

# Advancing the Art of Censorship Data Analysis

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

15 February 2023

#### 🛞 OONI

#### = (

\_



# 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

Internet censorship continues to advance, necessitating high-quality data



#### FREEDOM OF EXPRESSION

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

3 MARCH 2021 | 5:00 AM

### **State of Censorship Data**

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



#### 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



#### 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



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

OONI   Explorer		
	<b>! Anomaly</b> http://www.facebook.com DNS tampering	
Myanmar Country	AS58952 Network	February 17, 2021, 04:38 PM UTC Date & Time
🕄 OONI   Explorer		Search MAT Charts Circumvention Charts Cour
🕄 OONI   Explorer	<b>! Anomaly</b> http://www.facebook.com TCP/IP blocking	Search MAT Charts Circumvention Charts Cou

#### Challenges in Censorship Data Analysis: Data Limitations

#### Need to consider the data's

- Scale
- Coverage
- Continuity
- Protocols





#### Challenges in Censorship Data Analysis: **Data Limitations**

#### Need to consider the data's

- Scale
- Coverage
- Continuity
- Protocols

😵 OONI   Explorer			Search MAT Ch	arts Circumvention Charts Countries
		<b>! Anomaly</b> http://www.facebook.com DNS tampering		
DNS Querie	!S			
Resolver 116.206.	136.161 Loca	IDNS		
Query: IN A www. Engine: system	facebook.com		W	rong Address
Name	Class	TTL	Туре	DATA
@	IN		А	59.153.90.11
OONI   Explorer			Search MAT	Charts Circumvention Charts Countries
😧 OONI   Explorer		<b>! Anomaly</b> http://www.facebook.com TCP/IP blocking	Search MAT	Charts Circumvention Charts Countries
OONI Explorer	S	<b>! Anomaly</b> http://www.facebook.com TCP/IP blocking	Search MAT	Charts Circumvention Charts Countries
OONI  Explorer	s ma Google	! Anomaly http://www.facebook.com TCP/IP blocking	Search MAT	Charts Circumvention Charts Countries
OONI  Explorer DNS Querie: Resolver 172.253.2 Query: IN A www.f Engine: system	s m.3 Google	! Anomaly http://www.facebook.com TCP/IP blocking	search MAT	Charts Circumvention Charts Countries
OONI Explorer	S 111.3 Google facebook.com Class	! Anomaly http://www.facebook.com TCP/IP blocking CDNS	search MAT	Charts Circumvention Charts Countries

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	Censored Planet Measurements in Canada, September 2021
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	[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	Censored Planet Measurements in Canada, September 2021
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	[1] https://stats.labs.apnic.net/cgi-bin/
				aspop?c=ca

- 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

#### - CDN and hosting configurations

- DDoS/Bot protection
- Specific CDN behavior (e.g. Akamai edge)

# **Access Denied**

You don't have permission to access "/" on this server.

1 🕱 🔥

Reference #18.9872c17.1631203469.b24e5df9

#### - CDN and hosting configurations

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





- CDN and hosting configurations
- Internet Geoblocking



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

## Outline



#### Challenges in censorship data analysis

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





### **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



# 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 Identify Unexpected Responses

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



# 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

Data Collection Censored		Data Analysis Pipeling
Planet Raw Data	Certificate: MIIKeDCCCWCgA Common name: *.indeed.com Issuer: DigiCert TLS RSA SHA256 2020 CA1	Compare test and control Date: 2020-08-09 IP: 217.23.116.213 Domain: twitter.com Date: 2020-08-09 IP: 217.23.116.213 Domain: twitter.com Date: 2020-08-09 IP: 217.23.116.213 Domain: twitter.com Domain: twitter.com Domain
Сіtizen Lab Test List CAIDA Routeviews DBIP Database Censys Universal Internet	Add metadata     Domain     categories     IP metadata     IP metadata	Received: request canceled       Received: {template}         Fingerprint responses         Resolved IP: 86.57.224.82         Image: Second and the second



# Censored Planet Data Analysis Pipeline Map to Outcome

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



Stage	Outcome	Num. Measurements	% Measurements	Outcome Type
	expected/match	1,772,014,793	94.45	$\checkmark$
Expected Response	expected/hosting_provider(e.g. akamai)	61,943,574	3.30	$\checkmark$
	content/known_not_censorship	16,642,905	0.89	$\checkmark$
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	?

Stage	Outcome	Num. Measurements	% Measurements	Outcome Type
	expected/match	1,772,014,793	94.45	
Expected Response	expected/hosting_provider(e.g. akamai)	61,943,574	3.30	$\checkmark$
	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	?

Stage	Outcome	Num. Measurements	% Measurements	Outcome Type
	expected/match	1,772,014,793	94.45	$\checkmark$
Expected Response	expected/hosting_provider(e.g. akamai)	61,943,574	3.30	$\checkmark$
	content/known_not_censorship	16,642,905	0.89	$\checkmark$
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	?

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









# Censored Planet Data Analysis Pipeline Data Exploration

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



- Enables visualizations for longitudinal data
- Automatically updated after measurements
- Open to public https://dashboard.censoredplanet.org



Censored Planet

	(1) - N	letwork		- Su	bnetwork		- Site	Category	-	Domai	n	
Unexpected	outcome b	v domain	and netwo	rk							Ŧ	*
		,					Top 10 -	Network / Te	op 5 - Subnet	work / Probe	Count / Unex	pected R
							Republican Unita	ary Telecommur	ication Enterpris	e Beltelecom	Konstanta-Bre	st Ltd , Be
		A\$6697	AS6697 - Byfl	y Mogilev	AS6697 - Rel	iable Softw	AS6697	- Byfly Brest	AS6697 - Belt	elecom C	AS202324 - K	onstanta
] Top 50 - Domain	Probe C	Unexpe	Probe C	Unexpe	Probe C	Unexpe	Probe C	Unexpe	Probe C	Unexpe	Probe C	Unexpe
ww.svoboda.org	37	65%	13	0%	12	67%	14	64%	6	0%	7	(
ww.hrw.org	44	55%	15	0%	18	56%	16	81%	7	0%	7	3
iphon.ca	47	60%	15	0%	19	53%	15	73%	7	0%	7	
ww.europacasino.com	48	60%	15	0%	21	62%	11	55%	7	0%	7	
ww.partypoker.com	43	56%	15	0%	20	60%	16	63%	7	0%	7	
ww.betfair.com	45	56%	15	0%	17	47%	16	69%	7	0%	7	
ww.svaboda.org	38	63%	13	0%	11	64%	10	30%	6	0%	7	
ww.casinotropez.com	45	53%	15	0%	20	55%	16	81%	7	0%	7	
treon.com	46	57%	15	0%	16	44%	12	50%	7	0%	7	
ww.dw.com	42	50%	15	096	18	56%	10	30%	7	0%	7	
ww.change.org	45	56%	15	0%	19	53%	11	36%	7	0%	7	
ange.org	45	56%	15	0%	16	44%	13	62%	7	0%	7	
ww.goldenpalace.com	44	55%	15	0%	21	62%	11	36%	7	0%	7	
ww.currenttime.tv	44	52%	15	0%	19	53%	12	58%	7	0%	7	
ww.888casino.com	46	59%	15	0%	16	44%	14	50%	7	0%	7	
ww.pokerstars.com	43	53%	15	0%	20	60%	9	22%	7	0%	7	
irfshark.com	45	56%	15	0%	17	53%	11	36%	7	0%	7	
		1%	30.8K	0%	22.7K	1%	14.6K	1%	14.4K	0%	14.4K	
and total (2113) 🛈	57.5K											
Outcome Ti Outcome Ti 20match 40K	57.5K	ent/m ? dial/timeout	? content/st	/http.empty	F 1 I	0	utcome pe	r network		↑ ↓ match ? conter ? conter	년 구 t/mismatch t/status	*
and total (2113) © Outcome Ti © match 40K 30K	57.5K	ent/m ? dial/timeout	? content/st	iii = ? ( /http.empty	≂ ≁ : content/h	Repu	utcome pe blican Unitary T Konstanta-Bre	r network	tion Enterprise Beltelecom s ISP company	Conter Conter	tt/mismatch tt/status	*

# 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

- Controls
  - o data source
  - o date
  - country
  - $\circ$  network
  - $\circ$  subnetwork
  - $\circ$  site category
  - $\circ$  domain



🔮 Cens	sored	Plane	et 🛛	ata Source		-			Jan 3	0, 2023 - Feb	12, 2023	•
Country: Belerue	(1) <b>•</b> N	Jetwork			boetwork		• Site	Category		Doma	n	
oounity: Denirus		ictitoric			DITERMOTE		One	outegory		Doma		
Unexpected	outcome b	oy domain	and networ	k								*
								Network / To	p 5 - Subnew	Nork / Probe	Count / Unex	pected R
											AS202324 - K	
	Probe C	Unexpe	Probe C	Unexpe	Probe C	Unexpe	Probe C	Unexpe	Probe C	Unexpe	Probe C	Unexpe.
							14		6			
	44											
phon.ca	47				19							
w.europacasino.com	48											
w.partypoker.com	43						16					
	45											
									6			
w.casinotropez.com	45						16					
eon.com	46											
	42											
w.change.org	45				19							
nge.org	45											
w.goldenpalace.com	44											
w.currenttime.tv	44				19							
	46						14					
w.pokerstars.com	43						9					
	45											
ind total (2113) 🛈	57.5K	1%	30.8K	0%	22.7K	1%	14.6K	1%	14.4K	0%	14.4K	+(
IOUCCOME II match ioK	neout Cont	ent/m ? dial/timeout	? content/st ! read/h	ttp.empty	content/h	Repu	blican Unitary T	elecommunica		Conter Conter		
юк юк						Metwork	Konstanta-Bre	st Ltd , Belarus Mobile Tele				
									/ enterprise A1			
0 Jan 30, 2023 Feb	o 1, 2023 Feb	3, 2023 Feb	6, 2023 Feb 8,	2023 Feb	10, 2023							

• Unexpected outcome by domain and network



- Outcome Timeline
- Outcome per network





# **Belarus**

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

**J** JIGSAW



 $\oslash$ 

Country: Belarus (1) -

+ Top 50 - Domain

www.crazyshit.com

psiphon.ca

tutanota.com

www.betternet.co

www.privateinternetad

www.hidemyass.com

bridges.torproject.org

www.purevpn.com

google.pt

google.org

google.ch

google.com

twitter com

apple.com

vk.com

t.co

**Censored Planet** 

Network

Unexpe.

94%

75%

38%

24%

3496

28%

30%

13%

18%

0%

14%

10%

10%

26%

10%

Unexpected outcome by domain and network

AS6697 - Reliable Softw.

462

381

233

133

116

129

121

122

78

28

24

105

101

101

117

101

Probe C.

Data Source

Unexpe.

0%

78%

0%

36%

36%

36%

36%

30%

63%

36%

40%

29%

20%

36%

34%

36%

AS6697 - Byfly Mogilev

40

109

43

56

56

56

56

56

30

56

50

56

56

56

56

56

Probe C.

Subnetwork

Probe C...

79

107

40

63

63

62

68

68

40

48

41

63

63

64

55

59

August 2020

12 13 14

24 25 26 27 28

30 31 1 2 3 4 5

996

20%

21%

13%

45%

20%

22%

20%

20%

20%

16%

20%

Thu

19 20 21

29

10

10

10

10

8

2

2

10

10

10

10

10

30 3 6

Sun Mon Tue Wed

1

26

16 18

23

23

25

24

24

11

25

23

25

25

25

25

25

Reput

AS6697

63%

79%

0%

44%

46%

40%

49%

51%

70%

40%

37%

48%

44%

47%

36%

39%

Unexpe.

Fixed -

Fri

6 7 8

26 27 28 29

14

10

10

10

10

10

14

10

14

10

14

-\*

17 18 19 20 21 22

30 31 1 2 3 4 5

12 13 14 15

Sat

36%

0%

0%

0%

0%

0%

36%

0%

36%

0%

36%

1%

August 2020

10 11

CANCEL

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

16

23 24 25

# Which networks?

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





Q Censored	d Planet	Data Source	÷			Aug 1, 20	120 - Aug 31, 20	20		
Country: Belarus (1) +	Vetwork		twork	•	Site Category	•	Domain			•
Unexpected outcom	Q Type to search							Ŧ	*	1

# Visible ASNs: Customer Populations (Est.)

Which networks?



# 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





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





Can users in Belarus access social media?

Can users in Belarus access social media?

#### **Data Collection**

- Empirical Internet Measurements
- Qualitative Studies

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

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

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

#### - CDN and hosting configurations

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



Jul 2021

Jan 2022

Jul 2022

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

Jan 2021

Jul 2020

Jan 2023

- CDN and hosting configurations
- Internet Geoblocking

#### 403 ERROR

#### The request could not be satisfied.

The Amazon CloudFront distribution is configured to block access from your country. We can't connect to the server for this app or website at this time. There might be too much traffic or a config later, or contact the app or website owner.

If you provide content to customers through CloudFront, you can find steps to troubleshoot and help prevent this error by reviewing the CloudFront documentation.

Generated by cloudfront (CloudFront) Request ID: VhQy6f3LA699Q9-vVQFhRmI4ZOtHSRy3gD2tjv0fm7ChvoV1IwEJZA==

0 💽

**HTTP Geoblocking** 

I 🛡 🔺

- 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

Data Collection		Data Analysis Pipelin	Data Exploration	
Planet	🖆 Process raw fields	iii Identify unexpected responses	🛅 Map to outcome	Analyzed Data
	certificate: MIIKeDCCCWCgA	Compare test and control	dial tcp {IP:#}->{IP:#}: connect: dial/tcp.refused	-
	Issuer: DigiCert TLS RSA SHA256 2020 CA1	IP: 217.23.116.213 ↔ IP: 217.23.116.213	Get https://{IP}: remote error: tls/tls.failed	
		Received: request canceled	Get https://{IP}: write tcp {IP:#}->{IP:443}: write: write/tcp.reset	
	Add metadata	Fingerprint responses	Get https://194.158.196.43: EOF read/http.empty	
Citizen Lab Test List CAIDA Routeviews	categories IP metadata	Resolved IP: 86.57.224.82 Recover denset Identity assessed response for the process information approximation ap	Get http://{IP}: net/http: request canceled (Client.Timeout exceeded while awaiting headers) read/timeout	data analysis 🛛 🗕 🕨
DBIP Database	ASN: 28849 Org: JSC	Descent advection         Descent advection           Anne or support advectory advector	received_headers: ['Server: GHost'] expected/trusted_host:akamai	feedback loop
Lensys Universal Internet	T Globalonebel	$\blacksquare \Rightarrow f\_gen\_ru\_17\_satellite \blacksquare \Rightarrow a\_prod\_fortinet\_2$	error: None expected/match	

# **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 О 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

Stage	Outcome	Num. Measurements	% Measurements	Outcome Type
	expected/match	1,772,014,793	94.45	$\checkmark$
Expected Response	expected/hosting_provider(e.g. akamai)	61,943,574	3.30	$\checkmark$
	content/known_not_censorship	16,642,905	0.89	$\checkmark$
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	?

#### Challenges in Censorship Data Analysis:

### Accurate Metadata



Challenges in Censorship Data Analysis: Accurate Metadata

