Address-Based Website Fingerprinting

Christopher A. Wood, Nikita Borisov, Simran Patil

IETF 105 - Montreal - PEARG

Recent Trends

Focus shift from data confidentiality to privacy

- *What* resource are they accessing?
- *Who* is accessing a resource?

Data encryption in transit is growing in use

- DNS-over-HTTPS, DNS-over-TLS, TLS ESNI, etc.

Connection Privacy

Adversary: local and passive observer

Goal: learn information about a network connection and (optionally) link it to a specific client in an **open world** model

Features available:

- Network addresses
- Packet timing and sizes
- Cleartext information

Connection Privacy

Adversary: local and passive observer

Goal: learn information about a network connection and (optionally) link it to a specific client in an **open world** model

Features available:

- Network addresses
- Packet timing and sizes
- Cleartext information

Censors might use this information to block specific connections*

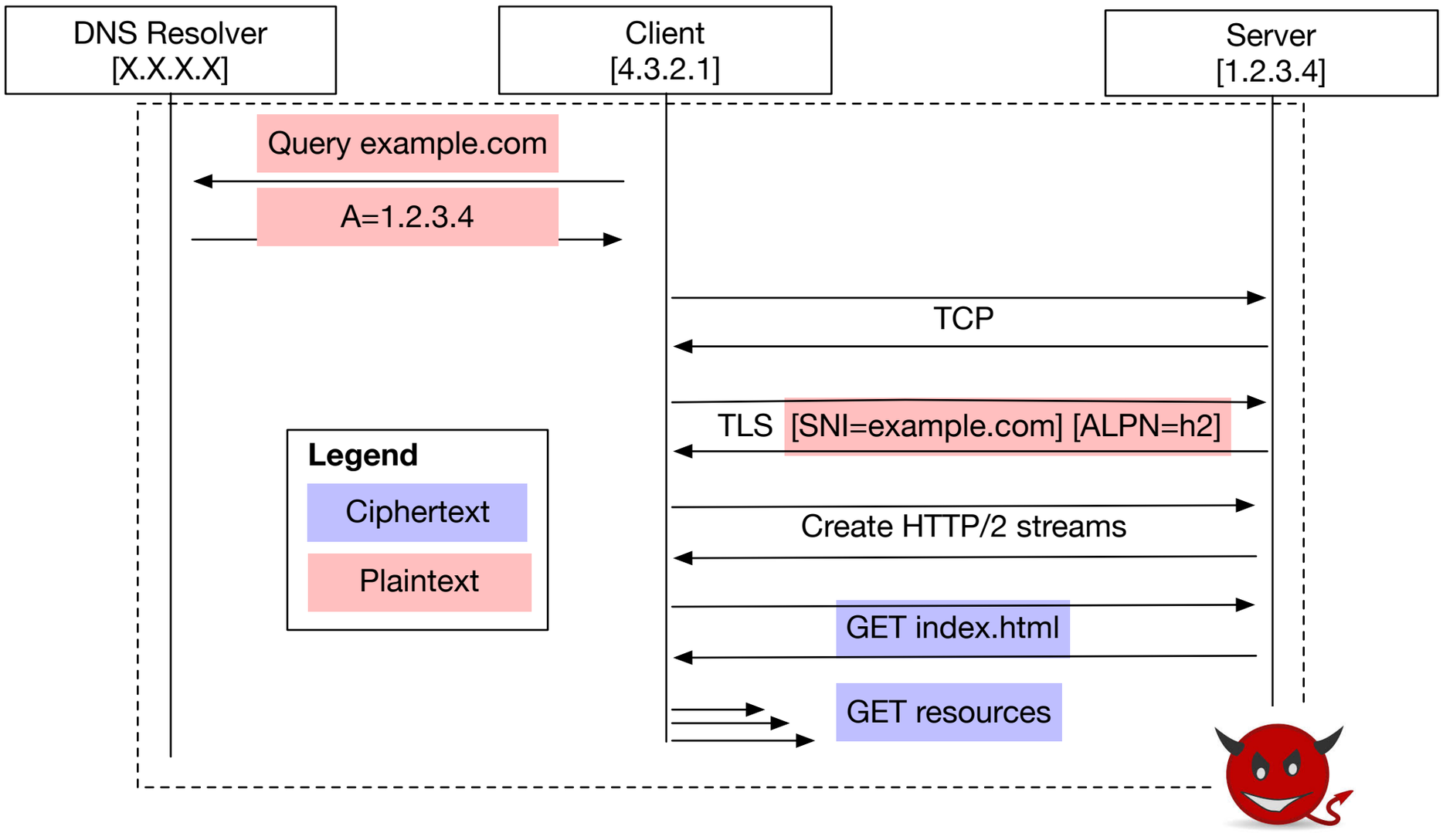
***https://www.bleepingcomputer.com/news/security/south-korea-is-censoring-the-internet-by-snooping-on-sni-traffic/**

Connection Fingerprinting

TLS (HTTPS) *ESNI* *DoT/DoH*

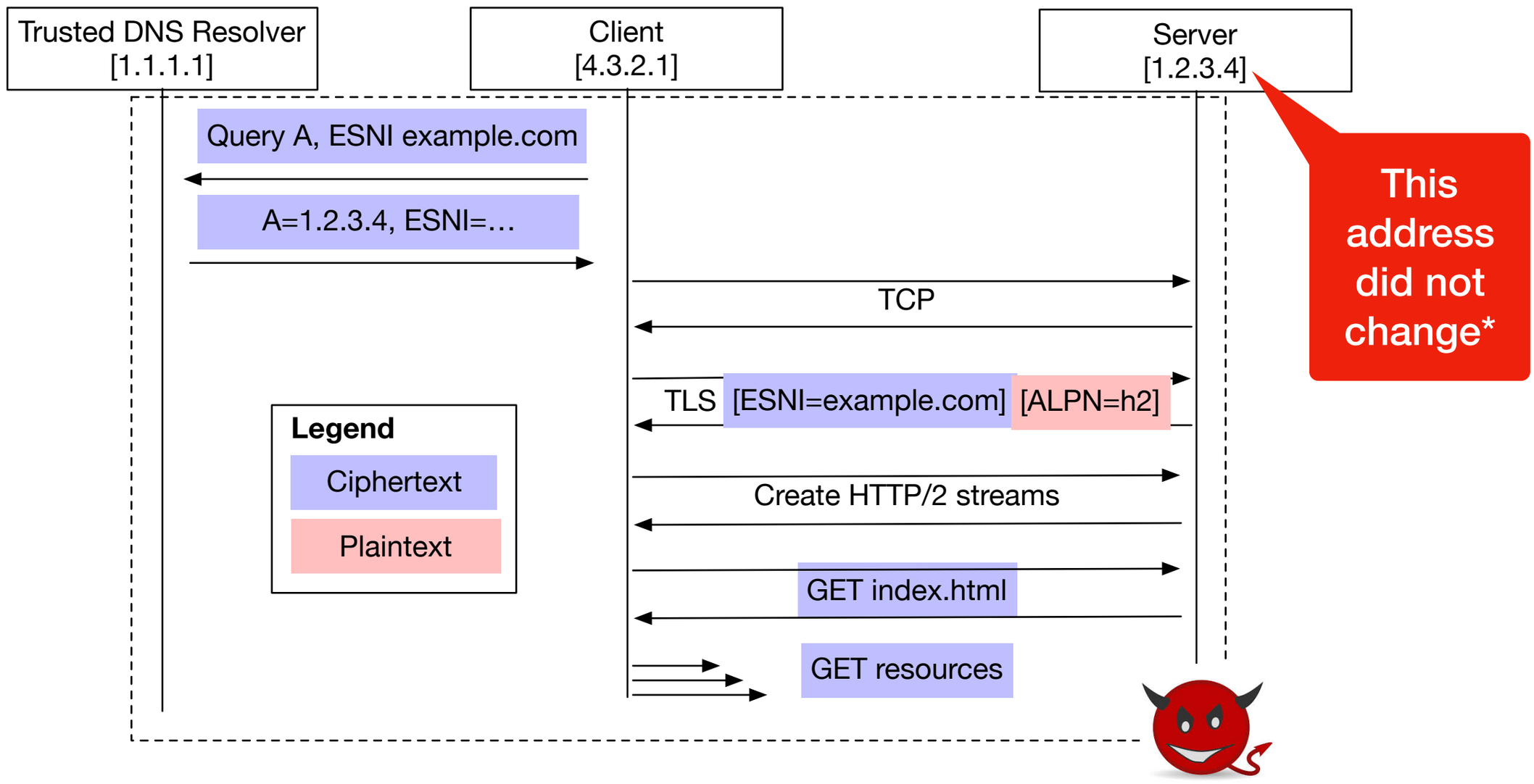
Connection Fingerprinting

Current State



Connection Fingerprinting

DoTH and ESNI



*Though it may with load balancers present

Experiments*

Setup

- MIDA Chromium-based web crawler using Alexa's top million domains (closed world)
- zDNS [1] for name resolution

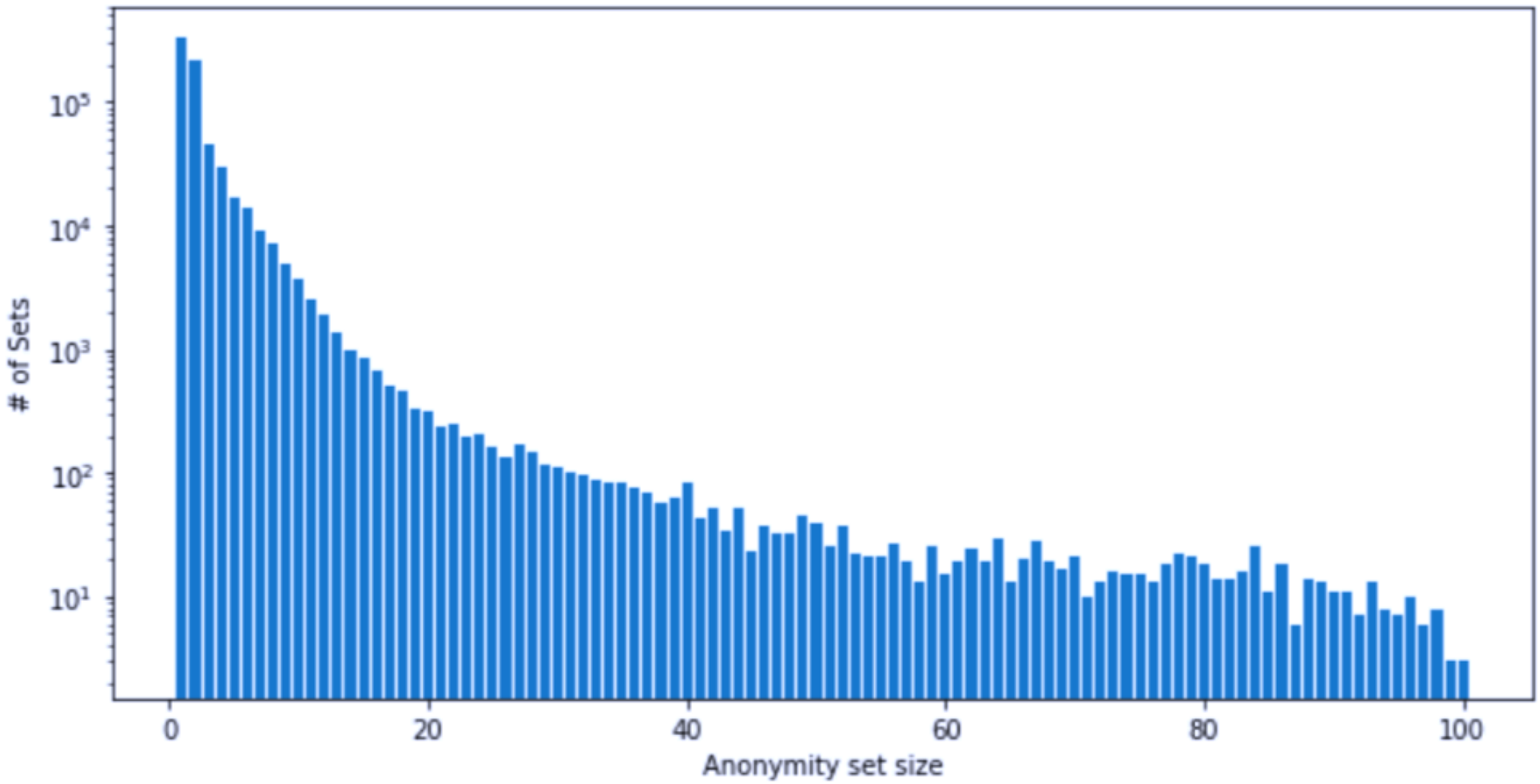
Data collection

- Load pages with MIDA and collect HAR-like traces
- Resolve domains with zDNS (bypassing stub cache)

[1] https://github.com/zmap/zdns

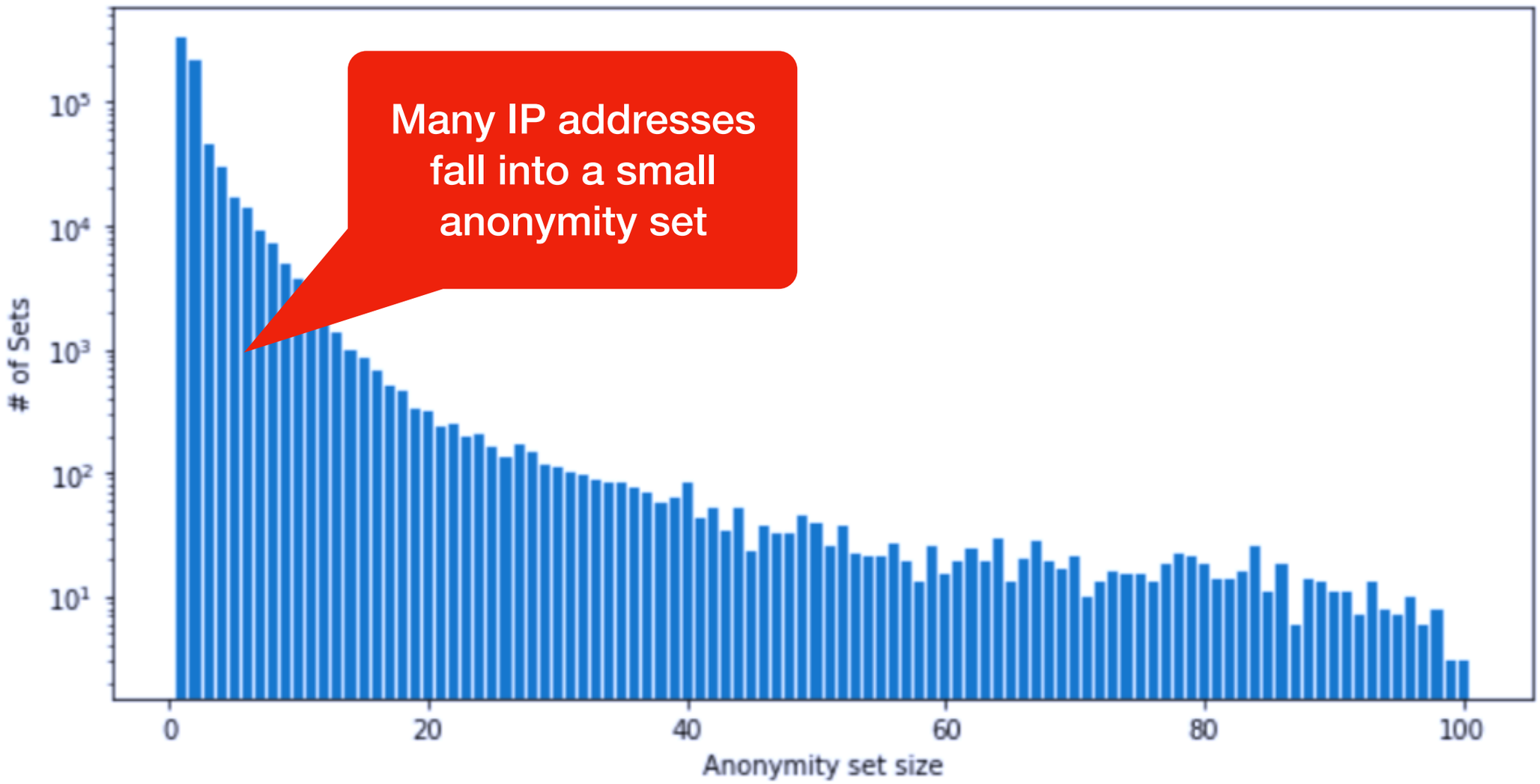
*** "What Can You Learn from an IP?" Simran Patil, Nikita Borisov, ANWR 2019**

Address Anonymity Set*



*** “What Can You Learn from an IP?” Simran Patil, Nikita Borisov, ANWR 2019**

Address Anonymity Set*



*** “What Can You Learn from an IP?” Simran Patil, Nikita Borisov, ANWR 2019**

Page Load Fingerprint (PLF)

Page load fingerprints (PLFs) contain the set of connections and their traffic associated with a page load event

- DNS query patterns
- TCP/TLS connection patterns

Even with all elements encrypted, the patterns are often uniquely identifying

Example: Loading https://nytimes.com in Safari

Page Load Privacy

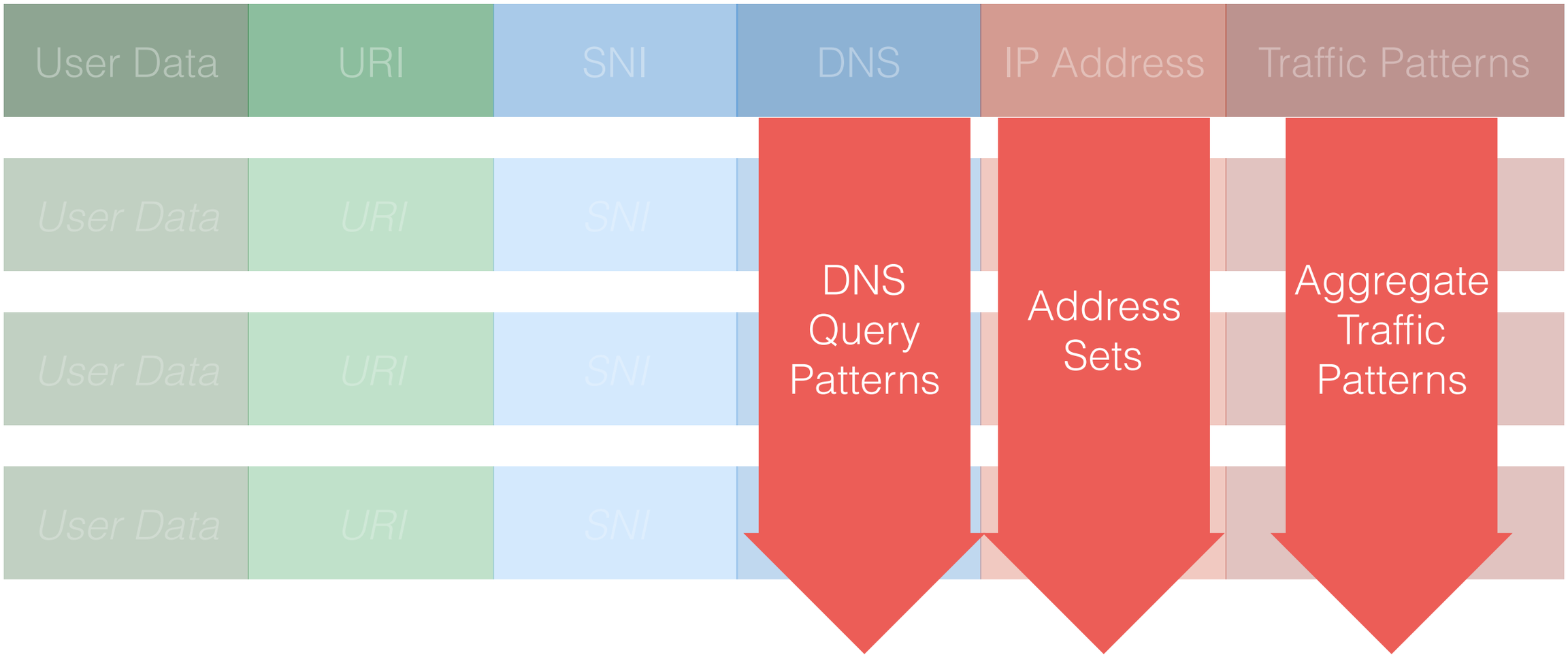
Adversary: local and passive observer

Goal: learn information about a page load and (optionally) link it to a specific client in an **open world** model

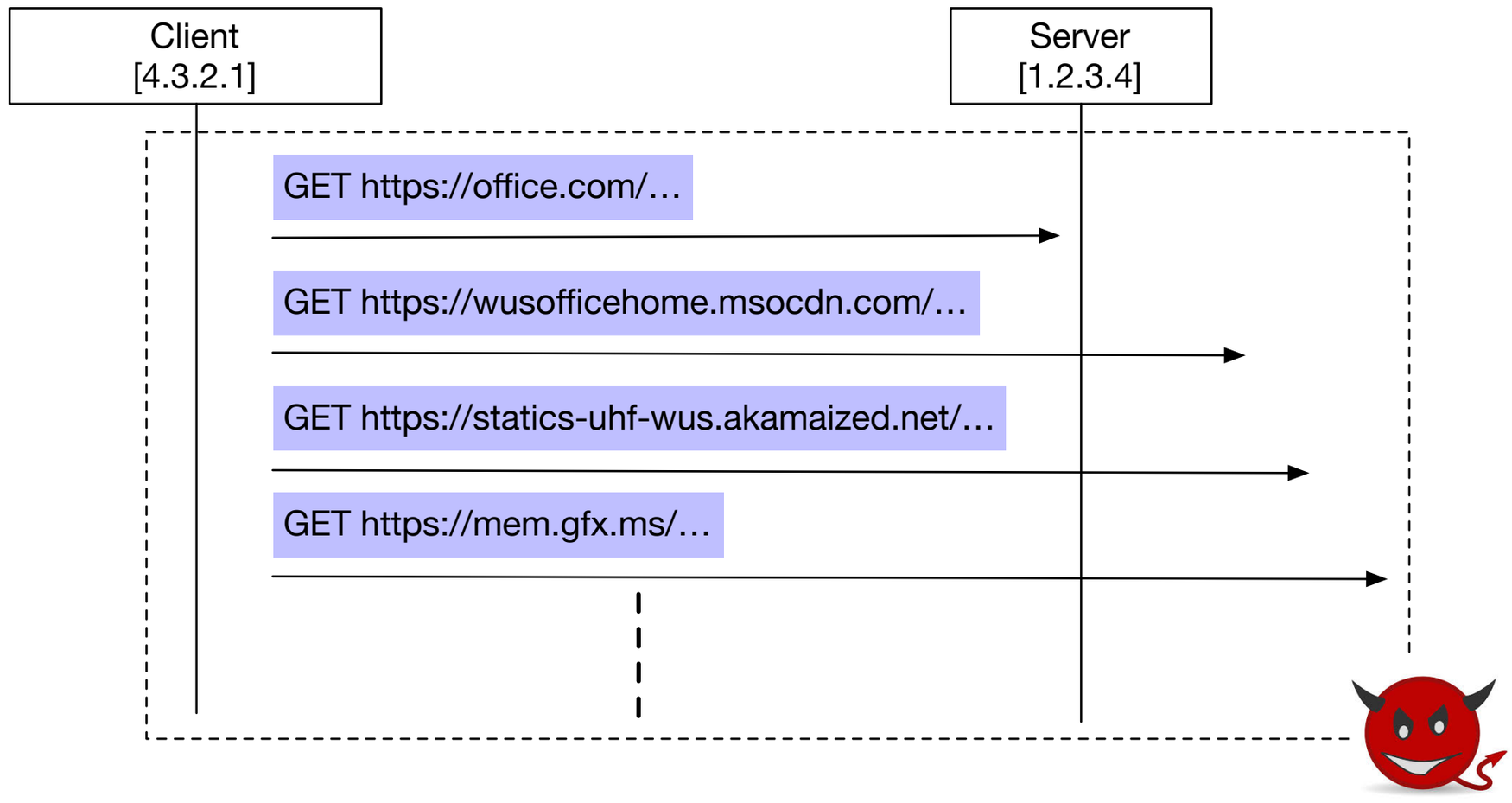
Features available:

- Connection fingerprints
- Traffic patterns

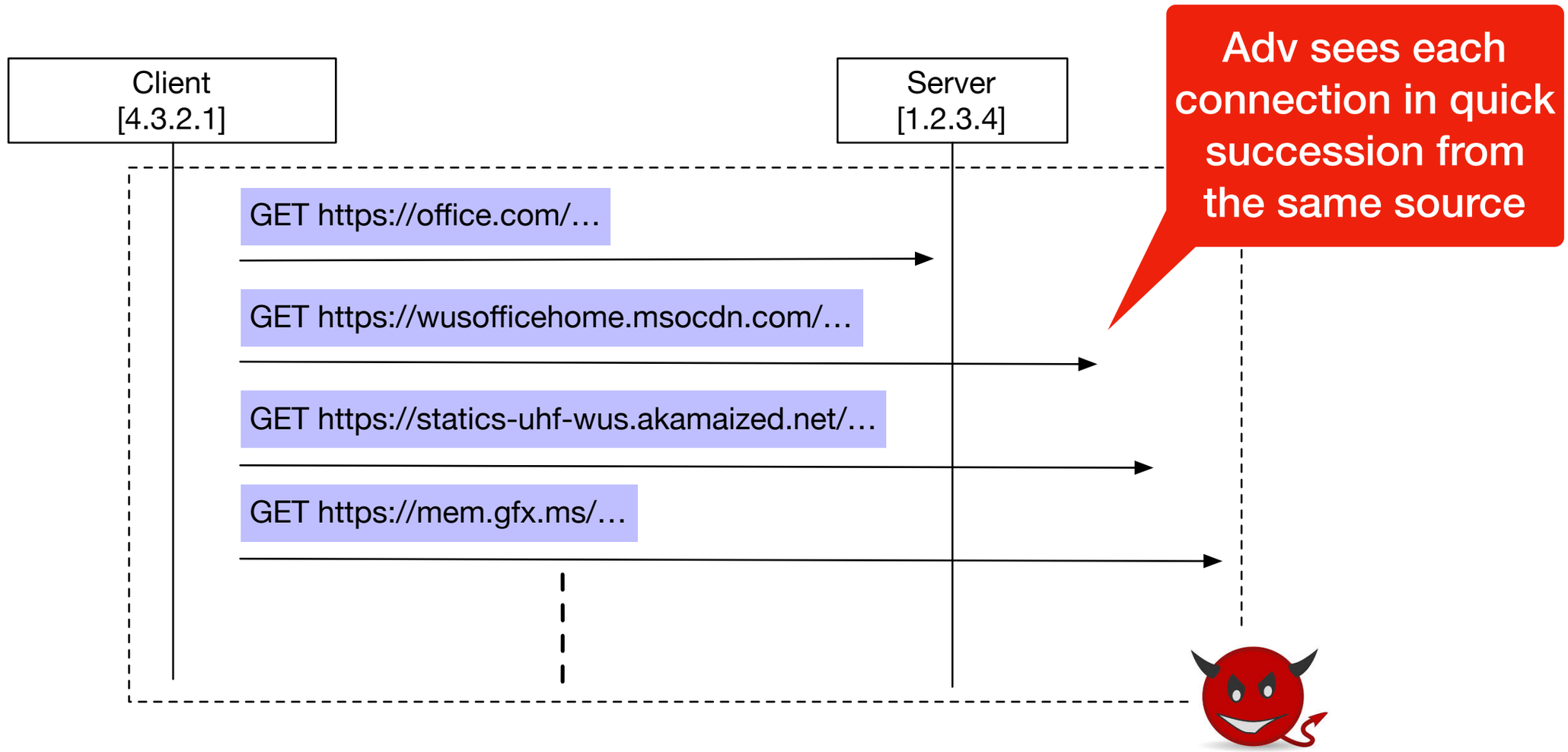
Page Load Fingerprinting



Page Loads


Client
[4.3.2.1]
Server
[1.2.3.4]
GET https://office.com/…
GET https://wusofficehome.msocdn.com/…
GET https://statics-uhf-wus.akamaized.net/…
GET https://mem.gfx.ms/…

Page Loads


Client
[4.3.2.1]
Server
[1.2.3.4]
Adv sees each connection in quick succession from the same source
GET https://office.com/…
GET https://wusofficehome.msocdn.com/…
GET https://statics-uhf-wus.akamaized.net/…
GET https://mem.gfx.ms/…

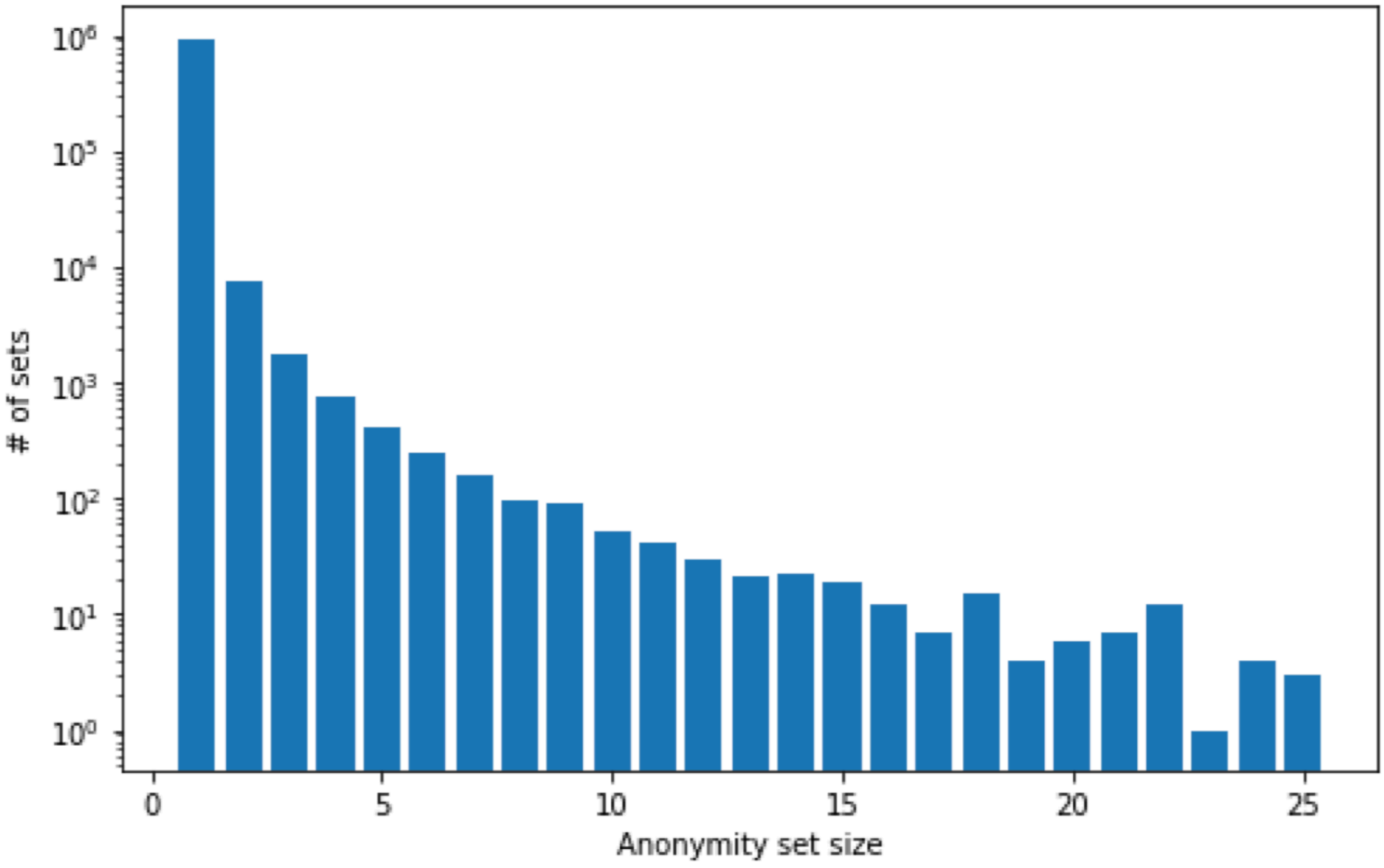
Page Load Statistics*

Upon loading the Alexa top 1 million pages, each page loaded on average

- 96 different URLs
- 16.5 different domains

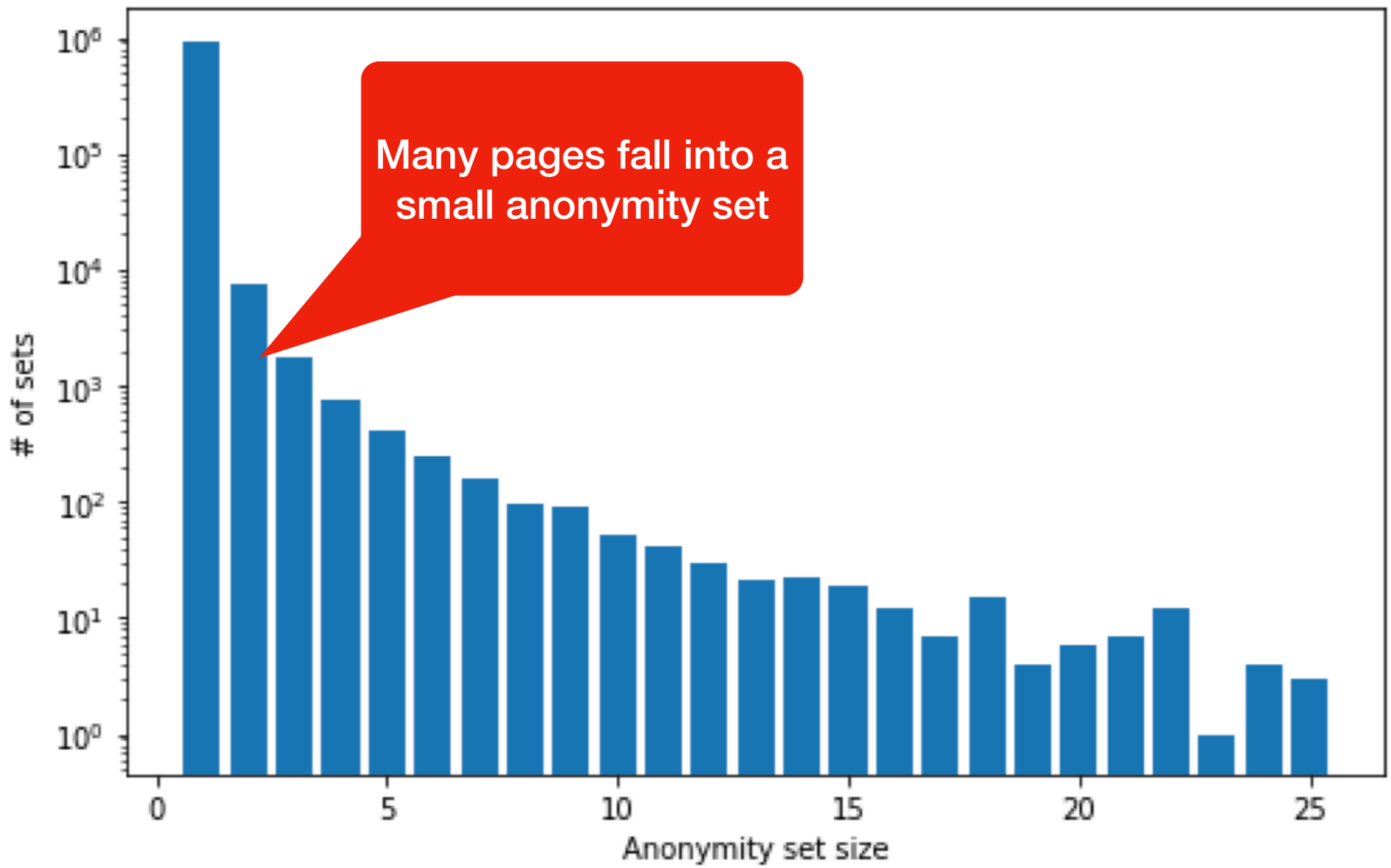
*** “What Can You Learn from an IP?” Simran Patil, Nikita Borisov, ANWR 2019**

PLF Anonymity Set*



*** “What Can You Learn from an IP?” Simran Patil, Nikita Borisov, ANWR 2019**

PLF Anonymity Set*



*** “What Can You Learn from an IP?” Simran Patil, Nikita Borisov, ANWR 2019**

Connection encryption

Connection privacy

Page load privacy

Related Issues

- Happy Eyeballs and connection racing may add more information to PLFs

+ Connection coalescing removes information from PLFs

+ CDN consolidation merges application PLFs

+ Proxies hide destination IP addresses

+ DNS-based load balancing may redirect clients to different servers, or even different providers

Website Fingerprinting

Address-based website fingerprinting may become harder with proxies, multiplexing, coalescing, etc.

Should focus shift towards traffic analysis?

Tor and academic research communities struggle with this problem

Next Steps

Call for research in website fingerprinting

Summarize and document existing research for posterity [1]

Work with academic community to develop and measure mitigations or countermeasures

[1] https://datatracker.ietf.org/doc/draft-wood-privsec-wfattacks/

Questions?