



Censored Planet



JIGSAW

Advancing the Art of Censorship Data Analysis

Ram Sundara Raman, Apurva Virkud, Sarah Laplante, Vinicius Fortuna, Roya Ensafi

15 February 2023



Technical multi-stakeholder report on Internet shutdowns: The case of Iran amid autumn 2022 protests

OONI, IODA, M-Lab, Cloudflare, Kentik, Censored Planet, ISOC, Article19, 2022-11-29

REPORT

Throttling of Twitter in Russia

accessNOW

OUR WORK

CAMPAIGNS

BLOG

NEWSROOM

ABOUT

HELPLINE

HOME / BLOG / INTERNET SHUTDOWNS...



FREEDOM OF EXPRESSION

Internet shutdowns report: Shattered dreams and lost opportunities — a year in the fight to #KeepItOn

3 MARCH 2021 | 5:00 AM

Internet censorship continues to advance, necessitating high-quality data

State of Censorship Data

- Active censorship measurement platforms with focus on achieving good coverage over:
 - Time
 - Networks
 - Countries
 - Domains
 - Censorship Methods



Censored Planet



OONI



Data collection is only *part* of the process

Parsing, analyzing, and exploring censorship measurements, **especially at large scale** is hard.

- Most previous studies have relied on ad-hoc analysis methods on case by case basis
- There is lack of ground-truth at scale
- The size of the Internet and the large number of stakeholders introduce many extraneous factors that can cause incorrect censorship characterization.

Outline

1

Challenges in censorship data analysis

1. Data limitations
2. Accurate Metadata
3. Unexpected Interference

2

Censored Planet data analysis pipeline

1. Design Goals
2. Workflow
3. Censored Planet dashboard

Outline

1

Challenges in censorship data analysis

1. Data limitations
2. Accurate Metadata
3. Unexpected Interference

2

Censored Planet data analysis pipeline

1. Design Goals
2. Workflow
3. Censored Planet dashboard

Challenges in Censorship Data Analysis: Data Limitations

Need to consider the data's

- Scale
- Coverage
- Continuity
- **Protocols**

OOONI | Explorer Search MAT Charts Circumvention Charts Countries

! Anomaly
http://www.facebook.com
DNS tampering

 Myanmar
Country

AS58952
Network

February 17, 2021, 04:38 PM UTC
Date & Time

OOONI | Explorer Search MAT Charts Circumvention Charts Countries

! Anomaly
http://www.facebook.com
TCP/IP blocking

 Myanmar
Country

AS58952
Network

February 17, 2021, 05:03 PM UTC
Date & Time

Challenges in Censorship Data Analysis:

Data Limitations

Need to consider the data's

- Scale
- Coverage
- Continuity
- **Protocols**

OOONI | Explorer Search MAT Charts Circumvention Charts Countries

! Anomaly
http://www.facebook.com
DNS tampering

 Myanmar
Country

AS58952
Network

February 17, 2021, 04:38 PM UTC
Date & Time

OOONI | Explorer Search MAT Charts Circumvention Charts Countries

! Anomaly
http://www.facebook.com
TCP/IP blocking

 Myanmar
Country

AS58952
Network

February 17, 2021, 05:03 PM UTC
Date & Time

Challenges in Censorship Data Analysis:

Data Limitations

Need to consider the data's

- Scale
- Coverage
- Continuity
- **Protocols**

OOONI | Explorer Search MAT Charts Circumvention Charts Countries

! Anomaly
http://www.facebook.com
DNS tampering

DNS Queries

Resolver: 116.206.136.161 Local DNS
Query: IN A www.facebook.com
Engine: system

Name	Class	TTL	Type	DATA
@	IN		A	59.153.90.11

Wrong Address

OOONI | Explorer Search MAT Charts Circumvention Charts Countries

! Anomaly
http://www.facebook.com
TCP/IP blocking

DNS Queries

Resolver: 172.253.211.3 Google DNS
Query: IN A www.facebook.com
Engine: system

Name	Class	TTL	Type	DATA
@	IN		A	69.171.250.35

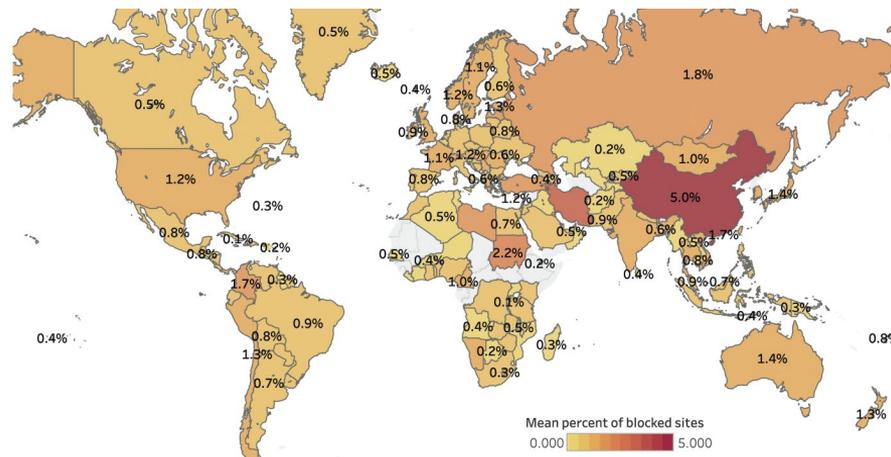
Right Address

Challenges in Censorship Data Analysis: Accurate Metadata

- IP metadata is key

Previous Work

- Country-level geolocation
- But country-level results can be an inaccurate estimate



Challenges in Censorship Data Analysis:

Accurate Metadata

ASN	AS Name	APNIC % of traffic [1]	Censored Planet Measurements
AS802	York University	0	698,037
AS5769	Videotron Ltee	10.48	579,096
AS31983	Queen's University	NA	405,519
AS812	Rogers Communications	14.42	270,483
AS6327	Shaw Communications	10.25	236,987

Censored Planet Measurements in Canada, September 2021

[1]
<https://stats.labs.apnic.net/cgi-bin/aspop?c=ca>

Challenges in Censorship Data Analysis:

Accurate Metadata

ASN	AS Name	APNIC % of traffic [1]	Censored Planet Measurements
AS802	York University	0	698,037
AS5769	Videotron Ltee	10.48	579,096
AS31983	Queen's University	NA	405,519
AS812	Rogers Communications	14.42	270,483
AS6327	Shaw Communications	10.25	236,987

Censored Planet
Measurements in
Canada, September 2021

[1]
<https://stats.labs.apnic.net/cgi-bin/aspop?c=ca>

Challenges in Censorship Data Analysis: Unexpected Interference

- **CDN and hosting configurations**
 - DDoS/Bot protection

• Access Denied - GoDaddy Website Firewall

If you are the site owner (or you manage this site), please whitelist your IP or if you think this block is an error please [open a support ticket](#) and make sure to include the block details (displayed in the box below), so we can assist you in troubleshooting the issue.

Block details:

Your IP:	141.212.121.192
URL:	hotmail.msn.com/
Your Browser:	Mozilla/5.0 quack/0.x
Block ID:	DDOS22
Block reason:	DDOS attempt was blocked.
Time:	2019-03-24 08:49:20
Server ID:	12014

Challenges in Censorship Data Analysis: **Unexpected Interference**

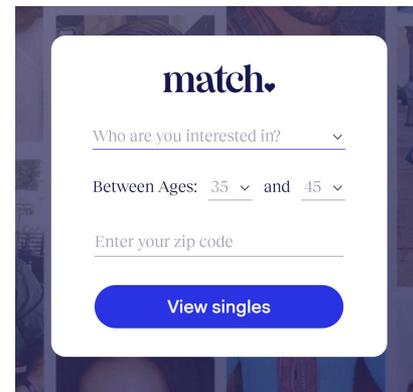
- **CDN and hosting configurations**
 - DDoS/Bot protection
 - Specific CDN behavior (e.g. Akamai edge)



Challenges in Censorship Data Analysis: Unexpected Interference

- **CDN and hosting configurations**
 - DDoS/Bot protection
 - Specific CDN behavior (e.g. Akamai edge)
 - Localization effects

match.com

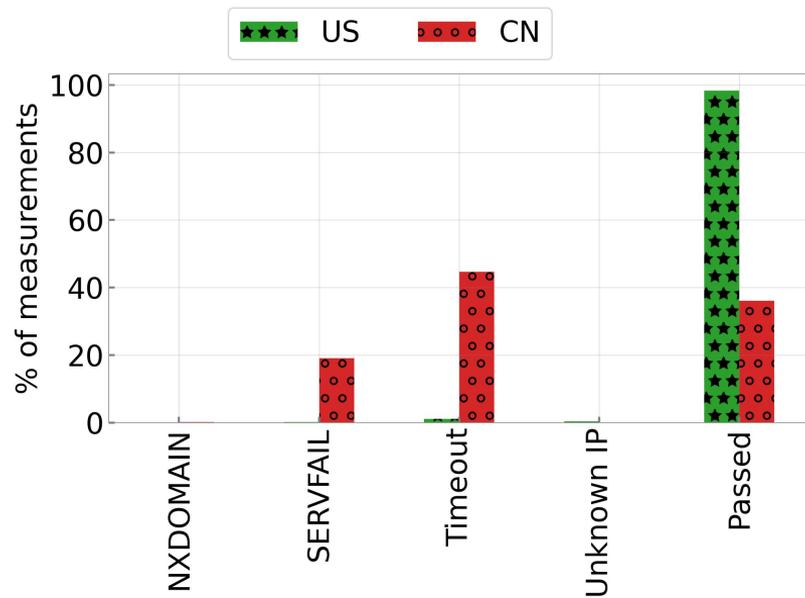


uk.match.com



Challenges in Censorship Data Analysis: Unexpected Interference

- CDN and hosting configurations
- **Internet Geoblocking**



DNS resolutions for 75 .gov and .mil domains in US and CN

Outline

1

Challenges in censorship data analysis

1. Data limitations
2. Accurate Metadata
3. Unexpected Interference

2

Censored Planet data analysis pipeline

1. Design Goals
2. Workflow
3. Censored Planet dashboard



Censored Planet Data Analysis Pipeline

Measurements vs Analysis:

Enables future data analysis improvements

Efficiency:

Process 13 TBs of compressed data over 4.5 years in < 24 hours

Modular:

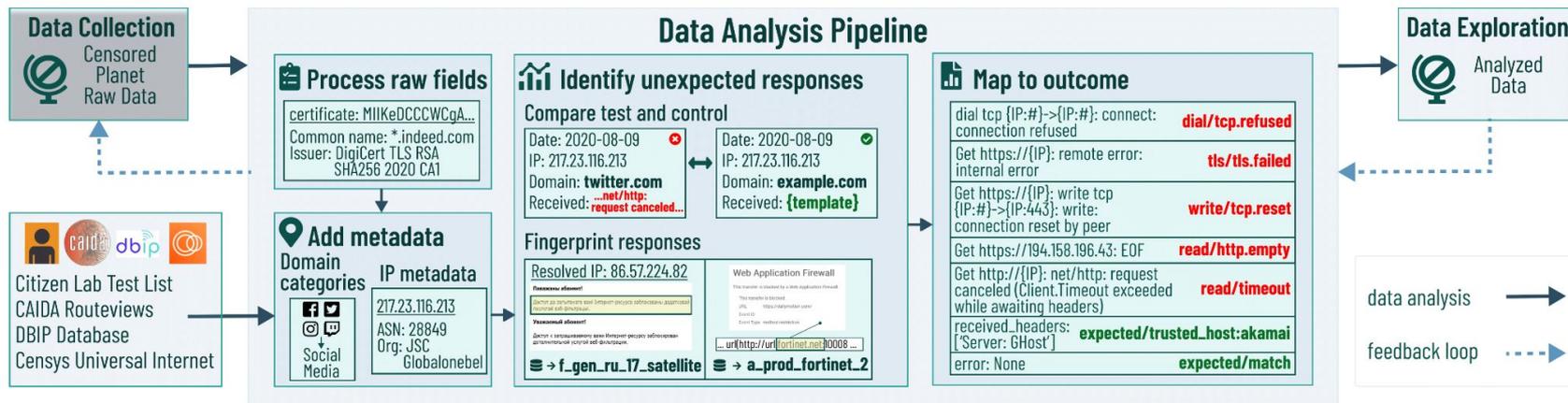
New analysis can be added easily and run on subset of the data

Fully Open Source:

<https://github.com/censoredplanet/censoredplanet-analysis>



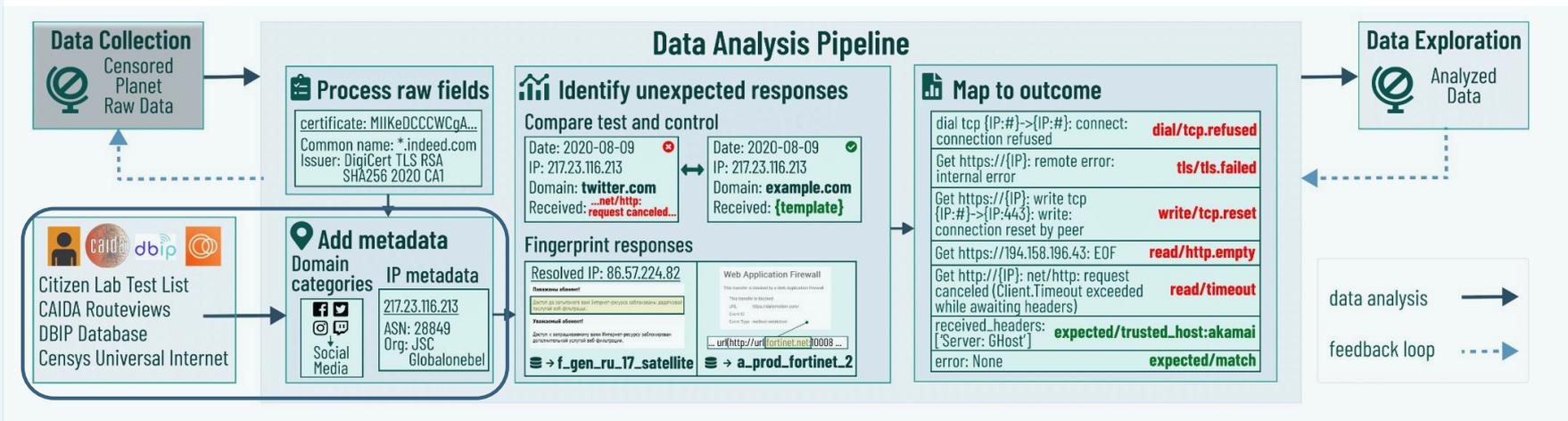
Censored Planet Data Analysis Pipeline



Censored Planet Data Analysis Pipeline

Add Metadata

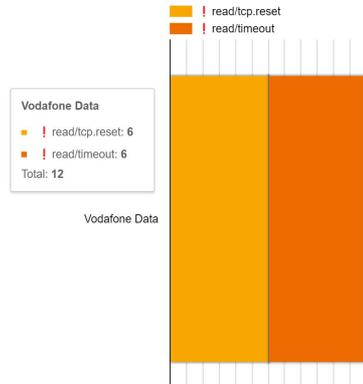
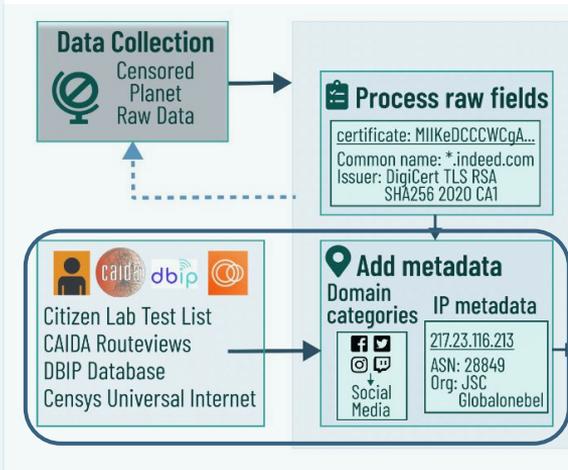
- Add domain metadata such as category, TLS certificates, HTTP Body
- Add IP metadata such as ASN, IP Organization



Censored Planet Data Analysis Pipeline

Add Metadata

- Add domain metadata such as category, TLS certificates, HTTP Body
- Add IP metadata such as ASN, IP Organization

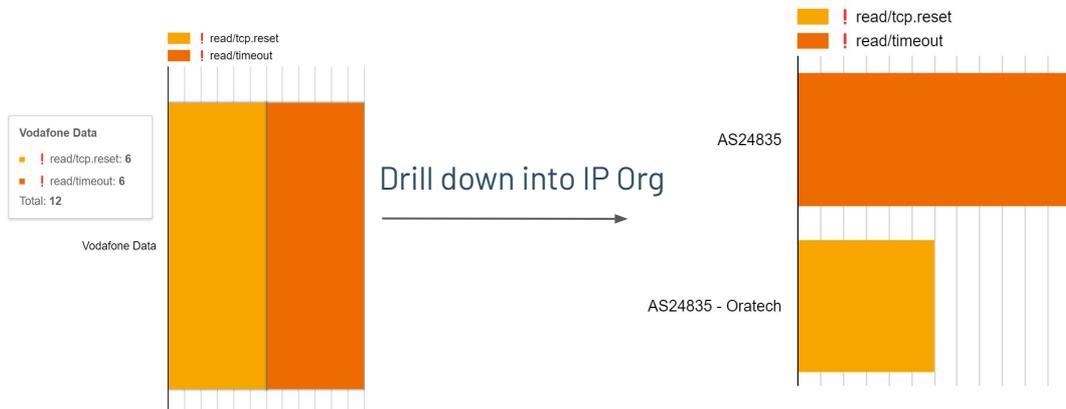
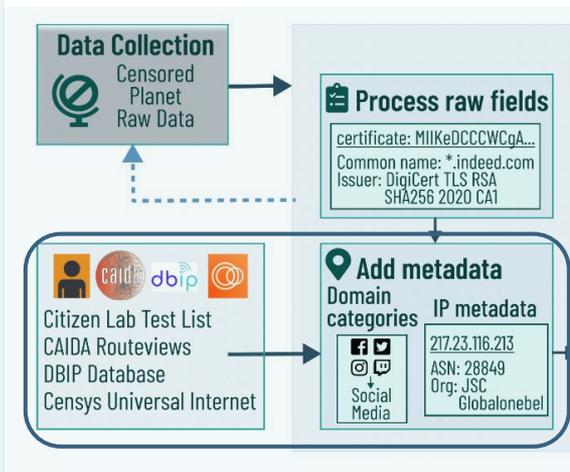


Blocking of www.hotspotshield.com in Egypt

Censored Planet Data Analysis Pipeline

Add Metadata

- Add domain metadata such as category, TLS certificates, HTTP Body
- Add IP metadata such as ASN, IP Organization

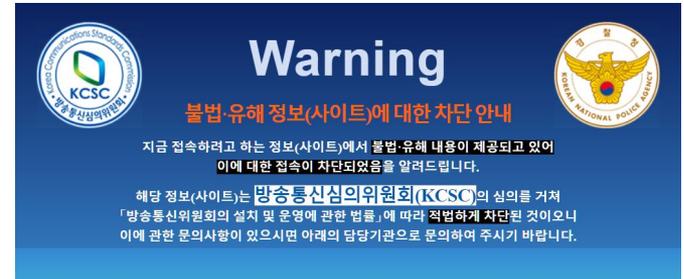
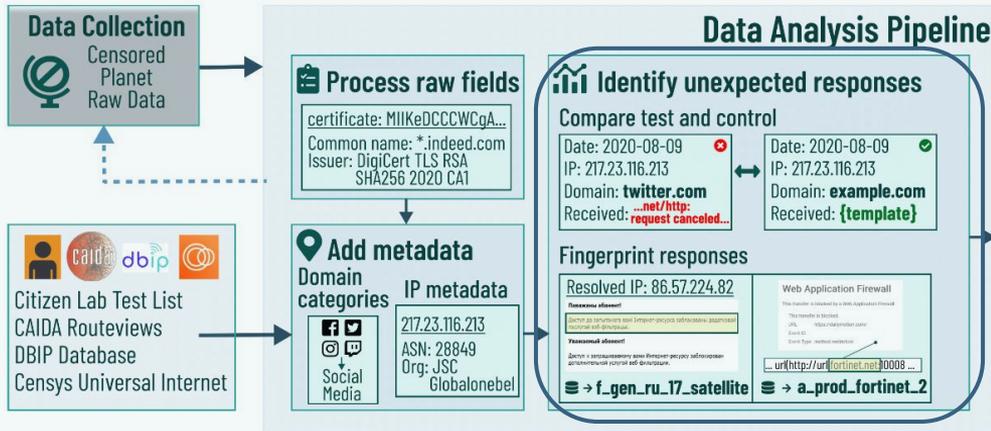


Blocking of www.hotspotshield.com in Egypt

Censored Planet Data Analysis Pipeline

Identify Unexpected Responses

- Compare with control measurements to identify measurements to look further into
- Check responses for indications of censorship

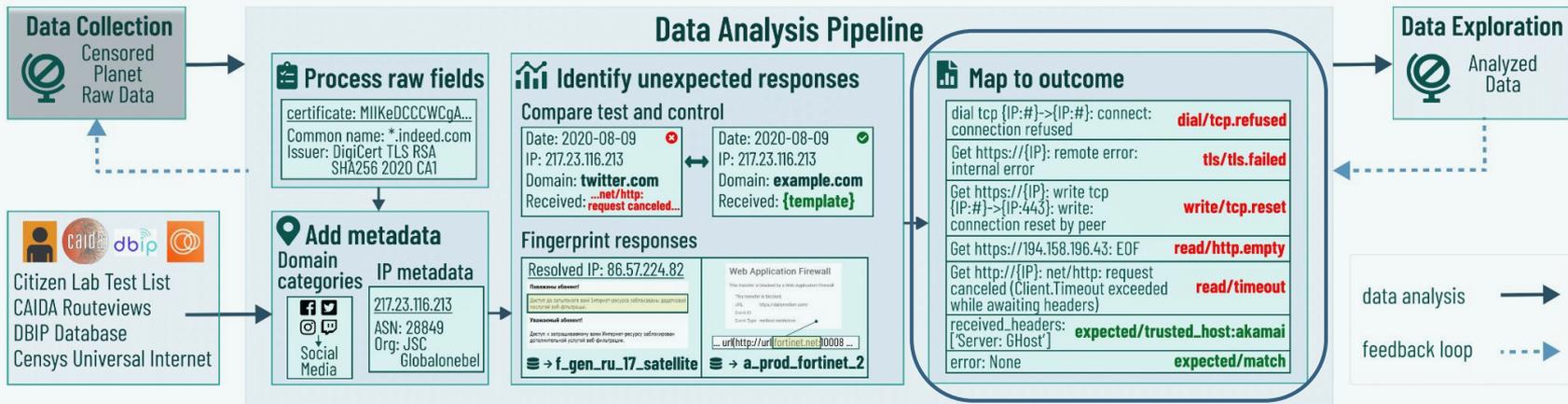


Blockpage in South Korea

Censored Planet Data Analysis Pipeline

Map to Outcome

- Map each measurement to human-readable outcome
- 53 distinct identifiers mapped to outcomes
- Iterative process



Outcomes in HTTP measurements

Stage	Outcome	Num. Measurements	% Measurements	Outcome Type
Expected Response	expected/match	1,772,014,793	94.45	✓
	expected/hosting_provider (e.g. akamai)	61,943,574	3.30	✓
Content Mismatch	content/known_not_censorship	16,642,905	0.89	✓
	content/status_mismatch	13,533,254	0.72	?
	content/known_blockpage	743,396	0.04	!
Read/Write Failure	read/timeout	6,356,637	0.34	!
	read/tcp.reset	4,273,880	0.23	!
Dial Failure	dial/ip.no_route_to_host	28,954	0.001	?

Outcomes in HTTP measurements

Stage	Outcome	Num. Measurements	% Measurements	Outcome Type
Expected Response	expected/match	1,772,014,793	94.45	✓
	expected/hosting_provider (e.g. akamai)	61,943,574	3.30	✓
	content/known_not_censorship	16,642,905	0.89	✓
Content Mismatch	content/status_mismatch	13,533,254	0.72	?
	content/known_blockpage	743,396	0.04	!
Read/Write Failure	read/timeout	6,356,637	0.34	!
	read/tcp.reset	4,273,880	0.23	!
Dial Failure	dial/ip.no_route_to_host	28,954	0.001	?

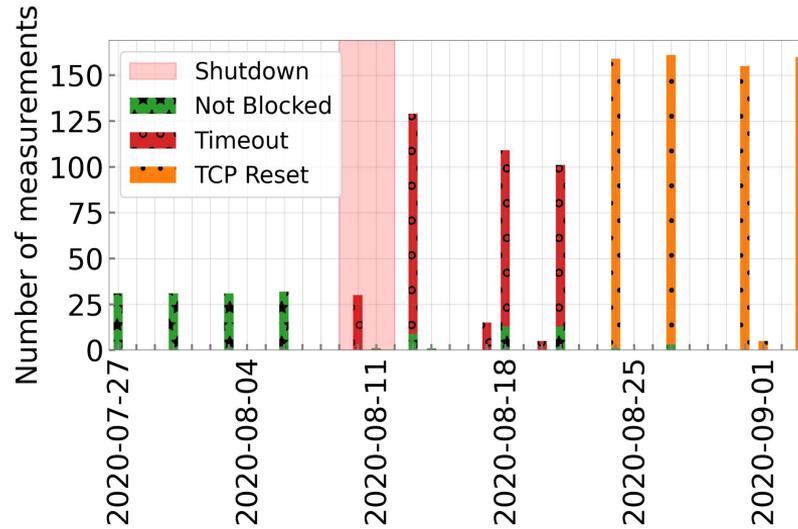
Outcomes in HTTP measurements

Stage	Outcome	Num. Measurements	% Measurements	Outcome Type
Expected Response	expected/match	1,772,014,793	94.45	✓
	expected/hosting_provider (e.g. akamai)	61,943,574	3.30	✓
Content Mismatch	content/known_not_censorship	16,642,905	0.89	✓
	content/status_mismatch	13,533,254	0.72	?
Read/Write Failure	content/known_blockpage	743,396	0.04	!
	read/timeout	6,356,637	0.34	!
	read/tcp.reset	4,273,880	0.23	!
Dial Failure	dial/ip.no_route_to_host	28,954	0.001	?

Outcomes in HTTP measurements

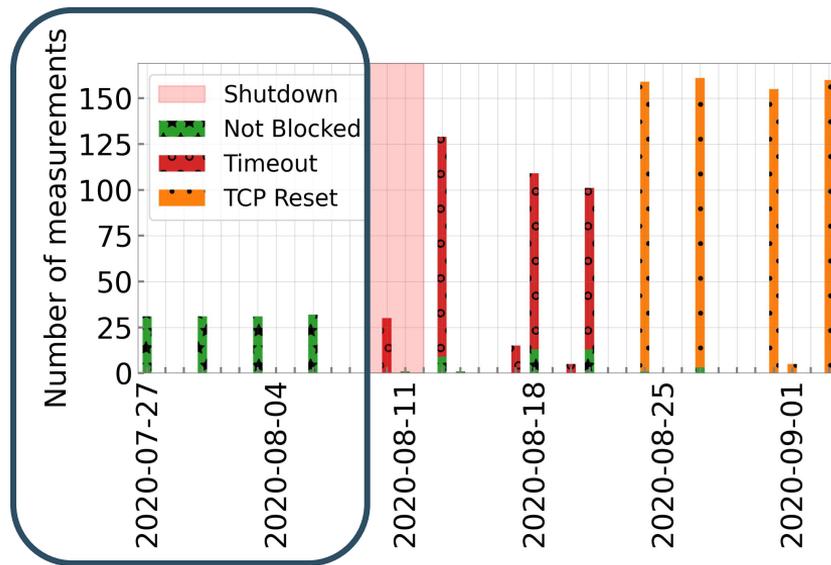
Stage	Outcome	Num. Measurements	% Measurements	Outcome Type
Expected Response	expected/match	1,772,014,793	94.45	✓
	expected/hosting_provider (e.g. akamai)	61,943,574	3.30	✓
	content/known_not_censorship	16,642,905	0.89	✓
Content Mismatch	content/status_mismatch	13,533,254	0.72	?
	content/known_blockpage	743,396	0.04	!
Read/Write Failure	read/timeout	6,356,637	0.34	!
	read/tcp.reset	4,273,880	0.23	!
Dial Failure	dial/ip.no_route_to_host	28,954	0.001	?

Value of Outcomes



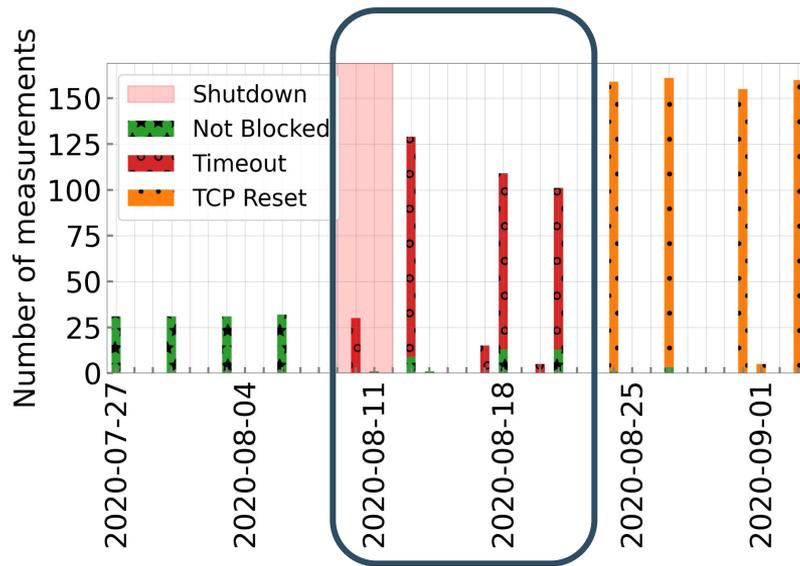
HTTPS measurements to psiphon.ca in Belarus

Value of Outcomes



HTTPS measurements to psiphon.ca in Belarus

Value of Outcomes



HTTPS measurements to psiphon.ca in Belarus

Value of Outcomes

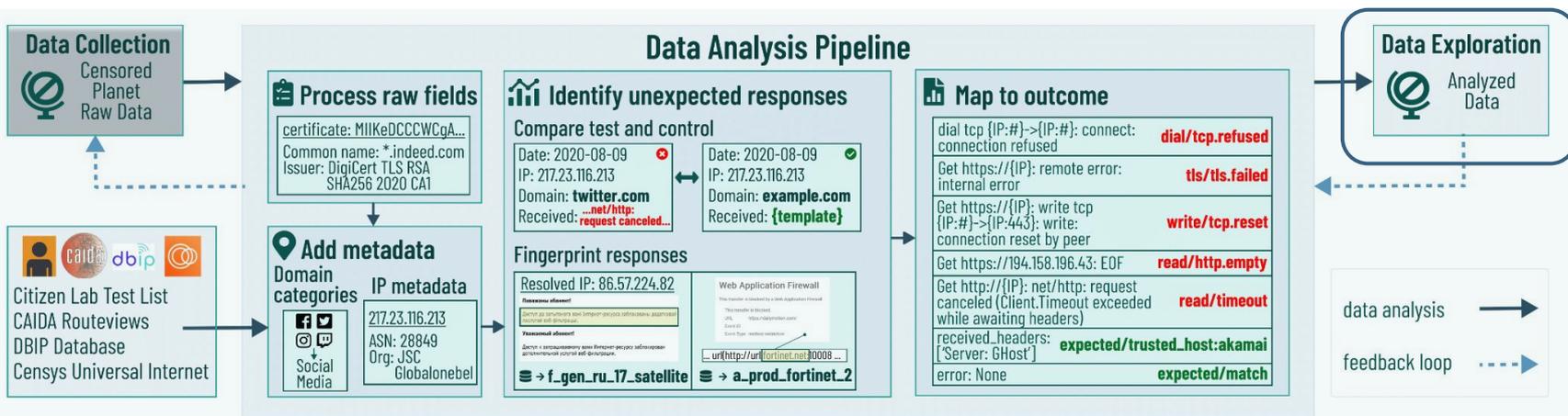


HTTPS measurements to psiphon.ca in Belarus

Censored Planet Data Analysis Pipeline

Data Exploration

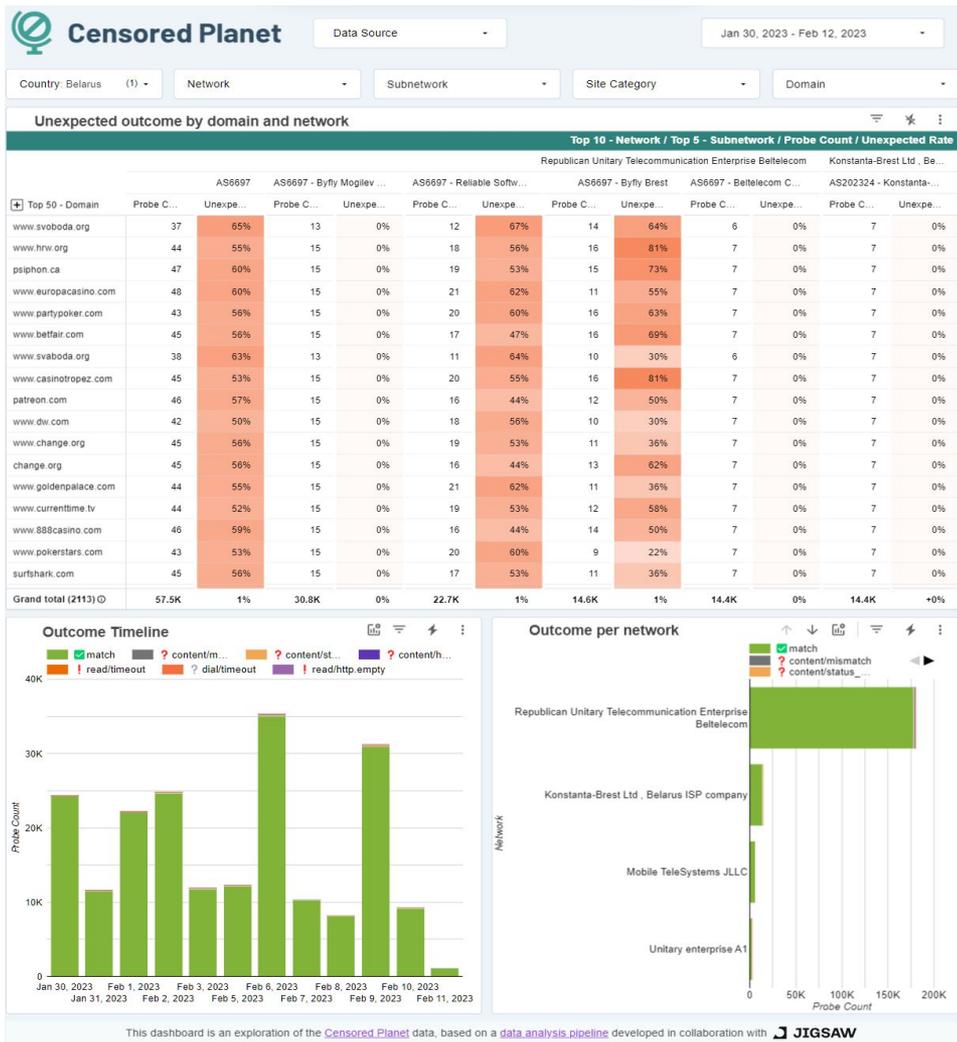
- Feedback loop from analysis to measurements
- Data exploration is key



Dashboard

- Enables visualizations for longitudinal data
- Automatically updated after measurements
- Open to public

<https://dashboard.censoredplanet.org>



Key Takeaways

- Censorship data analysis is **complex**, both due to the nature of Internet as well as censorship itself
- **Common challenges** - Data limitations, Accurate metadata availability, unexpected network interference
- We built a **censorship data analysis pipeline** to address many of these challenges

Thank you!

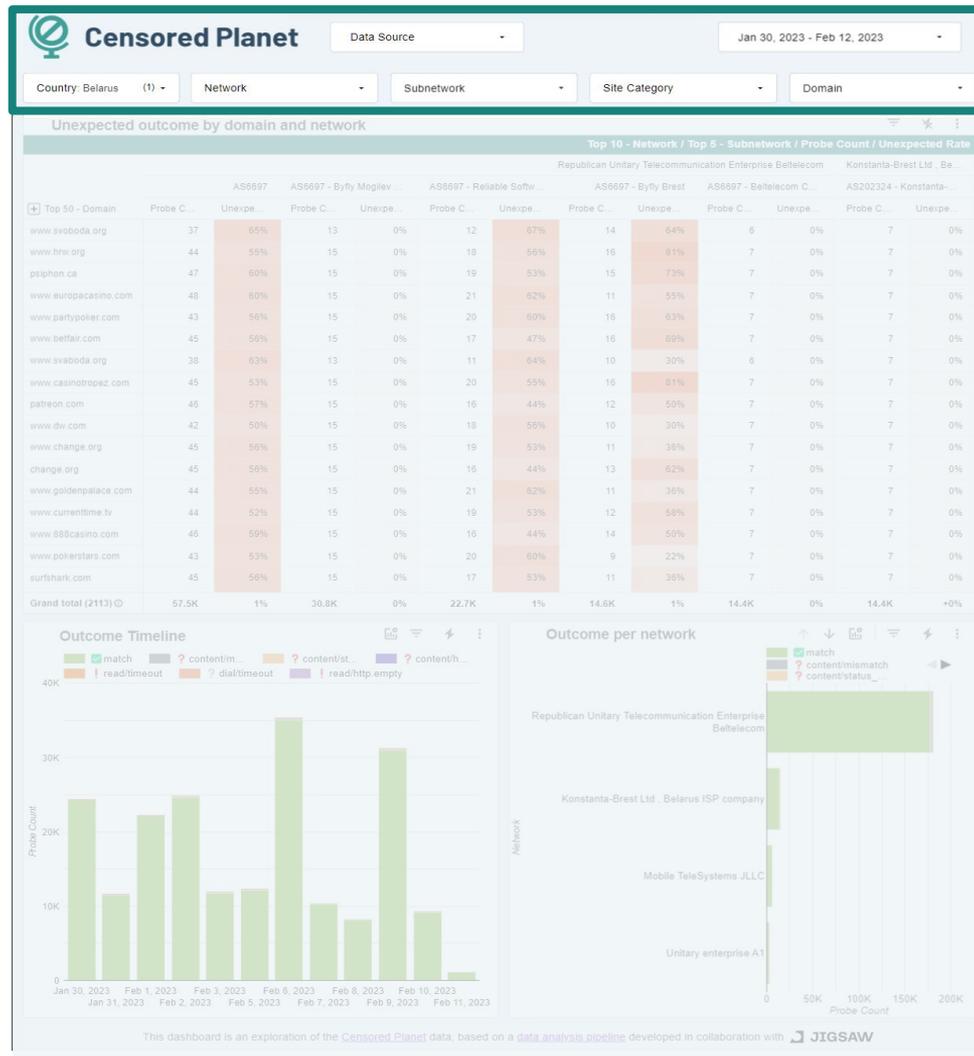
Questions?

Reach out to us at ramaks@umich.edu and censoredplanet-analysis@umich.edu

<https://censoredplanet.org>

Dashboard

- Controls
 - data source
 - date
 - country
 - network
 - subnetwork
 - site category
 - domain



Dashboard

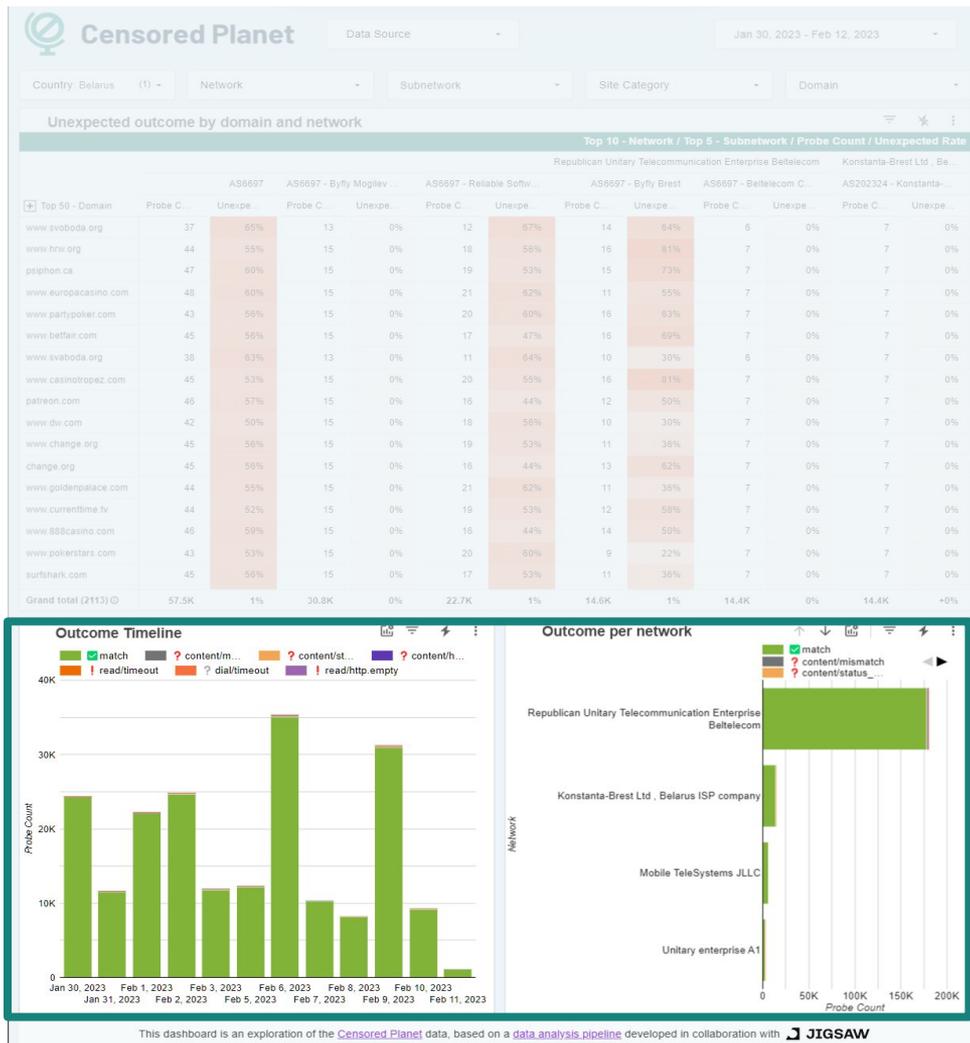
- Unexpected outcome by domain and network



This dashboard is an exploration of the Censored Planet data, based on a data analysis pipeline developed in collaboration with JIGSAW

Dashboard

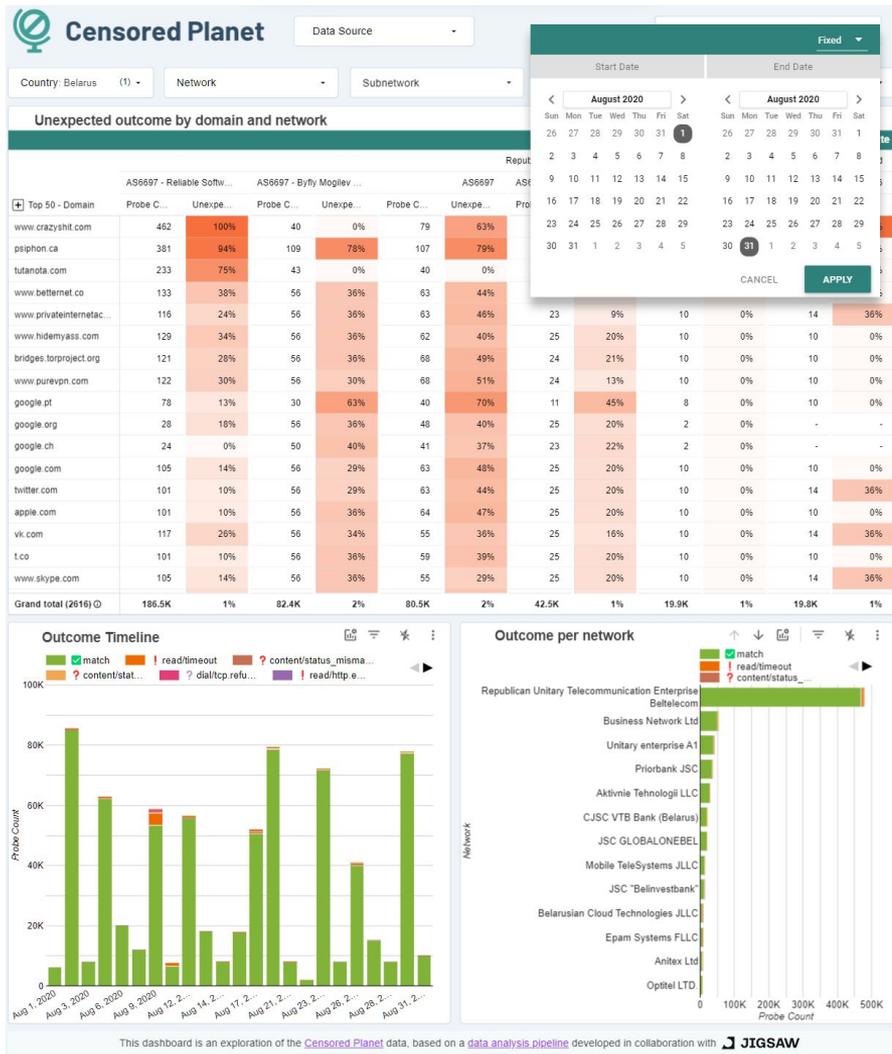
- Outcome Timeline
- Outcome per network



This dashboard is an exploration of the [Censored Planet](#) data, based on a [data analysis pipeline](#) developed in collaboration with [JIGSAW](#)

Belarus

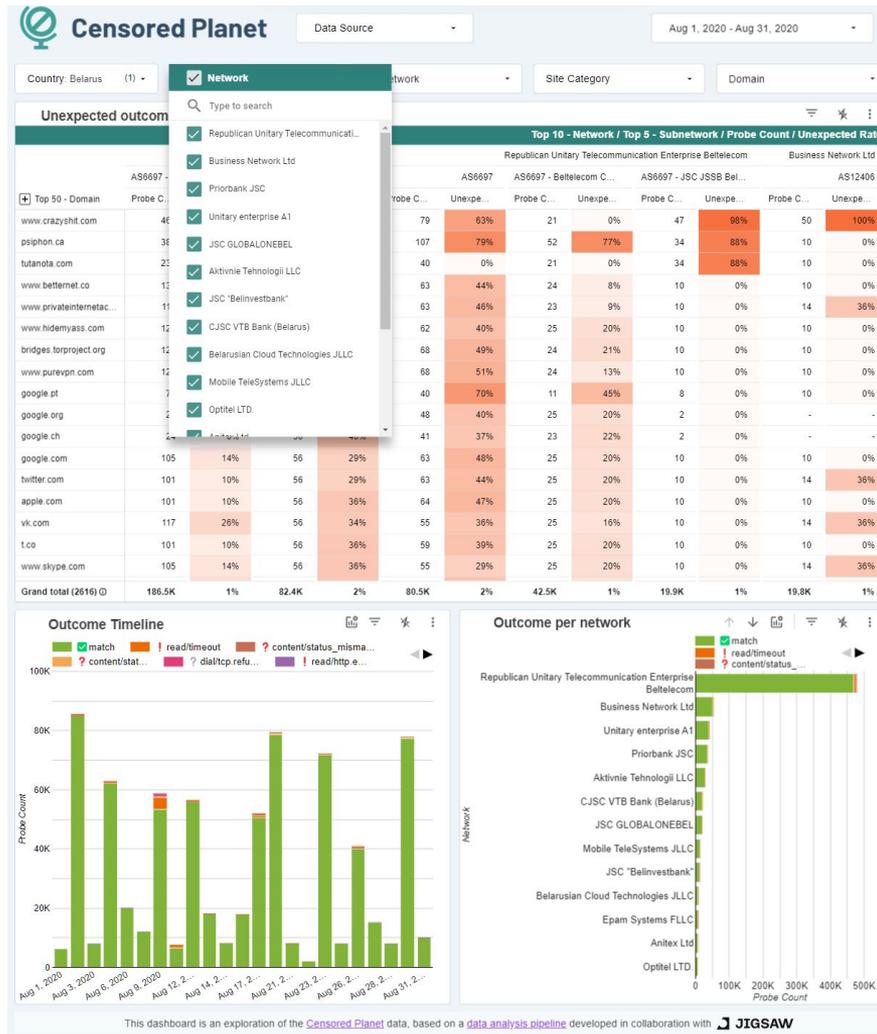
- Country: Belarus
- Election day: Aug. 9
- CP Data: Aug. 2020



This dashboard is an exploration of the [Censored Planet](#) data, based on a [data analysis pipeline](#) developed in collaboration with [JIGSAW](#)

Which networks?

- Country: Belarus
- Election day: Aug. 9
- CP Data: Aug. 2020

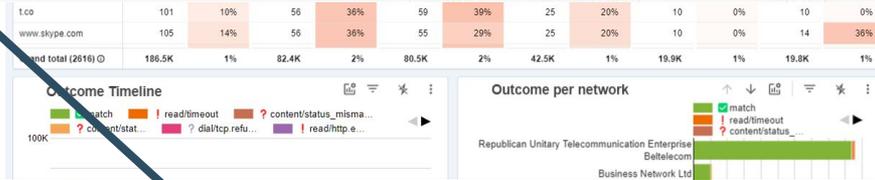


Which networks?

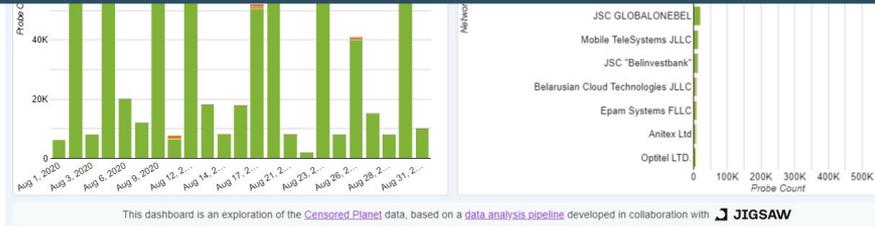
Censored Planet
Data Source: [v]
Aug 1, 2020 - Aug 31, 2020
Country: Belarus (1) | Network | Site Category | Domain
Unexpected outcom | Type to search

Visible ASNs: Customer Populations (Est.)

Rank	ASN	AS Name	CC	Users (est.)	% of country	% of Internet	Samples
1	AS6697	BELPAK-AS BELPAK	BY	3,843,654	49.56	0.09	1,715,377
2	AS25106	MTS BY-AS	BY	1,856,112	23.93	0.044	828,361
3	AS42772	A1-BY-AS	BY	1,407,196	18.15	0.033	628,015
4	AS44087	BEST-AS	BY	312,871	4.03	0.007	139,631
5	AS31143	COSMOSTV-AS	BY	85,433	1.1	0.002	38,128



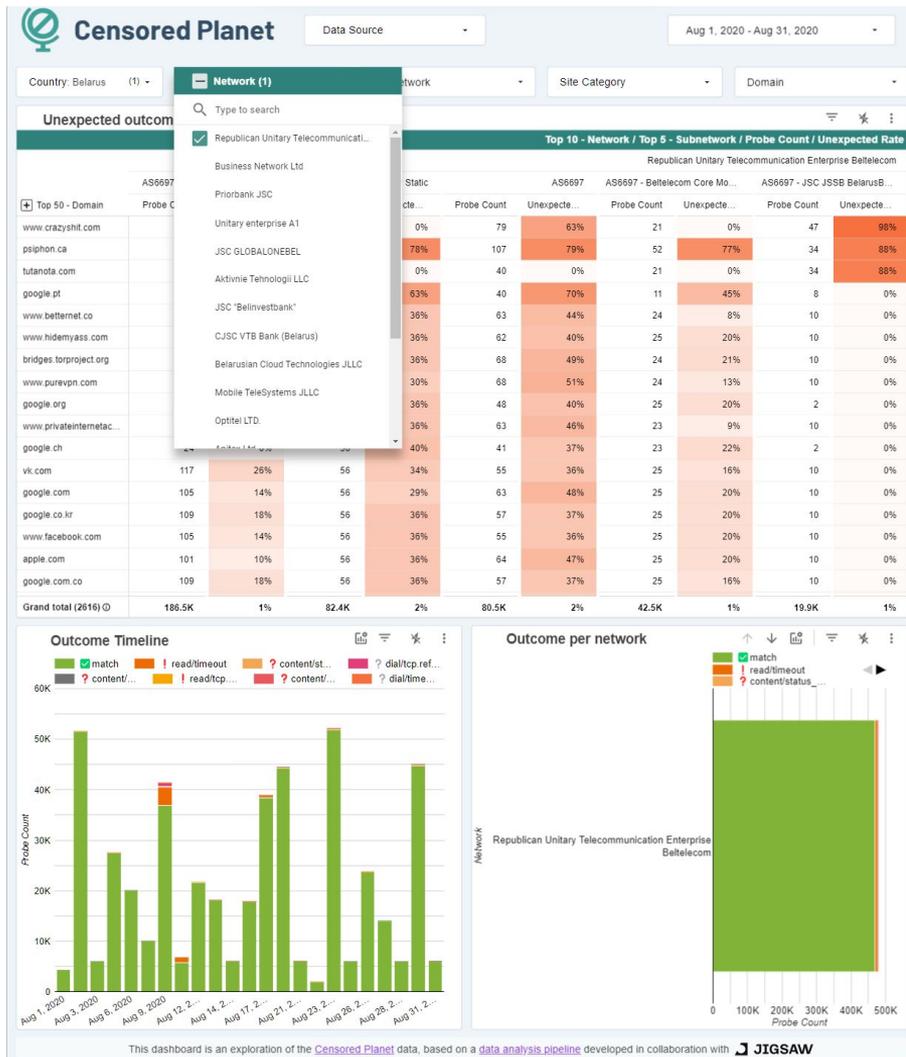
Republican Unitary Telecommunication Enterprise Beltelecom



This dashboard is an exploration of the Censored Planet data, based on a data analysis pipeline developed in collaboration with JIGSAW

Filter networks

- Country: Belarus
- Election day: Aug. 9
- CP Data: Aug. 2020
- AS6697: Beltelecom



Outcome/Subnetwork

- Country: Belarus
- Election day: Aug. 9
- CP Data: Aug. 2020
- AS6697: Beltelecom

Censored Planet
Data Source
Aug 1, 2020 - Aug 31, 2020

Country: Belarus (1)
Network (1)
Site Category
Domain

- Republican Unity Telecommunicati...
- Business Network Ltd
- Priorbank JSC
- Unitary enterprise A1
- JSC GLOBALONEBEL
- Aktivnie Tehnologii LLC
- JSC "Belinvestbank"
- CJSC VTB Bank (Belarus)
- Belarusian Cloud Technologies JLLC
- Mobile TeleSystems JLLC
- Optitel LTD.

Unexpected outcome

Top 50 - Domain	AS6697	Probe Count	Unexpected	AS6697 - Beltelecom Core Mo...	Probe Count	Unexpected	AS6697 - JSC JSSB Belarus B...	Probe Count	Unexpected	
www.crazyshit.com	AS6697	79	63%	21	0%	47	98%			
peiphon.ca	AS6697	107	78%	52	77%	34	88%			
tutanota.com	AS6697	40	0%	21	0%	34	88%			
google.pt	AS6697	40	63%	11	45%	8	0%			
www.belnet.by	AS6697	63	36%	24	8%	10	0%			
www.hide-my-ss.com	AS6697	62	36%	25	20%	10	0%			
bridges.torproject.org	AS6697	68	36%	24	21%	10	0%			
www.purevpn.com	AS6697	68	36%	24	13%	10	0%			
google.org	AS6697	48	35%	25	20%	2	0%			
www.privateinternet.com	AS6697	63	36%	23	9%	10	0%			
google.ch	AS6697	41	40%	23	22%	2	0%			
vk.com	AS6697	55	26%	25	16%	10	0%			
google.com	AS6697	63	14%	25	20%	10	0%			
google.co.kr	AS6697	57	18%	25	20%	10	0%			
www.facebook.com	AS6697	55	14%	25	20%	10	0%			
apple.com	AS6697	64	10%	25	20%	10	0%			
google.com.co	AS6697	57	18%	25	16%	10	0%			
Grand total (ESTR CD)	180.5K	1%	52.4K	2%	80.5K	2%	42.5K	1%	19.9K	1%

Top 10 - Network / Top 5 - Subnetwork / Probe Count / Unexpected Rate

Static	AS6697	AS6697 - Beltelecom Core Mo...	AS6697 - JSC JSSB Belarus B...
Static	AS6697	AS6697 - Beltelecom Core Mo...	AS6697 - JSC JSSB Belarus B...
Probe Count	79	21	47
Unexpected	63%	0%	98%
Static	AS6697	AS6697 - Beltelecom Core Mo...	AS6697 - JSC JSSB Belarus B...
Probe Count	107	52	34
Unexpected	78%	77%	88%
Static	AS6697	AS6697 - Beltelecom Core Mo...	AS6697 - JSC JSSB Belarus B...
Probe Count	40	21	34
Unexpected	0%	0%	88%
Static	AS6697	AS6697 - Beltelecom Core Mo...	AS6697 - JSC JSSB Belarus B...
Probe Count	40	11	8
Unexpected	63%	45%	0%
Static	AS6697	AS6697 - Beltelecom Core Mo...	AS6697 - JSC JSSB Belarus B...
Probe Count	63	24	10
Unexpected	36%	8%	0%
Static	AS6697	AS6697 - Beltelecom Core Mo...	AS6697 - JSC JSSB Belarus B...
Probe Count	62	25	10
Unexpected	36%	20%	0%
Static	AS6697	AS6697 - Beltelecom Core Mo...	AS6697 - JSC JSSB Belarus B...
Probe Count	68	24	10
Unexpected	36%	21%	0%
Static	AS6697	AS6697 - Beltelecom Core Mo...	AS6697 - JSC JSSB Belarus B...
Probe Count	68	24	10
Unexpected	36%	13%	0%
Static	AS6697	AS6697 - Beltelecom Core Mo...	AS6697 - JSC JSSB Belarus B...
Probe Count	48	25	2
Unexpected	35%	20%	0%
Static	AS6697	AS6697 - Beltelecom Core Mo...	AS6697 - JSC JSSB Belarus B...
Probe Count	63	23	10
Unexpected	36%	9%	0%
Static	AS6697	AS6697 - Beltelecom Core Mo...	AS6697 - JSC JSSB Belarus B...
Probe Count	41	23	2
Unexpected	40%	22%	0%
Static	AS6697	AS6697 - Beltelecom Core Mo...	AS6697 - JSC JSSB Belarus B...
Probe Count	55	25	10
Unexpected	36%	16%	0%
Static	AS6697	AS6697 - Beltelecom Core Mo...	AS6697 - JSC JSSB Belarus B...
Probe Count	63	25	10
Unexpected	14%	20%	0%
Static	AS6697	AS6697 - Beltelecom Core Mo...	AS6697 - JSC JSSB Belarus B...
Probe Count	57	25	10
Unexpected	18%	20%	0%
Static	AS6697	AS6697 - Beltelecom Core Mo...	AS6697 - JSC JSSB Belarus B...
Probe Count	55	25	10
Unexpected	14%	20%	0%
Static	AS6697	AS6697 - Beltelecom Core Mo...	AS6697 - JSC JSSB Belarus B...
Probe Count	64	25	10
Unexpected	10%	20%	0%
Static	AS6697	AS6697 - Beltelecom Core Mo...	AS6697 - JSC JSSB Belarus B...
Probe Count	57	25	10
Unexpected	18%	16%	0%

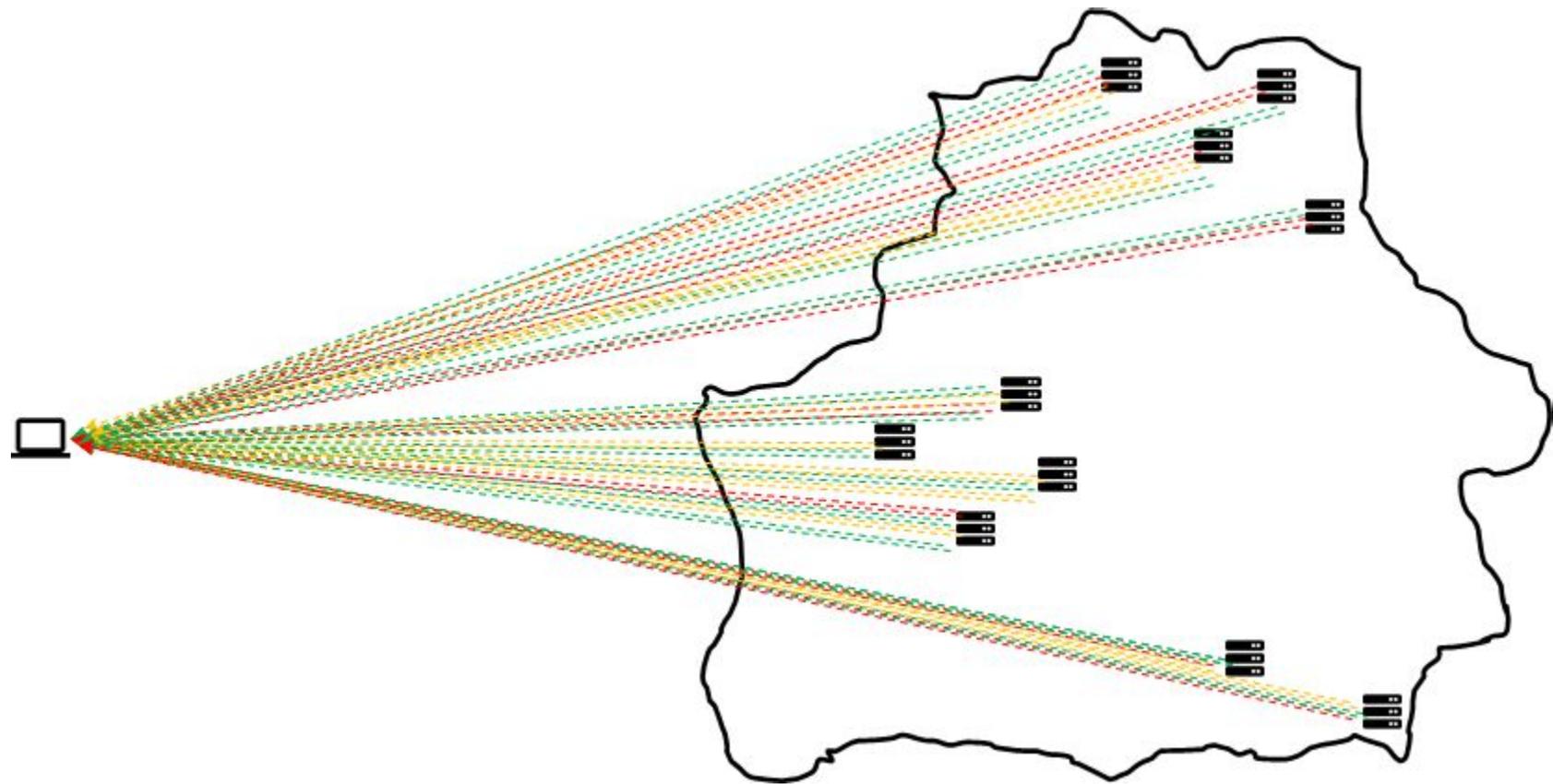
Outcome Timeline

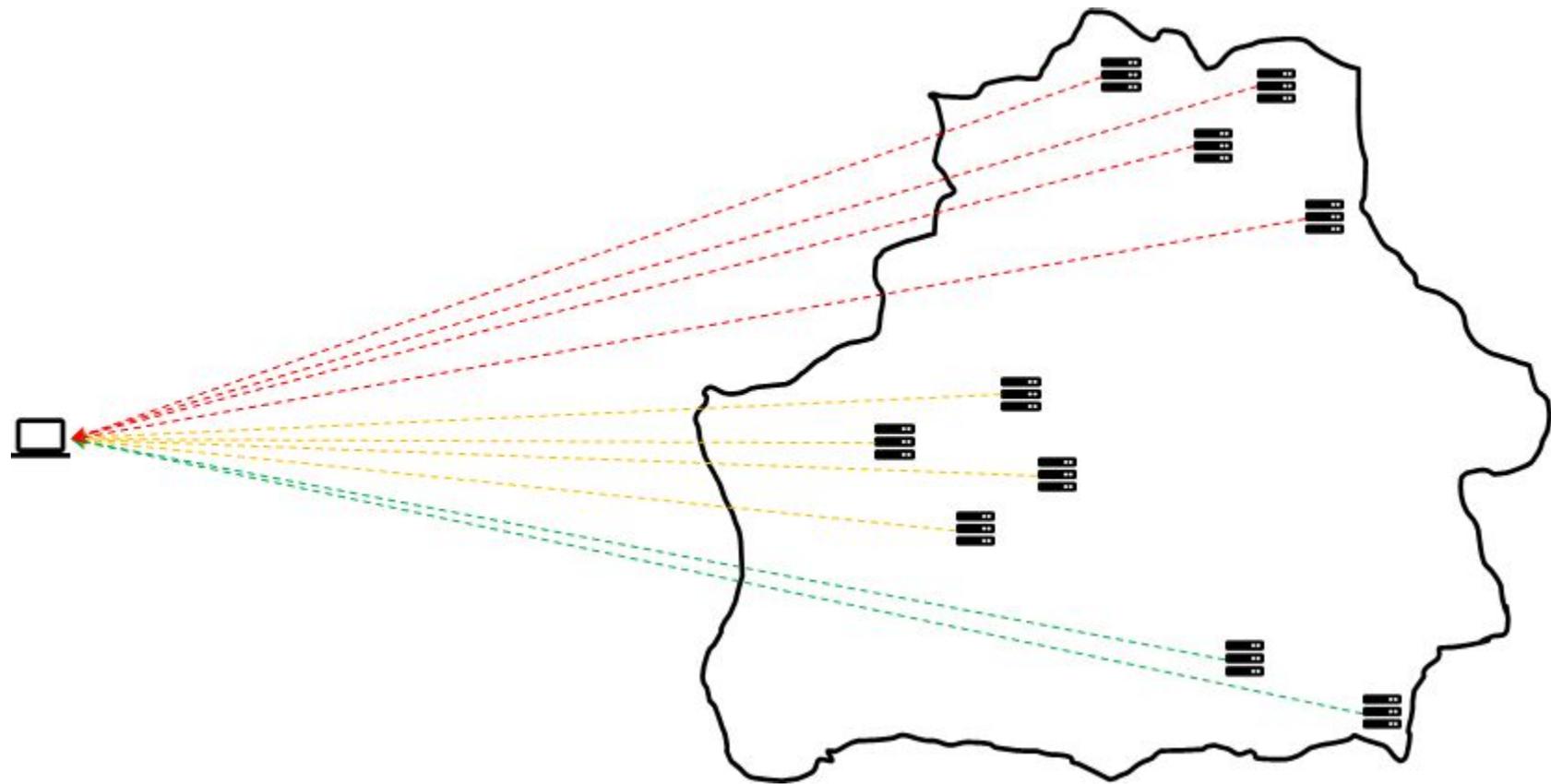
Outcome per network

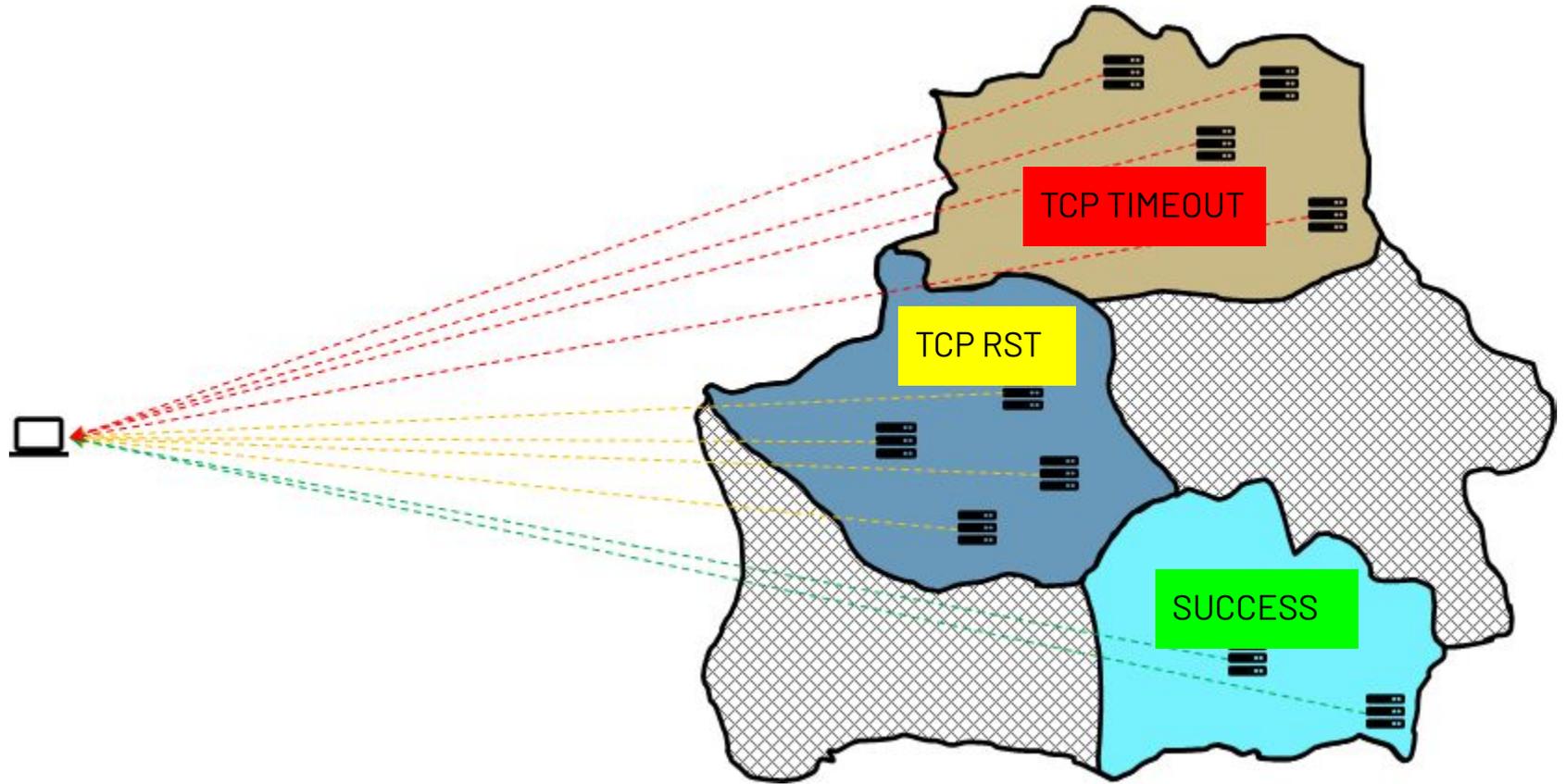
This dashboard is an exploration of the [Censored Planet](#) data, based on a [data analysis pipeline](#) developed in collaboration with [JIGSAW](#)

Measurements should display
consistent behavior when
correctly aggregated.

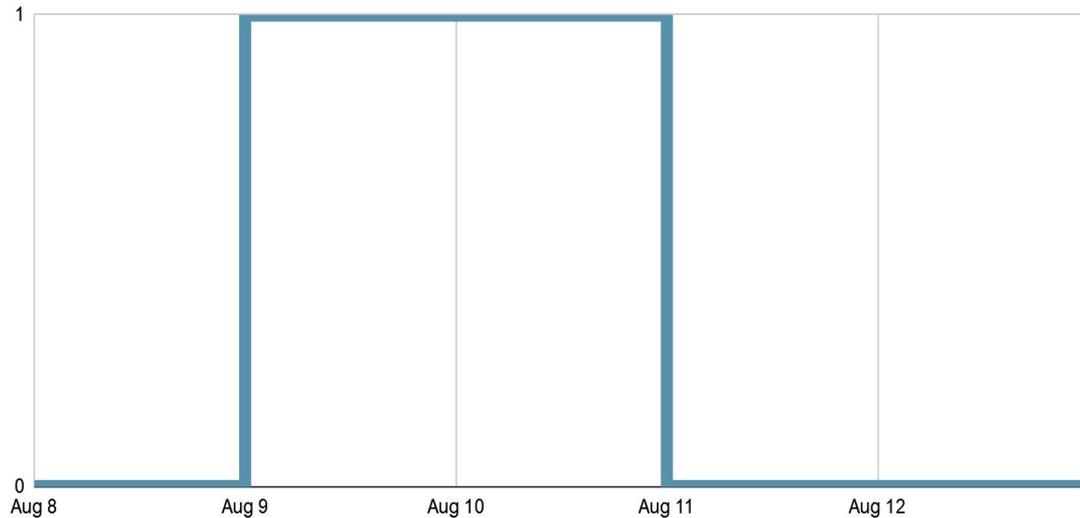
This translates to unexpected
outcome rates of ~0% or ~100%.



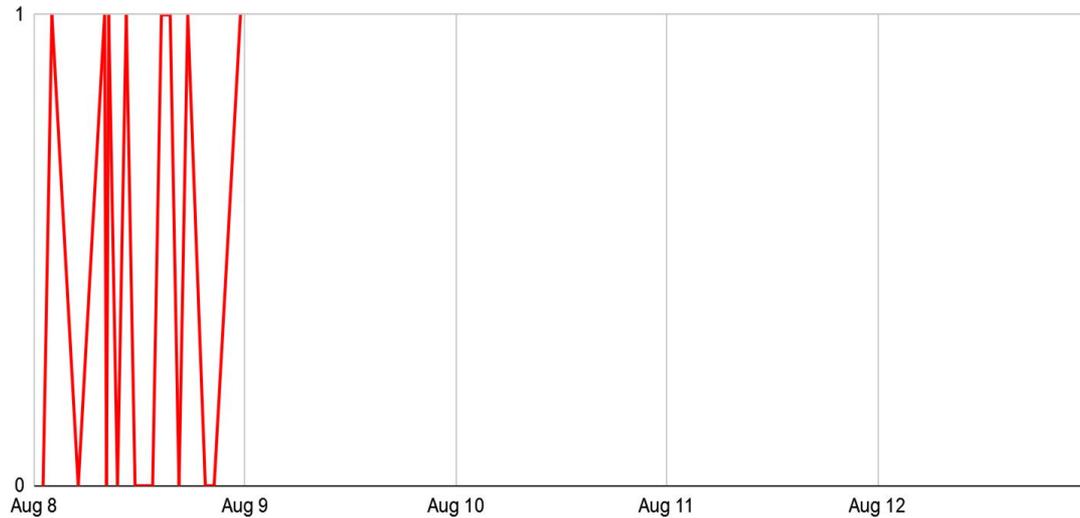




How do we identify longitudinal changes in censorship?

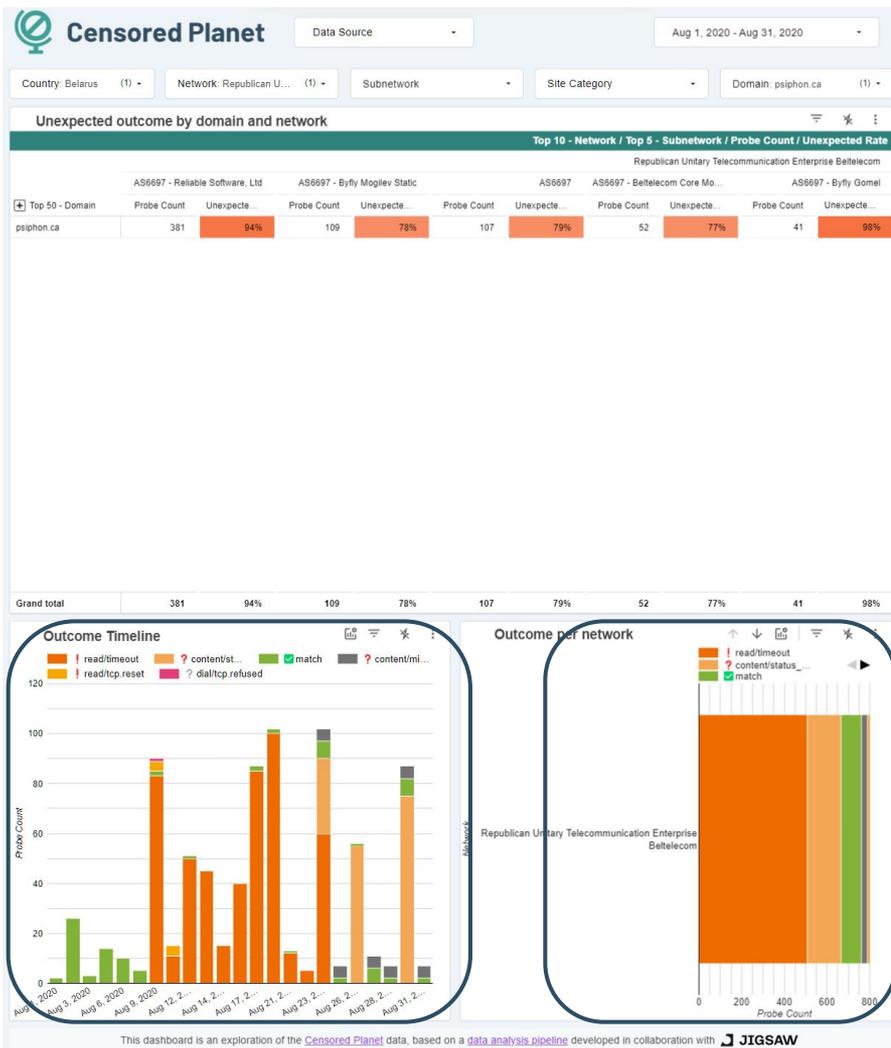


How do we identify longitudinal changes in censorship?



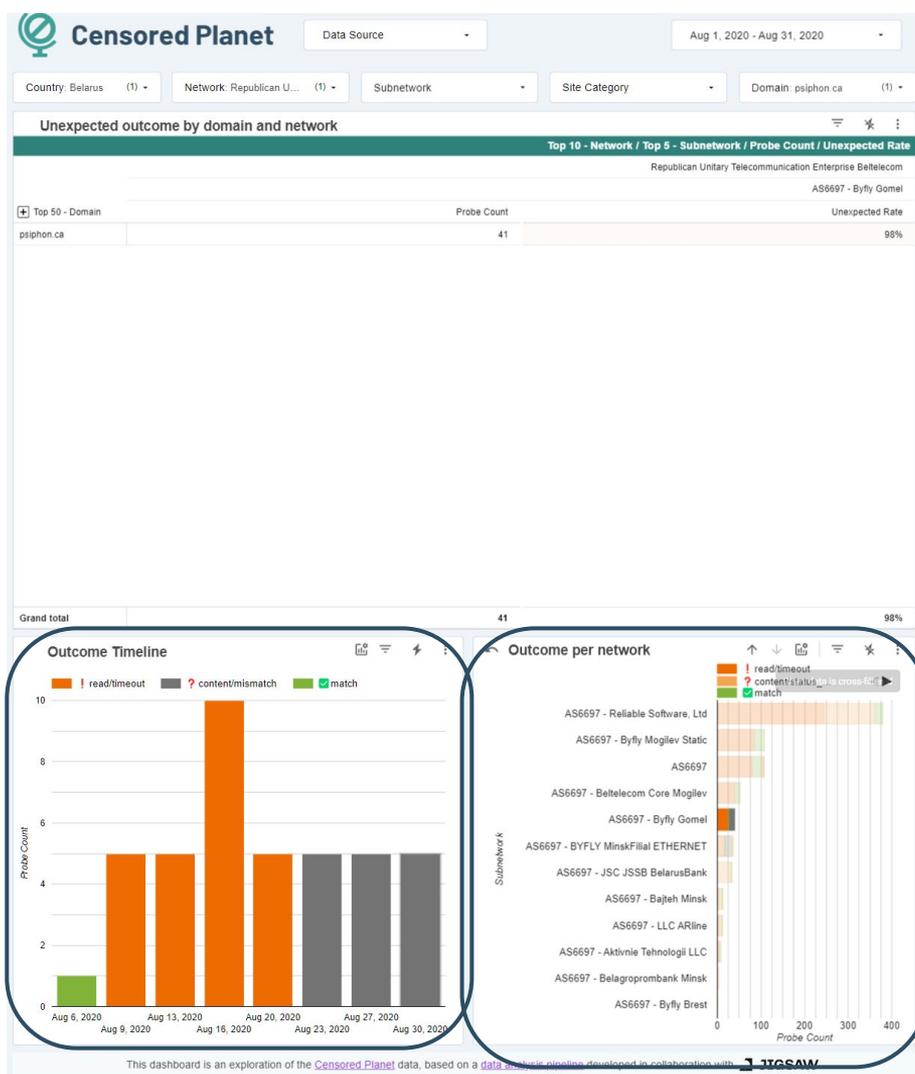
psiphon.ca

- Country: Belarus
- Election day: Aug. 9
- CP Data: Aug. 2020
- AS6697: Beltelecom



psiphon.ca

- Country: Belarus
- Election day: Aug. 9
- CP Data: Aug. 2020
- AS6697: Beltelecom
- BYFLY



Characterizing Censorship

Can users in Belarus access social media?



Characterizing Censorship

Can users in Belarus access social media?

Data Collection

- Empirical Internet Measurements
- Qualitative Studies

Characterizing Censorship

Can users in Belarus access social media?

Data Collection

- Empirical Internet Measurements
- Qualitative Studies

Data Analysis

- Processing big data
- Extending with metadata
- Identifying and classifying censorship

Characterizing Censorship

Can users in Belarus access social media?

Data Collection

- Empirical Internet Measurements
- Qualitative Studies

Data Analysis

- Processing big data
- Extending with metadata
- Identifying and classifying censorship

Data Exploration

- Drilling and expanding data through visualizations and metrics

Characterizing Censorship

Can users in Belarus access social media?

Data Collection

- Empirical Internet Measurements
- Qualitative Studies

Data Analysis

- Processing big data
- Extending with metadata
- Identifying and classifying censorship

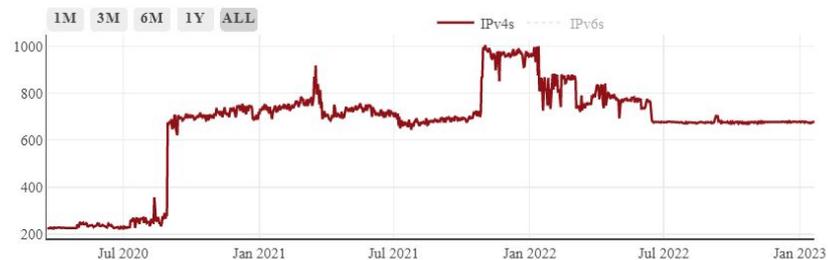
Data Exploration

- Drilling and expanding data through visualizations and metrics

Challenges in Censorship Data Analysis: Unexpected Interference

- **CDN and hosting configurations**
 - DDoS/Bot protection
 - Specific CDN behavior (e.g. Akamai edge)
 - Some censorship infrastructure is on CDNs

Number of forged IPv4s and IPv6s injected by the Great Firewall over time



Challenges in Censorship Data Analysis: **Unexpected Interference**

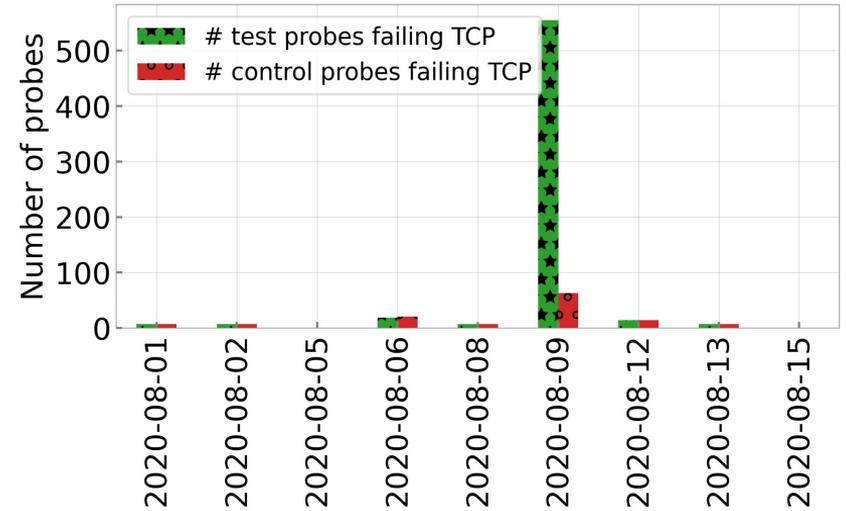
- CDN and hosting configurations
- **Internet Geoblocking**



HTTP Geoblocking

Challenges in Censorship Data Analysis: **Unexpected Interference**

- CDN and hosting configurations
- Internet Geoblocking
- **Internet Shutdowns**

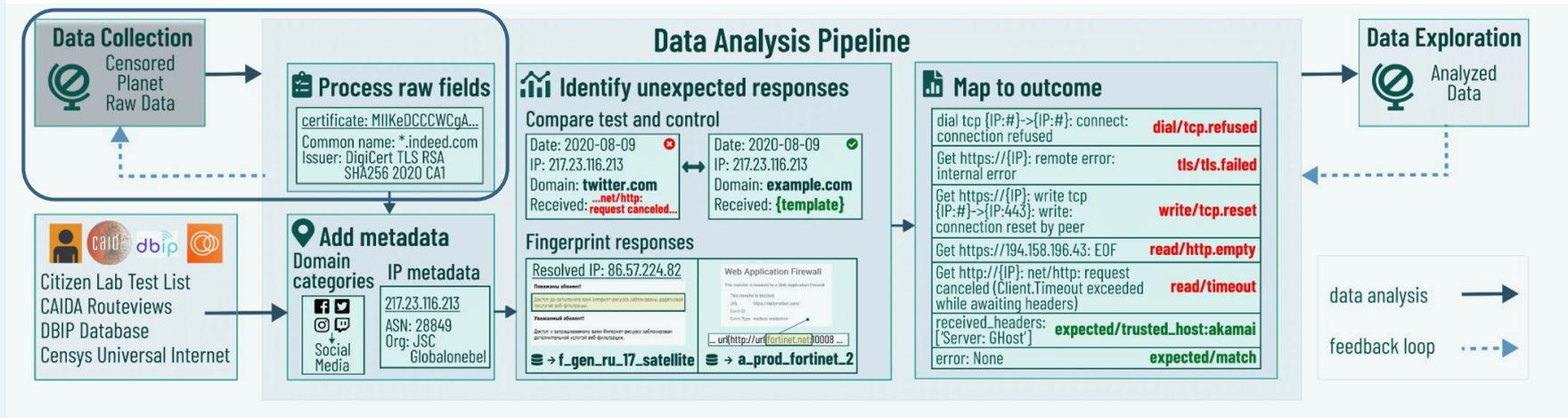


Increase in probe failures during Internet shutdown in Belarus, August 2020

Censored Planet Data Analysis Pipeline

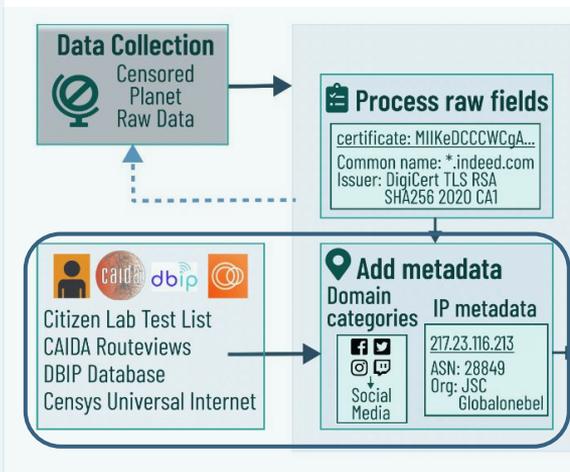
Process Raw Data

- Process specific to dataset
- Can be extended easily using a new module for other datasets



Censored Planet Data Analysis Pipeline

- Add domain metadata such as category, TLS certificates, HTTP Body



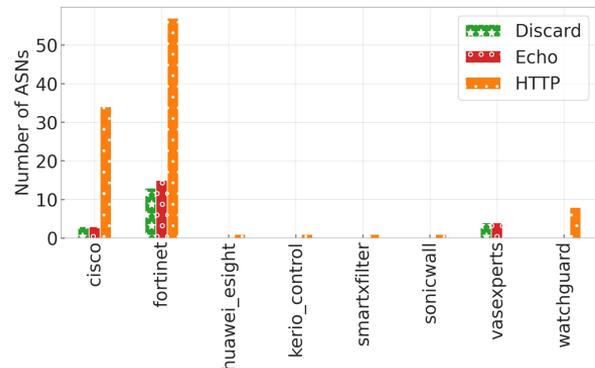
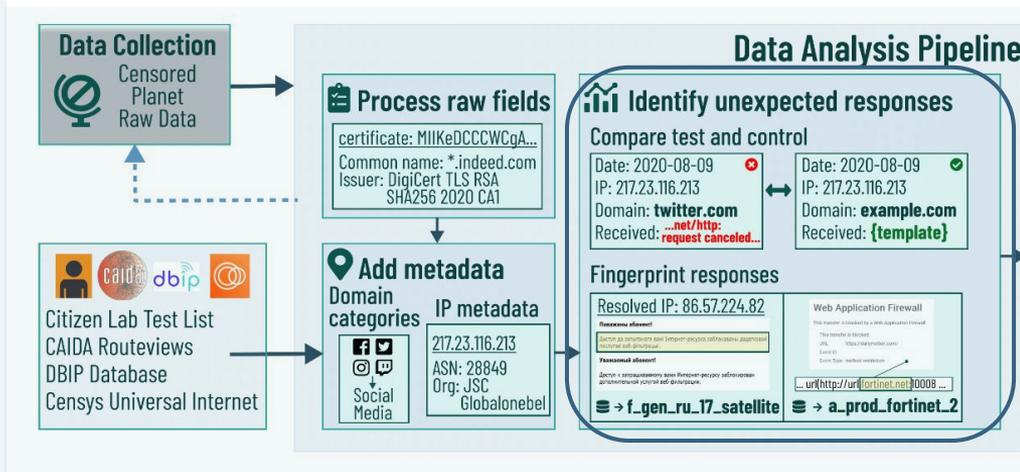
SkyDNS Root CA
Корневое бюро сертификации
Истекает: понедельник, 10 января 2028 г., 14:57:20 Екатеринбург, стандартное время
+ Данный сертификат помечен как надежный для этой учетной записи

Имя	Тип	Срок действия
-----	-----	---------------

Source: <https://www.skydns.ru/guides/tls-ca-setup/>

Censored Planet Data Analysis Pipeline

- Compare with control measurements to identify measurements to look further into
- Check responses for indications of censorship



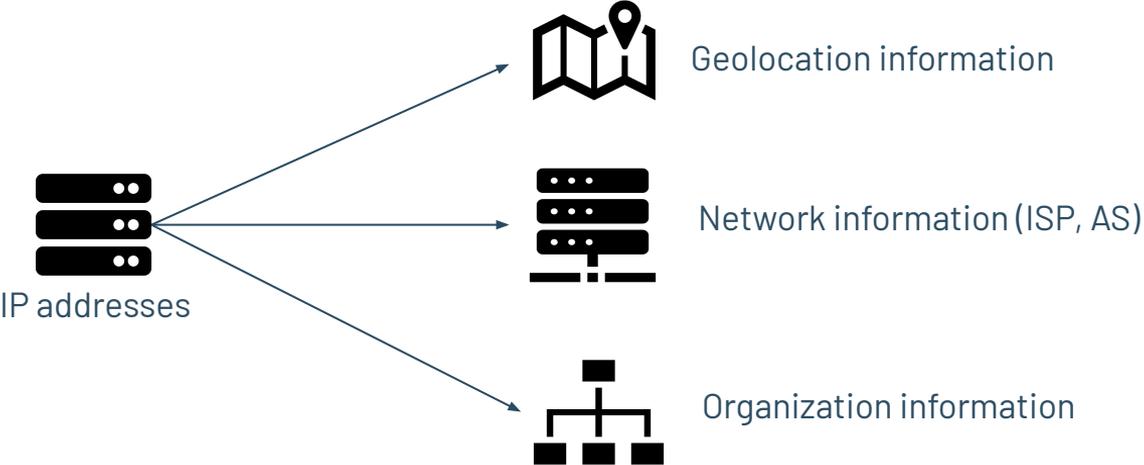
Number of ASNs with commercial firewalls in Censored Planet data sources

Outcomes in HTTP measurements

Stage	Outcome	Num. Measurements	% Measurements	Outcome Type
Expected Response	expected/match	1,772,014,793	94.45	✓
	expected/hosting_provider (e.g. akamai)	61,943,574	3.30	✓
Content Mismatch	content/known_not_censorship	16,642,905	0.89	✓
	content/status_mismatch	13,533,254	0.72	?
	content/known_blockpage	743,396	0.04	!
Read/Write Failure	read/timeout	6,356,637	0.34	!
	read/tcp.reset	4,273,880	0.23	!
Dial Failure	dial/ip.no_route_to_host	28,954	0.001	?

Challenges in Censorship Data Analysis:

Accurate Metadata



Challenges in Censorship Data Analysis:

Accurate Metadata

