MORE SPACE, LESS PRIVACY?

Effectiveness of Fingerprinting in IPv6-enabled Websites

Muhammad Sumeer Ahmad

HEXLAB

Stony Brook University

Background

DoT/DoH

IP-based
Fingerprinting

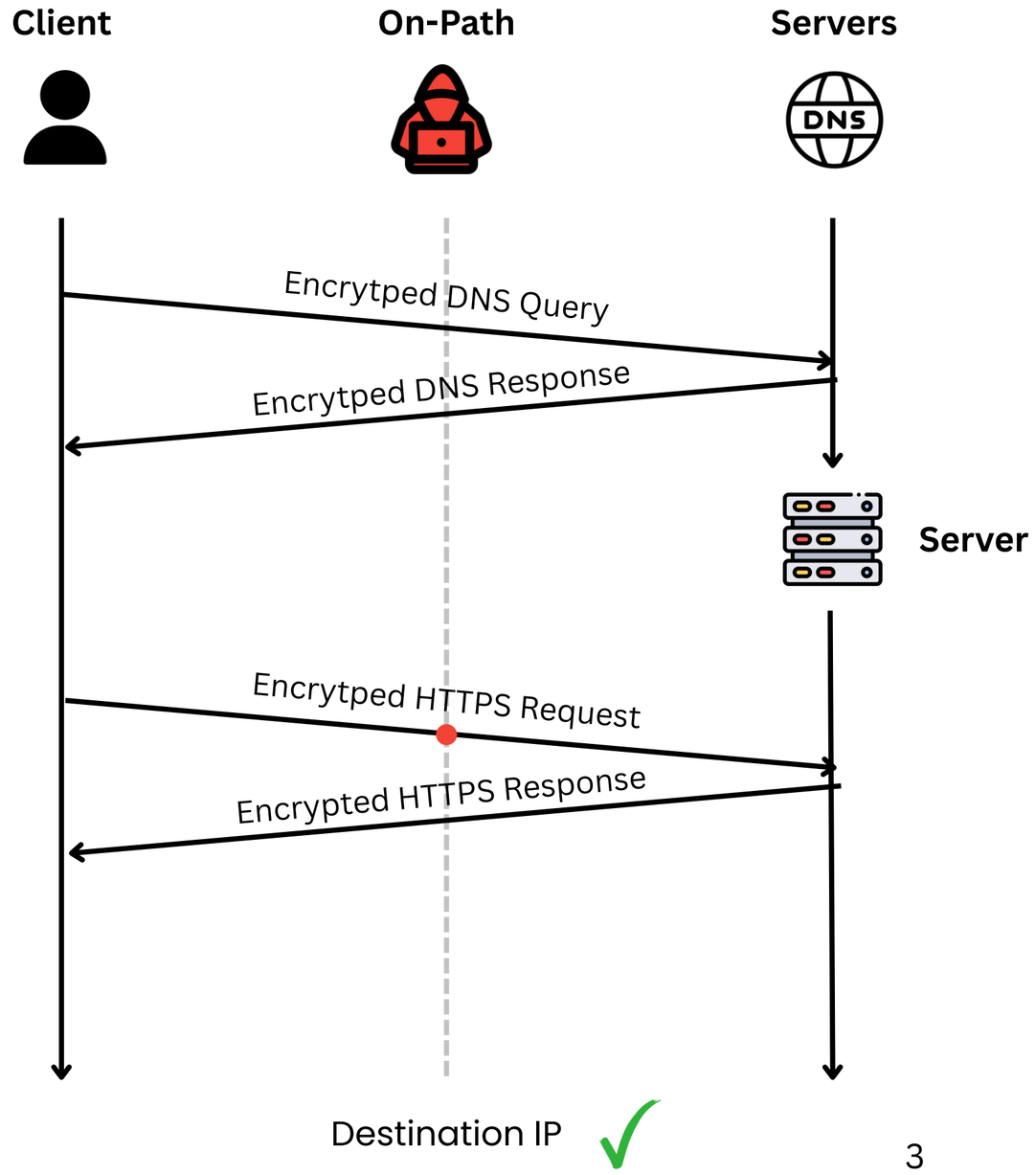
Website Fingerprinting

- DNS encryption only hides domain names
- WF still possible via metadata:
  1. Packet size & frequency
  2. Website oracles
  3. Sequence of IP connections

IP-based Fingerprinting

IPv4 ✓

IPv6 ✗

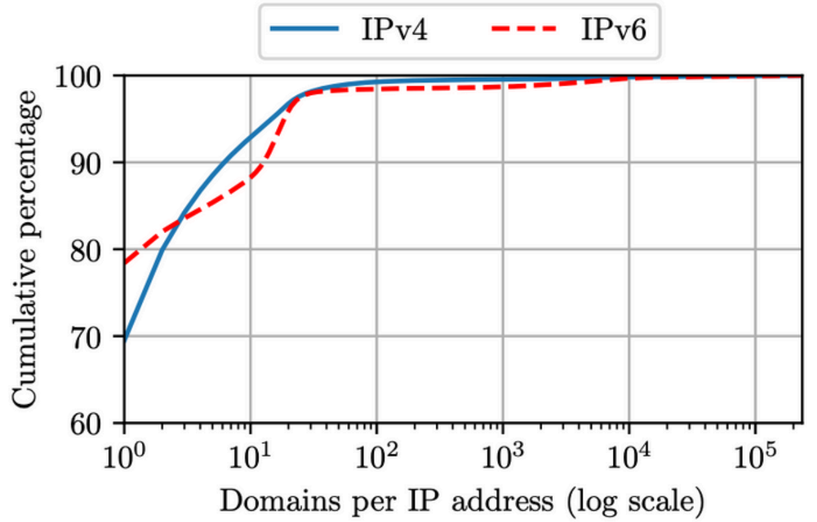


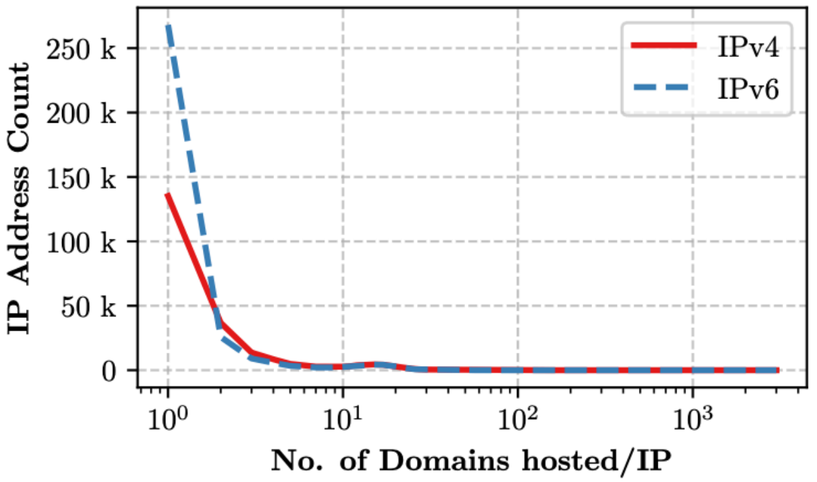
The IPv6 Landscape

- Huge address space → no need for virtual hosting
- End-to-end connectivity

Community Concern

- Most IPv6 addresses host a single domain as compared to IPv4. [1]
- IPv6 may provide better fingerprintability?





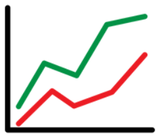
[1] https://www3.cs.stonybrook.edu/~mikepo/papers/esni.asiaccs20.pdf

Research Questions

RQ1: *What are the trends of IPv6 adoption in websites?*
- Across website categories, CDNs, resources

RQ2: *What is the effectiveness of IPv6-based WF?*
- Compare IPv6 vs IPv4 in IP-based WF (accuracy)

RQ3: *Is IPv6 fingerprinting stable and scalable?*
- Entropy-based, computationally efficient, longitudinal analysis



RQ4: *What are the patterns of IPv6 deployment accross providers?*
- Domain co-location patterns in IPv6 vs IPv4

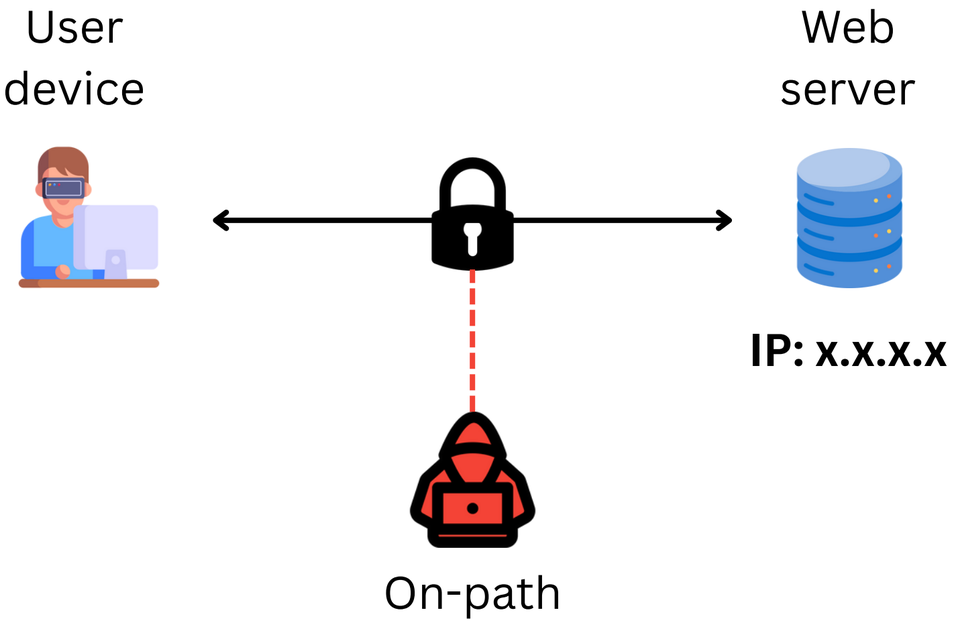
Threat Model

Adversary Capabilities:

- On-path observer (ISP, censor, network admin)
- Sees destination IP addresses per session
- Can capture sequential browsing sessions

Assumptions:

1. All DNS queries are encrypted (DoT/DoH)
2. Users mostly browse one site at a time (single tab)
3. No use of flow metadata, TLS fingerprints, or timing



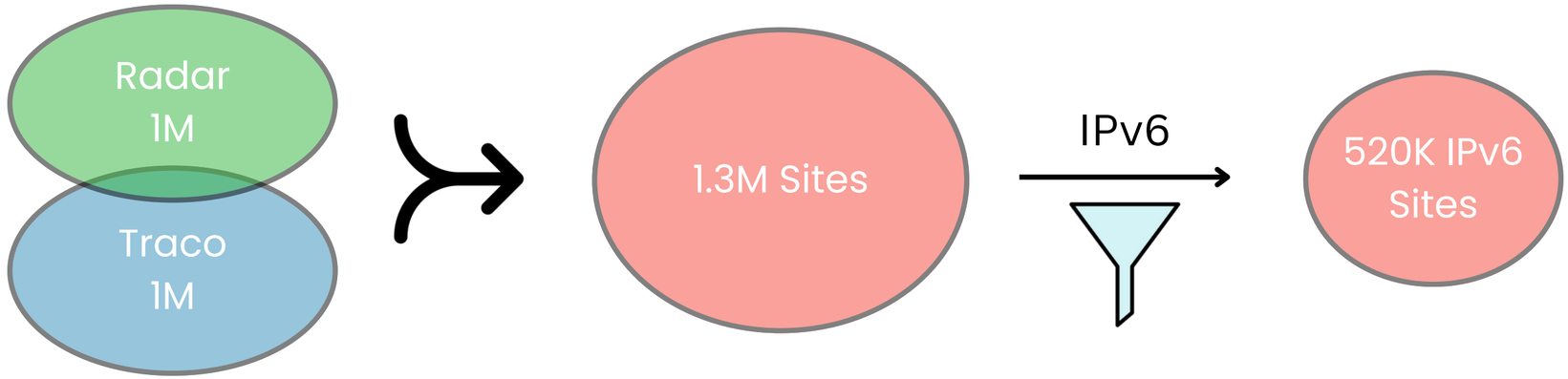
Target Websites

Sources:

a. Tranco (Top 1M)
b. Cloudflare Radar (Top 1M)

IPv6 Filtering:

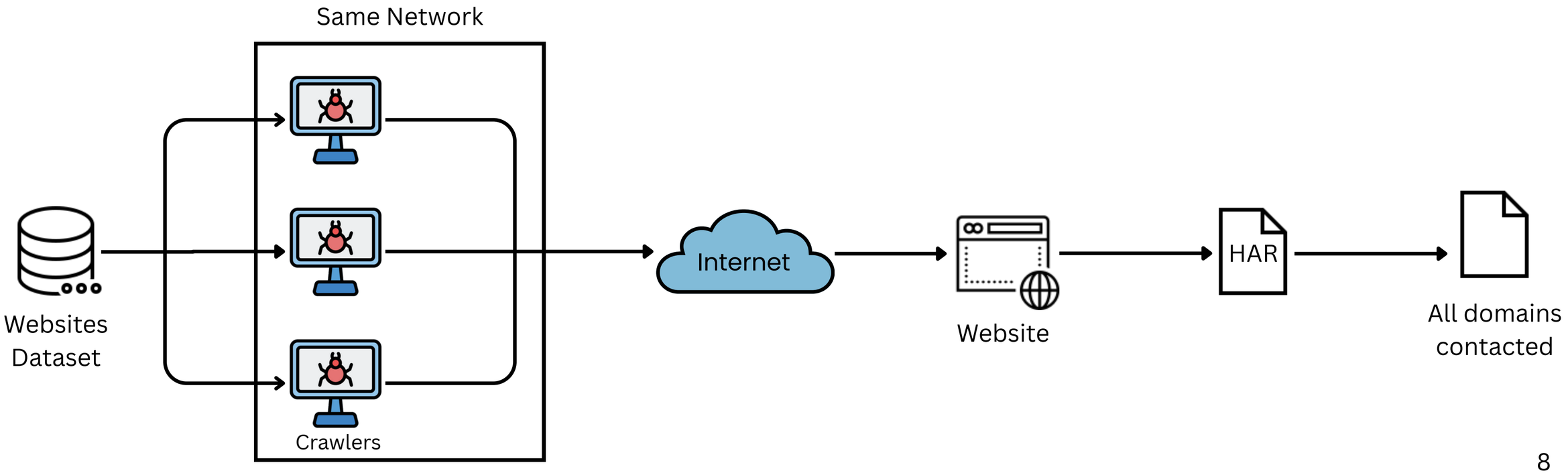
Filter domains with valid AAAA record on public DoH/DoT resolvers (1.1.1.1/8.8.8.8)



Measurement Set up

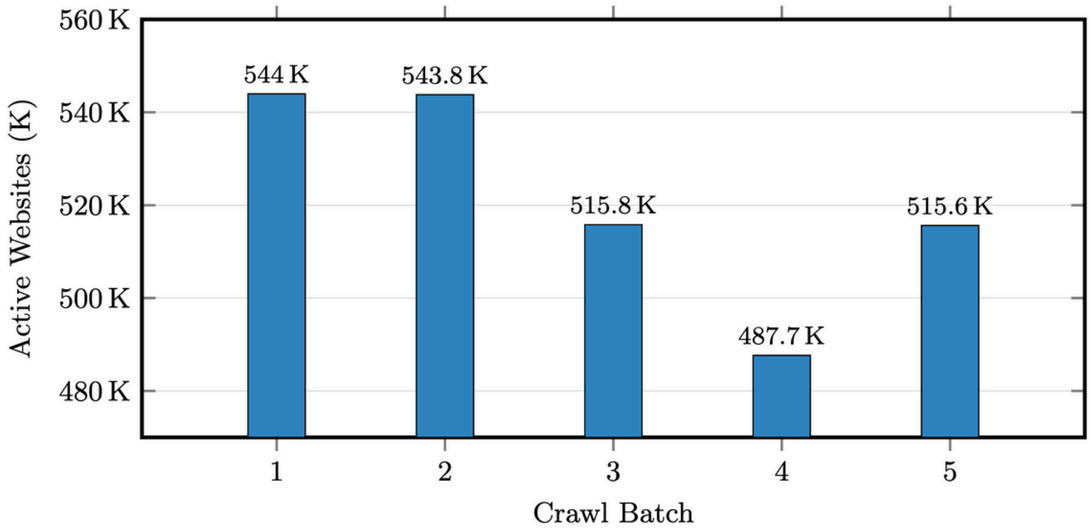
- Adversary is regionally colocated
- Crawlers in the same network

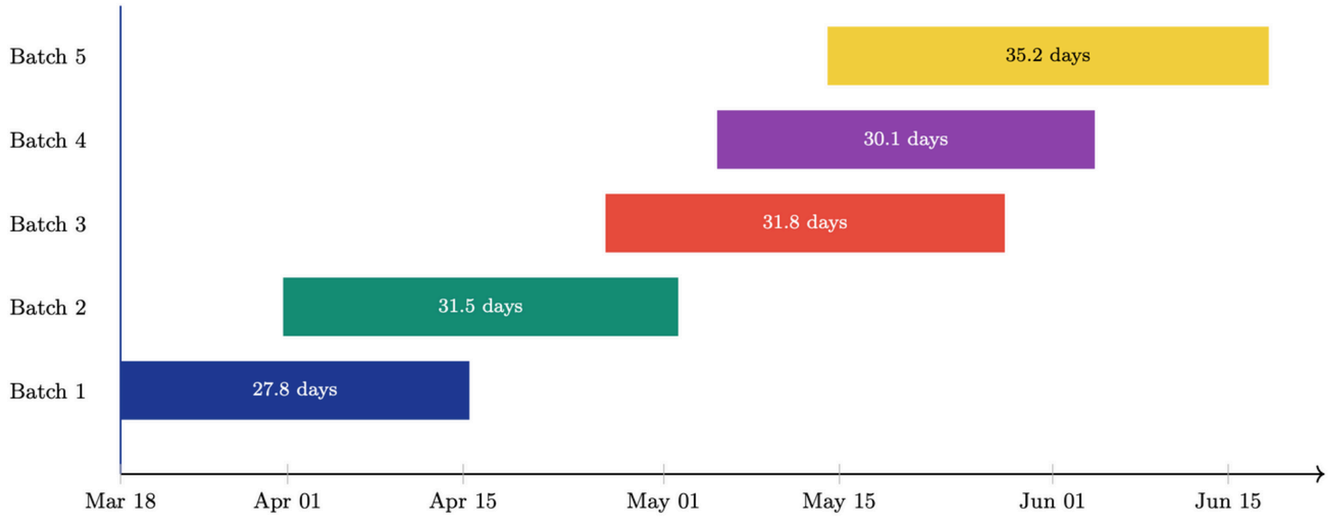
- Browser only affects loading sequence
- Capture HAR files → domains contacted



Crawling Process

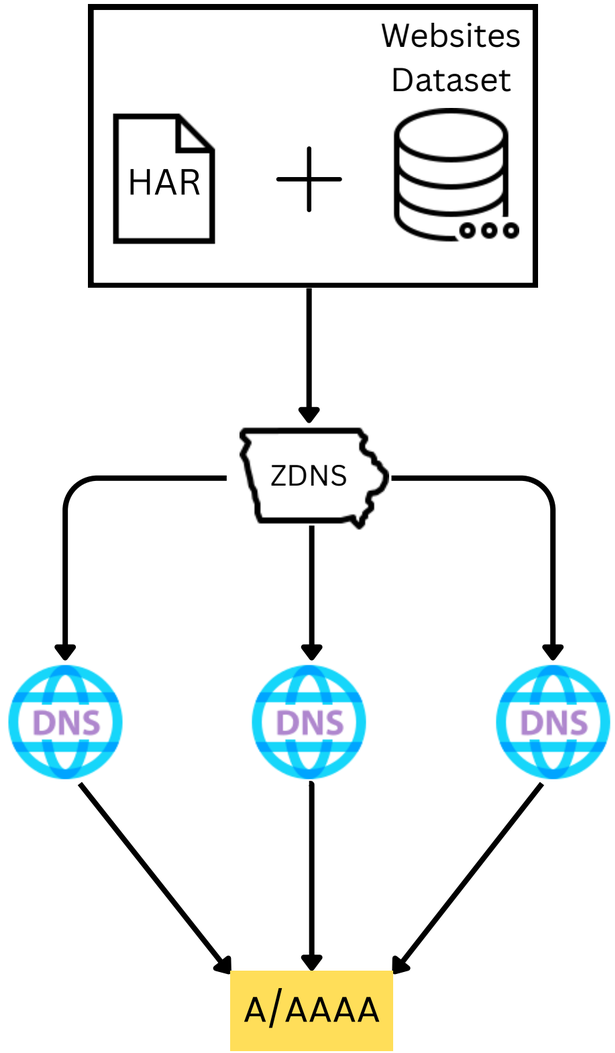
- Dataset crawled in batches over 3 months
- 5 crawl batches, ~15 days apart, ~1 month each
- Each site visited 5 times in total



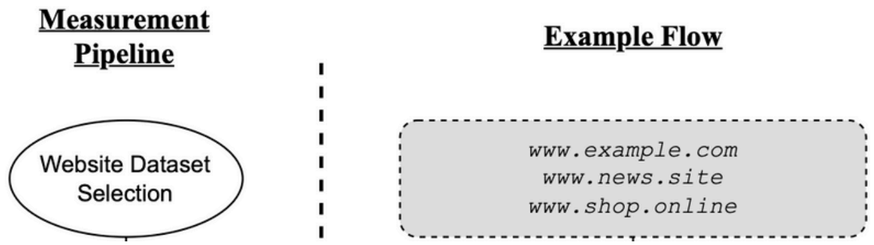


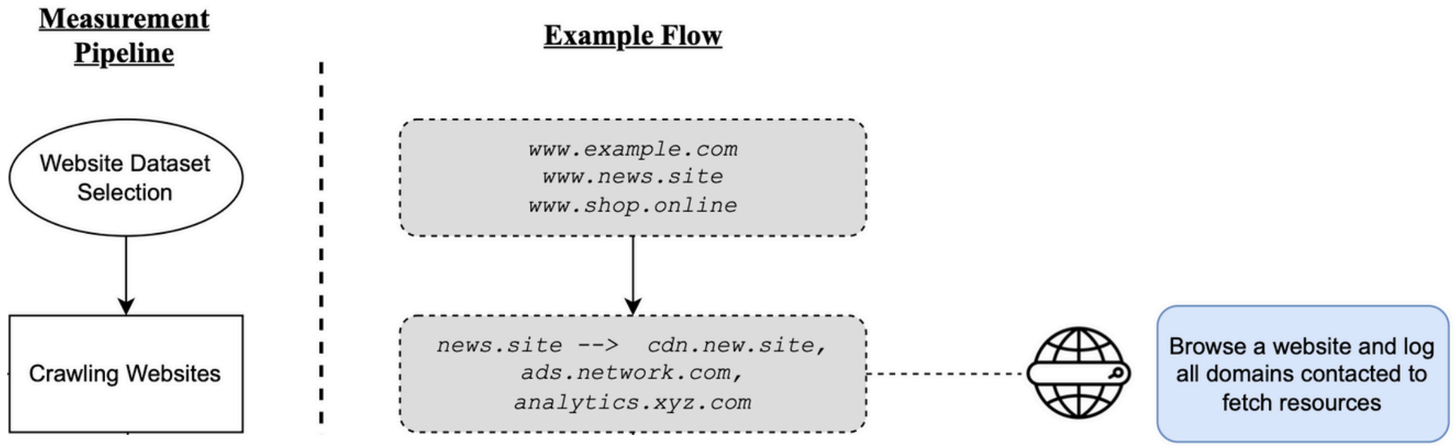
Active DNS Resolution

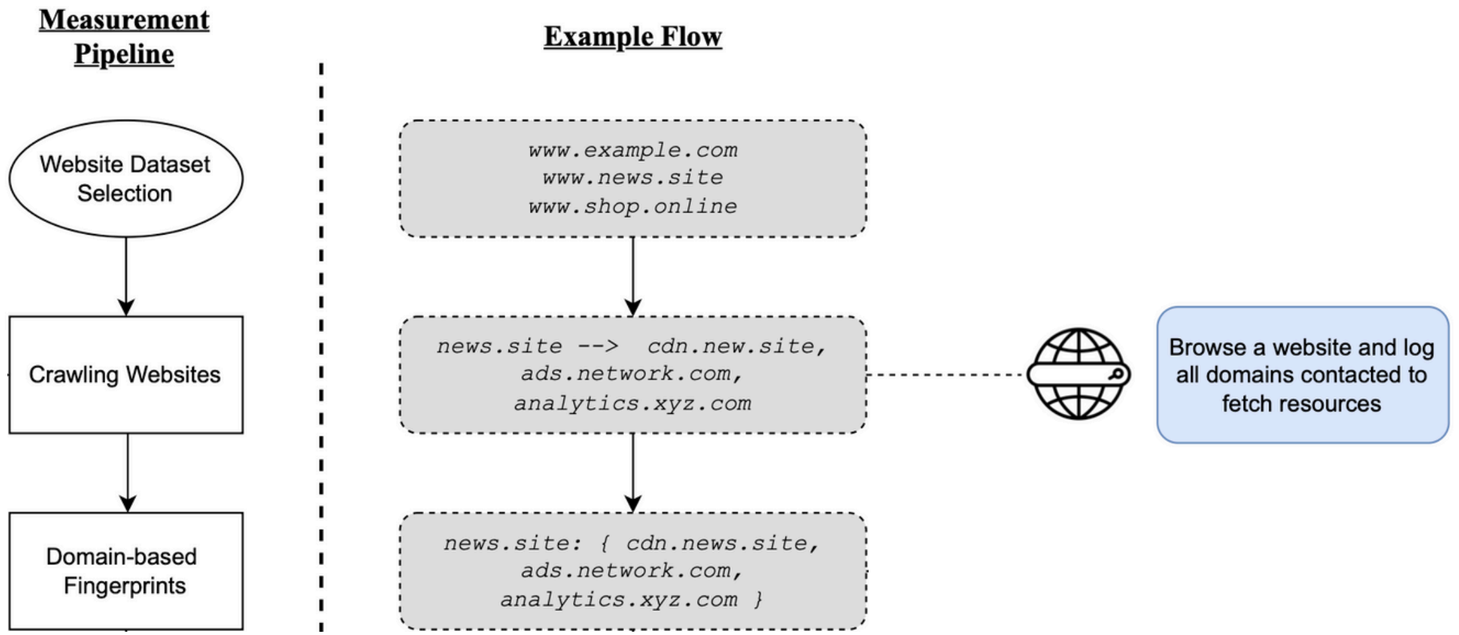
- Captured ~850K domains per batch
- Hourly DNS queries in a round-robin fashion
- Queried 6 encrypted public DNS resolvers:
  - Google (8.8.8.8)
  - Cloudflare (1.1.1.1)
  - OpenDNS
  - Quad9 (9.9.9.9)
  - Comodo
  - CleanBrowsing



METHODOLOGY

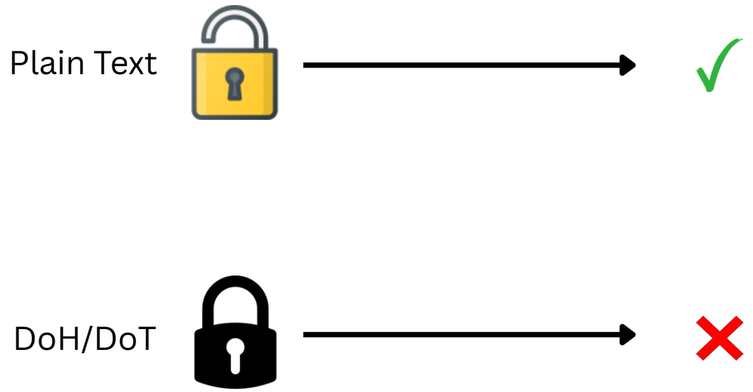

Measurement Pipeline
Example Flow
Website Dataset Selection
www.example.com
www.news.site
www.shop.online


Measurement Pipeline
Website Dataset Selection
Crawling Websites
Example Flow
www.example.com
www.news.site
www.shop.online
news.site --> cdn.new.site,
ads.network.com,
analytics.xyz.com
Browse a website and log all domains contacted to fetch resources


Measurement Pipeline
Website Dataset Selection
Crawling Websites
Domain-based Fingerprints
Example Flow
www.example.com
www.news.site
www.shop.online
news.site --> cdn.new.site,
ads.network.com,
analytics.xyz.com
news.site: { cdn.news.site,
ads.network.com,
analytics.xyz.com }
Browse a website and log all domains contacted to fetch resources

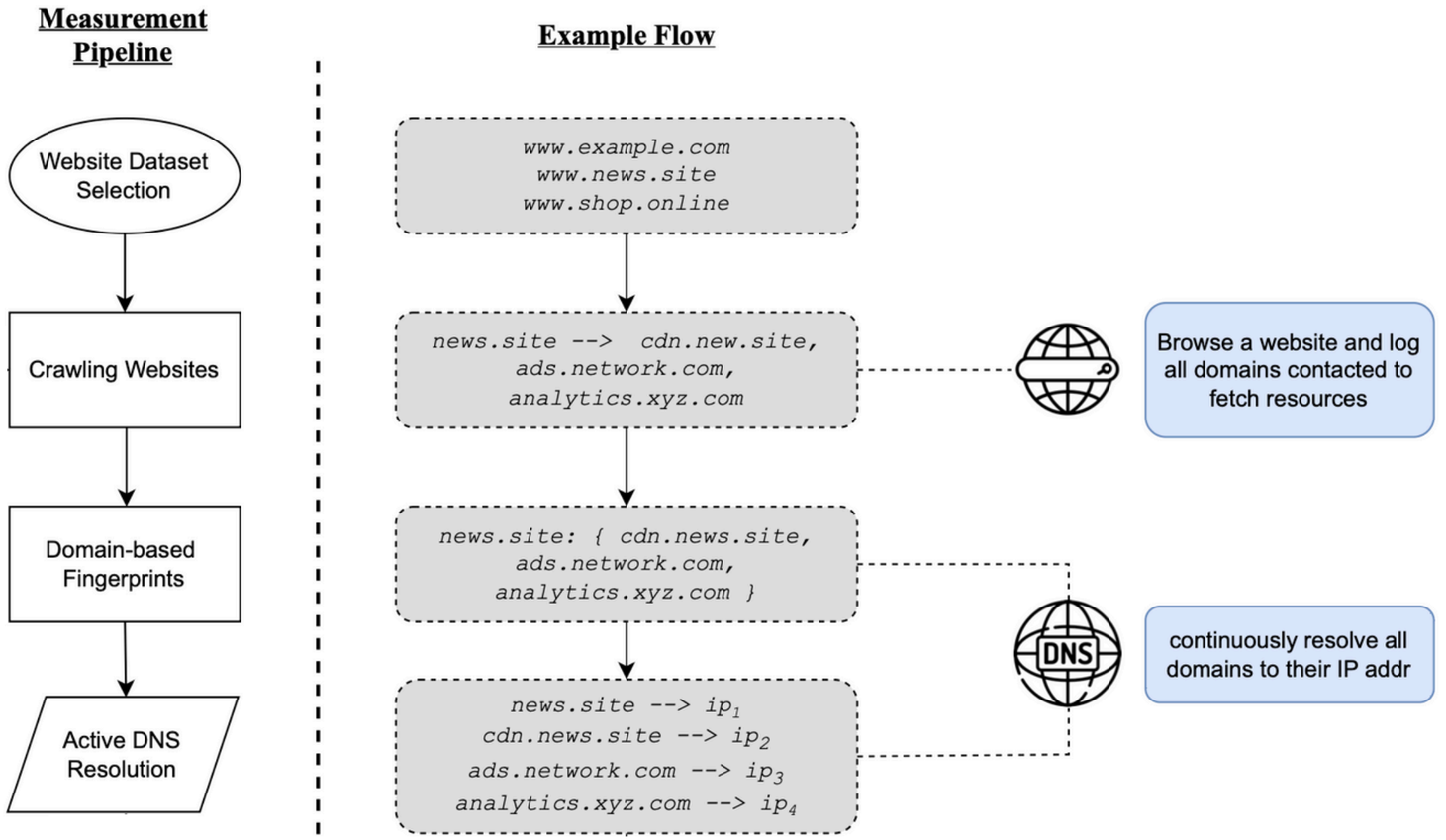
Domain-based Fingerprints

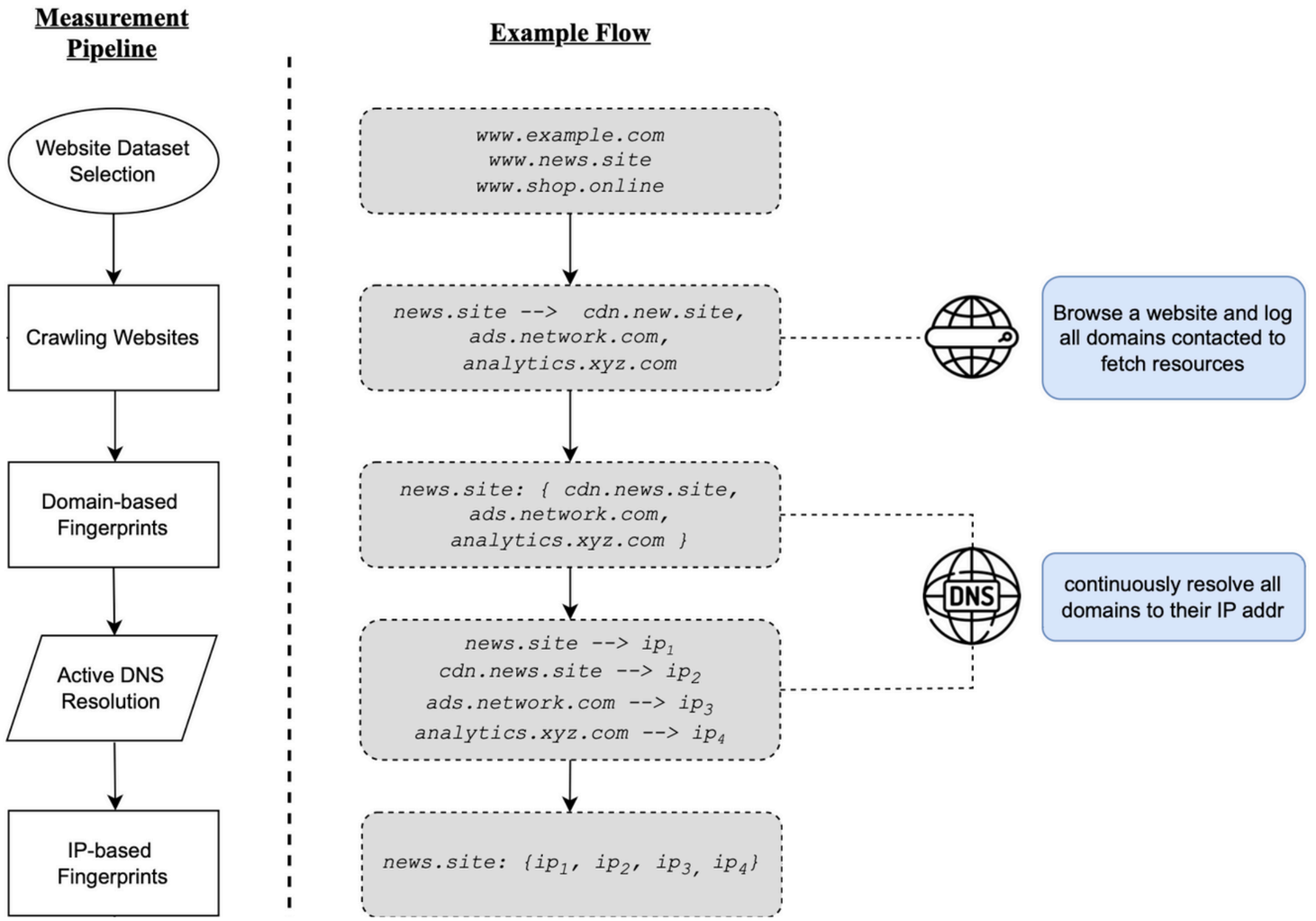
- *Primary domain* = website visited
- *Secondary domains* = additional resources (images, scripts, ads, etc.)
- Unordered sets, no sequences
- Order depends on:
  - Network configuration
  - Browser caching & speculative parsing
  - Resource prioritization



$$\mathcal{D}_w = d_p \cup \{d_{s_1}, d_{s_2}, \ldots, d_{s_n}\}$$

{0: {'stonybrook.edu'}, 1: {'s.adroll.com', 'script.crazyegg.com', 'eb2.3lift.com', 'analytics.twitter.com', 'a.us.silktide.com', 'ups.analytics.yahoo.com', 'tr.snapchat.com' ...}}


Measurement Pipeline
Website Dataset Selection
Crawling Websites
Domain-based Fingerprints
Active DNS Resolution
Example Flow
www.example.com
www.news.site
www.shop.online
news.site --> cdn.new.site,
ads.network.com,
analytics.xyz.com
Browse a website and log all domains contacted to fetch resources
news.site: { cdn.news.site,
ads.network.com,
analytics.xyz.com }
DNS
continuously resolve all domains to their IP addr
news.site --> ip1
cdn.news.site --> ip2
ads.network.com --> ip3
analytics.xyz.com --> ip4

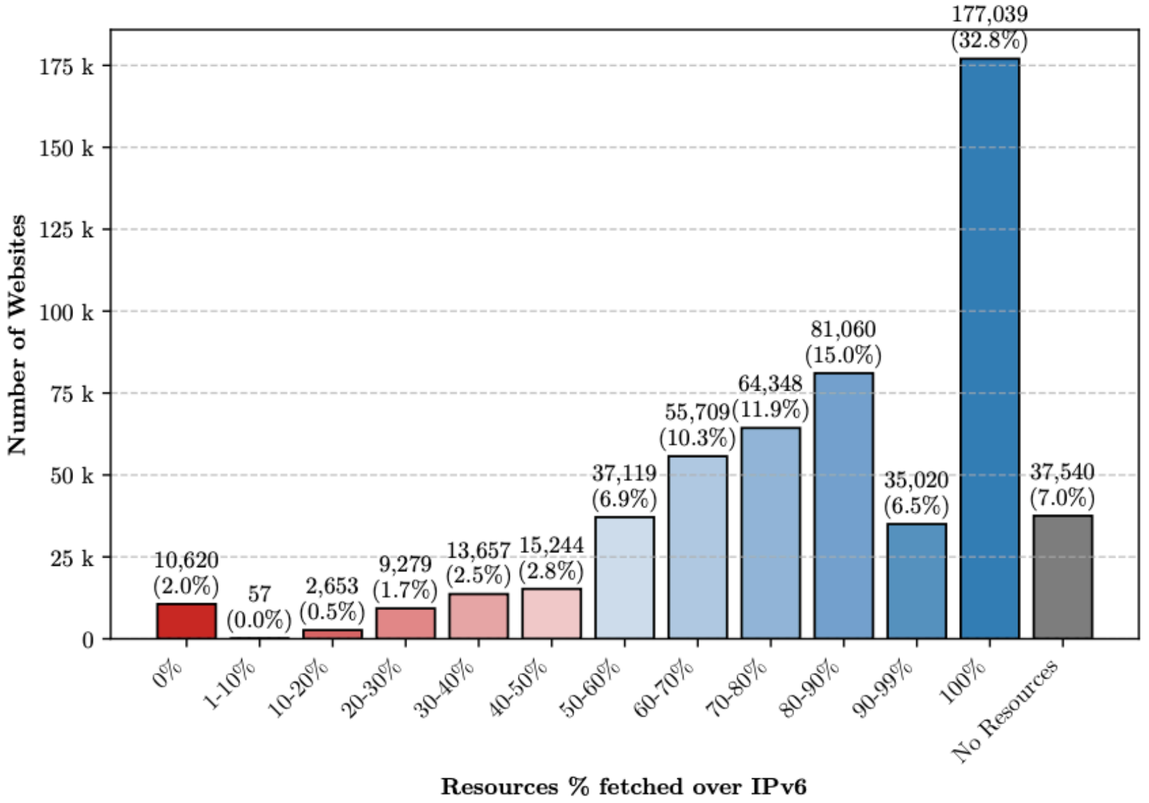

Measurement Pipeline
Website Dataset Selection
Crawling Websites
Domain-based Fingerprints
Active DNS Resolution
IP-based Fingerprints
Example Flow
www.example.com
www.news.site
www.shop.online
news.site --> cdn.new.site,
ads.network.com,
analytics.xyz.com
Browse a website and log all domains contacted to fetch resources
news.site: { cdn.news.site,
ads.network.com,
analytics.xyz.com }
DNS
continuously resolve all domains to their IP addr
news.site --> ip1
cdn.news.site --> ip2
ads.network.com --> ip3
analytics.xyz.com --> ip4
news.site: {ip1, ip2, ip3, ip4}

IP-based Fingerprints

1. Domain-based → IP-based using DNS.
2. Build two versions:
   a. IPv4 Fingerprint (Fv4): all A records
   b. IPv6 Fingerprint (Fv6): AAAA preferred, fallback to A

IPv6 adoption among resources:

1. 32% sites fetch all resources over IPv6
2. 7% IPv6 sites fetch no resources
3. IPv6 landing page only: 2%



IP-based Fingerprints

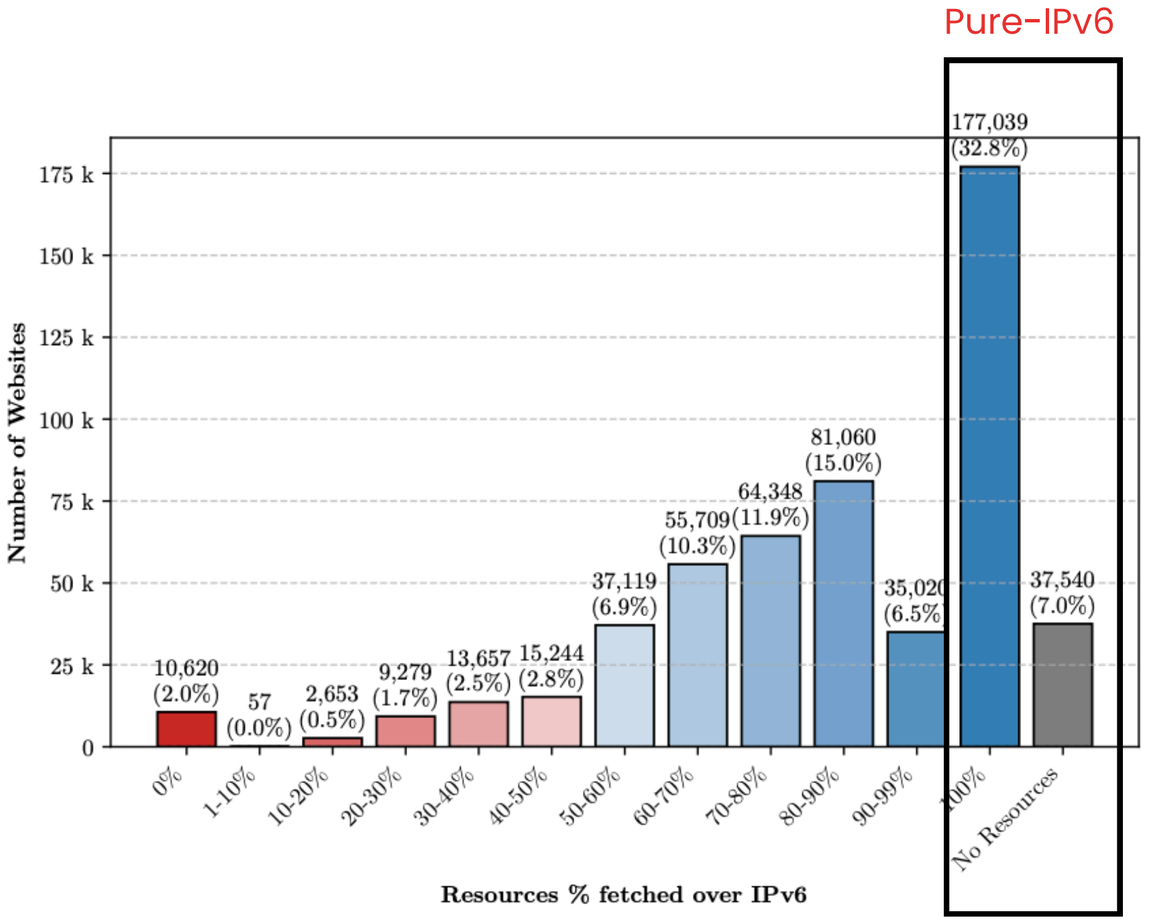
1. Domain-based → IP-based using DNS.
2. Build two versions:
   a. IPv4 Fingerprint (Fv4): all A records
   b. IPv6 Fingerprint (Fv6): AAAA preferred, fallback to A

IPv6 adoption among resources:

1. 32% sites fetch all resources over IPv6
2. 7% IPv6 sites fetch no resources
3. IPv6 landing page only: 2%

Categorization:

- *Pure-IPv6 sites (56.4%)*
- *Mixed-IPv4/v6 sites (43.6%)*



IP-based Fingerprints

1. Domain-based → IP-based using DNS.
2. Build two versions:
   a. IPv4 Fingerprint (Fv4): all A records
   b. IPv6 Fingerprint (Fv6): AAAA preferred, fallback to A

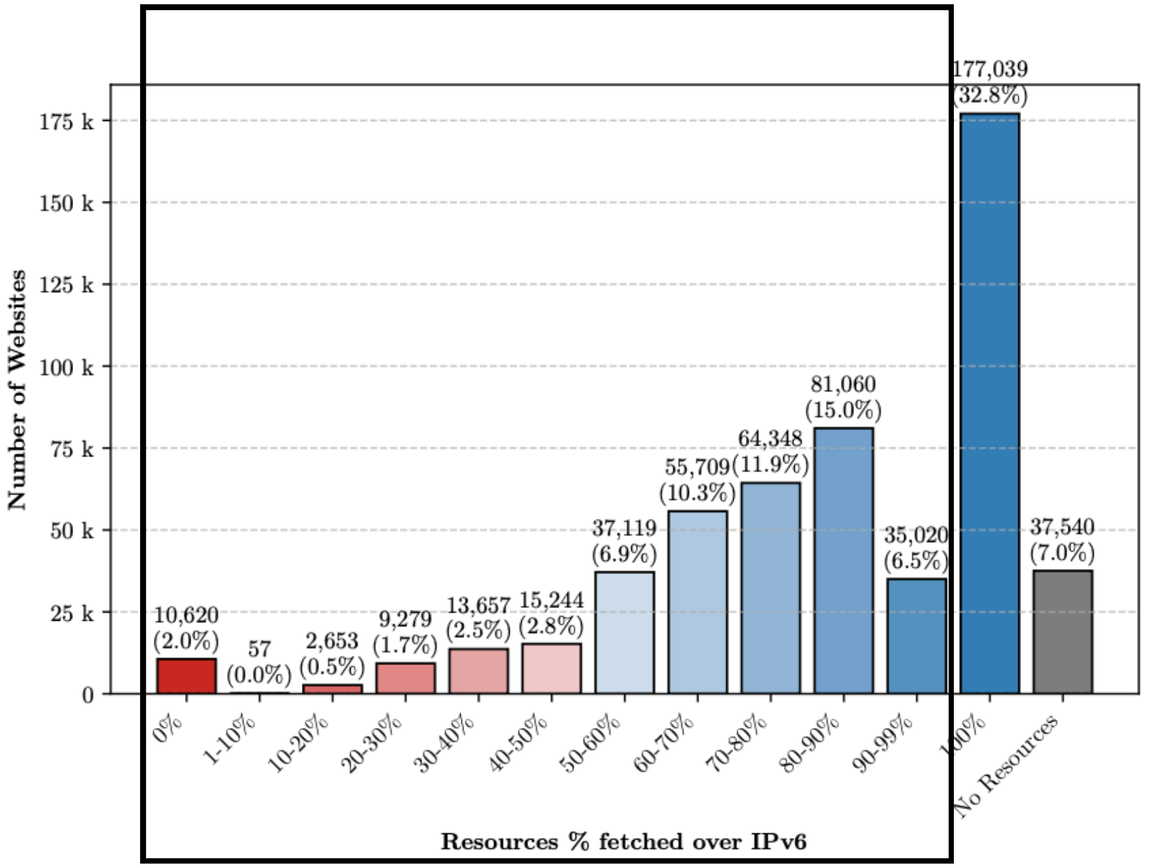
IPv6 adoption among resources:

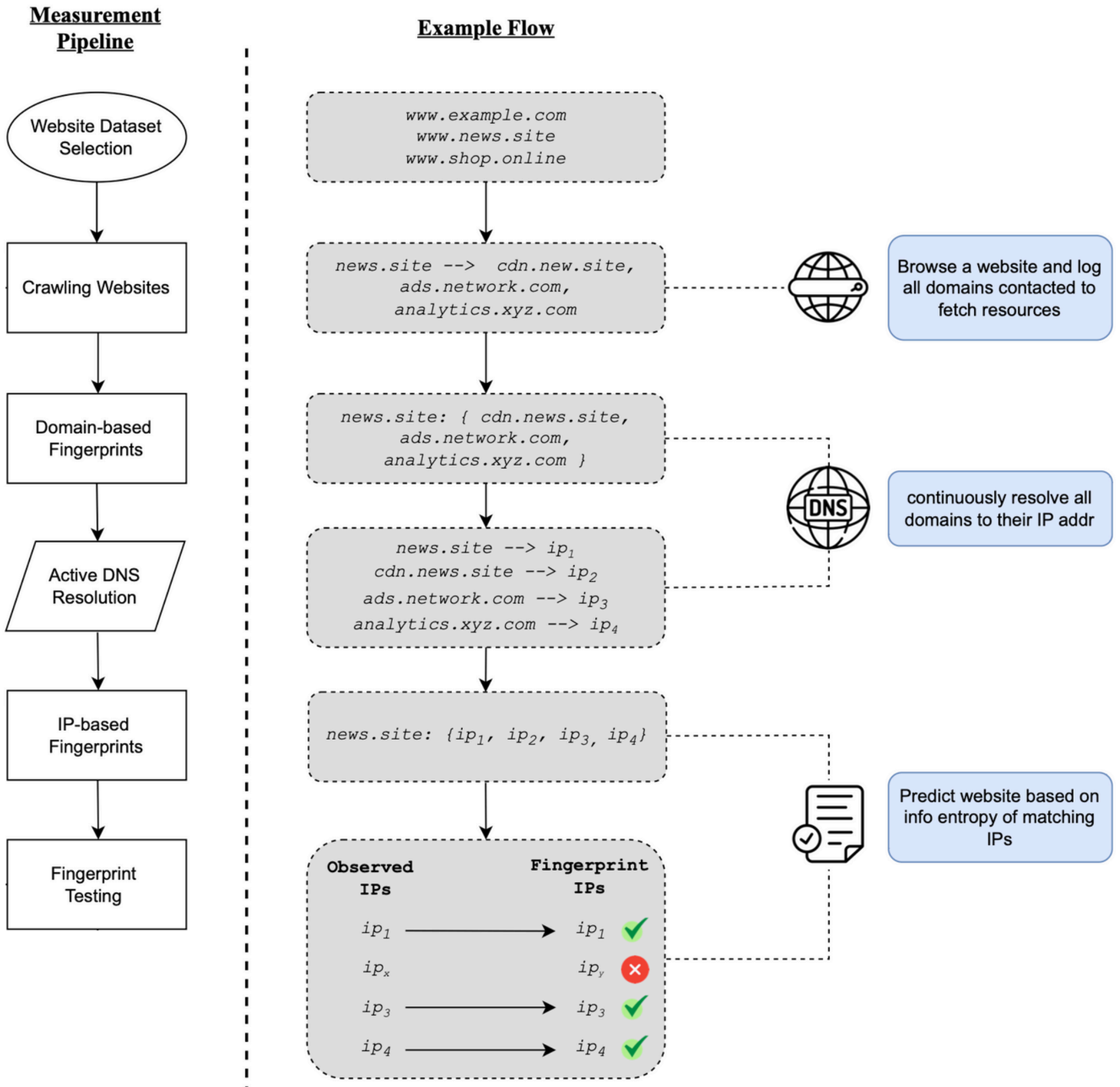
1. 32% sites fetch all resources over IPv6
2. 7% IPv6 sites fetch no resources
3. IPv6 landing page only: 2%

Categorization:

- *Pure-IPv6 sites (56.4%)*
- *Mixed-IPv4/v6 sites (43.6%)*

Mixed IPv4/v6 Sites




Measurement Pipeline
Website Dataset Selection
Crawling Websites
Domain-based Fingerprints
Active DNS Resolution
IP-based Fingerprints
Fingerprint Testing
Example Flow
www.example.com
www.news.site
www.shop.online
news.site --> cdn.new.site, ads.network.com, analytics.xyz.com
Browse a website and log all domains contacted to fetch resources
news.site: { cdn.news.site, ads.network.com, analytics.xyz.com }
DNS
continuously resolve all domains to their IP addr
news.site --> ip1
cdn.news.site --> ip2
ads.network.com --> ip3
analytics.xyz.com --> ip4
news.site: {ip1, ip2, ip3, ip4}
Predict website based on info entropy of matching IPs
Observed IPs
Fingerprint IPs
ip1
ip1
ipx
ipy
ip3
ip3
ip4
ip4

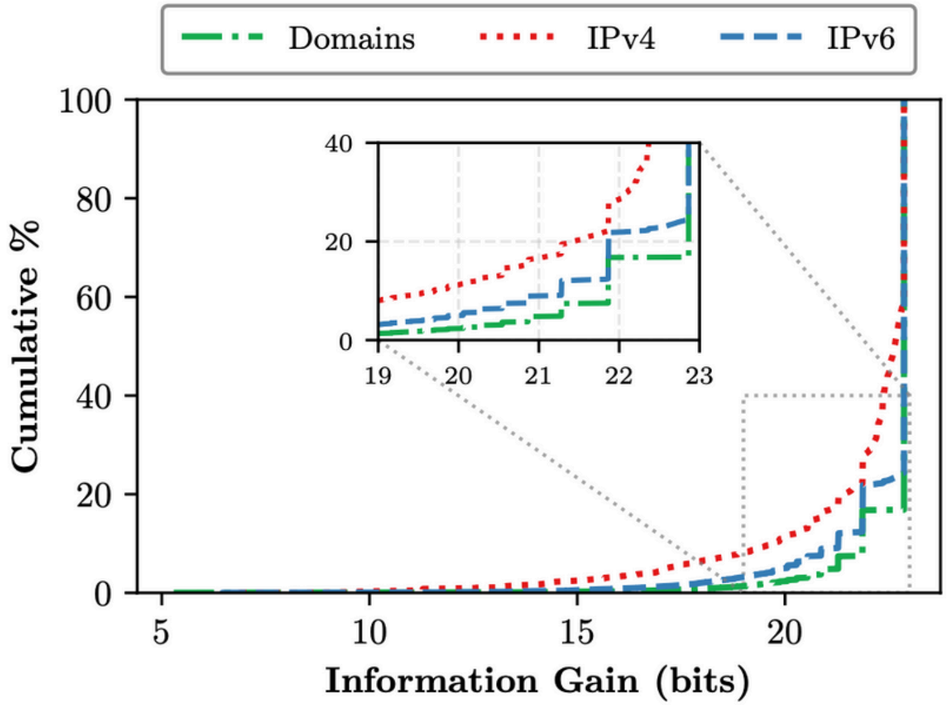
Fingerprint Matching

- Strong identifiers = domains appearing rarely across sites
- Weak identifiers = domains appearing everywhere
- *Entropy* to measure uniqueness:

$$Information\ Entropy = -\log_2 P(d)$$

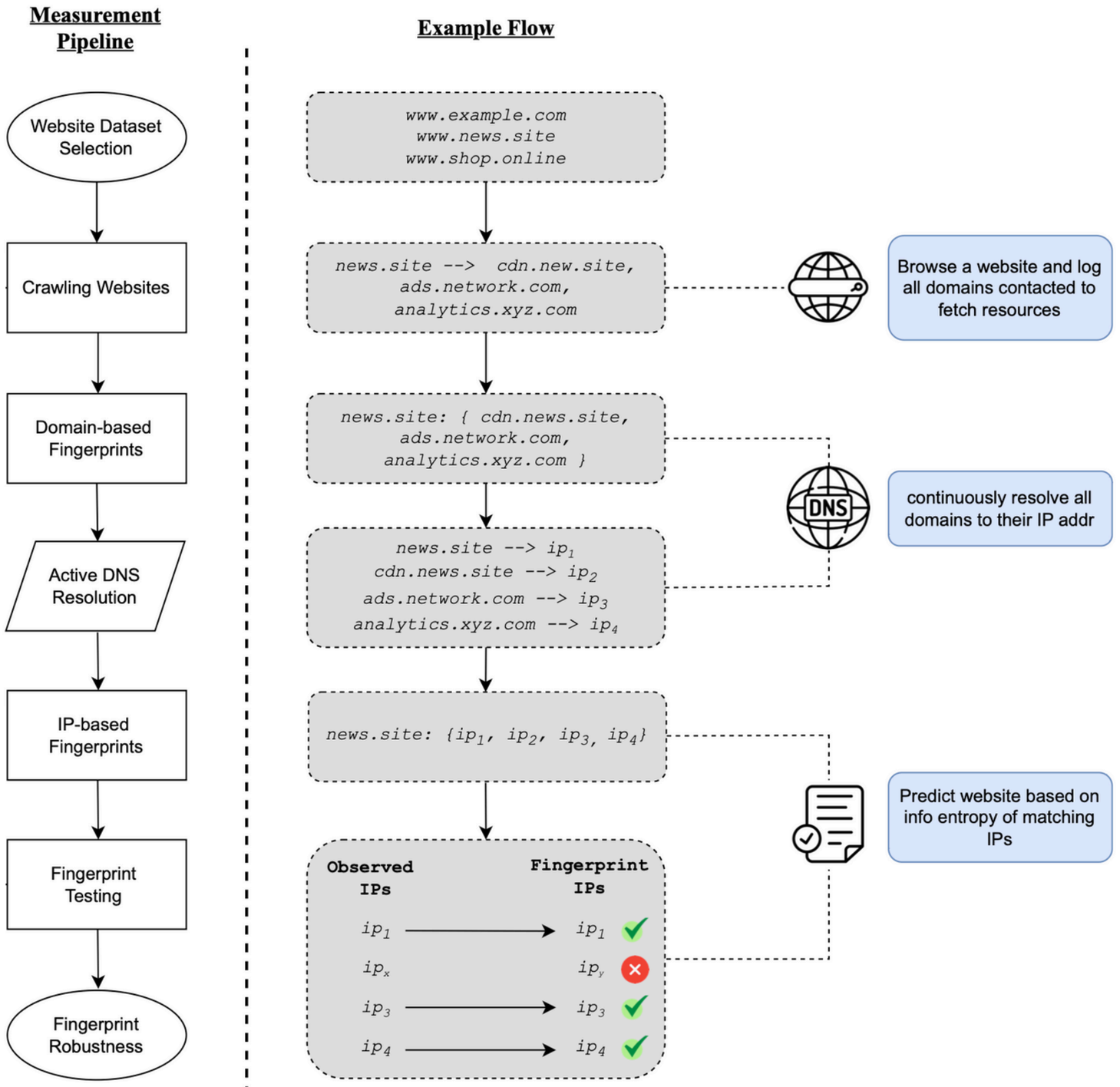
$P(d)$ = probability of a domain being contacted while visiting a given website

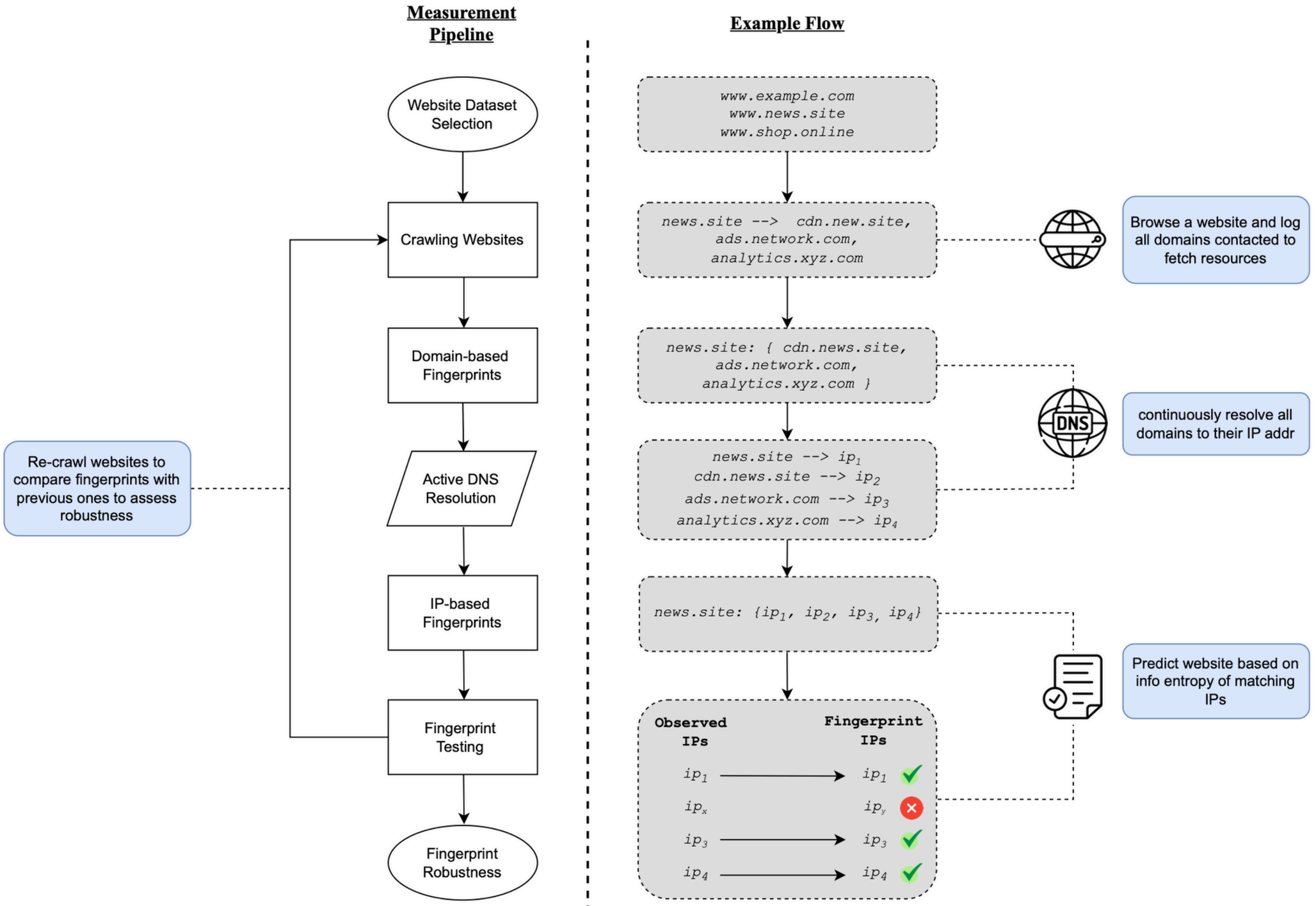
IP-entropy: avg entropy of domains they host



High Entropy → distinctive

Low Entropy → common


Measurement Pipeline
Website Dataset Selection
Crawling Websites
Domain-based Fingerprints
Active DNS Resolution
IP-based Fingerprints
Fingerprint Testing
Fingerprint Robustness
Example Flow
www.example.com
www.news.site
www.shop.online
news.site --> cdn.new.site,
ads.network.com,
analytics.xyz.com
Browse a website and log all domains contacted to fetch resources
news.site: { cdn.news.site,
ads.network.com,
analytics.xyz.com }
DNS
continuously resolve all domains to their IP addr
news.site --> ip1
cdn.news.site --> ip2
ads.network.com --> ip3
analytics.xyz.com --> ip4
news.site: {ip1, ip2, ip3, ip4}
Predict website based on info entropy of matching IPs
Observed IPs
Fingerprint IPs
ip1
ip1
ipx
ipy
ip3
ip3
ip4
ip4


Measurement Pipeline
Website Dataset Selection
Crawling Websites
Domain-based Fingerprints
Active DNS Resolution
IP-based Fingerprints
Fingerprint Testing
Fingerprint Robustness
Re-crawl websites to compare fingerprints with previous ones to assess robustness
Example Flow
www.example.com
www.news.site
www.shop.online
news.site --> cdn.new.site, ads.network.com, analytics.xyz.com
Browse a website and log all domains contacted to fetch resources
news.site: { cdn.news.site, ads.network.com, analytics.xyz.com }
DNS
continuously resolve all domains to their IP addr
news.site --> ip1
cdn.news.site --> ip2
ads.network.com --> ip3
analytics.xyz.com --> ip4
news.site: {ip1, ip2, ip3, ip4}
Predict website based on info entropy of matching IPs
Observed IPs
Fingerprint IPs
ip1
ip1
ipx
ipy
ip3
ip3
ip4
ip4

RESULTS

Accuracy Results

TABLE I
FINGERPRINTING ACCURACY IN IPV4 & IPV6 ACROSS MIXED-IPV4/V6 AND PURE-IPV6 SITES

Rankings	Mixed-IPv4/v6 Sites			Pure-IPv6 Sites		
	# Sites	IPv4 %	IPv6 %	# Sites	IPv4 %	IPv6 %
Top 100	39	89.74	89.74	61	29.51	24.59
Top 1000	540	90.37	90.19	460	56.96	49.35
Top 10K	6038	94.10	94.17	3962	64.84	54.72
Top 50k	29742	94.81	94.96	20258	65.33	56.64
Top 100K	56491	94.68	94.86	43509	62.24	53.29
Top 250K	119997	93.06	93.31	130003	56.37	44.94
Top 500K	219501	93.79	94.16	280499	56.09	44.80
All Websites	228108	93.78	94.20	295315	56.02	44.82

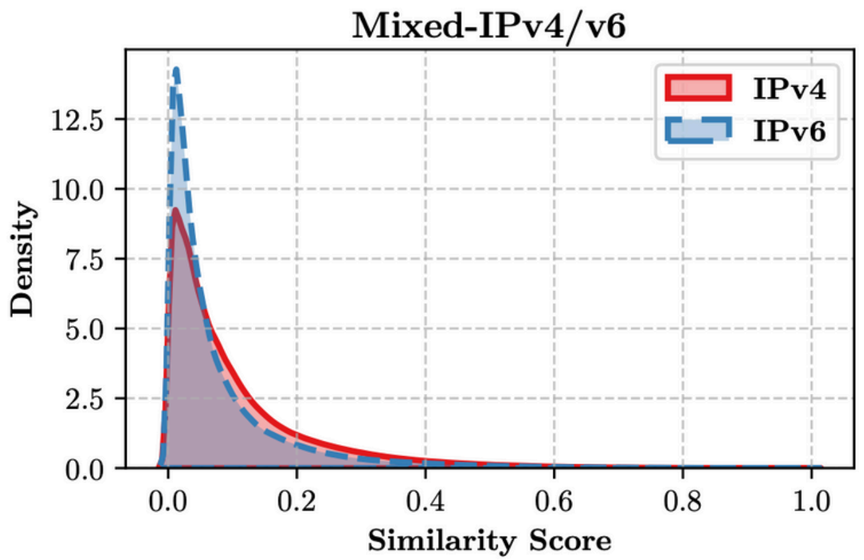
Mixed IPv4/v6 websites

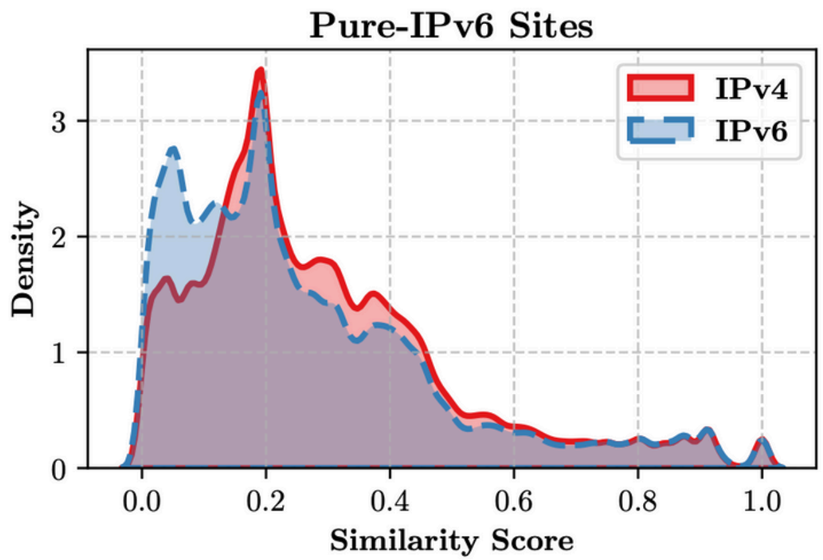
- High accuracy: 90–94% for both IPv4 & IPv6
- Negligible difference (<1%) b/w protocols

Pure IPv6 websites

- Accuracy drops significantly:
  - IPv4: ~56–66%
  - IPv6: often <45%
- Cause: ? →

Fingerprint Uniqueness





$$\mathbf{J}_{i,j} = \frac{|\mathcal{F}_i \cap \mathcal{F}_j|}{|\mathcal{F}_i \cup \mathcal{F}_j|}$$

0 = unique
1 = identical

Mixed IPv4/v6 websites

- Almost no overlap between sites
- Highly heterogeneous deployments (distinct prefixes, providers).

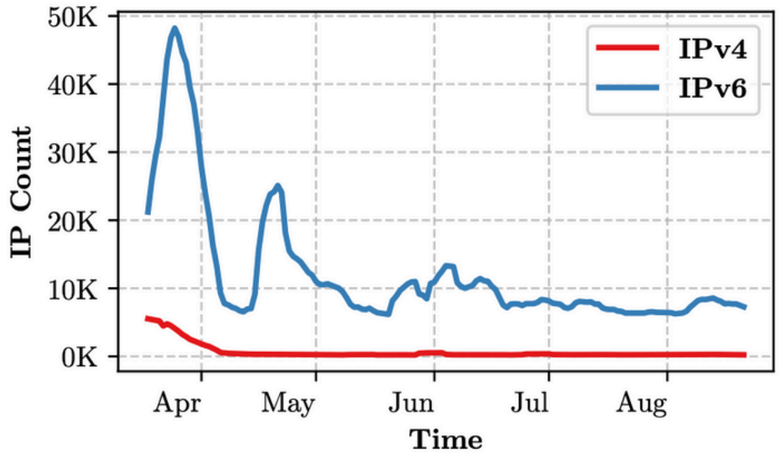
Pure-IPv6

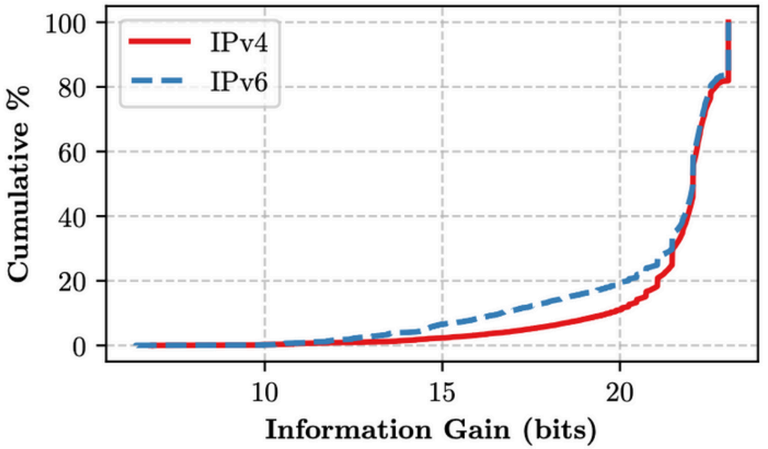
- A multi-modal pattern with moderate overlap.
- Many sites share overlapping IPv6 pools.

Pure-IPv6 sites share large provider prefixes, creating tight clusters of overlapping IPv6 addrs.

More Space, Not More Information

- IPv6 entropy is inflated by heavy address churn, creating the illusion of uniqueness
- Naive entropy overcounts these ephemeral addresses
- Realized Information Gain = only IPs actually used
- 40% IPv6 addr are less distinctive than IPv4





Realized entropy shows IPv6 fingerprints are not richer than IPv4

Hosting Infrastructure

- Fingerprinting risk → how hosting providers deploy IPv6/IPv4.
- **Observation:**
  - Pure IPv6 sites → lower accuracy,
  - Mixed IPv4/v6 → high accuracy (~94%).
- **Why?**
  - Hosting practices and co-location of domains/IP.

Co-location Degree:

- Avg # domains hosted on a single IP by a provider.

$$\textit{Co-location Degree} = \frac{1}{N}\sum_{i=1}^{N}(\#\ \textit{domains hosted on}\ ip_i)$$

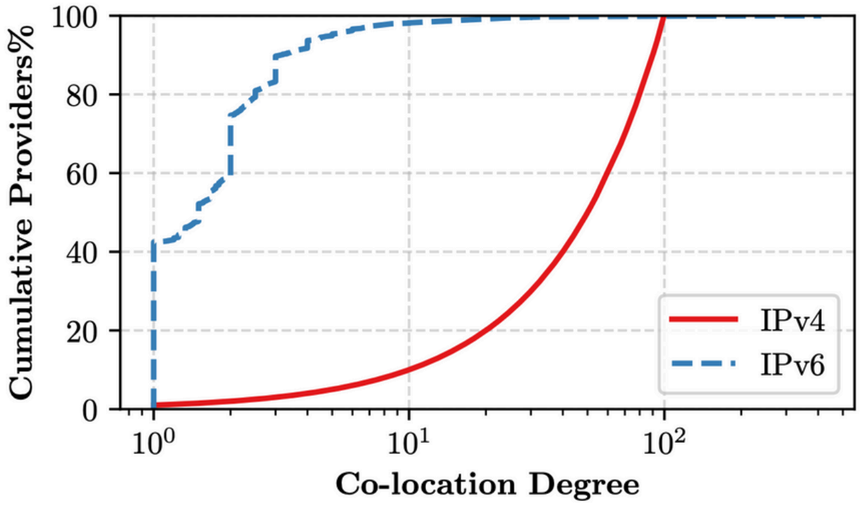


TABLE III
TOP IPV6 HOSTING PROVIDERS AND THEIR COLOCATION DEGREE

ASN	Provider Name	Colocation
AS15169	GOOGLE	65.74
AS54113	FASTLY	59.41
AS13335	CLOUDFLARENET	12.37
AS9123	JSC Timeweb	9.76
AS8560	IONOS SE	7.40
AS396982	GOOGLE-GCP	6.88
AS209242	Cloudlfare London LLC	6.14
AS16276	OVH SAS	5.80
AS20940	Akamai Int'l B.V.	5.43
AS14061	DIGITALOCEAN	4.83
AS24940	Hetzner Online GmbH	2.30
AS47583	Hostinger Int'l Ltd	1.16
AS16509	AMAZON	1.13

Provider Specific Trends

TABLE II
TOP 10 IPV6 HOSTING PROVIDERS IN MIXED-IPV4/IPV6 & PURE-IPV6 SITES

Rank	mixed-IPv4/v6 Sites			Pure-IPv6 Sites		
	Provider	# Sites	WF %	Provider	# Sites	WF %
1	CLOUDFLARENET	146115	97.0	CLOUDFLARENET	212195	43.6
2	Hostinger Int'l Ltd	4255	97.6	Hostinger Int'l Ltd	4775	79.5
3	AMAZON	4006	84.2	Hetzner Online GmbH	3892	48.5
4	FASTLY	3344	71.4	Cloudflare London, LLC	3626	25.6
5	DIGITALOCEAN	3195	14.6	OVH SAS	2865	49.6
6	OVH SAS	1985	86.1	AMAZON	2508	62.8
7	Hetzner Online GmbH	1896	93.4	IONOS SE	2321	50.7
8	Akamai Int'l B.V.	1848	78.9	FASTLY	1706	22.5
9	IONOS SE	1717	93.8	GOOGLE-GCP	1629	9.8
10	JSC Timeweb	1607	88.1	GOOGLE	1601	16.3

ASN	Provider Name	Colocation
AS15169	GOOGLE	65.74
AS54113	FASTLY	59.41
AS13335	CLOUDFLARENET	12.37
AS9123	JSC Timeweb	9.76
AS8560	IONOS SE	7.40
AS396982	GOOGLE-GCP	6.88
AS209242	Cloudlfare London LLC	6.14
AS16276	OVH SAS	5.80
AS20940	Akamai Int'l B.V.	5.43
AS14061	DIGITALOCEAN	4.83
AS24940	Hetzner Online GmbH	2.30
AS47583	Hostinger Int'l Ltd	1.16
AS16509	AMAZON	1.13

- Pure: Major providers use large shared IPv6 pools → weaker fingerprints: ↑ co-location, ↓ accuracy
- Mixed: Co-location patterns are less predictive; dominated by IPv4 uniqueness
- Cloudflare, Fastly, Google ⇒ dense IPv6 pooling unlike Amazon

"Provider deployment practices—not IPv6 itself—dictate fingerprintability."

Provider Specific Trends

TABLE II
TOP 10 IPV6 HOSTING PROVIDERS IN MIXED-IPV4/IPV6 & PURE-IPV6 SITES

Rank	mixed-IPv4/v6 Sites			Pure-IPv6 Sites		
	Provider	# Sites	WF %	Provider	# Sites	WF %
1	CLOUDFLARENET	146115	97.0	CLOUDFLARENET	212195	43.6
2	Hostinger Int'l Ltd	4255	97.6	Hostinger Int'l Ltd	4775	79.5
3	AMAZON	4006	84.2	Hetzner Online GmbH	3892	48.5
4	FASTLY	3344	71.4	Cloudflare London, LLC	3626	25.6
5	DIGITALOCEAN	3195	14.6	OVH SAS	2865	49.6
6	OVH SAS	1985	86.1	AMAZON	2508	62.8
7	Hetzner Online GmbH	1896	93.4	IONOS SE	2321	50.7
8	Akamai Int'l B.V.	1848	78.9	FASTLY	1706	22.5
9	IONOS SE	1717	93.8	GOOGLE-GCP	1629	9.8
10	JSC Timeweb	1607	88.1	GOOGLE	1601	16.3

ASN	Provider Name	Colocation
AS15169	GOOGLE	65.74
AS54113	FASTLY	59.41
AS13335	CLOUDFLARENET	12.37
AS9123	JSC Timeweb	9.76
AS8560	IONOS SE	7.40
AS396982	GOOGLE-GCP	6.88
AS209242	Cloudlfare London LLC	6.14
AS16276	OVH SAS	5.80
AS20940	Akamai Int'l B.V.	5.43
AS14061	DIGITALOCEAN	4.83
AS24940	Hetzner Online GmbH	2.30
AS47583	Hostinger Int'l Ltd	1.16
AS16509	AMAZON	1.13

- Pure: Major providers use large shared IPv6 pools → weaker fingerprints: ↑ co-location, ↓ accuracy
- Mixed: Co-location patterns are less predictive; dominated by IPv4 uniqueness
- Cloudflare, Fastly, Google ⇒ dense IPv6 pooling unlike Amazon

"Provider deployment practices—not IPv6 itself—dictate fingerprintability."

Provider Specific Trends

TABLE II
TOP 10 IPV6 HOSTING PROVIDERS IN MIXED-IPV4/IPV6 & PURE-IPV6 SITES

Rank	mixed-IPv4/v6 Sites			Pure-IPv6 Sites		
	Provider	# Sites	WF %	Provider	# Sites	WF %
1	CLOUDFLARENET	146115	97.0	CLOUDFLARENET	212195	43.6
2	Hostinger Int'l Ltd	4255	97.6	Hostinger Int'l Ltd	4775	79.5
3	AMAZON	4006	84.2	Hetzner Online GmbH	3892	48.5
4	FASTLY	3344	71.4	Cloudflare London, LLC	3626	25.6
5	DIGITALOCEAN	3195	14.6	OVH SAS	2865	49.6
6	OVH SAS	1985	86.1	AMAZON	2508	62.8
7	Hetzner Online GmbH	1896	93.4	IONOS SE	2321	50.7
8	Akamai Int'l B.V.	1848	78.9	FASTLY	1706	22.5
9	IONOS SE	1717	93.8	GOOGLE-GCP	1629	9.8
10	JSC Timeweb	1607	88.1	GOOGLE	1601	16.3

ASN	Provider Name	Colocation
AS15169	GOOGLE	65.74
AS54113	FASTLY	59.41
AS13335	CLOUDFLARENET	12.37
AS9123	JSC Timeweb	9.76
AS8560	IONOS SE	7.40
AS396982	GOOGLE-GCP	6.88
AS209242	Cloudlfare London LLC	6.14
AS16276	OVH SAS	5.80
AS20940	Akamai Int'l B.V.	5.43
AS14061	DIGITALOCEAN	4.83
AS24940	Hetzner Online GmbH	2.30
AS47583	Hostinger Int'l Ltd	1.16
AS16509	AMAZON	1.13

- Pure: Major providers use large shared IPv6 pools → weaker fingerprints: ↑ co-location, ↓ accuracy
- Mixed: Co-location patterns are less predictive; dominated by IPv4 uniqueness
- Cloudflare, Fastly, Google ⇒ dense IPv6 pooling unlike Amazon

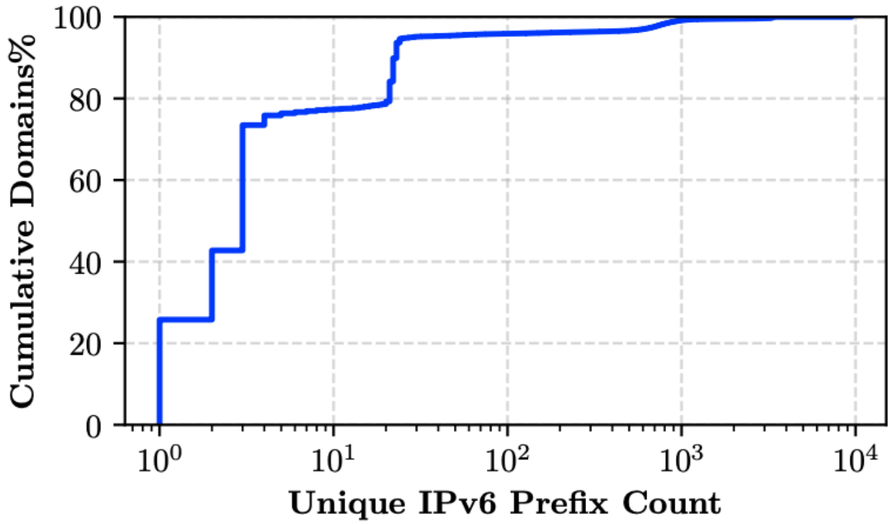
Provider deployment practices—not IPv6 itself—dictate fingerprintability.
Residual unique IPv4-domain mapping are still responsible for most fingerprinting

Mitigation & Evasion

- IP-based WF works: domains → stable, predictable IPs.
- If mappings vary, fingerprints break.
- *Goal*: introduce uncertainty & variability into IP assignments.

Strategies:

- Privacy extensions
- Server side prefix rotation
- More co-location b/w unrelated sites
- Mindful use of IPv6 address space



Pros: harder for adversaries, fingerprints decay quickly.
Cons: adds complexity to DNS, load-balancing, and config.

QUESTIONS?

Fingerprint Stability

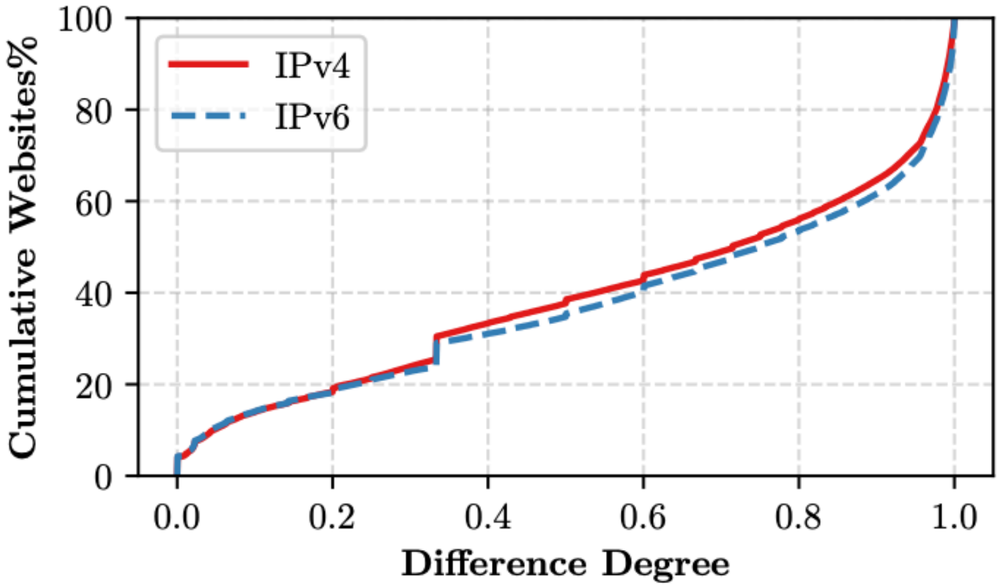
Stability depends on:

1. Difference degree → how much domain/IP sets change
2. IP Churn → how often addrs mapped to a domain change.

Difference Degree:

- 0 = identical fingerprints (perfect stability)
- 1 = completely different fingerprints
- Similar pattern for IPv4 & IPv6
- ~40% have avg difference degree < 0.5 after 15 days

$$Difference\ degree = \frac{|(\mathcal{D}_{t_0} \cup \mathcal{D}_{t_1}) - (\mathcal{D}_{t_0} \cap \mathcal{D}_{t_1})|}{|(\mathcal{D}_{t_0} \cup \mathcal{D}_{t_1})|}$$

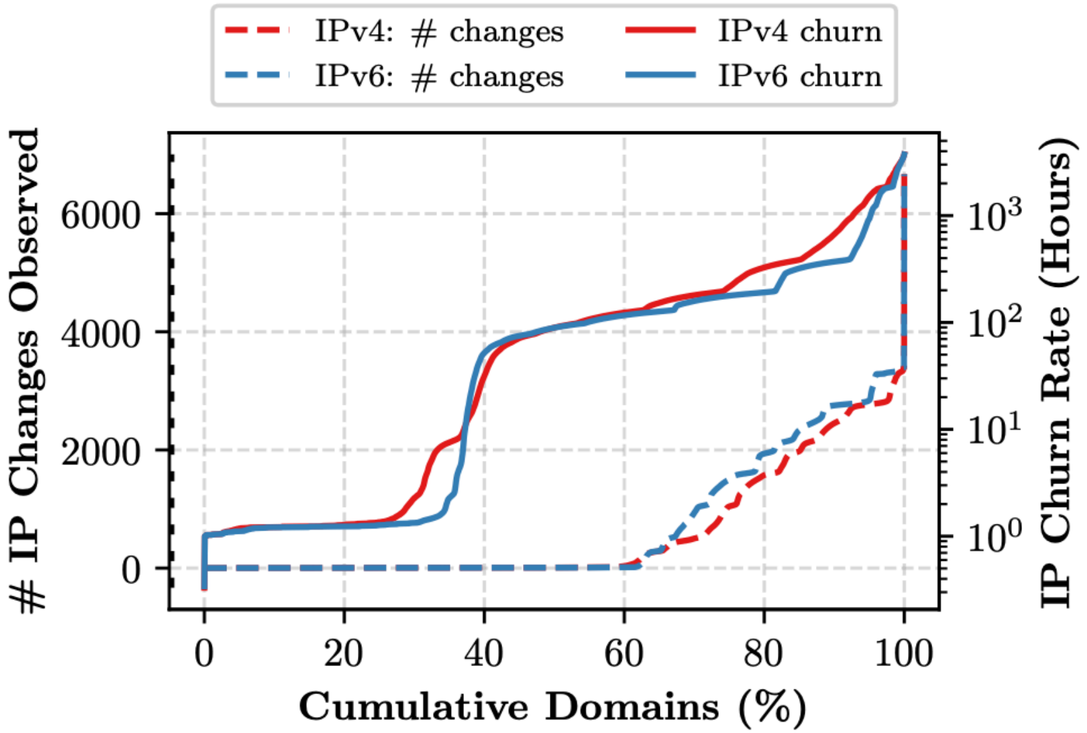


IP-Churn/Rotation:

Average time interval between IP changes

$$IP\text{-}Churn\ Interval = \frac{1}{k-1}\sum_{i=1}^{k-1}\left(t_{c_{i+1}} - t_{c_i}\right)$$

- ~60% of domains = no IP changes
- Among remaining: IPv6 >= IPv4
- ~38% of domains IP-churn <10 hrs
  - 28% domains < 1 hr (IPv4 & IPv6)

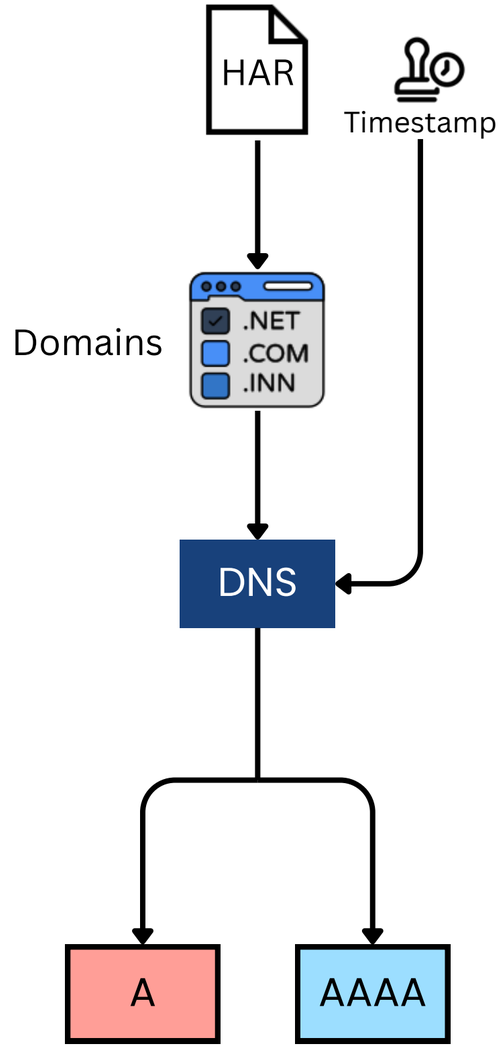


Overall: IPv6 churn ≈ IPv4 churn

Fingerprinting Testing

IP connections:
Need to have realistic IP connections to test against our fingerprints.

- Reconstruct connections from DNS data
- First IP returned closest to crawl timestamp
- Mimic browser with Happy Eyeballs



Appendix

B. Domain Stability

- Both categories show frequent domain churn
- Slightly higher stability in mixed-IPv4/v6 websites
- Can be refreshed with repeated crawling