U.S. NAVAL
RESEARCH
LABORATORY

A Measurement of Genuine Tor Traces for Realistic Website Fingerprinting

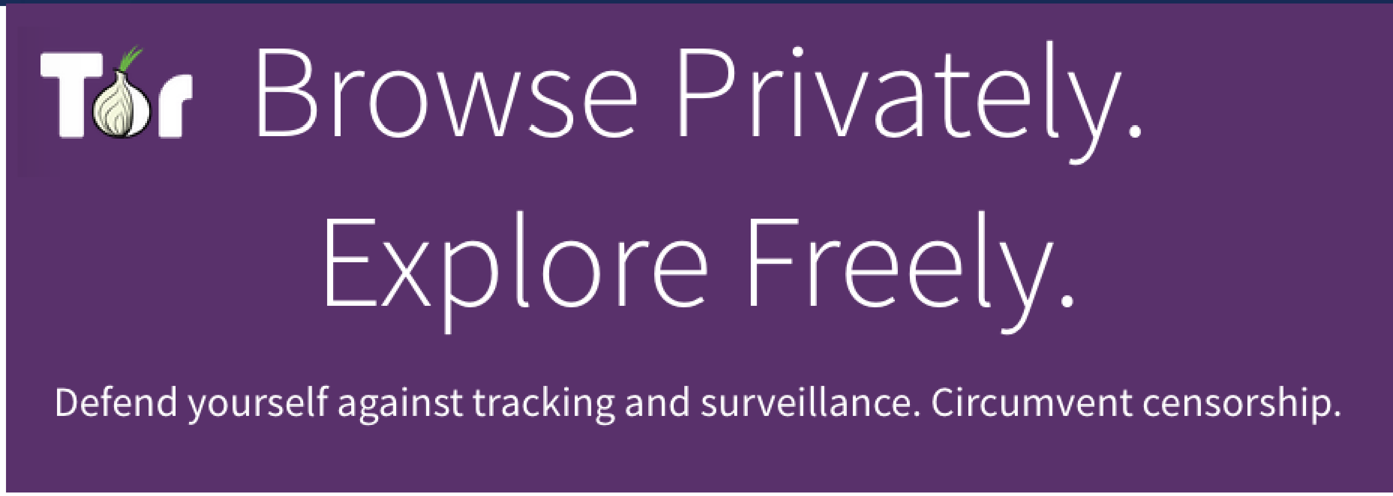
Rob Jansen, U.S. Naval Research Laboratory
Ryan Wails, Georgetown University
Aaron Johnson, U.S. Naval Research Laboratory

Rob Jansen, PhD
Computer Scientist
Center for High Assurance Computer Systems
U.S. Naval Research Laboratory

The Passive and Active Measurement Conference 2026
Virtual Event
March 25th, 2026

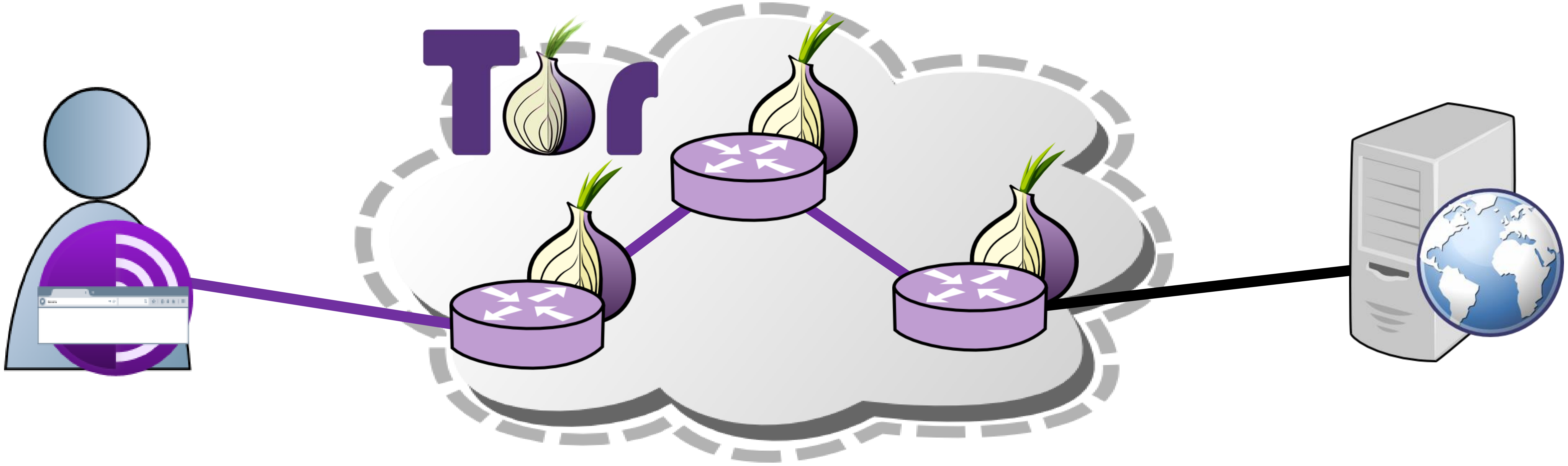
Anonymous Communication with Tor

- Separates *identification* from *routing*
- Provides unlinkable communication
- Promotes user safety and privacy online



Browse Privately.
Explore Freely.

Defend yourself against tracking and surveillance. Circumvent censorship.

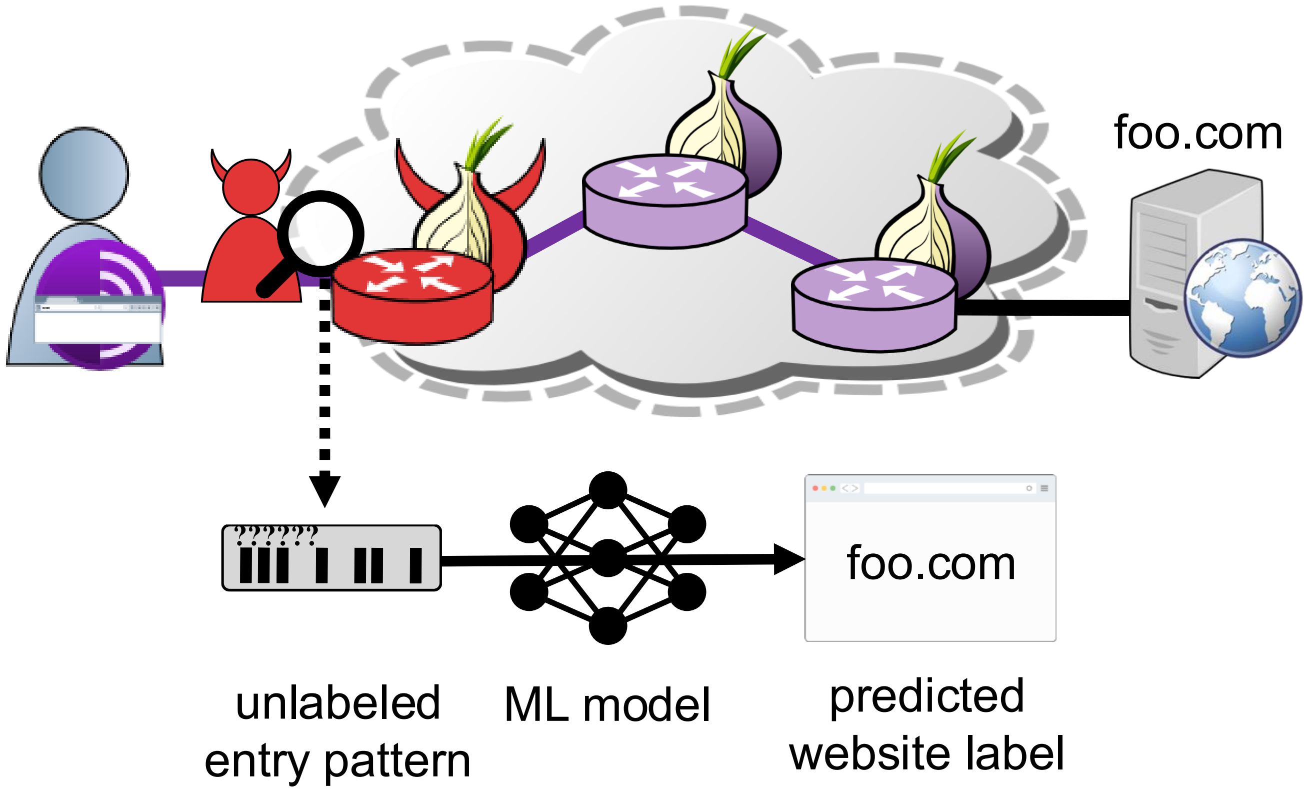


Website Fingerprinting Adversary Model

Website fingerprinting (WF) attacks can **link a source to its destination**, breaking Tor’s anonymity

Adversary can:

- Obtain entry-side vantage point
- Observe traffic patterns
- Predict website visited by user using **a trained ML model**

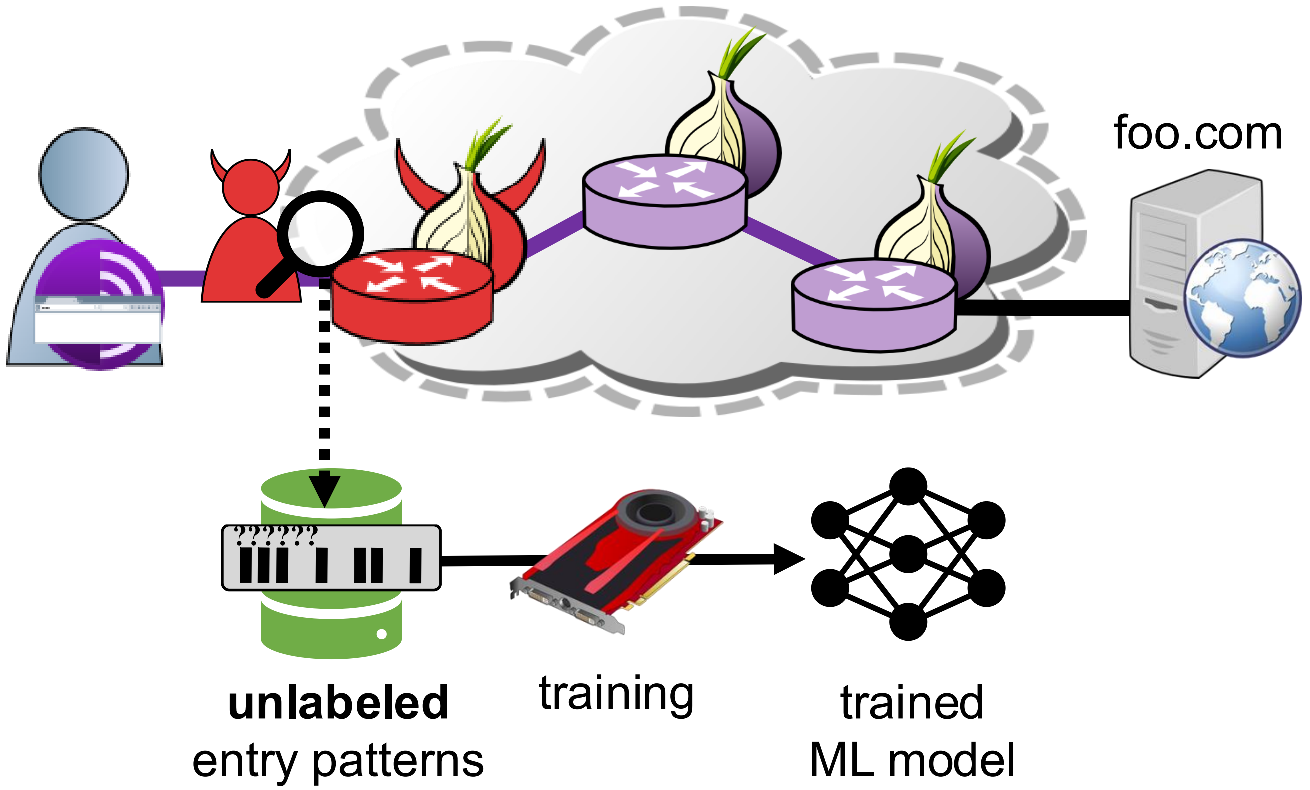


How Might Adversary Train ML Models?

Non-option: use **entry examples**

- Need **labeled** examples of patterns
- **Onion-encryption** hides labels
- Observed examples are **unlabeled**

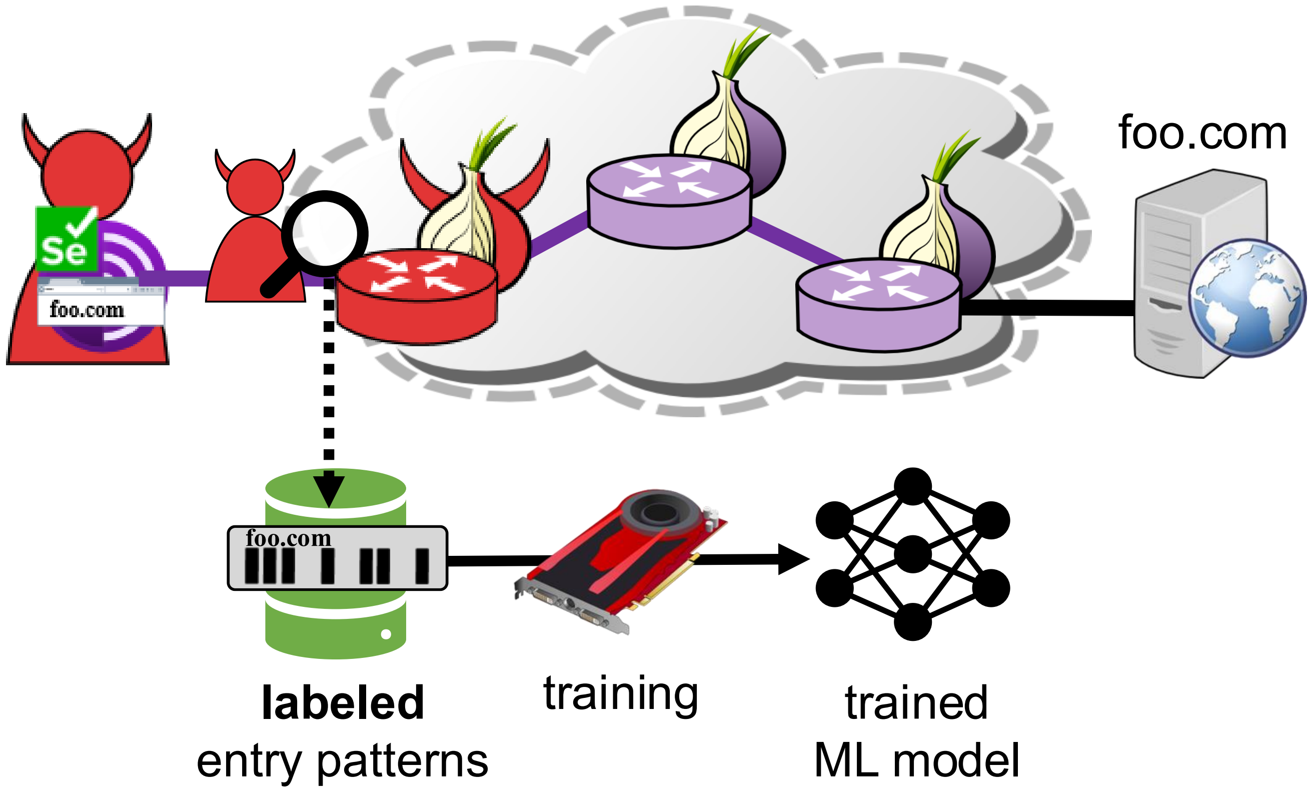
☒



How Might Adversary Train ML Models?

Option 1: **synthetic data**

- Use **automated browser** (selenium) to **replicate users'** behavior/diversity
  - Usually by crawling frontpages of top sites…



Problems

- Modeling WF with synthetic user data **oversimplifies the ML task** [CCS'14, USENIX'22, PoPETs'23]
  - Browser version, configuration
  - URL choice, fetch order, usage of tabs
  - # of sites/pages, world size dynamics
  - Geo-location, concept drift
  - Tor network variation: relay churn, software versions, congestion, location…

How Might Adversary Train ML Models?

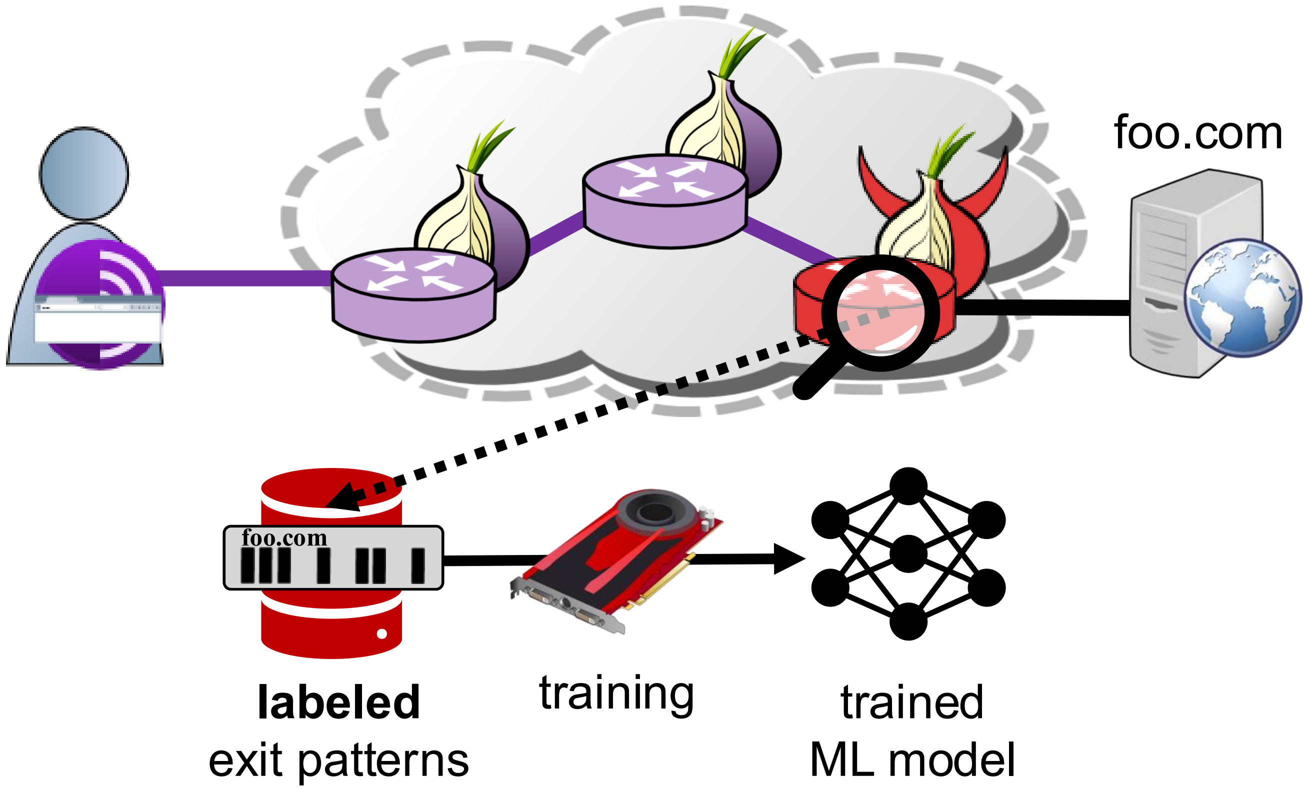
Option 2: **real Tor user data**

- Run **exit relay**, observe traffic [USENIX'22]
  - Traffic patterns from **real Tor users**
  - Website **labels** observed in DNS requests

Problems

- Study done in **online** setting [USENIX'22]
  - Data was **not persistently stored**
  - Results cannot be replicated
  - Difficult to build on the methodology without **new measurements**

- Exit-entry position mismatch
  - **Train** on **exit** side, **predict** on **entry** side
  - Position "distortion" reduces performance
    - 5-18% [USENIX'22]
    - 17% median, 93% worse-case [WPES'24]

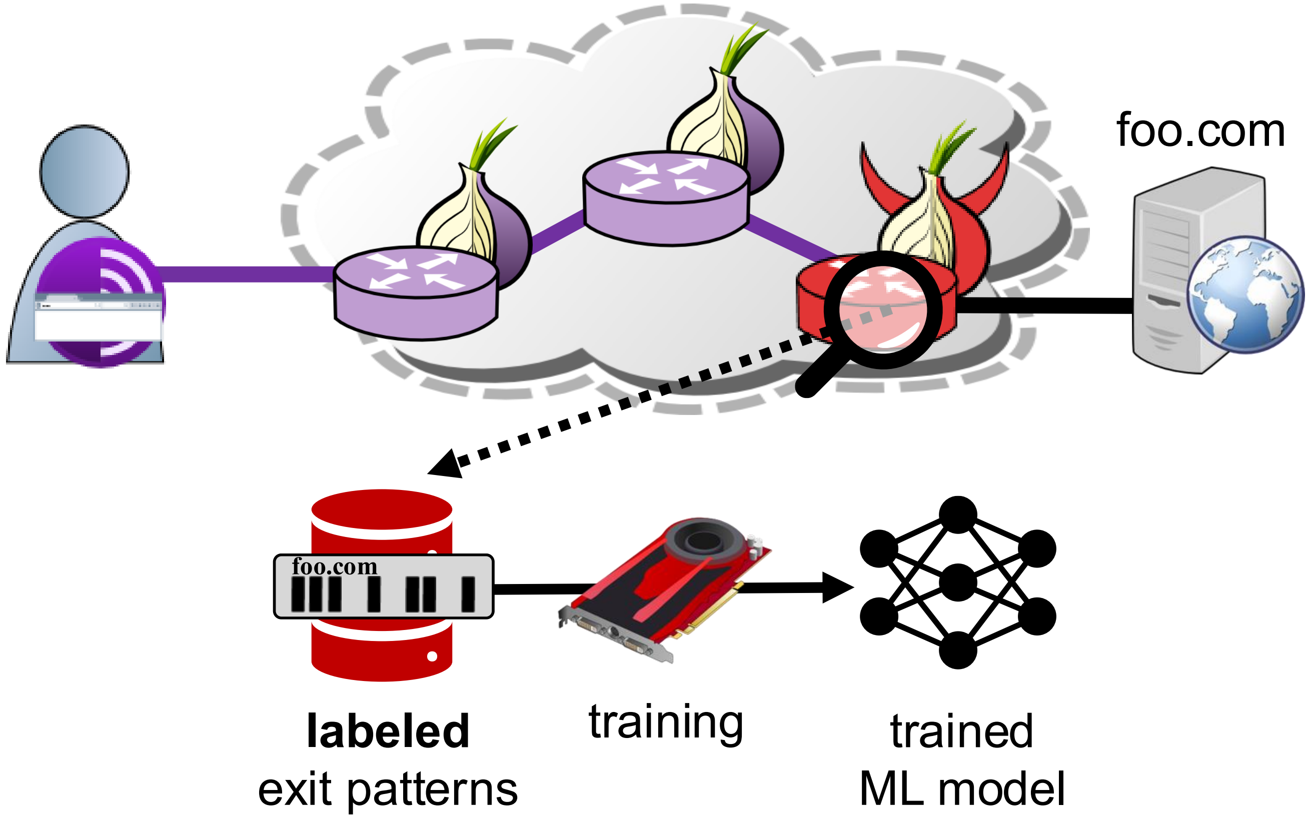


Our Research Direction: Key Insights

Key insights:

- If adversary would **test** on real user data, they should **train** on it too
- The **real network** is the best place to get traffic patterns of **real users**
- Easier to **mitigate** entry-exit distortion than **accurately replicate users**

	Synthetic client	Real Tor exit
Ground-truth labels	✓	✓
Entry-side examples	✓	✗
Real Tor user data	✗	✓

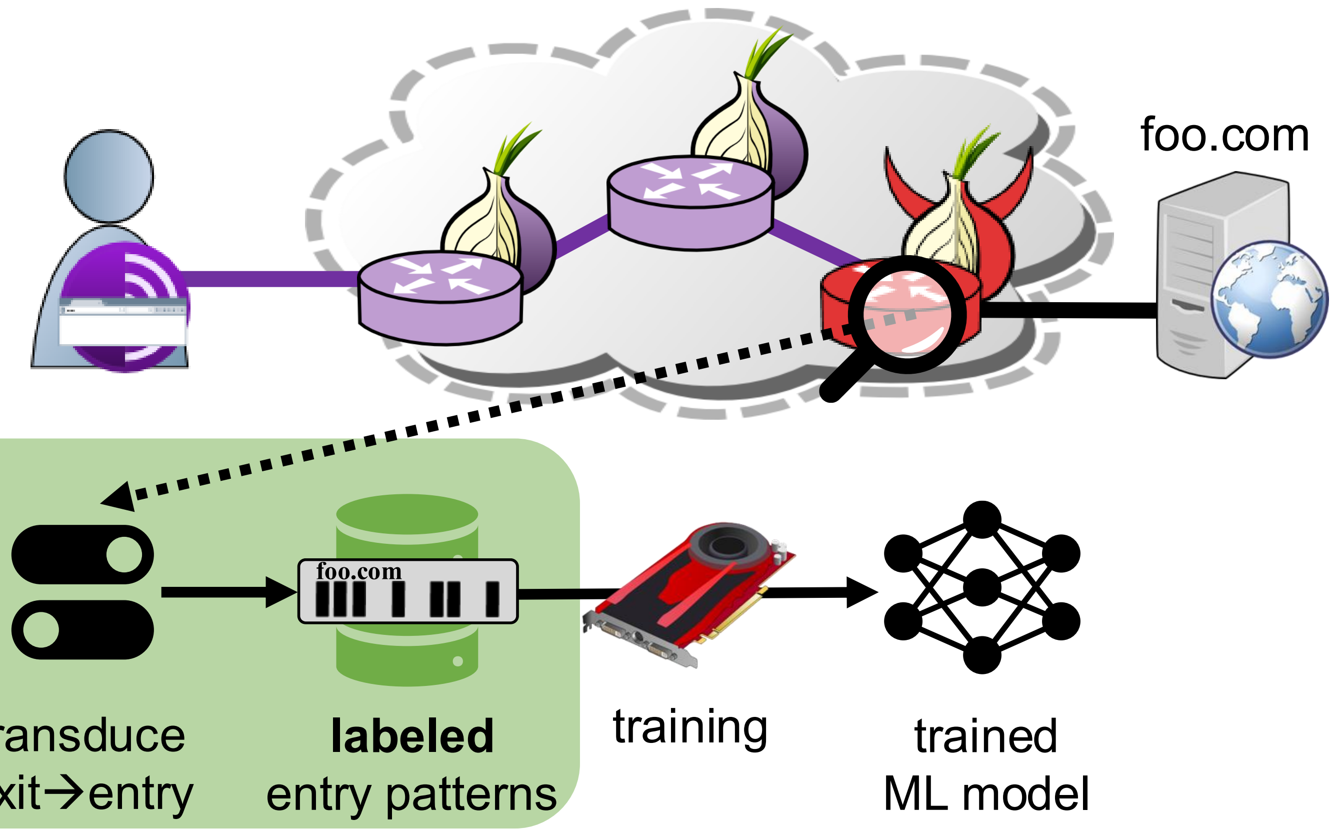


Our Research Direction: Key Insights

Key insights:

- If adversary would **test** on real user data, they should **train** on it too
- The **real network** is the best place to get traffic patterns of **real users**
- Easier to **mitigate** entry-exit distortion than **accurately replicate users**
  - **Now mitigated** with exit→entry **transducers**:
    - **Retracer**: using network simulation [WPES'24]
    - **CellShift**: using cell RTTs and math [NDSS'26]

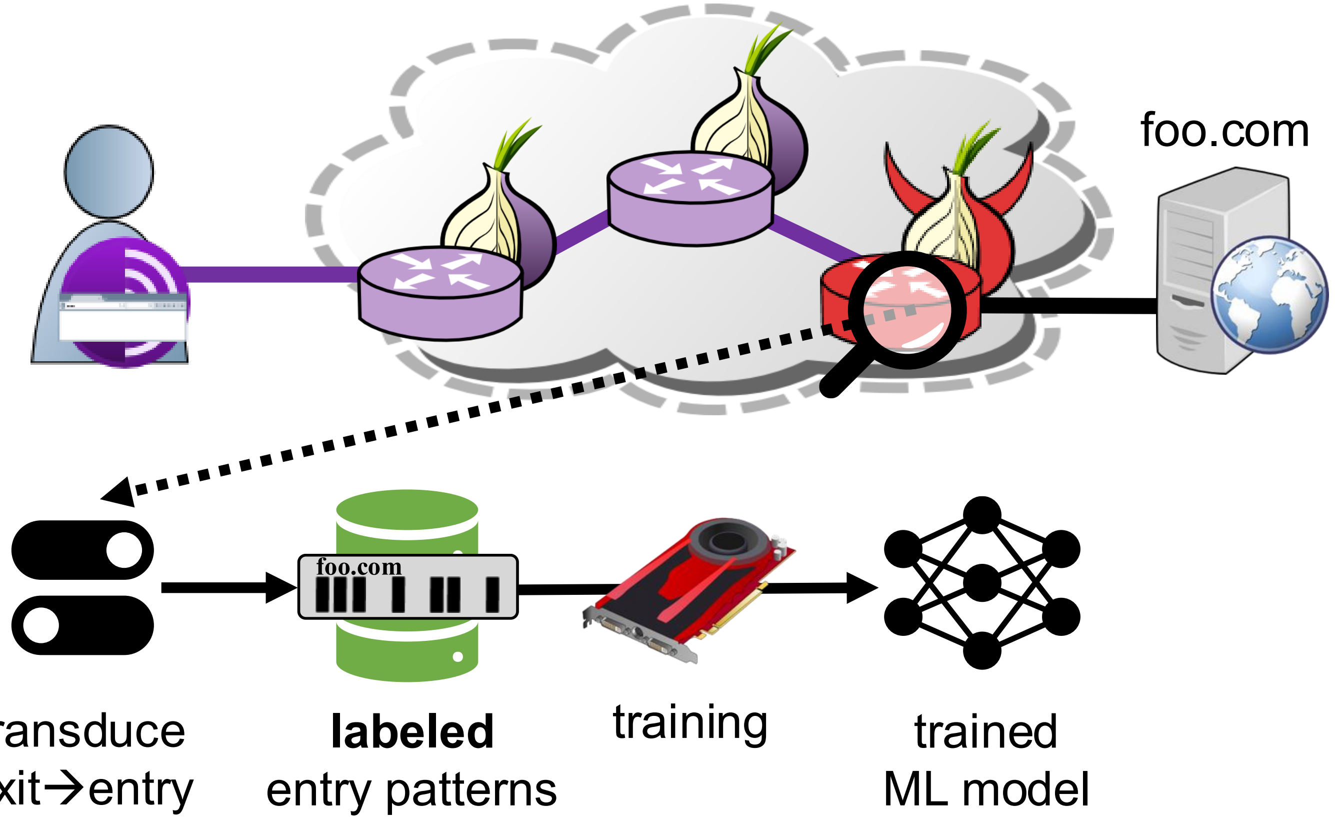
	Synthetic client	Real Tor exit
Ground-truth labels	✓	✓
Entry-side examples	✓	✗ → ✓
Real Tor user data	✗	✓



Our Goals in this Work

- Goals:
  - Create a **persistent dataset** to improve the study of real-world WF
  - Understand **disparities** between **synthetic** datasets and **real** data
  - Inform/prioritize WF **defenses**
- Non-Goals:
  - Develop new WF attacks
  - Improve attacks to benefit adversary

	Synthetic client	Real Tor exit
Ground-truth labels	✓	✓
Entry-side examples	✓	✗ → ✓
Real Tor user data	✗	✓

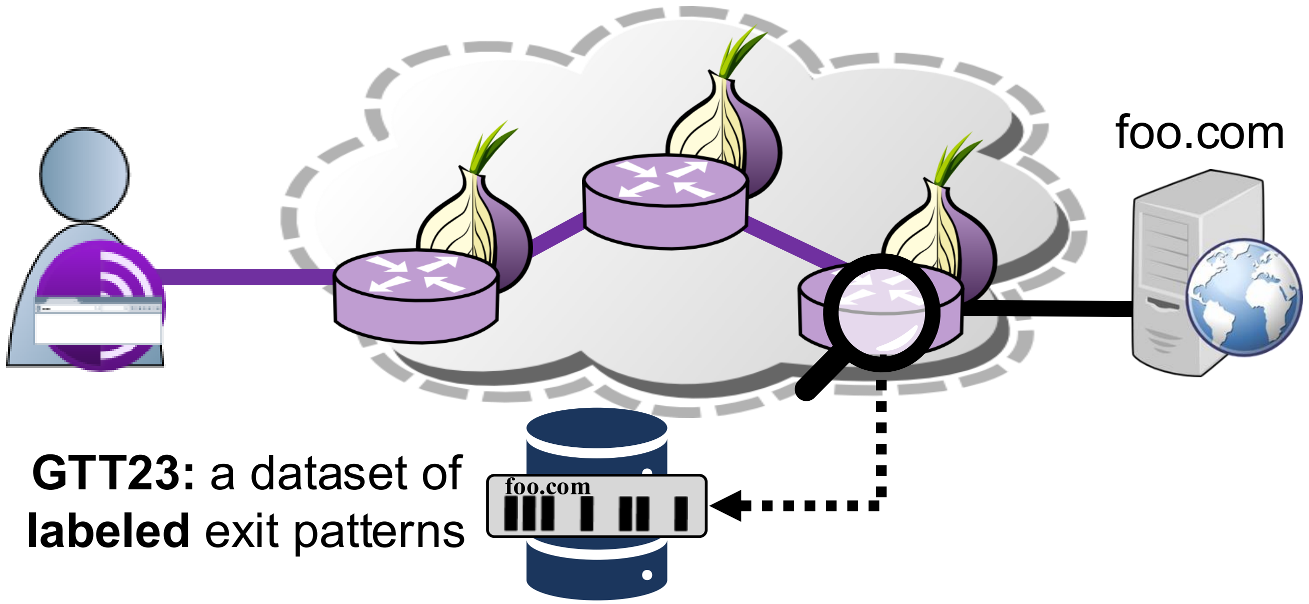


GTT23: A dataset of Genuine Tor Traces

Introducing **GTT23**

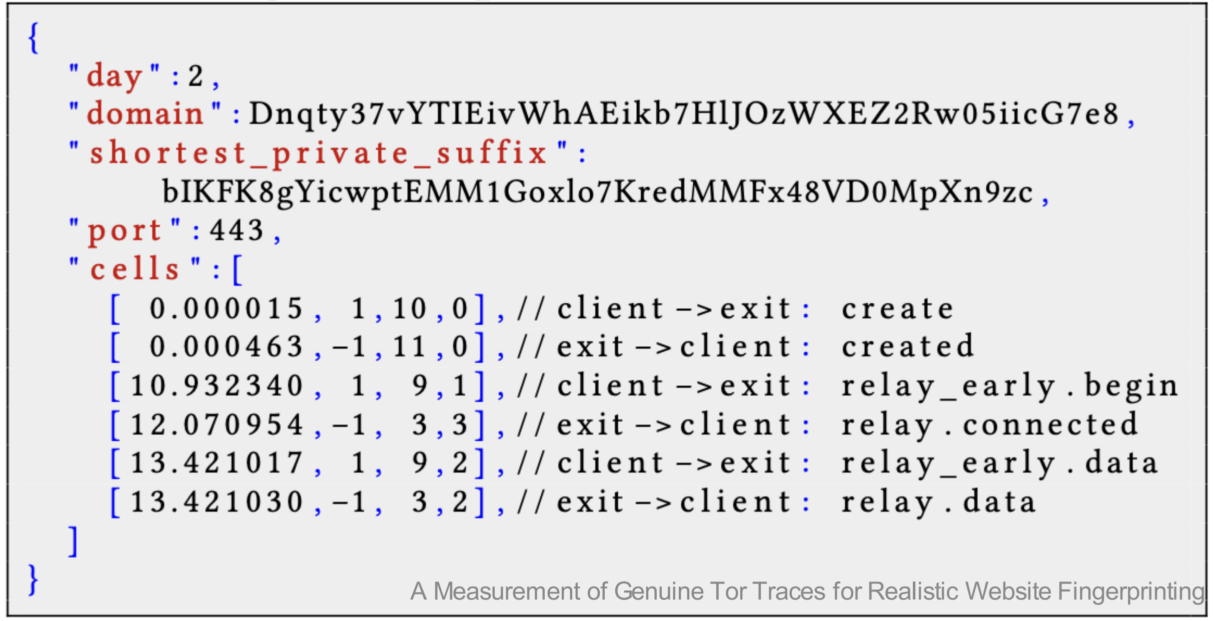
- A dataset of **genuine Tor traces** of real Tor user traffic patterns
    - **13.9M** traces to **1.1M** unique domains measured over **13 weeks**

- Measured from Tor **exit relays**
    - Traffic **traces** of normal activities
        - The first **5,000 cells** on a Tor circuit
        - **Natural**, real-world website **base rates**
        - No **PII** is recorded, website labels are **protected** with HMAC
    - Measurement plan **reviewed** by IRB, Tor Research Safety Board



GTT23: a dataset of **labeled** exit patterns

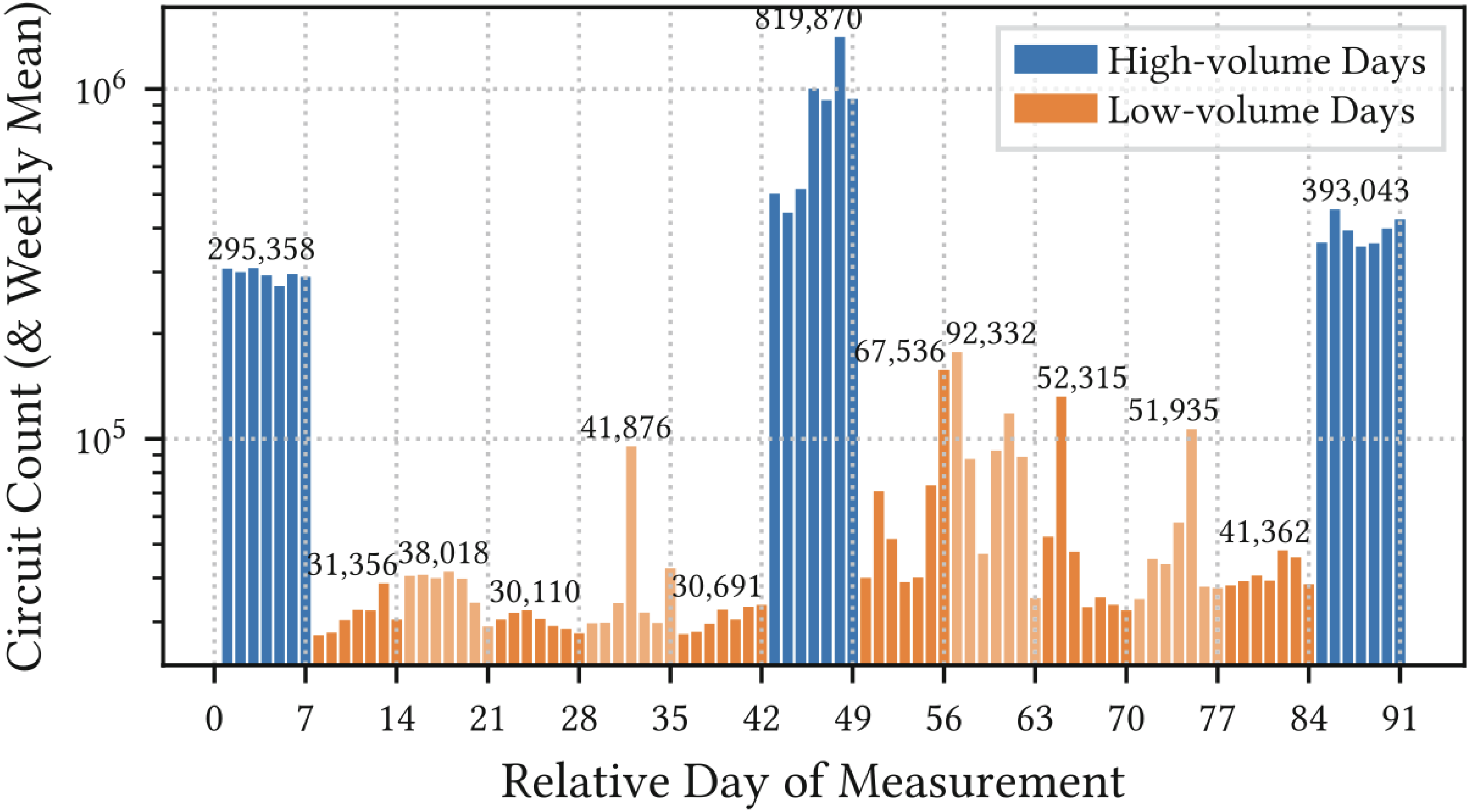
Listing 1: Example circuit metadata record.



```
{
  "day":2,
  "domain":Dnqty37vYTIEivWhAEikb7HlJOzWXEZ2Rw05iicG7e8,
  "shortest_private_suffix":
       bIKFK8gYicwptEMM1Goxlo7KredMMFx48VD0MpXn9zc,
  "port":443,
  "cells":[
    [ 0.000015, 1,10,0],//client->exit: create
    [ 0.000463,-1,11,0],//exit->client: created
    [10.932340, 1, 9,1],//client->exit: relay_early.begin
    [12.070954,-1, 3,3],//exit->client: relay.connected
    [13.421017, 1, 9,2],//client->exit: relay_early.data
    [13.421030,-1, 3,2],//exit->client: relay.data
  ]
}
```

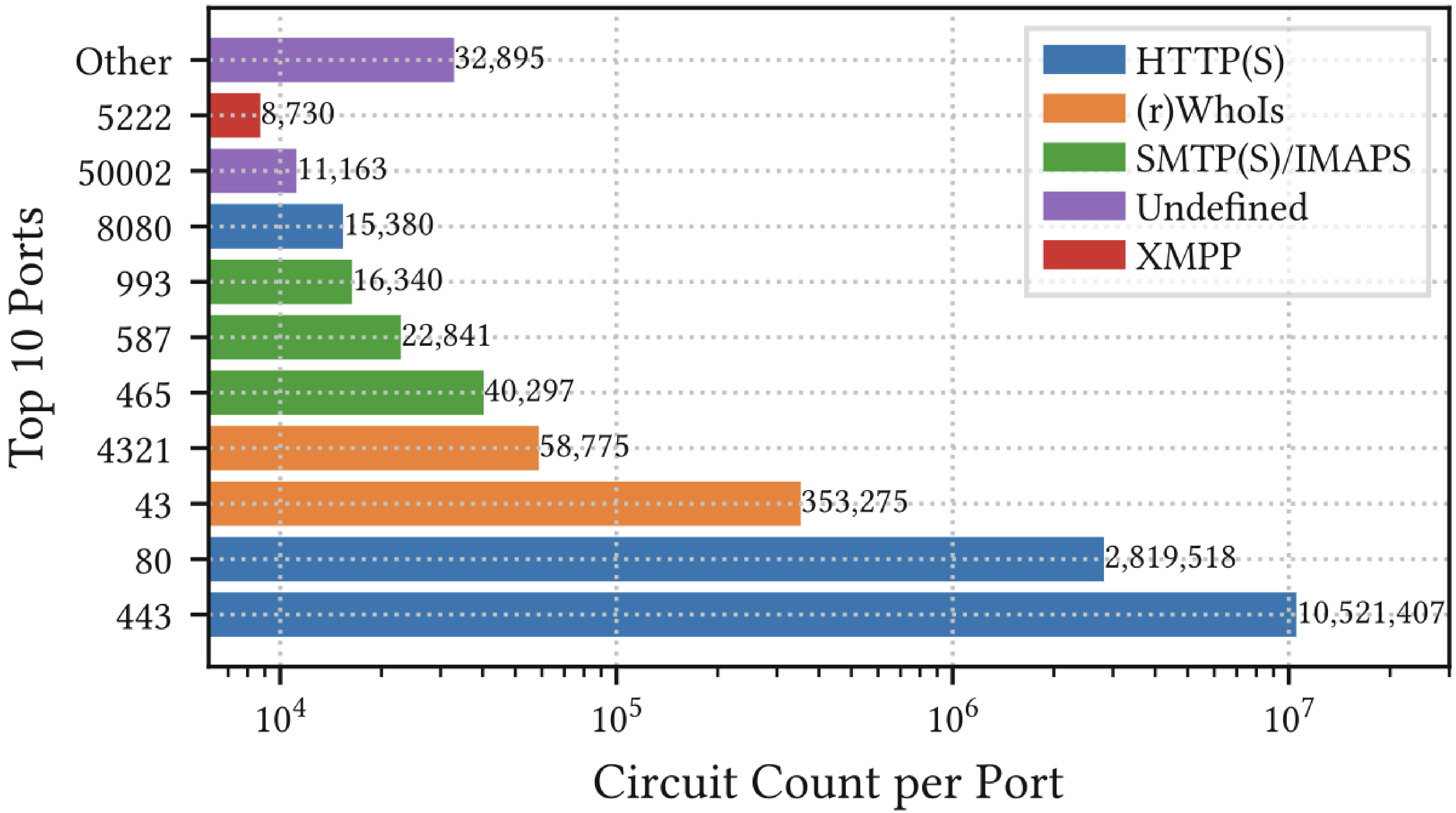
Temporal Composition of GTT23

13.9M traces across the **13 week** measurement



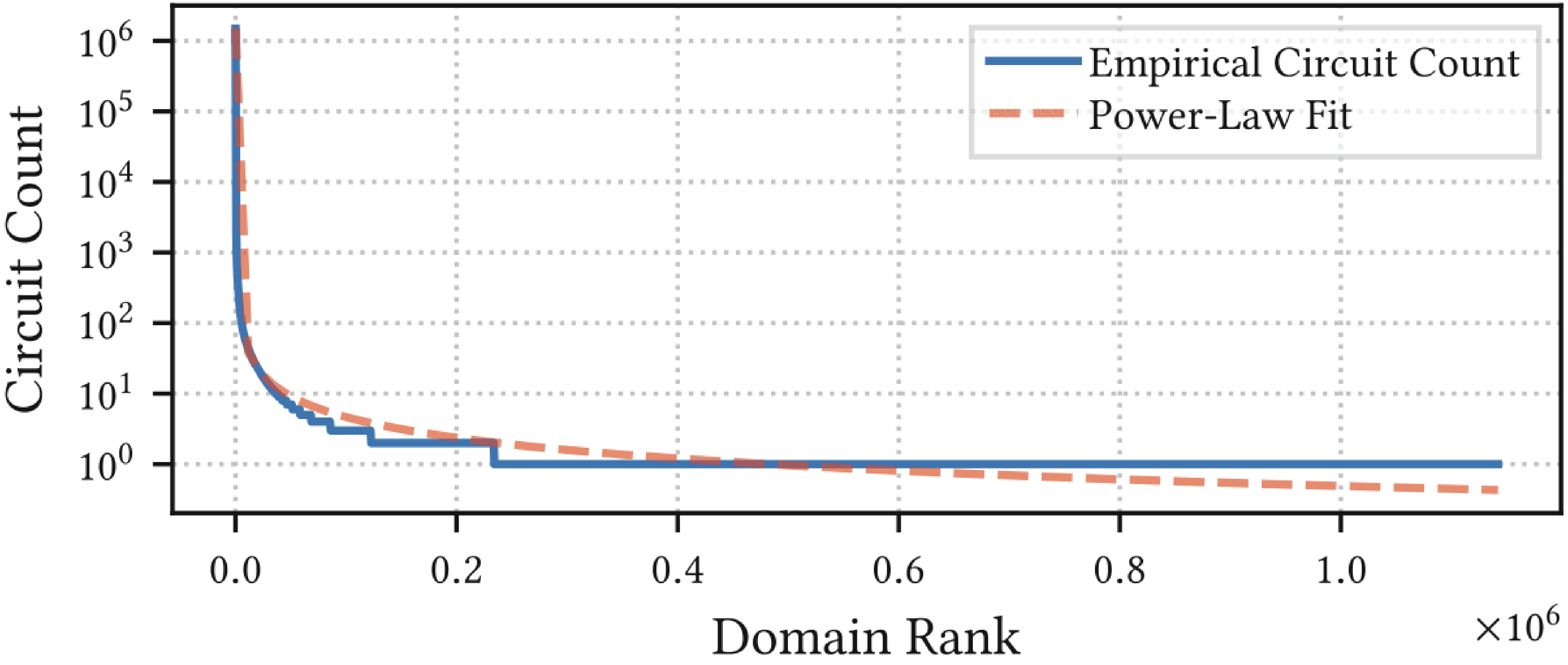
IANA-assigned Service (Port) Composition

68 unique destination service **ports**



Circuit Count per Domain is Similar to Internet Traffic

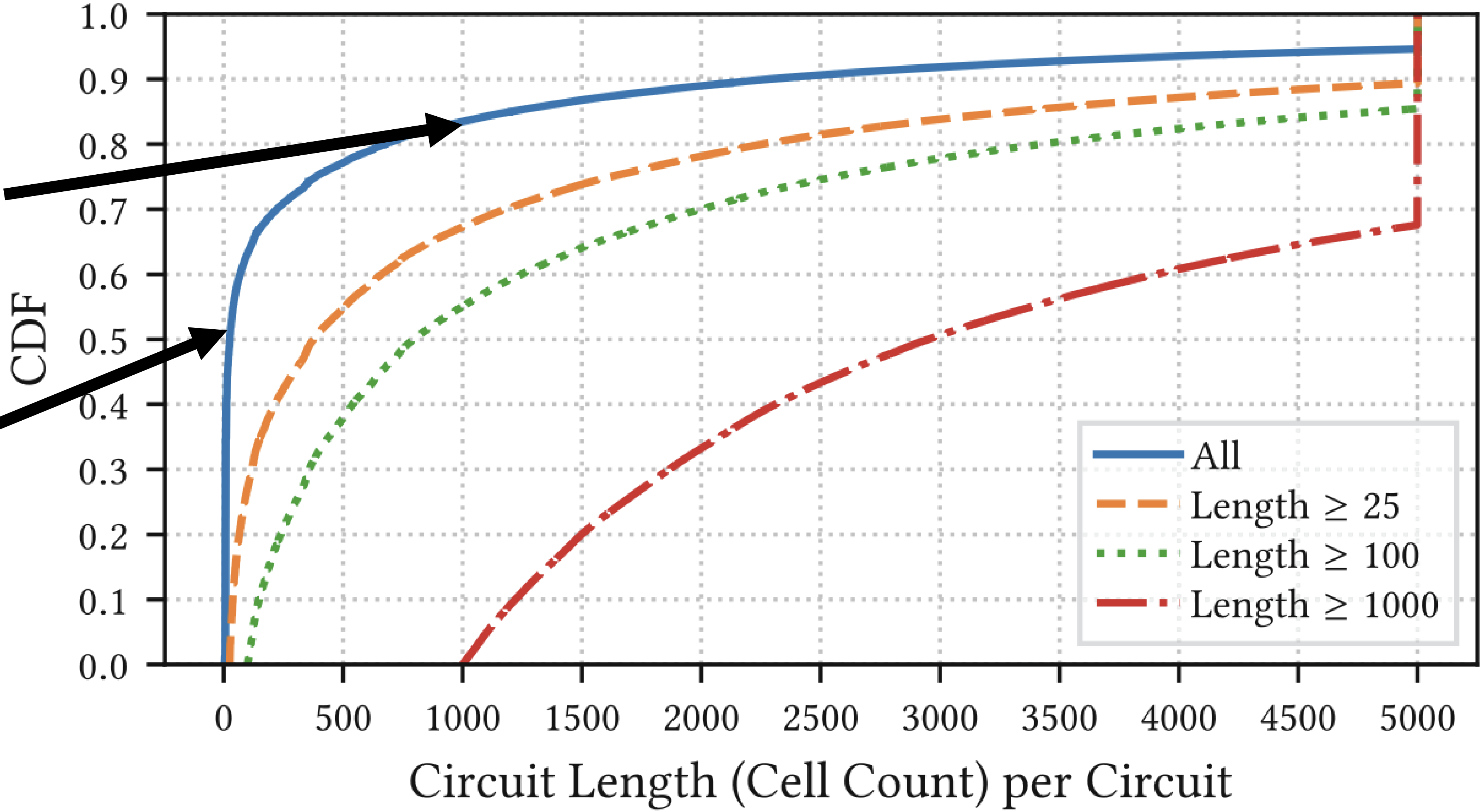
1.1M unique destination **domains**



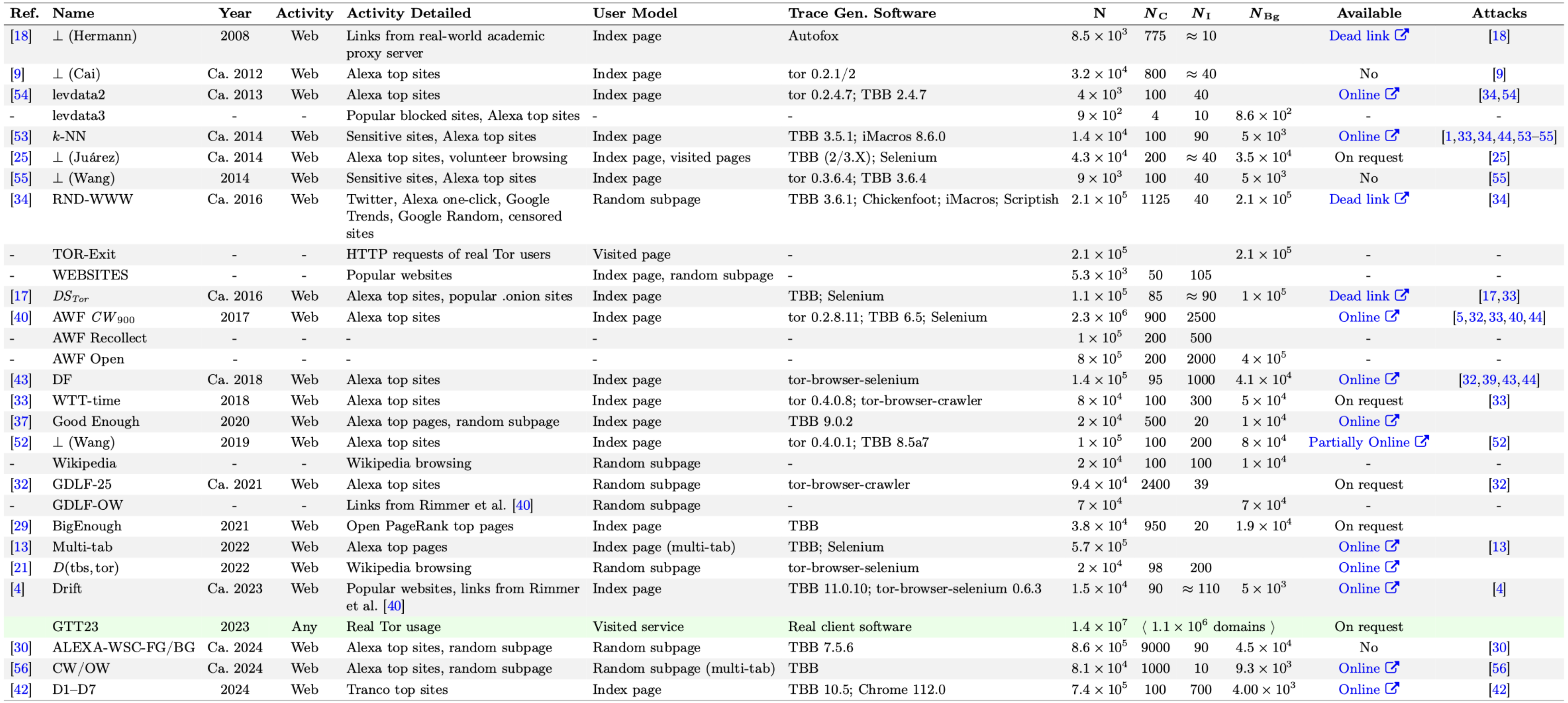
Most Circuits are Much Shorter than Expected

HTTP Archive:
90% of webpages are > **450 KB**

Median circuit is **10.5 KB** (25 cells)



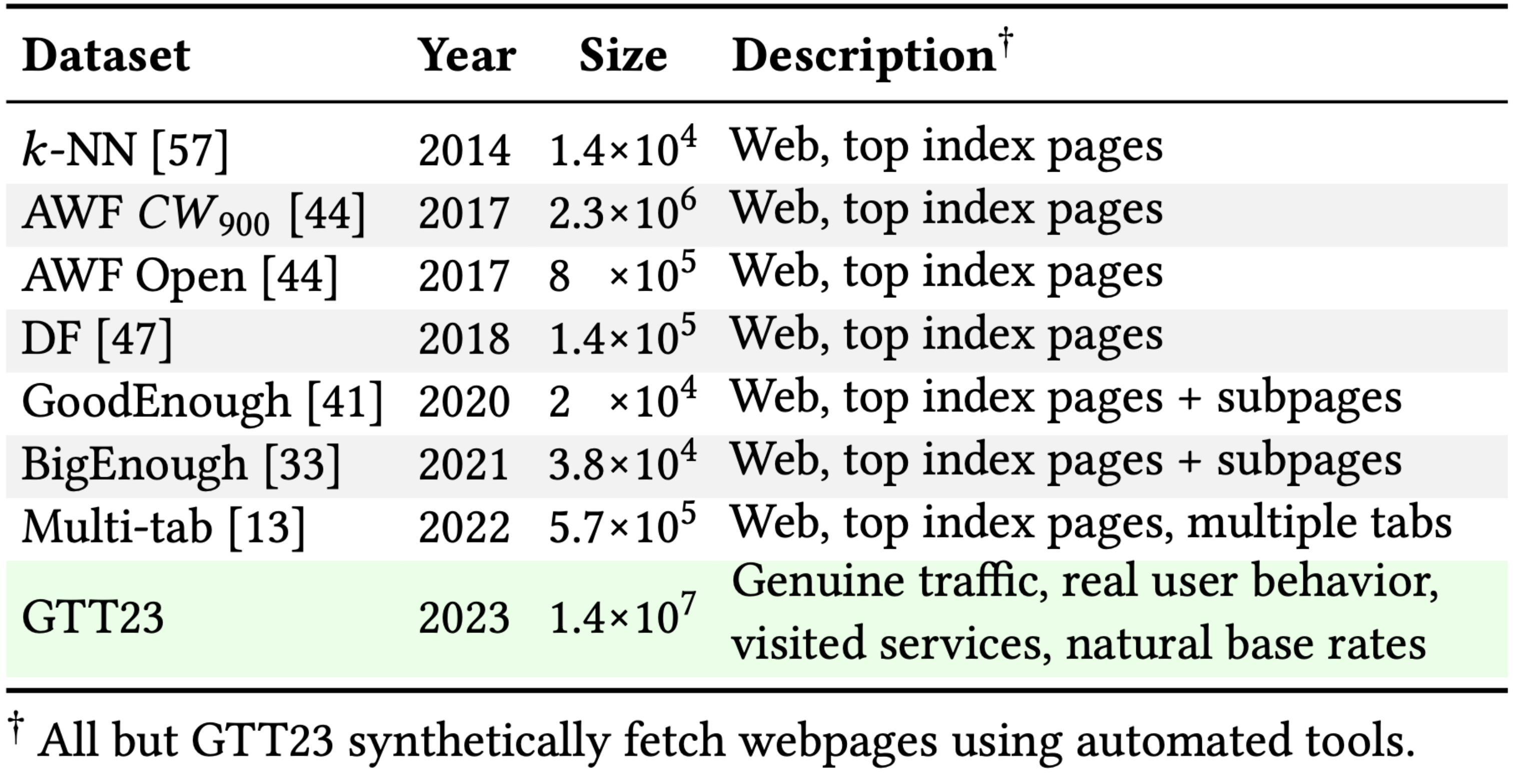
We surveyed 28 WF datasets published since 2008



Ref.	Name	Year	Activity	Activity Detailed	User Model	Trace Gen. Software	N	N_{C}	N_{I}	N_{Bg}	Available	Attacks
[18]	⊥ (Hermann)	2008	Web	Links from real-world academic proxy server	Index page	Autofox	8.5×10^3	775	≈ 10		Dead link ↗	[18]
[9]	⊥ (Cai)	Ca. 2012	Web	Alexa top sites	Index page	tor 0.2.1/2	3.2×10^4	800	≈ 40		No	[9]
[54]	levdata2	Ca. 2013	Web	Alexa top sites	Index page	tor 0.2.4.7; TBB 2.4.7	4×10^3	100	40		Online ↗	[34,54]
-	levdata3	-	-	Popular blocked sites, Alexa top sites	-	-	9×10^2	4	10	8.6×10^2	-	-
[53]	*k*-NN	Ca. 2014	Web	Sensitive sites, Alexa top sites	Index page	TBB 3.5.1; iMacros 8.6.0	1.4×10^4	100	90	5×10^3	Online ↗	[1,33,34,44,53–55]
[25]	⊥ (Juárez)	Ca. 2014	Web	Alexa top sites, volunteer browsing	Index page, visited pages	TBB (2/3.X); Selenium	4.3×10^4	200	≈ 40	3.5×10^4	On request	[25]
[55]	⊥ (Wang)	2014	Web	Sensitive sites, Alexa top sites	Index page	tor 0.3.6.4; TBB 3.6.4	9×10^3	100	40	5×10^3	No	[55]
[34]	RND-WWW	Ca. 2016	Web	Twitter, Alexa one-click, Google Trends, Google Random, censored sites	Random subpage	TBB 3.6.1; Chickenfoot; iMacros; Scriptish	2.1×10^5	1125	40	2.1×10^5	Dead link ↗	[34]
-	TOR-Exit	-	-	HTTP requests of real Tor users	Visited page	-	2.1×10^5			2.1×10^5	-	-
-	WEBSITES	-	-	Popular websites	Index page, random subpage	-	5.3×10^3	50	105		-	-
[17]	DS_{Tor}	Ca. 2016	Web	Alexa top sites, popular .onion sites	Index page	TBB; Selenium	1.1×10^5	85	≈ 90	1×10^5	Dead link ↗	[17,33]
[40]	AWF CW_{900}	2017	Web	Alexa top sites	Index page	tor 0.2.8.11; TBB 6.5; Selenium	2.3×10^6	900	2500		Online ↗	[5, 32, 33, 40, 44]
-	AWF Recollect	-	-	-	-	-	1×10^5	200	500		-	-
-	AWF Open	-	-	-	-	-	8×10^5	200	2000	4×10^5	-	-
[43]	DF	Ca. 2018	Web	Alexa top sites	Index page	tor-browser-selenium	1.4×10^5	95	1000	4.1×10^4	Online ↗	[32,39,43,44]
[33]	WTT-time	2018	Web	Alexa top sites	Index page	tor 0.4.0.8; tor-browser-crawler	8×10^4	100	300	5×10^4	On request	[33]
[37]	Good Enough	2020	Web	Alexa top pages, random subpage	Index page	TBB 9.0.2	2×10^4	500	20	1×10^4	Online ↗	
[52]	⊥ (Wang)	2019	Web	Alexa top sites	Index page	tor 0.4.0.1; TBB 8.5a7	1×10^5	100	200	8×10^4	Partially Online ↗	[52]
-	Wikipedia	-	-	Wikipedia browsing	Random subpage	-	2×10^4	100	100	1×10^4	-	-
[32]	GDLF-25	Ca. 2021	Web	Alexa top sites	Random subpage	tor-browser-crawler	9.4×10^4	2400	39		On request	[32]
-	GDLF-OW	-	-	Links from Rimmer et al. [40]	Random subpage	-	7×10^4			7×10^4	-	-
[29]	BigEnough	2021	Web	Open PageRank top pages	Index page	TBB	3.8×10^4	950	20	1.9×10^4	On request	
[13]	Multi-tab	2022	Web	Alexa top pages	Index page (multi-tab)	TBB; Selenium	5.7×10^5				Online ↗	[13]
[21]	$D(\mathrm{tbs}, \mathrm{tor})$	2022	Web	Wikipedia browsing	Random subpage	tor-browser-selenium	2×10^4	98	200		Online ↗	
[4]	Drift	Ca. 2023	Web	Popular websites, links from Rimmer et al. [40]	Index page	TBB 11.0.10; tor-browser-selenium 0.6.3	1.5×10^4	90	≈ 110	5×10^3	Online ↗	[4]
	GTT23	2023	Any	Real Tor usage	Visited service	Real client software	1.4×10^7	$\langle$ 1.1×10^6 domains $\rangle$			On request	
[30]	ALEXA-WSC-FG/BG	Ca. 2024	Web	Alexa top sites, random subpage	Random subpage	TBB 7.5.6	8.6×10^5	9000	90	4.5×10^4	No	[30]
[56]	CW/OW	Ca. 2024	Web	Alexa top sites, random subpage	Random subpage (multi-tab)	TBB	8.1×10^4	1000	10	9.3×10^3	Online ↗	[56]
[42]	D1–D7	2024	Web	Tranco top sites	Index page	TBB 10.5; Chrome 112.0	7.4×10^5	100	700	4.00×10^3	Online ↗	[42]

Summary of Comparisons to Synthetic Datasets

- Most synthetic datasets contain traces of **index pages** fetched with automated tools
- GTT23 is **larger** than any existing WF dataset by an **order of magnitude**
- No other WF dataset contains **genuine** traces of **real Tor user behavior**

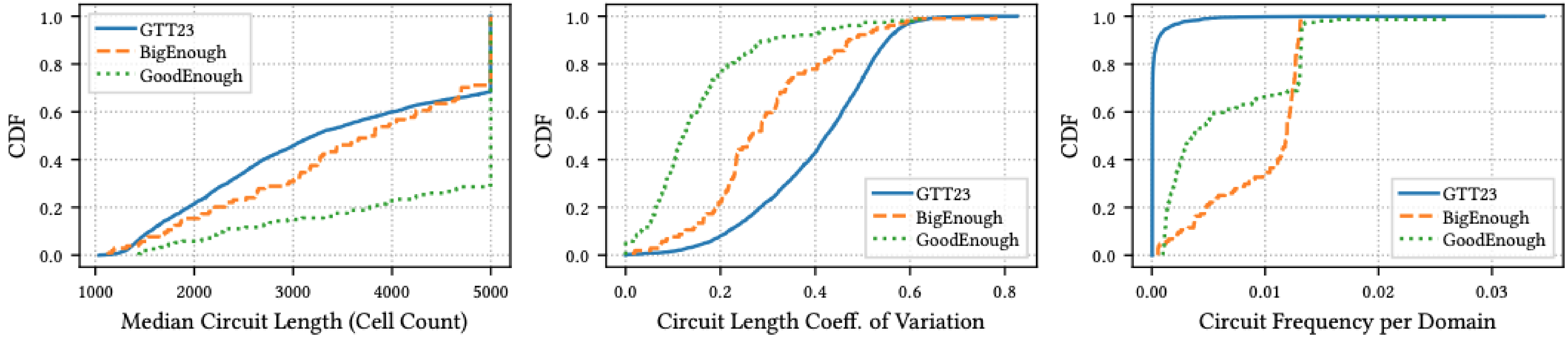


Dataset	Year	Size	Description[†]
k-NN [57]	2014	1.4×10^4	Web, top index pages
AWF CW_{900} [44]	2017	2.3×10^6	Web, top index pages
AWF Open [44]	2017	$8\ \times10^5$	Web, top index pages
DF [47]	2018	1.4×10^5	Web, top index pages
GoodEnough [41]	2020	$2\ \times10^4$	Web, top index pages + subpages
BigEnough [33]	2021	3.8×10^4	Web, top index pages + subpages
Multi-tab [13]	2022	5.7×10^5	Web, top index pages, multiple tabs
GTT23	2023	1.4×10^7	Genuine traffic, real user behavior, visited services, natural base rates

[†] All but GTT23 synthetically fetch webpages using automated tools.

Disparities between GTT23 and Synthetic Datasets

- BigEnough and GoodEnough have 10 pages per website
  - Highest web*site* **diversity** among synthetic datasets
- Compared to GTT23, data is still too homogeneous
  - **Chosen domains** are over-represented, **traffic variation** is still too low



A Measurement of Genuine Tor Traces for Realistic Website Fingerprinting

Contributions – GTT23

- The first dataset of labeled genuine Tor traces
  - **13.9M** traces, **1.1M** unique domains, **68** unique ports
  - An order-of-magnitude **larger** than existing WF datasets
- Analysis of its statistical **properties**
- Analysis of **disparities** between characteristics of **genuine** and **synthetic** datasets

Impact and Future Work

- Already has ~30 approved users
  - E.g., used to study **trace transduction** [WPES'24, NDSS'26]
- Genuine traces could inform the study of:
  - **WF** attacks/defenses, **end-to-end correlation** attacks/defenses, Tor **user** patterns, Tor **performance** characteristics, …

Read the Paper!

Access the Dataset!

Contact:
robert.g.jansen7.civ@us.navy.mil
robgjansen.com