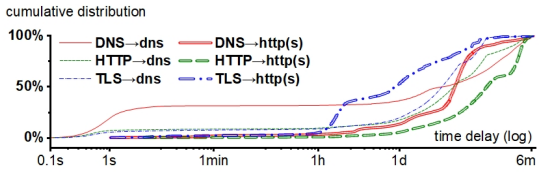

Fig. 3: Cumulative distribution of the time delay between server-received requests and the initial decoy (Phase One)

networks, even if unable to localize traffic observers, can still provide valuable supplementary insights to data center network measurement results.

IV. Phase I

The scale of the experiment is in Table XII. In Phase I, we send traffic decoys from 4 VPSs to open servers on the Internet.

A. Landscape

1) Traffic shadowing affects three protocols of our decoys: 1.7% DNS, 0.6% HTTP, and 0.5% TLS.
2) Traffic decoys may trigger the same protocol of unsolicited requests (1.7% DNS-dns, 0.08% HTTP-http, and 0.11% TLS-https), and different protocols of unsolicited requests (0.4% DNS-http, 0.2% DNS-https, 0.6% HTTP-dns, 0.02% HTTP-https, 0.5% TLS-dns, and 0.04% TLS-http).
3) Web requests triggered by HTTP/TLS decoys tend to be the same protocol as the initial decoy.

To detect traffic shadowing, we need to identify unsolicited requests among the many requests our honeypots receive. These include 1) legitimate requests triggered by the initial decoy, 2) unsolicited requests related to traffic shadowing, and 3) unsolicited requests resulting from other forms of traffic manipulation (e.g., replication of DNS queries [40]). The key characteristic of traffic shadowing is that the arrival time of the unsolicited request is significantly delayed relative to the initial decoy, a delay not attributable to network congestion, timeouts, retransmissions, or server overload. Figure 3 illustrates the cumulative distribution of these time delays. Nearly one third of the dns requests triggered by DNS decoy occur within 1.5s, while most other incoming requests occur after more than one hour. This can be explained by the destination address of our traffic decoy. Recall our traffic decoy destinations: servers with UDP port 53 open and servers with TCP ports 80/443 open. The former are typically open resolvers, which perform recursive resolution upon receiving our DNS decoy (usually within seconds); the latter are generally open HTTP/HTTPS servers or open reverse proxy servers, which do not typically access our honeypot. We set one hour as the threshold to distinguish unsolicited requests from those caused by normal mechanisms or other traffic manipulations, excluding any requests with delays not exceeding one hour. The delay between unsolicited requests and the initial traffic lure ranges from 1 hour to 6 months, indicating that traffic shadowing is a slow, data-exploitation process rather than a software feature with real-time feedback.

TABLE III: The percentage of affected decoys in Phase I

decoy	% of decoys that trigger unsolicited dns	http	https	any
DNS	1.7	0.4	0.2	1.7
HTTP	0.6	0.08	0.02	0.6
TLS	0.5	0.04	0.11	0.5

TABLE IV: Normalized TTL of traffic observers on the path

decoy	Hops* from VP 1-3	4	5	6	7	8	9	10
DNS (%)	0	0	0	0	0	0.01	0.06	99.92
HTTP (%)	0	0.02	0.41	1.6	3.9	10	15	69
TLS (%)	0	0.04	0.62	2.80	9	16	18	54

* Hops are normalized into 1 (source) to 10 (destination).

After selecting unsolicited requests based on the delay, we assess the scale and scope of traffic shadowing. The proportion of DNS decoys affected is approximately 1.7%, higher than that of HTTP and TLS decoys (around 0.5%). From another perspective, we find traffic decoys sent to more than 6.1% DNS destinations, 2.8% HTTP destinations, and 3.9% TLS destinations are affected. For the decoys that are affected, we distinguish the types of unsolicited requests, as presented in Table III. Unsolicited dns requests are more than unsolicited http and https requests. This is because unsolicited http and https requests also require preceding DNS queries. In addition, unsolicited web requests are usually of the same protocol as the initial traffic decoy: HTTP-http $>$ HTTP-https but TLS-http $<$ TLS-https. Unsolicited visitors may infer our honeypot's network infrastructure based on the type of traffic decoy, sending targeted unsolicited probes.

B. Traffic Observers

1) DNS decoys are primarily observed at the destination (99% cases), while HTTP and TLS are observed either by on-path sniffers (31% and 46%) or at the destination (69% and 54%).
2) The traffic sniffing device is often deployed at the entry or exit points of its AS, which suggests that they are likely installed by the administrators of the ASes.

Traffic observation can occur either before or after user traffic reaches its destination. In the former case, the traffic observer is an on-path sniffer recording the passing packets; in the latter case, the traffic observer is a server that records the incoming user data. We use a traceroute-based method (Figure 2) to localize traffic observers, shown in Table IV[4]. 99.92% of the traffic observers against DNS are located at the destination (TTL=10), while a significant proportion of traffic observers targeting HTTP or TLS are distributed in the

[4]Due to noise (e.g., ICMP no reply), only 87% paths can localize observers.

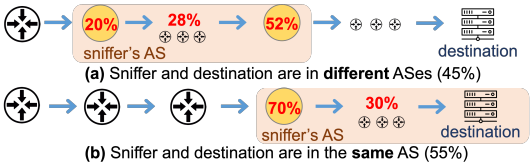

Fig. 4: On-path sniffers distribution in its AS.

TABLE V: Top ten open ports of sniffers

port	sniffers	protocol
TCP/179	234	BGP [53]
TCP/646	20	LDP (distribute MPLS among routers) [8]
TCP/443	19	HTTPS [54]
TCP/80	15	HTTP [24]
TCP/22	14	SSH [71]
TCP/161	9	SNMP (network management) [14]
TCP/23	8	TELNET [48]
TCP/830	7	NETCONF-SSH (network management) [61]
TCP/21	6	FTP [49]
TCP/639	5	MSDP (IP multicast) [23]

TABLE VI: The proportion of open servers affected by traffic shadowing in each country (only the top 5 countries are shown)

DNS		HTTP		TLS	
CN	34%	CN	9.3%	CN	10.1%
RU	11%	CL	1.4%	TH	0.6%
KZ	9%	MD	1.4%	NZ	0.5%
TR	8%	TH	1.1%	ID	0.5%
GR	6.8%	ID	0.9%	BE	0.4%

middle of the path (e.g., TTL=7, 8, 9). This indicates that traffic observers design different data collection methods for different protocols. For DNS, they may prefer to use the well-established passive DNS scheme [20, 18] to collect data rather than deploying the high-cost on-path sniffing devices.

On-path sniffers. Most detected on-path sniffers are located in China (85%) and the US (6.9%). To characteristic traffic sniffing devices, we perform port scans on the on-path sniffers, shown in Table V. Most open ports are related to routing protocols (e.g., bgp and ldp) while some ports are for remote controlling (e.g., ssh and telnet). Figure 4 shows that the sniffing devices tend to be located at the entry hop or exit hop, indicating they are deployed by the administrators of the ASes rather than the users in the ASes. This deployment strategy of sniffer devices ensures that as much incoming and outgoing traffic as possible can be sniffed.

Open servers. Table VI shows the rate of open servers affected by traffic shadowing per country (the impact of on-path sniffers has been eliminated). Among the three protocols, open DNS servers (open resolvers) are particularly affected, which may be attributed to the widespread use of passive DNS. For each protocol, open servers in China are most likely to be affected.

C. Unsolicited Visitors

1) User data is stored for a long time (up to 6 months) and used multiple times.
2) Unsolicited requests come from a large number of ASes (unsolicited dns from 976 ASes, http from 136 ASes, and https from 115 ASes), indicating that traffic shadowing is not an isolated case.
3) The unsolicited requests' content is mainly for reconnaissance, with no obvious evidence of malicious intent, but 42% and 27% of the source IP addresses for unsolicited http and https requests are listed on the Spamhaus IP blocklist [2].

During the experiment, we receive unsolicited requests from over a thousand ASes, shown in Table VII. Unsolicited dns visitors are the most common, distributed across 976 ASes. Unsolicited http and https visitors come from over one hundred ASes. However, most unsolicited requests are concentrated in the top ASes. The top 5 ASes take up 58% unsolicited dns, 90% unsolicited HTTP, and 81% unsolicited https. These ASes are mainly cloud service providers (e.g., AS15169 Google LLC and AS398823 PEG TECH INC) and ISPs (e.g., AS9808 China Mobile and AS4134 Chinanet-backbone). Unsolicited visitors may host their businesses in the cloud or rent ISP's networks, or hide their identities using residential proxies (the residential proxy service usually lies in ISP networks[69]). We match the IP addresses of unsolicited visitors with the Spamhaus IP blocklist [2] (a famous IP blocklist widely used in previous works [64, 6]). Marked source IPs of unsolicited http and https account for 42% and 27%, respectively, suggesting unsolicited visitors may be involved in other Internet scanning activities. Marked source addresses of unsolicited dns are only 0.22%. We further verify the maliciousness of the IP addresses using VirusTotal [4] and find that none of the IP addresses are listed in its blacklist.

Access intensity. Table VIII shows the number of unsolicited requests triggered by each traffic decoy. More than half of the decoys typically trigger two or more unsolicited requests. Nearly 10% of DNS decoys and 30% of HTTP and TLS decoys trigger more than 10 unsolicited requests. As shown in Figure 3 in §IV-A, the time delay between unsolicited requests and the initial decoy is typically more than one hour and can extend to several hours or days. The latest unsolicited request appears six months after the initial decoy was sent. However, the six-month period is just the time span of our data analysis, not the maximum duration for which unsolicited visitors may retain the traffic decoy. The unsolicited visitor keeps the traffic decoy for a long time and repeatedly accesses our honeypot, indicating their eagerness to probe its content and understand how it evolves.

Request content. The dns queries sent to our authoritative server are mainly of type A, with the intent of obtaining the IP address of our honeypot website. The http/https requests sent to our honeypot website contain various paths. Our honeypot website is built using the WordPress framework and Nginx as a reverse proxy. The resources on the website include two categories: those from the WordPress framework (such as some *.js* files) and the web pages and images that we created and inserted ourselves. The path of unsolicited http/https requests includes both of the two categories of resources. After visiting the website's homepage, unsolicited visitors can further explore the site using depth-first or breadth-first algorithms. The requested content is sufficient to reconstruct our honeypot website. Additionally, they enumerate potential pages based on the website's page numbering pattern. For example, we have */p1, /p3, /p4, ..., /p10* on our website, and they will request */p2*. In the unsolicited requests, we find no malicious paths or payloads, such as SQL injection attacks or cross-site scripting (XSS) attacks. Our findings indicate that unsolicited visitors are attempting to scan and probe the contents of our honeypot website as deeply as possible, rather than attempting to execute malicious attacks.

Device characteristics. We analyze the user-agent field of

TABLE VII: IP and AS of unsolicited visitors

	unsolicited dns	unsolicited http	unsolicited https
# of IP	16,887	9,260	23,152
# of AS	976	136	115
top 5 AS	AS15169 Google LLC (25%) AS4134 Chinanet-backbone (17%) AS137798 Digital-Guangdong (8%) AS43832 MSK-IX (5%) AS4812 China Telecom (3%)	AS9808 China Mobile (46%) AS137687 Henan, China (25%) AS396982 Google LLC (10%) AS45102 Alibaba US (6%) AS24139 Huashu media (3%)	AS398823 PEG TECH INC (37%) AS24139 Huashu media (16%) AS396982 Google LLC (11%) AS32097 WholeSale Internet (9%) AS37963 Alibaba, Hangzhou (8%)
spamhaus	0.22%	42%	27%

%	number of unsolicited requests								
	dns			http			https		
decoy	1	2-10	>10	1	2-10	>10	1	2-10	>10
DNS	32	57	11	23	67	10	21	77	1.6
HTTP	9.7	62	29	42	35	24	1.7	94	4.1
HTTPS	12	56	32	20	51	28	8.1	66	26

TABLE VIII: The proportion of decoy triggering different numbers of unsolicited requests.

unsolicited HTTP/HTTPS triggered by **DNS** decoy	
29.5%	Expanse, a Palo Alto Networks company, searches across the global IPv4 ⋯ to identify customers ⋯
26.6%	Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/x.x.x.x Safari/537.36 (65 versions)
24.4%	-
5.6%	fasthttp
2.0%	curl/8.6.0
unsolicited HTTP/HTTPS triggered by **HTTP** decoy	
88.4%	Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/58.0.3029.110 Safari/537.3
7.6%	Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/x.x.x.x Safari/537.36 (80 versions)
1.2%	fasthttp
unsolicited HTTP/HTTPS triggered by **TLS** decoy	
55.1%	Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/x.x.x.x Safari/537.36 (76010 versions)
21.3%	Mozilla/5.0 (Linux; Android 9; ASUS_X00TD; Flow) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/359.0.0.288 Mobile Safari/537.36
7.2%	fasthttp

TABLE IX: User-agent distribution of unsolicited web requests triggered by three different traffic decoys

unsolicited web requests to characterize the device features sending them. The result is shown in Table IX. Five device types make up the bulk of unsolicited web requests:

- **Expanse**. This accounts for 29.5% unsolicited web requests triggered by our DNS decoys. Expanse is a Palo Alto Networks company that searches across the global IPv4 space multiple times per day to identify customers. The unsolicited web requests from this company indicate that it is extracting domain names from our decoys sent to open DNS resolvers.
- **Chrome on Windows 10**. The pattern is *Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/x.x.x.x Safari/537.36*. It appears in unsolicited requests triggered by all three types of decoys. Unsolicited visitors use various Chrome versions to hide their device identity: 65, 80, and 76010, respectively. However, user-agents with two specific Chrome versions account for 88.4% and 21.3% of unsolicited web requests triggered by HTTP/TLS decoys, indicating that two relatively unified organizations are collecting and exploiting the information. The typo (Safari/537.3, should be Safari/537.36) in the first user-agent in unsolicited web requests triggered by HTTP decoys may be a mistake in their scanning software.
- **None**. This account from 24.4% unsolicited web requests triggered by DNS decoys, indicating the purpose of hiding the scanner's identity.
- **fasthttp**. It is a high-performance HTTP client library in Go. Scanners can use it to perform high-speed scanning.
- **curl/8.6.0**. An http web client on Linux.

D. Data Flow Between Traffic Observers and Unsolicited Visitors.

We use "data flow" to describe the observation where our decoy data is observed by devices in one AS but eventually embedded in unsolicited requests from ASes of different organizations. By correlating the decoys and unsolicited requests captured by our honeypot, this section describes the flow and characteristics of sniffed user data, particularly the relations between traffic observers to unsolicited visitors. However, from collected data, we may not determine why user data has been transferred between organizations, or whether they are reuslts of malicious activities. Data flow is also not necessarily the result of privacy leakage: as discussed in [64], one anecdotal presumption of this outcome is data sharing between networks for operational considerations, or sharing between subsidiary organizations due to business choices. As a result, this section does not make claims for the underlying reasons.

1) Data flow is prevalent: User data flows from 130 organizations equipped with on-path sniffers to 26 organizations with unsolicited visitors, and from 892 organizations hosting open servers to 98 organizations with unsolicited visitors.
2) Data flows across organizations: 83% on-path sniffers and 81% open servers are located in organizations distinct from those hosting unsolicited visitors.
3) The organizations from which data flows out are mainly

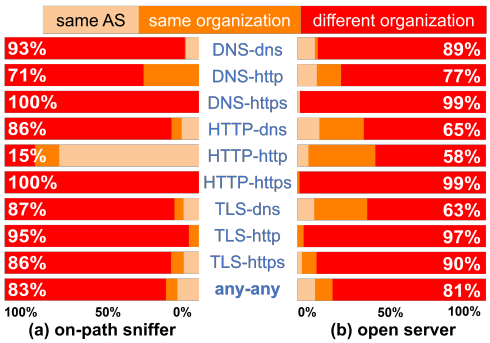

Fig. 5: The relationship between traffic observers (on-path sniffers and open servers) and unsolicited visitors.

ISPs and data centers, while the organizations receiving data are primarily ISPs, cloud service providers, and security companies.

4) Data flows from CN to the largest number of countries, while the US receives data from the largest number of countries.
5) We find various types of data flow, with three detailed cases analyzed.

1) Characteristics of data flow: Analyzing traffic observers and unsolicited visitors separately is insufficient to fully reveal the traffic shadowing process, as it overlooks the migration of user data from observation to exploitation, which largely determines the privacy risks. For the unsolicited requests received by our honeypot, we determine where the initial decoy is observed using the method shown in Figure 2, and analyze the relationships between different unsolicited visitors and various traffic observers.

After traffic observers acquire user data, they may directly exploit it from the data collection devices (on-path sniffing devices and open servers logging requests), sending unsolicited requests. Alternatively, they might aggregate user data from various collection devices and send unsolicited requests through separate systems. They may also share the user data with other organizations. To determine which of the above scenarios is most common, we analyze the relationship [5] between traffic observers and unsolicited visitors, which is shown in Figure 5. 83% on-path sniffers and 81% open servers are in different organizations with unsolicited visitors, indicating user data sharing across organizations is prevalent.

Data flow from on-path sniffers to unsolicited visitors. We find data flow from **130** organizations to **26** organizations. The sources of over 90% of data flow are ISPs, which can access users' Internet traffic easily. The destinations of over 70% of data flow are also ISPs, which may be due to security scans taken by themselves or someone exploiting ISP's network. Approximately 20% of data flow destinations are cloud service providers, indicating that unsolicited visitors are hosting their businesses in the cloud.

Data flow from open servers to unsolicited visitors. We find data flows from **892** organizations to **98** organizations.

[5]We look up the AS and organization of the IP address from a public IP database provider *Ipinfo* [9], which is supported and used by many previous works [41, 58, 43].

TABLE X: The number of source or destination countries for data ingress or egress of each country.

on-path sniffer				open server			
egress		ingress		egress		ingress	
CN	10	US	22	CN	10	US	57
US	4	CN	4	RU	10	BE	16
JP	2	SG	2	US	8	CN	12
SG	1	JP	1	AT	7	DE	4
ZA	1	VN	1	IN	6	SG	4

The sources of data flow include ISPs (38%), data centers (21%), cloud service providers (10%) [6] , and others. The open servers on these networks accept requests from Internet users and may provide logs to other organizations. The destinations of data flow include ISPs (42%), security companies (18%), cloud service providers (17%), and others. The flow of data to security companies may be to support threat intelligence collection.

Furthermore, we analyze data ingress and egress from a national perspective. Table X shows the number of source/destination countries for each country's data ingress/egress. Data observed by on-path sniffers in China flows to the largest number of countries, including the US, UK, SG, and others. Data observed by on-path sniffers in the largest number of countries, including SG, JP, AU, CN, and others, flows to the US. The situation is similar for data observed by open servers: data from CN and RU flows out to the largest number of countries, while the US receives data from the largest number of countries. Please note that the countries involved in data flows in and out are merely macro indicators and do not reflect the actions of the respective governments.

2) Case Study: **a. Flows from *one* on-path sniffer to *one* unsolicited visitor.** We show two examples: user data shared from Telefónica to Oracle and from Gulfnet International to Amazon. Telefónica is a multinational telecommunications company headquartered in Madrid. Gulfnet International is an international IT product supplier headquartered in the UAE. Oracle and Amazon are both cloud service providers. This one-to-one relationship indicates that traffic observers exercise caution when handling user data, sharing it only with a single partner or simply using cloud services to access our honeypot themselves.

b. Flows from *many* on-path sniffers to *one* unsolicited visitor. We found that sniffed data across 56 distinct networks was linked to unsolicited requests originating from a single entity: CenturyLink Communications. CenturyLink Communications is a U.S. telecommunications provider. These 56 networks include, among others: 1) Global Telecom & Technology, a global enterprise communication services provider headquartered in Arlington; 2) Telstra Limited, a global telecommunications and media company headquartered in Australia; 3) Universidade Federal de Santa Catarina, a public university in Brazil; 4) Nerviano Medical Sciences, a biopharmaceutical company based in Italy; 5) Internet Initiative Japan, an internet service provider in Japan. While there is no

[6]The difference between a cloud service provider and a data center is whether they provide VPS rental services to end users or serve solely as enterprise networks for large companies.

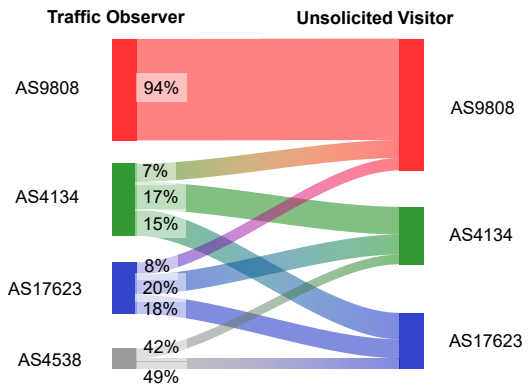

Fig. 6: Data flow case: Several organizations share user data with each other

evidence that CenturyLink Communications actively collects user data from these networks, the aggregation of such user data is nonetheless anomalous.

c. Flows from *many* on-path sniffers and *many* unsolicited visitors. We observe intensive user data sharing among some ASes, shown in Figure 6. Except for AS9808, all other organizations share user data with nearly all others. Except for AS4538, all other organizations utilize user data, sending unsolicited requests to our honeypot. AS4538 is not a commercial entity but rather a network primarily consisting of universities and other educational institutions. As such, it has stricter limitations on outbound traffic.

E. Comparison with the Previous Work [64]

Path Diversity. This paper uses the same decoy-and-honeypot measurement logic of [64], but the variation of the vantage points of this paper is carefully designed to capture traffic shadowing behaviors by their intrinsic characteristics, which is more than only increasing the measurement scale. [64] used VPN probing points to send decoys to a few top sites, which are also located in datacenters. As a result, the paths covered by their work, though substantial in number (see first column in Table II), remain limited because they are linking datacenters where traffic shadowing behaviors are known to occur less often. In this work, by varying destination IPs instead of probing vantages, we effectively increased not only numbers but also types of networks that our decoys are transmitted through (we select public IPv4 servers with ports open, regardless of their popularity), overcoming the limitation of [64]. This design also limits the number of source IPs to avoid excessive measurement overhead on the platform.

Landscape. The previous work and this paper both find nine types of traffic shadowing: $\{DNS|HTTP|TLS\}$ - $\{dns|http|https\}$. This paper shows the proportion of affected decoys: 1.7% DNS, 0.6% HTTP, and 0.5% TLS. The results in [64] are similar for HTTP and TLS: except for the paths with destinations in China, the proportion of HTTP and TLS affected by traffic shadowing is below 1%. The DNS measurement results in [64] cannot be directly compared with those in this paper, as they only measure a few public resolvers and authoritative servers, and the results vary significantly across different DNS servers.

Unsolicited Visitors. This paper and [64] both find that user data may be stored for extended periods and utilized multiple times, with no clear evidence of malicious intent. This paper also analyzes the characteristics of devices that send unsolicited web requests (Table IX).

Traffic observers. This paper and [64] both find DNS observers are mainly at the destination, while a significant proportion of HTTP/TLS observers are in the middle of the path. This paper analyzes the on-path sniffers' deployment strategy and the working characteristics of open servers, but [64] does not. This paper presents detailed detection results of open ports of sniffing devices, whereas [64] only provides brief results.

Data flow. This part is critical for accurately assessing the privacy risks of traffic shadowing, which is analyzed in this paper but not in [64].

V. PHASE II

In phase II, we integrate vantage points in Chinese residential networks. Due to the limitations of the vantage point capabilities, the decoys in this phase are specially designed (§III-C), differing from those in Phase I.

1) Traffic shadowing in Chinese residential networks (over 10%) is more prevalent than the global average measured in data center networks (nearly 1%).
2) User data is stored for extended periods (up to 6 months) and exploited multiple times, yet no obvious malicious actions have been detected.
3) Case study: A Chinese IT company collects user traffic sent from its network and makes unsolicited requests to the domain names in the data, rotating IP addresses to avoid detection.

A. Landscape

We calculate the delay between the requests received by our honeypots and the initial decoy to identify unsolicited requests, which is shown in Figure 7. Requests with a delay of more than 1 hour are considered shadowing traffic, while others are ignored. Similar to Phase I, the traffic decoy may be retained for several months before triggering unsolicited requests. The traffic decoys affected by traffic shadowing are shown in Table XI. More than 10% DNS(CName) decoys trigger unsolicited dns queries, while a few proportion also trigger unsolicited http/https requests. In contrast, although the DNS(NoAnswer) is also a purely DNS decoy, only 6% trigger unsolicited DNS queries. This may be because some traffic observers, upon detecting that the initial decoy is rejected by the authoritative server, discard these decoys. The DNS+HTTP and DNS+TLS decoys each contain two protocols, as we cannot send only web requests without performing a DNS query, using the advertising service. Unsolicited requests triggered by pure HTTP and TLS can be inferred by comparing these two decoys with the DNS(CName) decoy: subtract the first row from the third and fourth rows. The unsolicited dns queries they trigger are similar to, or fewer than, those triggered by DNS. But the unsolicited http/https requests they

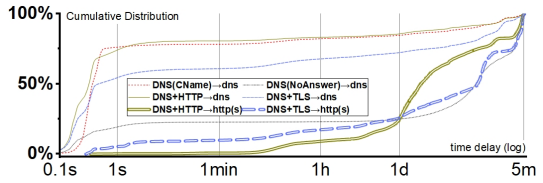

Fig. 7: Cumulative distribution of the time delay between server-received requests and the initial decoy (Phase II)

TABLE XI: The percentage of affected decoys in Phase II.

decoy	% of decoys that trigger unsolicited dns	http	https	any
DNS(CName)	12	0.3	0.1	12
DNS(NoAnswer)	6	0	0	6
DNS+HTTP	12	12	2	12
DNS+TLS	11	3	8	11

trigger are significantly more than those triggered by DNS. A phenomenon similar to Phase I is that unsolicited visitors tend to access our honeypots using the same protocol as the decoy: DNS+HTTP triggers more http than https while DNS+TLS triggers more https than http. By comparing Table XI with the parallel results in Phase I (Table III), it is evident that traffic shadowing is more prevalent in Chinese residential networks than the global average — almost all types of traffic shadowing are at least one order of magnitude higher.

B. Characteristics

Due to the limitations of the advertising service (only high-level HTTP-API and HTTPS-API calls are possible, while IP-layer TTL cannot be adjusted), we are unable to implement the traceroute-based method (Figure 2) to identify traffic observers, so we only analyse unsolicited visitors here. **Access intensity.** Figure 7 has shown the distribution of time delay between incoming requests and initial decoys. Unsolicited requests may occur hours or even days after the initial decoy is sent, suggesting that user data is stored for an extended period. Long-term storage of user data implies multiple exploits: when analyzing the number of unsolicited requests triggered by a single decoy, we find that most decoys trigger more than one request, with approximately 35% of traffic decoys triggering more than ten. **Request content.** We analyze the content of unsolicited requests and, similar to Phase I, find no malicious access attempts; instead, we observe a large number of requests aimed at scanning and probing. Unsolicited visitors crawl our honeypot website and perform path enumeration. We find that 19% and 17% of source IP addresses of unsolicited http and https requests are flagged as malicious by Spamhaus IP blocklist [2], indicating that they may be associated with other Internet scanning activities.

C. Case Study

We select a special case to illustrate the behavior pattern of unsolicited visitors. In Phase II, our vantage points are distributed across 104 ASes in China. Among them, we find that a significant proportion of traffic decoys originating from *AS55960 Beijing Guanghuan Xinwang* trigger unsolicited

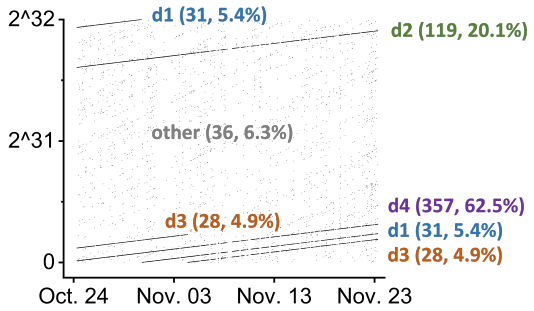

Fig. 8: The TCP Timestamp Value against the real timestamp of unsolicited requests from AS55960. The lines in the figure form from the densely distributed data points.

requests. That is 86% DNS(CName), 82% DNS(NoAnswer), 83% DNS+HTTP and 86% DNS+TLS. This is significantly higher than the typical level seen in residential networks in China (Table XI). The triggered unsolicited requests come from 571 different IPs in AS55960 itself. This phenomenon suggests that the operator of this network employs an aggressive strategy to observe and exploit user data.

Inspired by another study on traffic manipulation devices in China [5], we cluster the unsolicited visitors' devices by analyzing the TCP timestamp fingerprints of unsolicited http/https requests. Timestamp value is a 32-bit integer field in the TCP header, representing a relative timestamp. It helps both sides of the TCP communication calculate round-trip delay and optimize transmission parameters. This value increases linearly with the actual timestamp, and both its growth rate and initial value depend on the device, which can be used as a distinct fingerprint of the device — TCP timestamp fingerprint. We plot the values of timestamp value from TCP connections originating from AS55960 against the actual timestamps on a scatter plot: Figure 8. For each TCP connection's handshake request from each source IP address, we extract the timestamp value and combine it with the actual timestamp to form a data point. Then, we plot all the data points on the same graph. We use the least squares method to fit the four lines formed by the data points, resulting in four timestamp value fingerprints (d1, d2, d3, d4). For each IP address, we again use the least squares method to fit its TCP timestamp value, then classify it into one of the four lines based on the slope and intercept of the fitted line, or mark it as not belonging to any line. Although the unsolicited requests come from 571 different IP addresses, we only identify four distinct TCP timestamp fingerprints (i.e. the four straight lines in the figure): d1, d2, d3, d4. Each fingerprint represents a distinct device or a virtual machine with an independent TCP stack. Most IP addresses (357, 62.5%) belong to the fourth device — d4, then d2 (119, 20.1%), d1 (31, 5.4%), and d3 (28, 4.9%). Only 6.3% IP addresses are not hosted in the 4 devices. These experimental results indicate that the number of devices actually used by AS55960 to send unsolicited requests may be only around 4, much smaller than the number of IP addresses it uses. This phenomenon is similar to the results of the previous study on Chinese traffic manipulation [5], and the underlying reason may be that unsolicited visitors rotate their IP addresses to avoid being blocked. This suggests that they may have realized

that the unsolicited requests they are sending are unwelcome, and are attempting to bypass IP address-based restrictions of the server.

The *AS55960* is owned by the Chinese IT company *Guanghuan Xinwang*. The company provides services such as cloud computing, network access, and IT consulting. Headquartered in Beijing, it serves numerous large tech companies across China. Although our findings show that the company implements traffic shadowing on user data, their intentions may be benign, including detecting malicious traffic, monitoring internal employees, and so on.

VI. Ethics

In this paper, we use a web advertising service provider to obtain vantage points in residential networks, which may raise ethical concerns. **We communicate the explicit purpose of our study to the advertising provider, and they provide a banner that allows users to opt out of our study.** The ads we publish are blank pages containing only JavaScript code. The ads we submit have been reviewed by the web advertising service provider to ensure they do not harm users viewing them or the websites displaying them. The links accessed by our JavaScript point to the domain of our honeypot website, which was specifically registered for this experiment. It has not been flagged as a malicious domain by any security companies. Our honeypot website is hosted on a newly rented cloud server, and its IP address has not been listed as malicious on any IP blacklists. Therefore, the traffic generated by our JavaScript code in users' browsers is unlikely to be identified as malicious by any security software, and it should not cause any issues for users. **There is a detailed censorship system in China, but we do not cause any trouble to the users who watch our advertisements**: Our website is specifically established for experimental purposes and does not contain any topics such as pornography, violence, politics, or news that might draw particular attention from internet censors.

Because our school lacks an IRB, we follow best practices such as [10, 38, 45, 50] in network measurement to minimize the negative impact on others when traffic decoys are deployed. When setting the destinations for traffic decoys, we sample IP addresses at 1% and limit each AS to no more than 5 IP addresses to avoid overloading a single network. The decoys we send use common internet protocols—DNS, HTTP, and TLS. The devices sending the decoys and our honeypot are physical or cloud servers that we own or rent. Our actions comply with the terms of service of the cloud provider that rents us the VPSs. Although the traffic decoys we send to open servers differ from regular user access requests, we limit the rate of traffic decoy delivery. We publish a webpage on our website explaining the purpose of this experiment and providing our contact information so stakeholders can opt out. Throughout the experiment, we did not receive any complaints. We purchased and own the domain names used in the experiment, which resolve to our honeypots, so unsolicited requests will not be directed to others' servers or bother them.

VII. Discussion

We perform a large-scale measurement of traffic shadowing affecting DNS, HTTP, and TLS. User's traffic is observed by on-path sniffing devices or at the destination servers. Then, they flow to unsolicited visitors, are stored for a long time, and are utilized multiple times. The unsolicited requests exhibit no obvious malicious intention but may come from potentially abusive networks. More alarmingly, we observe data migration between large numbers of organizations, some part of which may be due to the selling of user data or the extensive and aggressive collection of user privacy.

The traffic shadowing studied in this paper relates to the pervasive monitoring addressed in RFC 7258 [21], which established the consensus that pervasive monitoring is a technical attack. So, although we haven't found any evidence of malicious intent, this phenomenon—an unauthorized manipulation of user privacy—should be considered and addressed by those who care about privacy.

In the years following RFC 7258 [21], many research studies [36, 22, 46] have mentioned the privacy leakage risks of network traffic and network protocols, and the technology community has proposed many **encryption protocols** and **oblivious solutions**, which are two effective ways to avoid traffic observation. Encryption protocols combat on-path sniffers by implementing end-to-end encryption, such as TLS 1.3 [55], DNS over TLS [34], DNS over HTTPS [31]. Oblivious solutions, such as OHTTP [57] and ODoH [37], prevent service providers from simultaneously accessing both the message content and the client's visibility, thereby preventing certain open resources (e.g., open resolvers) from acquiring user privacy. However, many people lack the motivation to promote the deployment of them because they are unclear about the answers to these two questions: 1. *How prevalent is on-path sniffing?* and 2. *To what extent would user privacy leak on the server side?* This paper provides empirical results through large-scale measurements, reminding people of the importance of encryption protocols and oblivious solutions.

Another method to mitigate the traffic shadowing phenomenon is to impose restrictions on traffic observers and unsolicited visitors. Note that most of on-path sniffers we find in our experiment are in ISP networks. We believe that ISPs should understand the risks of traffic shadowing and implement detection mechanisms to identify unknown traffic shadowing entities within their networks. For cases where open servers collect and share user information with other organizations, security organizations should regularly measure and report their practices to the public. Unsolicited visitors mainly originate from ISPs, cloud service providers, and data center networks. The administrators of these networks should monitor outbound request traffic and block illegal traffic.

VIII. Data Sharing and Reproducibility

To help readers reproduce our experiments, we provide a comprehensive table summarizing all measurement parameters, shown in Table XII. We also make all of the raw data available. Due to the large total volume, we share a one-day sample of data (collected on October 30, 2024) via a

TABLE XII: Experiment parameters for results' replication

	phase I	**phase II**
start	Oct.12, 2024	Oct.24, 2024
end	Apr.12, 2025	Mar.24, 2025
period	6 months	5 months
decoy source	4 VPS (US, UK, AU, ZA)	73,682 residential IP
decoy destination	open IPv4 scanned DNS: 61,369 HTTP: 50,865 TLS: 50,011	DNS: local resolvers HTTP: 1 honeypot in US TLS: 1 honeypot in US
# of decoys	DNS: 81,175,531 HTTP: 115,846,499 TLS: 112,767,450	DNS(CName): 2,793,299 DNS(NoAnswer): 2,557,428 DNS+HTTP: 1,770,160 DNS+HTTPS: 2,023,139

publicly accessible (read-only) MongoDB database: mongodb://ton_paper_reader:123456@tondatashare.queryrecord.com:29879/?authSource=TrafficShadowing_one_day. The complete dataset is available upon request by emailing the authors.

IX. Conclusion

In this paper, we perform a global measurement of traffic shadowing, a covert but less-studied form of traffic manipulation. We use the decoy-based method to detect traffic shadowing, and make 2 improvements on the previous work. We show the landscape and analyze the traffic observers, unsolicited visitors, and especially the data flow between them, and discuss the possible underlying causes and countermeasures.

Appendix A
Vantage Points Distribution

In Phase I, we send traffic decoys from 4 VPS to 1.6×10^5 open servers on the Internet. All open servers are in data center networks. Open servers per country are shown in Table XIII. DNS, HTTP, and TLS open servers are distributed across 20,267 ASes, 20376 ASes, and 20643 ASes, respectively, with at most 5 IP addresses per protocol per AS. In Phase II, using an AD service, we send traffic decoys from 73,682 residential IP addresses in China to our own honeypot in a data center network in the US. The residential IP addresses per province are shown in Table XIV. The residential IP addresses per AS are shown in Table XV.

TABLE XIII: Open servers per country or region in phase I; only the top 30 are shown. We treat each unique IP as one unique server, and each service per server is counted separately.

DNS		HTTP		TLS	
US	7318	US	10264	US	11293
RU	6817	RU	3509	RU	3106
BR	6279	BR	2861	DE	2390
ID	4260	DE	2090	BR	2106
UA	2368	IN	1632	GB	1732
IN	2093	GB	1474	NL	1366
PL	1962	NL	1273	IN	1283
BD	1957	PL	1216	CA	1251
AR	1121	CA	1132	FR	1245
GB	1083	IT	1113	PL	1119
DE	981	FR	1102	IT	1114
IT	942	ID	1094	ID	1092
CA	908	CN	1056	CN	1047
AU	781	UA	947	JP	910
JP	780	JP	928	AU	896
CN	779	KR	859	HK	783
FR	778	TR	795	TR	774
BG	768	HK	792	UA	706
TR	705	AU	779	KR	698
HK	703	ES	767	ES	689
CZ	697	BG	693	CH	586
IR	690	CZ	627	CZ	567
ZA	689	ZA	579	SE	544
KR	680	AR	554	ZA	514
ES	674	IR	525	AT	508
NL	657	CH	523	IR	500
MX	545	SE	496	RO	437
PK	438	RO	489	TH	437
CO	413	VN	447	AR	425
SE	408	TH	439	VN	399

TABLE XIV: Vantage points per province in Phase II, only top 20 are shown.

Province	VP num	Province	VP num
Guangdong	12646	Zhejiang	5555
Shandong	5497	Jiangsu	5349
Beijing	4691	Henan	4322
Hebei	4048	Fujian	3057
Shanghai	2957	Sichuan	2731
Liaoning	2625	Hunan	2373
Hubei	2067	Anhui	1941
Shaanxi	1682	Shanxi	1543
Guangxi	1444	Chongqing	1302
Tianjin	1267	Heilongjiang	1139

TABLE XV: Vantage points per AS in Phase II, only top 10 are shown.

AS	VP num	AS	VP num
AS4134	24152	AS4837	15073
AS9808	11370	AS56041	2241
AS24444	2130	AS24445	2028
AS56046	1991	AS56040	1695
AS24547	1062	AS4812	919

References

[1] 2020. *Internet Society: Blocking TLS 1.3 in China Makes the Internet Less Secure*.

[2] 2024. Spamhaus. https://www.spamhaus.org/.

[3] 2025. *YaoFaGuangGao*. https://www.zsj18.com/

[4] 2026. VirusTotal. https://www.virustotal.com/gui/home/upload.

[5] Alice, Bob, Carol, Jan Beznazwy, and Amir Houmansadr. 2020. How China Detects and Blocks Shadowsocks. In *Internet Measurement Conference*. ACM, 111–124.

[6] Sumayah A. Alrwais, Xiaojing Liao, Xianghang Mi, Peng Wang, Xiaofeng Wang, Feng Qian, Raheem A.

Beyah, and Damon McCoy. 2017. Under the Shadow of Sunshine: Understanding and Detecting Bulletproof Hosting on Legitimate Service Provider Networks. In *2017 IEEE Symposium on Security and Privacy*. IEEE Computer Society, 805–823.

[7] Abderrahmen Amich, Birhanu Eshete, Vinod Yegneswaran, and Nguyen Phong Hoang. 2023. DeResistor: Toward Detection-Resistant Probing for Evasion of Internet Censorship. In *32nd USENIX Security Symposium*, Joseph A. Calandrino and Carmela Troncoso (Eds.). USENIX Association, 2617–2633.

[8] L Andersson, I Minei, and B Thomas. 2007. RFC 5036: LDP Specification.

[9] IP Geolocation API. 2024. Ipinfo. https://ipinfo.io/.

[10] Michael D. Bailey, David Dittrich, Erin Kenneally, and Douglas Maughan. 2012. The Menlo Report. *IEEE Secur. Priv.* 10, 2 (2012), 71–75.

[11] Abhishek Bhaskar and Paul Pearce. 2022. Many Roads Lead To Rome: How Packet Headers Influence DNS Censorship Measurement. In *31st USENIX Security Symposium*, Kevin R. B. Butler and Kurt Thomas (Eds.). USENIX Association, 449–464.

[12] Rui Bian, Lin Jin, Shuai Hao, Haining Wang, and Chase Cotton. 2024. Silent Observers Make a Difference: A Large-scale Analysis of Transparent Proxies on the Internet. In *Proceedings of the IEEE Conference on Computer Communications (INFOCOM)*.

[13] Kevin Bock, George Hughey, Xiao Qiang, and Dave Levin. 2019. Geneva: Evolving Censorship Evasion Strategies. In *ACM SIGSAC Conference on Computer and Communications Security*, Lorenzo Cavallaro, Johannes Kinder, XiaoFeng Wang, and Jonathan Katz (Eds.). ACM, 2199–2214.

[14] J Case, K McCloghrie, M Rose, and S Waldbusser. 1996. RFC1901: Introduction to Community-based SNMPv2.

[15] Taejoong Chung, David R. Choffnes, and Alan Mislove. 2016. Tunneling for Transparency: A Large-Scale Analysis of End-to-End Violations in the Internet. In *Internet Measurement Conference*, Phillipa Gill, John S. Heidemann, John W. Byers, and Ramesh Govindan (Eds.). ACM, 199–213.

[16] Ryan Craven, Robert Beverly, and Mark Allman. 2014. A middlebox-cooperative TCP for a non end-to-end internet. In *ACM SIGCOMM 2014 Conference*, Fabián E. Bustamante, Y. Charlie Hu, Arvind Krishnamurthy, and Sylvia Ratnasamy (Eds.). ACM, 151–162.

[17] Gregory Detal, Benjamin Hesmans, Olivier Bonaventure, Yves Vanaubel, and Benoit Donnet. 2013. Revealing middlebox interference with tracebox. In *Internet Measurement Conference*, Konstantina Papagiannaki, P. Krishna Gummadi, and Craig Partridge (Eds.). ACM, 1–8.

[18] domain tools. 2024. Domain Tools. https://www.domaintools.com/products/farsight-dnsdb/.

[19] Korian Edeline and Benoit Donnet. 2017. A First Look at the Prevalence and Persistence of Middleboxes in the Wild. In *29th International Teletraffic Congress*, Raffaele Bolla and Florin Ciucu (Eds.). IEEE, 161–168.

[20] efficient IP. 2024. What is Passive DNS. https://efficientip.com/glossary/passive-dns/.

[21] Stephen Farrell and Hannes Tschofenig. 2014. RFC 7258: Pervasive Monitoring Is an Attack.

[22] Yebo Feng, Jun Li, Jelena Mirkovic, Cong Wu, Chong Wang, Hao Ren, Jiahua Xu, and Yang Liu. 2025. Unmasking the internet: A survey of fine-grained network traffic analysis. *IEEE Communications Surveys & Tutorials* (2025).

[23] B Fenner and D Meyer. 2003. RFC3618: Multicast Source Discovery Protocol (MSDP).

[24] Roy Fielding, Jim Gettys, Jeffrey Mogul, Henrik Frystyk, Larry Masinter, Paul Leach, and Tim Berners-Lee. 1999. RFC2616: Hypertext Transfer Protocol–HTTP/1.1.

[25] Arturo Filastò and Jacob Appelbaum. 2012. OONI: Open Observatory of Network Interference. In *2nd USENIX Workshop on Free and Open Communications on the Internet*, Roger Dingledine and Joss Wright (Eds.). USENIX Association.

[26] Michael Harrity, Kevin Bock, Frederick Sell, and Dave Levin. 2022. GET /out: Automated Discovery of Application-Layer Censorship Evasion Strategies. In *31st USENIX Security Symposium*, Kevin R. B. Butler and Kurt Thomas (Eds.). USENIX Association, 465–483.

[27] Benjamin Hesmans, Fabien Duchene, Christoph Paasch, Gregory Detal, and Olivier Bonaventure. 2013. Are TCP extensions middlebox-proof?. In *Proceedings of the 2013 workshop on Hot topics in middleboxes and network function virtualization*, Felipe Huici and Vyas Sekar (Eds.). ACM, 37–42.

[28] Fahad Hilal and Oliver Gasser. 2023. Yarrpbox: Detecting Middleboxes at Internet-Scale. *PACMNET* 1, CoNEXT1 (2023), 4:1–4:23.

[29] Fahad Hilal and Oliver Gasser. 2023. Yarrpbox: Detecting Middleboxes at Internet-Scale. *PACMNET* 1, CoNEXT1 (2023), 4:1–4:23.

[30] Nguyen Phong Hoang, Arian Akhavan Niaki, Jakub Dalek, Jeffrey Knockel, Pellaeon Lin, Bill Marczak, Masashi Crete-Nishihata, Phillipa Gill, and Michalis Polychronakis. 2021. How Great is the Great Firewall? Measuring China's DNS Censorship. In *30th USENIX Security Symposium*, Michael D. Bailey and Rachel Greenstadt (Eds.). USENIX Association, 3381–3398.

[31] Paul Hoffman and Patrick McManus. 2018. Rfc 8484: Dns queries over https (doh).

[32] Ralph Holz, Johannes Amann, Olivier Mehani, Matthias Wachs, and Mohamed Ali Kaafar. 2019. The era of TLS 1.3: Measuring deployment and use with active and passive methods. *arXiv preprint* 1907.12762 (2019).

[33] Michio Honda, Yoshifumi Nishida, Costin Raiciu, Adam Greenhalgh, Mark Handley, and Hideyuki Tokuda. 2011. Is it still possible to extend TCP?. In *Internet Measurement Conference*, Patrick Thiran and Walter Willinger (Eds.). ACM, 181–194.

[34] Zi Hu, Liang Zhu, John Heidemann, Allison Mankin, Duane Wessels, and Paul Hoffman. 2016. RFC 7858: Specification for DNS over transport layer security (TLS).

[35] Geoff Huston. 2016. DNS Zombies. https://blog.apnic.net/2016/04/04/dns-zombies/.

[36] Aminollah Khormali, Jeman Park, Hisham Alasmary, Afsah Anwar, Muhammad Saad, and David Mohaisen. 2021. Domain name system security and privacy: A contemporary survey. *Computer Networks* 185 (2021), 107699.

[37] Eric Kinnear, Patrick McManus, Tommy Pauly, Tanya Verma, and Christopher A Wood. 2022. RFC 9230: Oblivious DNS over HTTPS.

[38] Tadayoshi Kohno, Yasemin Acar, and Wulf Loh. 2023. Ethical Frameworks and Computer Security Trolley Problems: Foundations for Conversations. *arXiv preprint arXiv:2302.14326* (2023).

[39] Marc Kührer, Thomas Hupperich, Jonas Bushart, Christian Rossow, and Thorsten Holz. 2015. Going Wild: Large-Scale Classification of Open DNS Resolvers. In *Internet Measurement Conference*, Kenjiro Cho, Kensuke Fukuda, Vivek S. Pai, and Neil Spring (Eds.). ACM, 355–368.

[40] Baojun Liu, Chaoyi Lu, Haixin Duan, Ying Liu, Zhou Li, Shuang Hao, and Min Yang. 2018. Who Is Answering My Queries: Understanding and Characterizing Interception of the DNS Resolution Path. In *27th USENIX Security Symposium*, William Enck and Adrienne Porter Felt (Eds.). 1113–1128.

[41] Ioana Livadariu, Kevin Vermeulen, Maxime Mouchet, and Vasilis Giotsas. 2024. Geofeeds: Revolutionizing IP Geolocation or Illusionary Promises? *Proc. ACM Netw.* 2, CoNEXT3 (2024), 1–21.

[42] Alberto Medina, Mark Allman, and Sally Floyd. 2004. Measuring interactions between transport protocols and middleboxes. In *Internet Measurement Conference*, Alfio Lombardo and James F. Kurose (Eds.). ACM, 336–341.

[43] Oliver Michel, Satadal Sengupta, Hyojoon Kim, Ravi Netravali, and Jennifer Rexford. 2022. Enabling passive measurement of zoom performance in production networks. In *Internet Measurement Conference*, Chadi Barakat, Cristel Pelsser, Theophilus A. Benson, and David R. Choffnes (Eds.). ACM, 244–260.

[44] Ariana Mirian, Alisha Ukani, Ian D. Foster, Gautam Akiwate, Taner Halicioglu, Cynthia T. Moore, Alex C. Snoeren, Geoffrey M. Voelker, and Stefan Savage. 2023. In the Line of Fire: Risks of DPI-triggered Data Collection. In *Cyber Security Experimentation and Test Workshop*. ACM, 57–63.

[45] Arian Akhavan Niaki, Shinyoung Cho, Zachary Weinberg, Nguyen Phong Hoang, Abbas Razaghpanah, Nicolas Christin, and Phillipa Gill. 2020. ICLab: A Global, Longitudinal Internet Censorship Measurement Platform. In *IEEE Symposium on Security and Privacy*. IEEE, 135–151.

[46] Pavlos Papadopoulos, Nikolaos Pitropakis, William J Buchanan, Owen Lo, and Sokratis Katsikas. 2020. Privacy-preserving passive dns. *Computers* 9, 3 (2020), 64.

[47] Paul Pearce, Ben Jones, Frank Li, Roya Ensafi, Nick Feamster, Nicholas Weaver, and Vern Paxson. 2017. Global Measurement of DNS Manipulation. In *26th USENIX Security Symposium*, Engin Kirda and Thomas Ristenpart (Eds.). USENIX Association, 307–323.

[48] Jon Postel and JK Reynolds. 1983. RFC0854: Telnet Protocol Specification.

[49] Jon Postel and JK Reynolds. 1985. RFC0959: File transfer protocol.

[50] Ram Sundara Raman, Prerana Shenoy, Katharina Kohls, and Roya Ensafi. 2020. Censored Planet: An Internet-wide, Longitudinal Censorship Observatory. In *ACM SIGSAC Conference on Computer and Communications Security*, Jay Ligatti, Xinming Ou, Jonathan Katz, and Giovanni Vigna (Eds.). 49–66.

[51] Ram Sundara Raman, Mona Wang, Jakub Dalek, Jonathan R. Mayer, and Roya Ensafi. 2022. Network measurement methods for locating and examining censorship devices. In *Proceedings of the 18th International Conference on emerging Networking EXperiments and Technologies*, Giuseppe Bianchi and Alessandro Mei (Eds.). ACM, 18–34.

[52] 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 *Internet Measurement Conference*, Dave Levin, Alan Mislove, Johanna Amann, and Matthew Luckie (Eds.). ACM, 390–397.

[53] Yakov Rekhter, Tony Li, and Susan Hares. 2006. RFC 4271: A border gateway protocol 4 (BGP-4).

[54] Eric Rescorla. 2000. Rfc2818: Http over tls.

[55] Eric Rescorla. 2018. RFC 8446: The Transport Layer Security (TLS) Protocol Version 1.3.

[56] Valentin Thirion, Korian Edeline, and Benoit Donnet. 2015. Tracking Middleboxes in the Mobile World with TraceboxAndroid. In *Traffic Monitoring and Analysis - 7th International Workshop (Lecture Notes in Computer Science, Vol. 9053)*, Moritz Steiner, Pere Barlet-Ros, and Olivier Bonaventure (Eds.). Springer, 79–91.

[57] Martin Thomson and Christopher A. Wood. 2022. *Oblivious HTTP draft-thomson-http-oblivious-02*. Technical Report. Internet draft.[Online]. Available:https://datatracker.ietf.org/doc/draft-thomson-http-oblivious/.

[58] Elisa Tsai, Ram Sundara Raman, Atul Prakash, and Roya Ensafi. 2024. Modeling and Detecting Internet Censorship Events. In *31st Annual Network and Distributed System Security Symposium*. The Internet Society.

[59] Giorgos Tsirantonakis, Panagiotis Ilia, Sotiris Ioannidis, Elias Athanasopoulos, and Michalis Polychronakis. 2018. A Large-scale Analysis of Content Modification by Open HTTP Proxies. In *25th Annual Network and Distributed System Security Symposium*. The Internet Society.

[60] Zhaoguang Wang, Zhiyun Qian, Qiang Xu, Zhuoqing Morley Mao, and Ming Zhang. [n. d.]. An untold story of middleboxes in cellular networks. In *Proceedings of the ACM SIGCOMM 2011 Conference on Applications Technologies, Architectures, and Protocols for Computer Communications*, Srinivasan Keshav, Jörg Liebeherr, John W. Byers, and Jeffrey C. Mogul (Eds.). ACM, 374–385.

[61] M Wasserman and T Goddard. 2006. RFC 4742: Using the NETCONF Configuration Protocol over Secure SHell (SSH).
[62] Mingkui Wei. 2021. Domain Shadowing: Leveraging Content Delivery Networks for Robust Blocking-Resistant Communications. In *30th USENIX Security Symposium*, Michael D. Bailey and Rachel Greenstadt (Eds.). USENIX Association, 3327–3343.
[63] Mingshi Wu, Jackson Sippe, Danesh Sivakumar, Jack Burg, Peter Anderson, Xiaokang Wang, Kevin Bock, Amir Houmansadr, Dave Levin, and Eric Wustrow. 2023. How the Great Firewall of China Detects and Blocks Fully Encrypted Traffic. In *32nd USENIX Security Symposium*, Joseph A. Calandrino and Carmela Troncoso (Eds.). USENIX Association, 2653–2670.
[64] Yunpeng Xing, Chaoyi Lu, Baojun Liu, Haixin Duan, Junzhe Sun, and Zhou Li. 2024. Yesterday Once More: Global Measurement of Internet Traffic Shadowing Behaviors. In *Internet Measurement Conference*. 230–240.
[65] Diwen Xue, Michalis Kallitsis, Amir Houmansadr, and Roya Ensafi. 2024. Fingerprinting Obfuscated Proxy Traffic with Encapsulated TLS Handshakes. In *33st USENIX Security Symposium*. USENIX Association.
[66] Diwen Xue, Benjamin Mixon-Baca, ValdikSS, Anna Ablove, Beau Kujath, Jedidiah R. Crandall, and Roya Ensafi. 2022. TSPU: Russia's decentralized censorship system. In *Internet Measurement Conference*, Chadi Barakat, Cristel Pelsser, Theophilus A. Benson, and David R. Choffnes (Eds.). ACM, 179–194.
[67] Diwen Xue, Reethika Ramesh, Arham Jain, Michalis Kallitsis, J. Alex Halderman, Jedidiah R. Crandall, and Roya Ensafi. 2022. OpenVPN is Open to VPN Fingerprinting. In *31st USENIX Security Symposium*, Kevin R. B. Butler and Kurt Thomas (Eds.). USENIX Association, 483–500.
[68] Diwen Xue, Reethika Ramesh, Valdik S. S, Leonid Evdokimov, Andrey Viktorov, Arham Jain, Eric Wustrow, Simone Basso, and Roya Ensafi. 2021. Throttling Twitter: an emerging censorship technique in Russia. In *Internet Measurement Conference*, Dave Levin, Alan Mislove, Johanna Amann, and Matthew Luckie (Eds.). ACM, 435–443.
[69] Mingshuo Yang, Yunnan Yu, Xianghang Mi, Shujun Tang, Shanqing Guo, Yilin Li, Xiaofeng Zheng, and Haixin Duan. [n. d.]. An extensive study of residential proxies in China. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*. 3049–3062.
[70] Mingshuo Yang, Yunnan Yu, Xianghang Mi, Shujun Tang, Shanqing Guo, Yilin Li, Xiaofeng Zheng, and Haixin Duan. 2022. An Extensive Study of Residential Proxies in China. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, Heng Yin, Angelos Stavrou, Cas Cremers, and Elaine Shi (Eds.). ACM, 3049–3062.
[71] Tatu Ylonen. 2006. RFC 4251: The secure shell (SSH) protocol architecture.
[72] Mingming Zhang, Baojun Liu, Chaoyi Lu, Jia Zhang, Shuang Hao, and Hai-Xin Duan. 2018. Measuring Privacy Threats in China-Wide Mobile Networks. In *8th USENIX Workshop on Free and Open Communications on the Internet*, Lex Gill and Rob Jansen (Eds.).

Yunpeng Xing received the BE degree from Beihang University, in 2023. He is currently working toward the master's degree with the Institute for Network Sciences and Cyberspace, Tsinghua University, China. His research interests include network Security and Internet Measurement.

Chaoyi Lu received the PhD degree from Tsinghua University, in 2022. He is currently an associate research fellow with Zhongguancun Laboratory. His research interests include network security and Internet measurement. He is a recipient of the IRTF Applied Networking Research Prize (ANRP) and a current member of the Security and Stability Advisory Committee (SSAC) of ICANN.

Baojun Liu (Member, IEEE) received the PhD degree from Tsinghua University, in 2020. He is an associate professor with Tsinghua University. He has been a visiting research scholar with the International Computer Science Institute (ICSI), UC Berkeley. His research covers a range of topics in network security, Internet measurement, and data analysis. He is currently a caucus member of ICANN Root Server System Advisory Committee (RSSAC).

Ruixuan Li received the M.S. degree from Zhejiang Gongshang University, China, in 2024. He is currently working toward the Ph.D. degree with the Institute for Network Sciences and Cyberspace, Tsinghua University, China. His research interests include network security and Internet measurement.

Haixin Duan (Member,IEEE) received the PhD degree in computer science from Tsinghua University, and then became a faculty member with Tsinghua University. He has been a visiting scholar with UC Berkeley and a senior scientist with International Computer Science Institute (ICSI). He focuses his research on network security, including security of network protocols (DNS, Web, HTTP, and HTTPS). Most of his papers are published in the top security conferences (Oakland S&P, USENIX Security, CCS, and NDSS).