Lightweight Fingerprint Attack and Encrypted Traffic Analysis on News Articles

David Hasselquist, Linköping University & Sectra Communications, Sweden

Martin Lindblom, Linköping University, Sweden

Niklas Carlsson, Linköping University, Sweden

Proc. IFIP Networking, Catania, Italy, June 2022

LINKÖPING UNIVERSITY

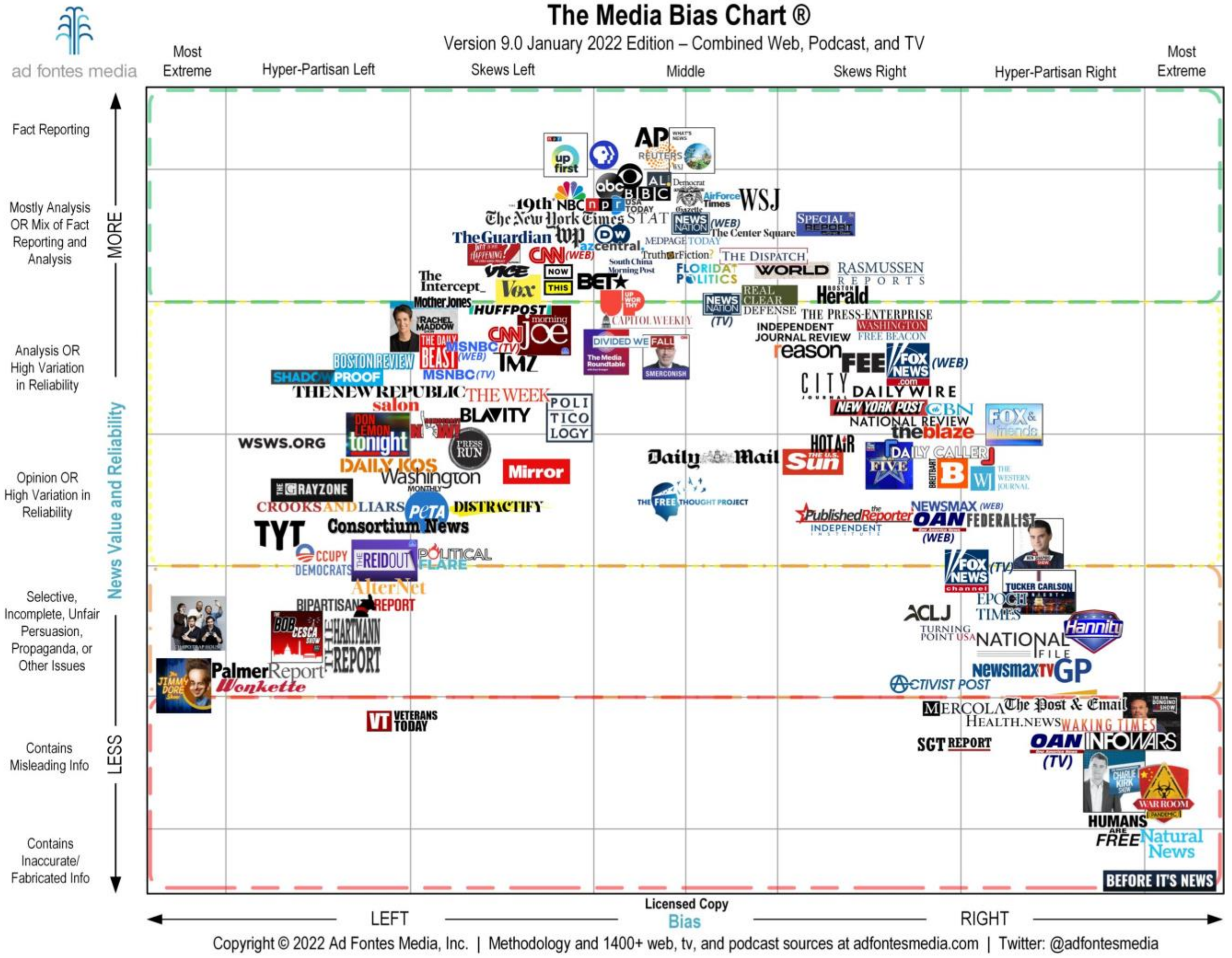
WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

SECTRA

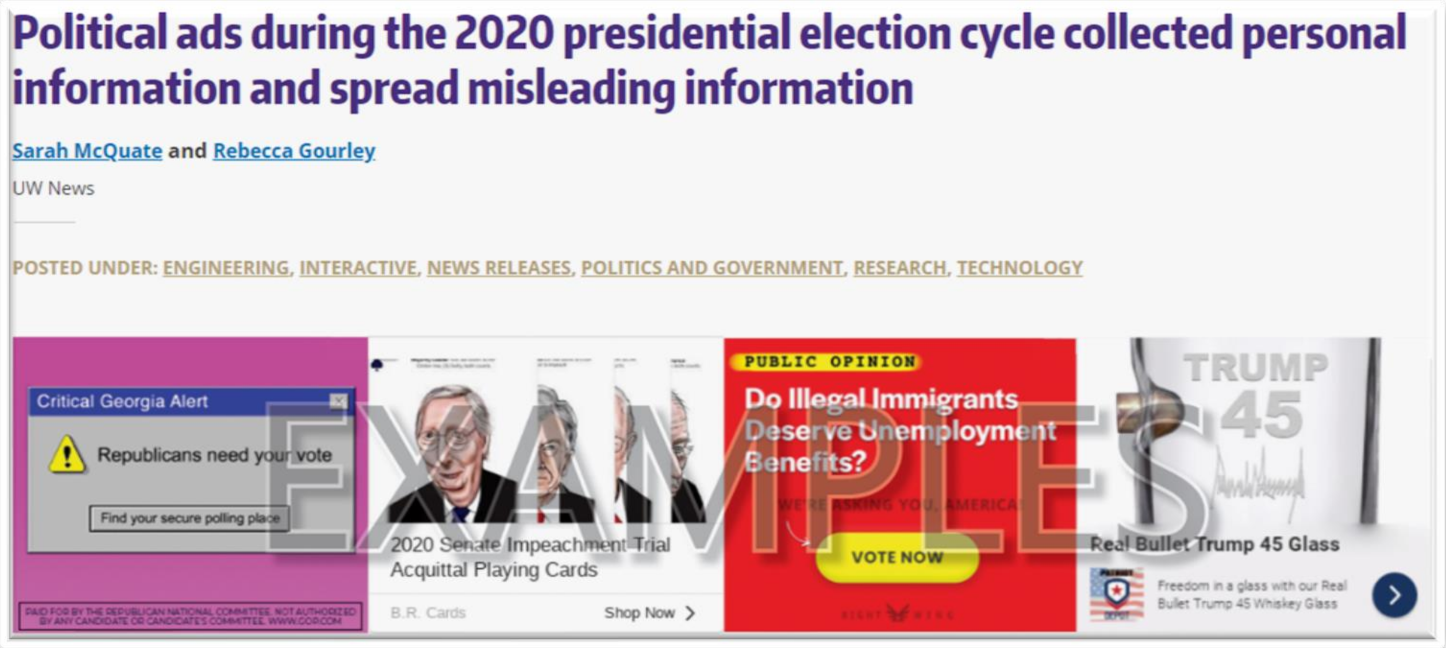
Motivation

- » Most of our news obtained online today
- » The news we read can reveal much about us
- » Users should be able to obtain independent news without adversary monitoring or control
- » An adversary capable of extracting small fraction of our obtained news presents a privacy threat

Example: news bias



Examples: political misinformation



Political ads during the 2020 presidential election cycle collected personal information and spread misleading information

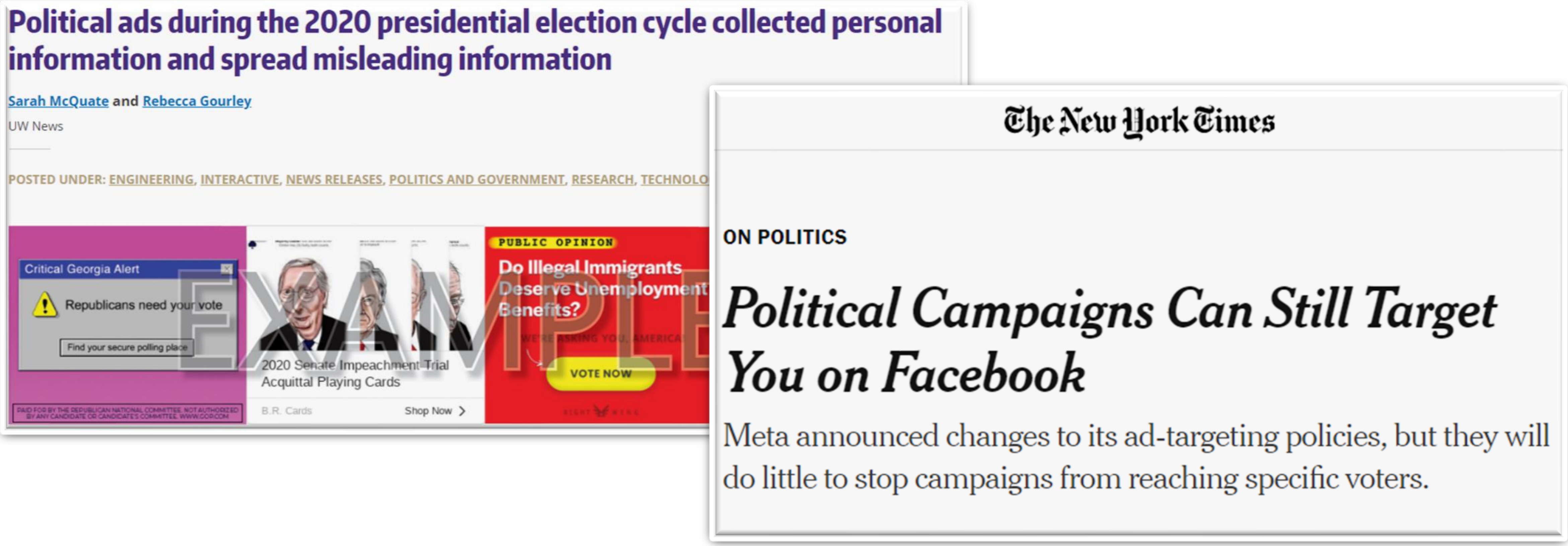
Sarah McQuate and Rebecca Gourley

UW News

POSTED UNDER: ENGINEERING, INTERACTIVE, NEWS RELEASES, POLITICS AND GOVERNMENT, RESEARCH, TECHNOLOGY

EXAMPLES
Critical Georgia Alert
Republicans need your vote
Find your secure polling place
PAID FOR BY THE REPUBLICAN NATIONAL COMMITTEE. NOT AUTHORIZED BY ANY CANDIDATE OR CANDIDATE'S COMMITTEE. WWW.GOP.COM
2020 Senate Impeachment Trial Acquittal Playing Cards
B.R. Cards
Shop Now
PUBLIC OPINION
Do Illegal Immigrants Deserve Unemployment Benefits?
WE'RE ASKING YOU, AMERICA!
VOTE NOW
RIGHT WING
TRUMP 45
Real Bullet Trump 45 Glass
Freedom in a glass with our Real Bullet Trump 45 Whiskey Glass

Examples: political misinformation



Political ads during the 2020 presidential election cycle collected personal information and spread misleading information

Sarah McQuate and Rebecca Gourley

UW News

POSTED UNDER: ENGINEERING, INTERACTIVE, NEWS RELEASES, POLITICS AND GOVERNMENT, RESEARCH, TECHNOLO

Critical Georgia Alert
Republicans need your vote
Find your secure polling place
PAID FOR BY THE REPUBLICAN NATIONAL COMMITTEE. NOT AUTHORIZED BY ANY CANDIDATE OR CANDIDATE'S COMMITTEE. WWW.GOP.COM
EXAMPLE
2020 Senate Impeachment Trial Acquittal Playing Cards
B.R. Cards
Shop Now
PUBLIC OPINION
Do Illegal Immigrants Deserve Unemployment Benefits?
WE'RE ASKING YOU, AMERICA
VOTE NOW
RIGHT WING

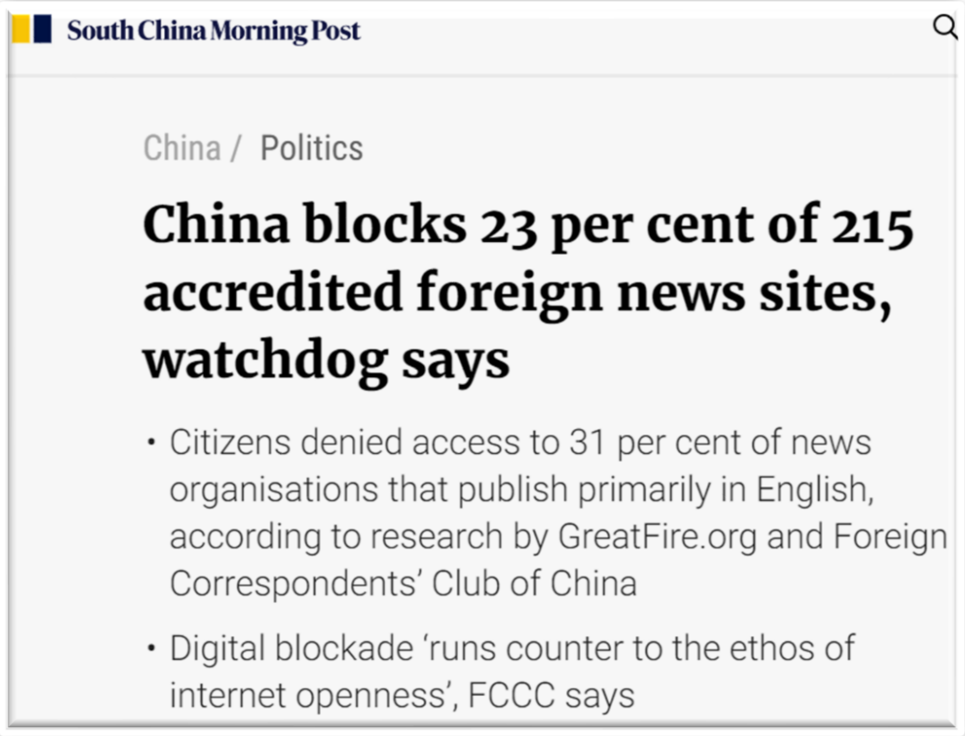
The New York Times

ON POLITICS

Political Campaigns Can Still Target You on Facebook

Meta announced changes to its ad-targeting policies, but they will do little to stop campaigns from reaching specific voters.

Examples: news filtering

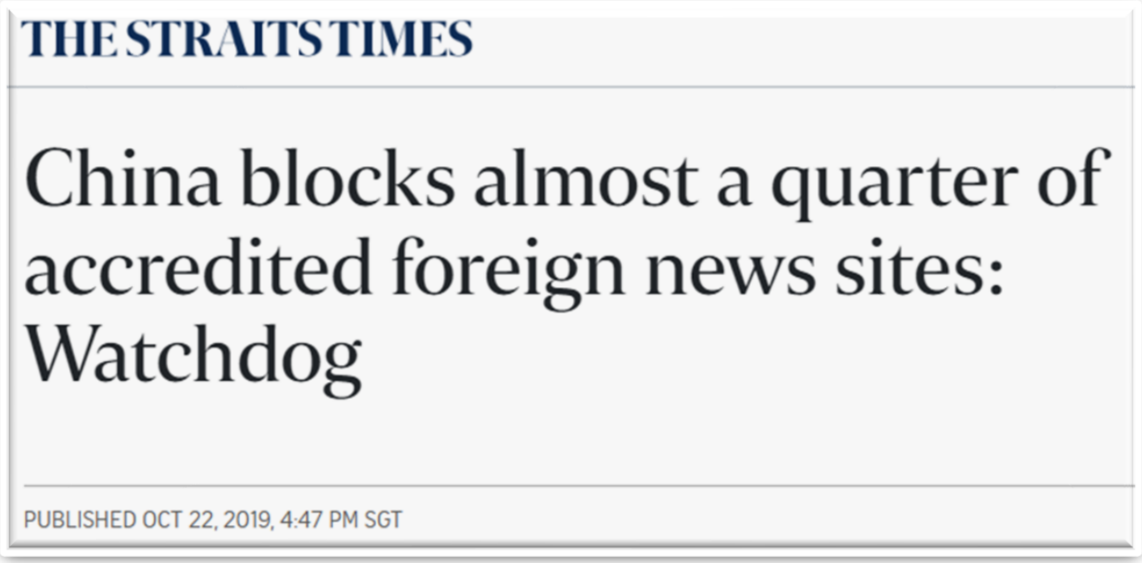


South China Morning Post

China / Politics

China blocks 23 per cent of 215 accredited foreign news sites, watchdog says

- Citizens denied access to 31 per cent of news organisations that publish primarily in English, according to research by GreatFire.org and Foreign Correspondents' Club of China
- Digital blockade 'runs counter to the ethos of internet openness', FCCC says

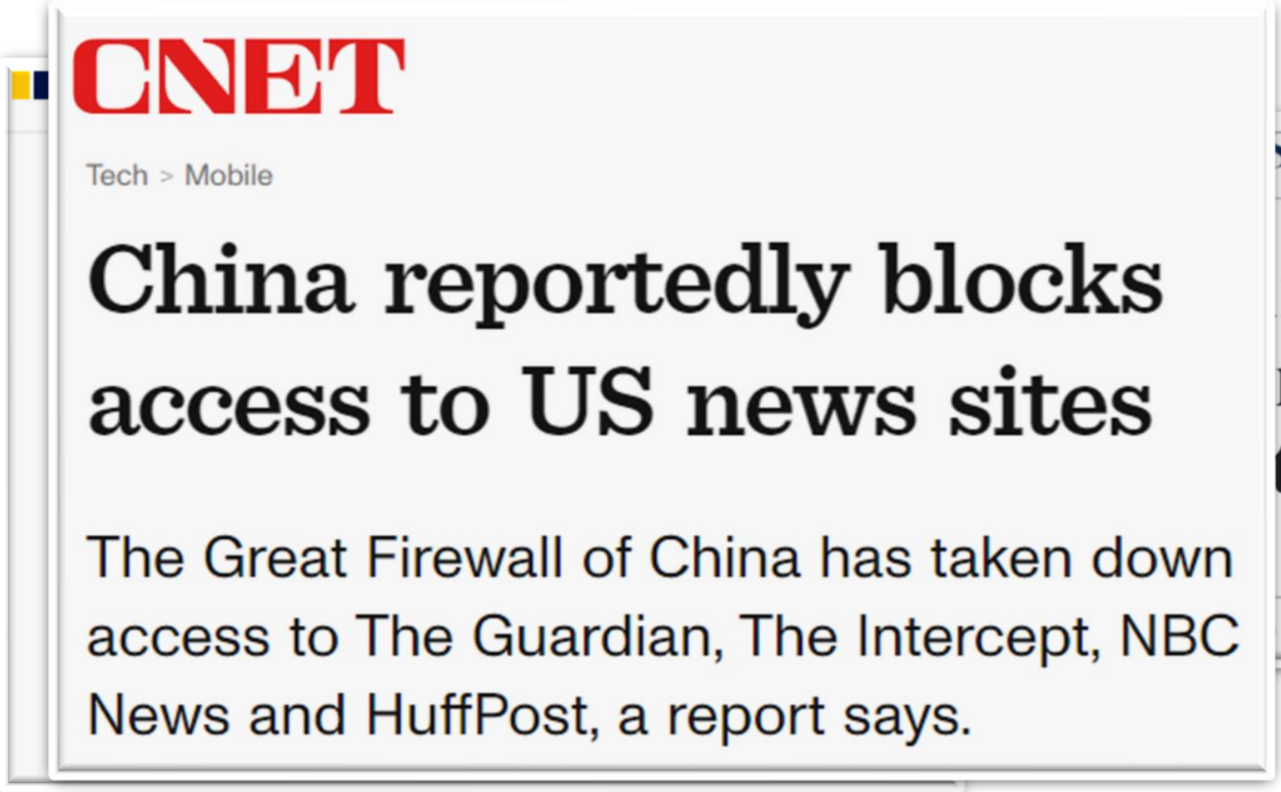


THE STRAITS TIMES

China blocks almost a quarter of accredited foreign news sites: Watchdog

PUBLISHED OCT 22, 2019, 4:47 PM SGT

Examples: news filtering

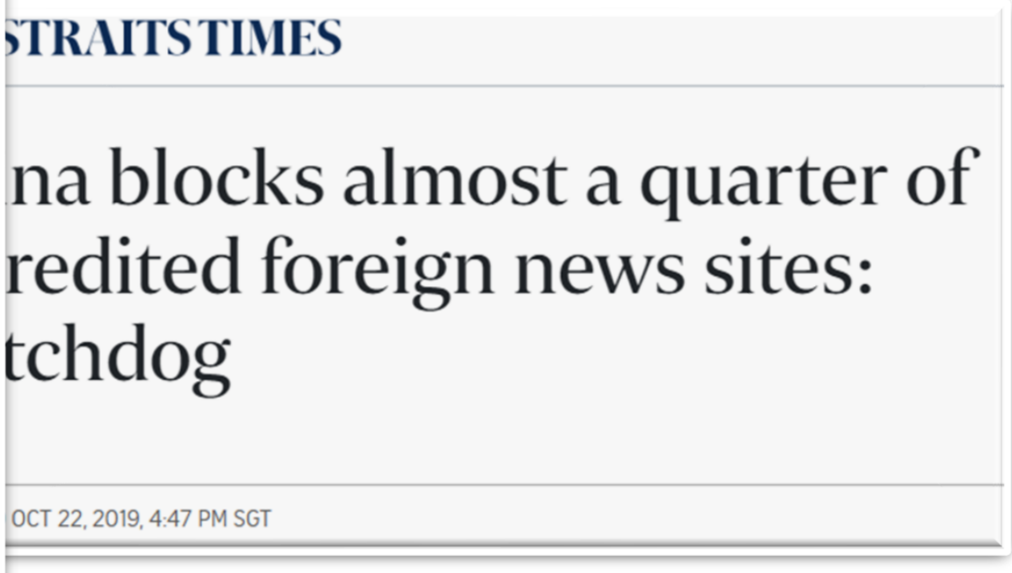


CNET

Tech > Mobile

China reportedly blocks access to US news sites

The Great Firewall of China has taken down access to The Guardian, The Intercept, NBC News and HuffPost, a report says.

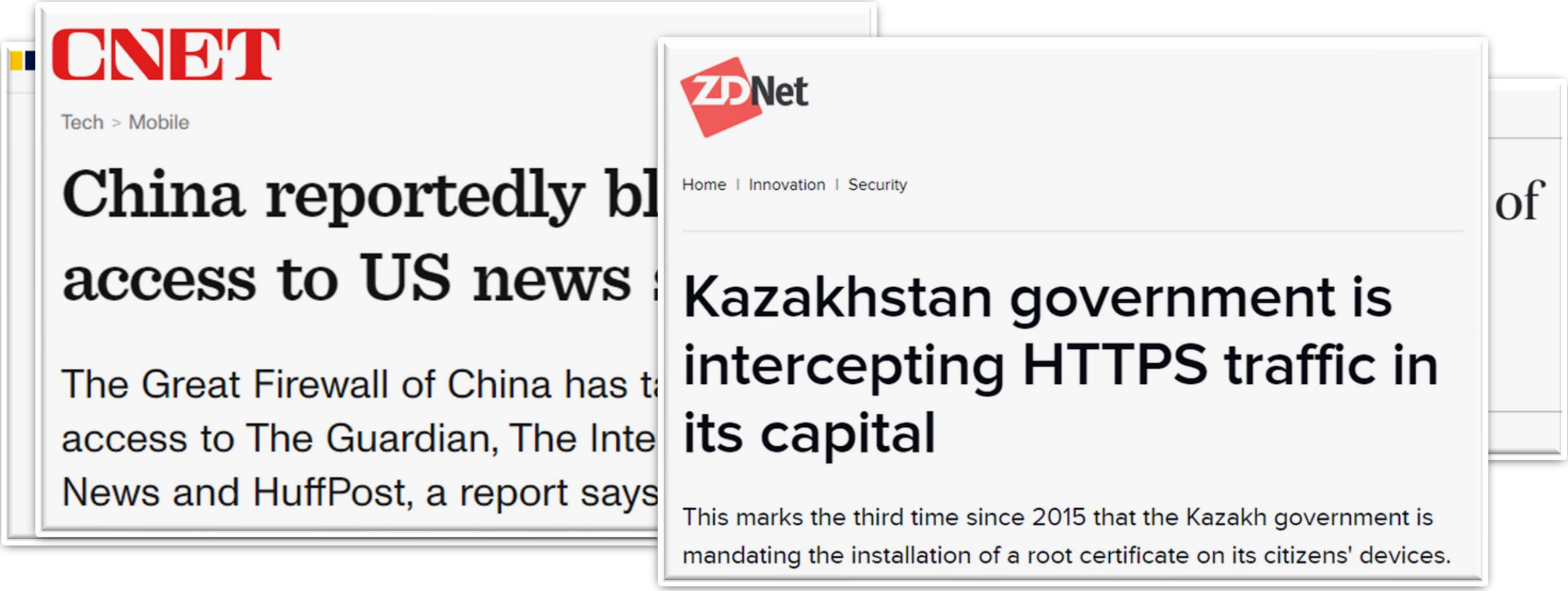


STRAITS TIMES

na blocks almost a quarter of redited foreign news sites: tchdog

OCT 22, 2019, 4:47 PM SGT

Examples: news filtering



CNET

Tech > Mobile

China reportedly bl

access to US news

The Great Firewall of China has t

access to The Guardian, The Inte

News and HuffPost, a report says

of

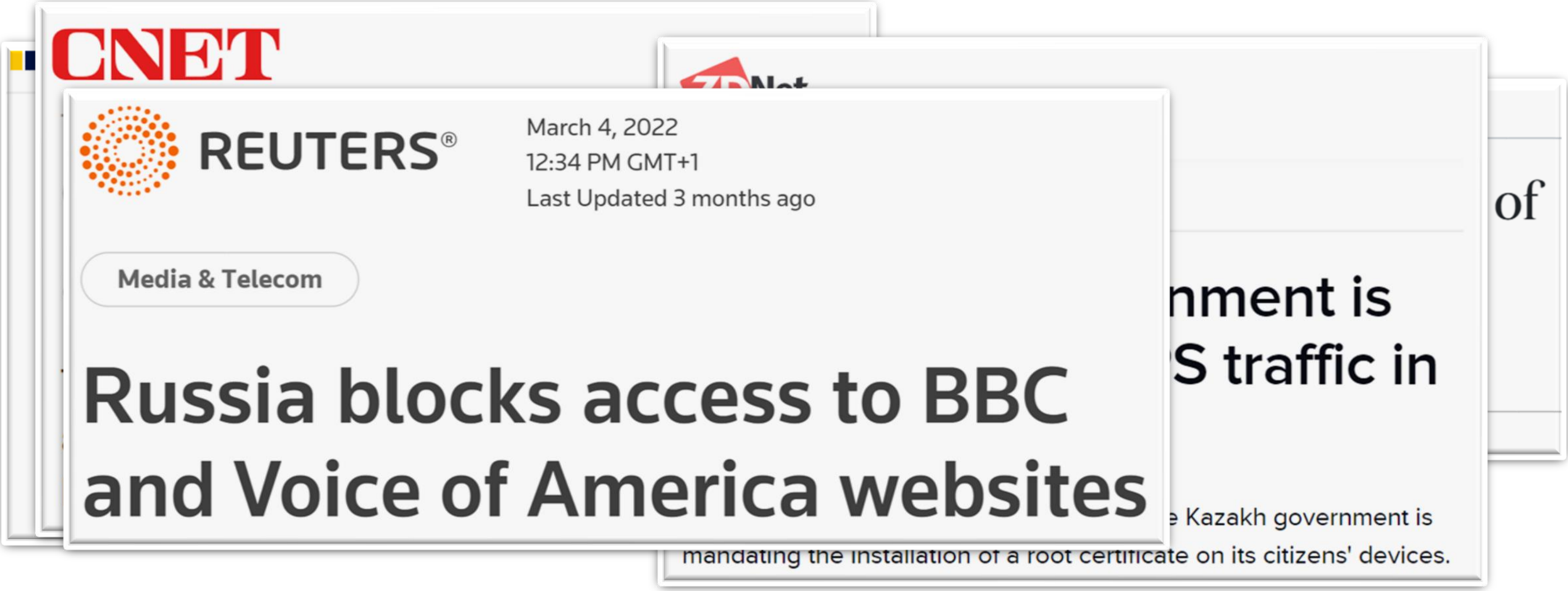
ZDNet

Home | Innovation | Security

Kazakhstan government is intercepting HTTPS traffic in its capital

This marks the third time since 2015 that the Kazakh government is mandating the installation of a root certificate on its citizens' devices.

Examples: news filtering



CNET

ZDNet

REUTERS®

March 4, 2022
12:34 PM GMT+1
Last Updated 3 months ago

Media & Telecom

Russia blocks access to BBC and Voice of America websites

nment is
S traffic in

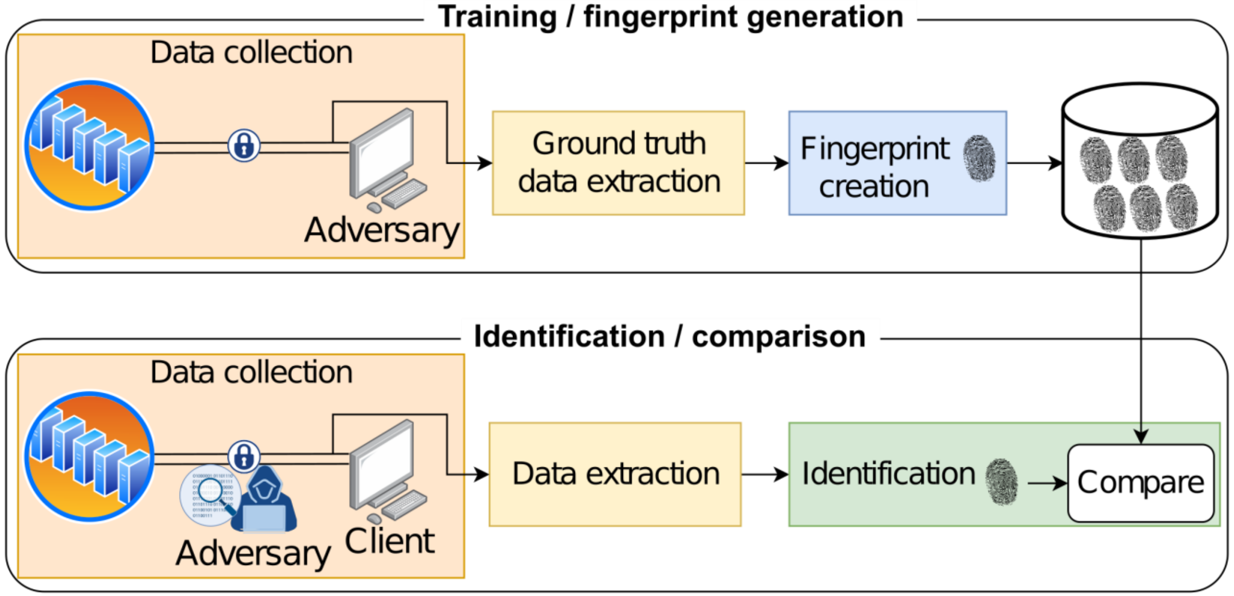
e Kazakh government is mandating the installation of a root certificate on its citizens' devices.

of

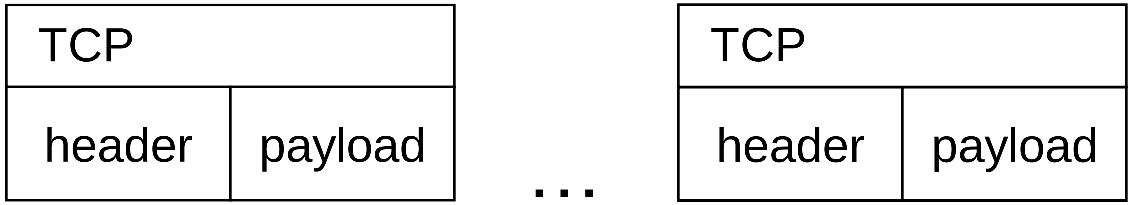
Contributions

- Design and evaluation of lightweight framework
  - Identify individual browsed news articles (internal pages) despite encryption
  - Separate between articles delivered over same infrastructure (e.g., CDN)
- Demonstrate that naive use of HTTPS is not enough to protect users' privacy
  - X.509 certificate size (encrypted with TLS 1.3)
  - Web document size
- Provide insights into why websites are more/less resilient to the attack
- Real-world scenario using Twitter
- Provide insights for websites and users to better protect their privacy

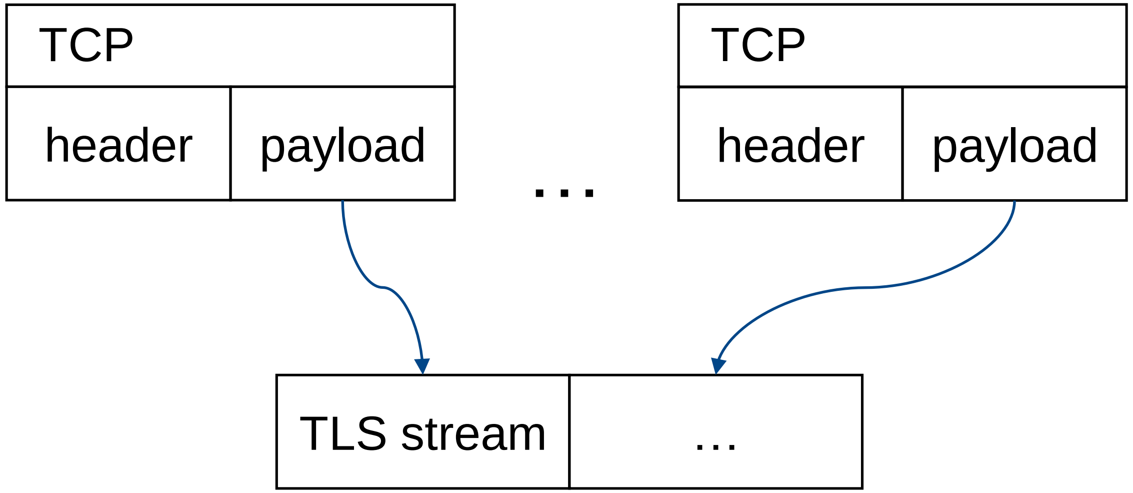
System overview



TLS record size extraction



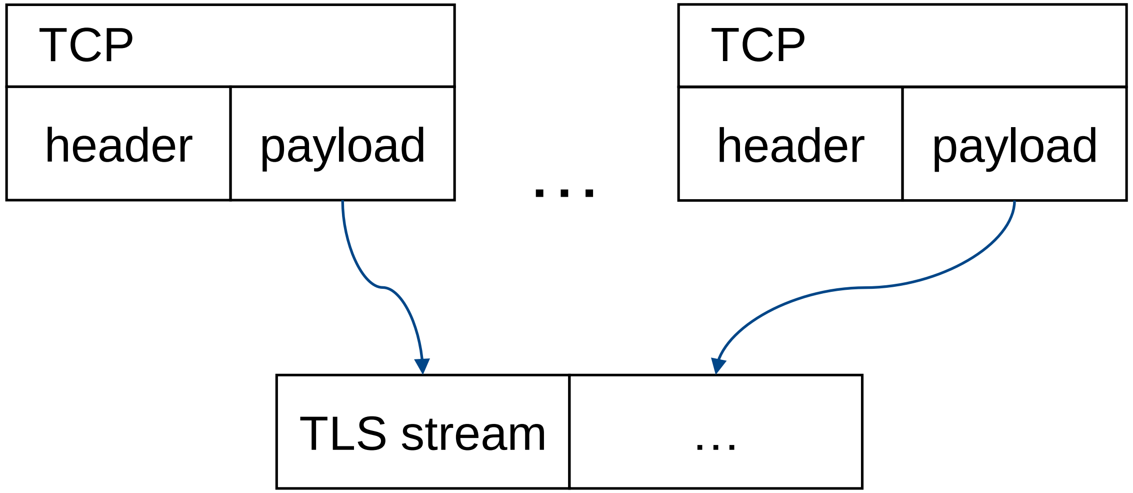
TLS record size extraction



TLS record size extraction

- Handshake: [0x16, 0x03, *m*]
- Application data: [0x17, 0x03, m_a]

$$m \in \{0x00, 0x01, 0x02, 0x03\}$$



TCP

header

payload

…

TCP

header

payload

TLS stream

…

TLS Record

Byte	+0	+1	+2	+3
0	Content type			
1..4	Version		Length	
5..n	Payload			
n..	…			

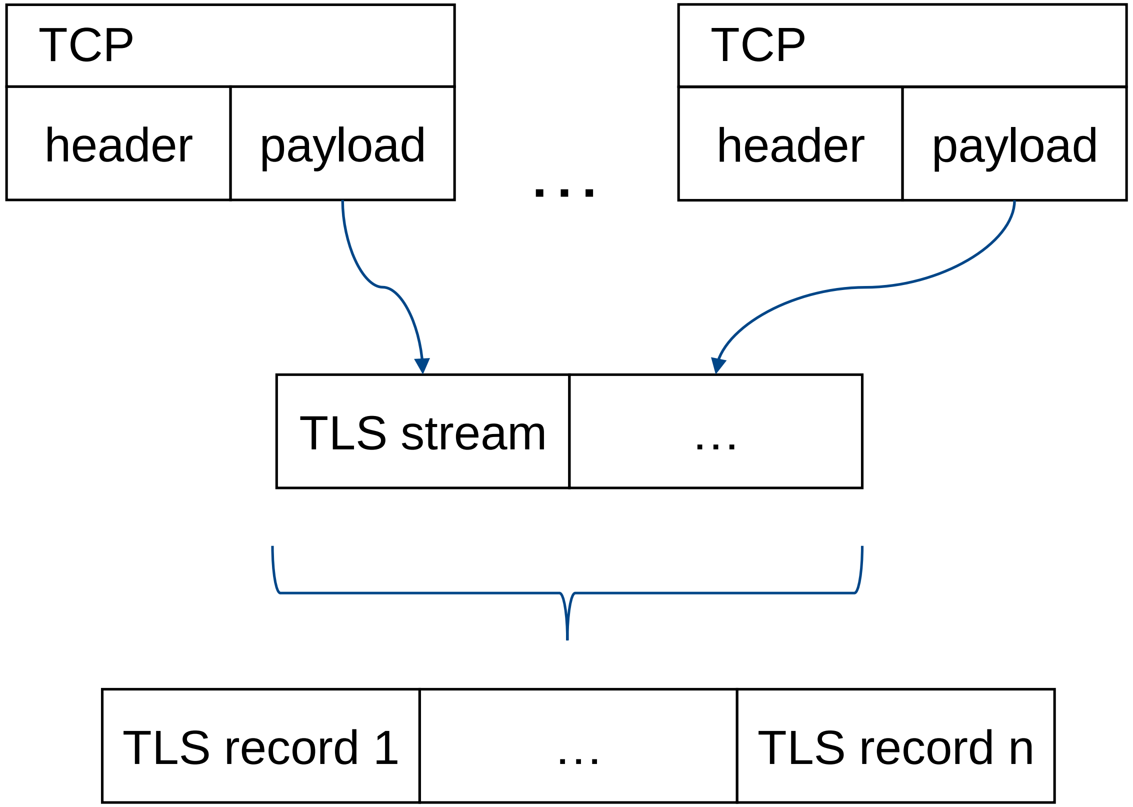
TLS record size extraction

- Handshake: [0*x*16, 0*x*03, *m*]
- Application data: [0*x*17, 0*x*03, m_a]

$$m \in \{0x00, 0x01, 0x02, 0x03\}$$

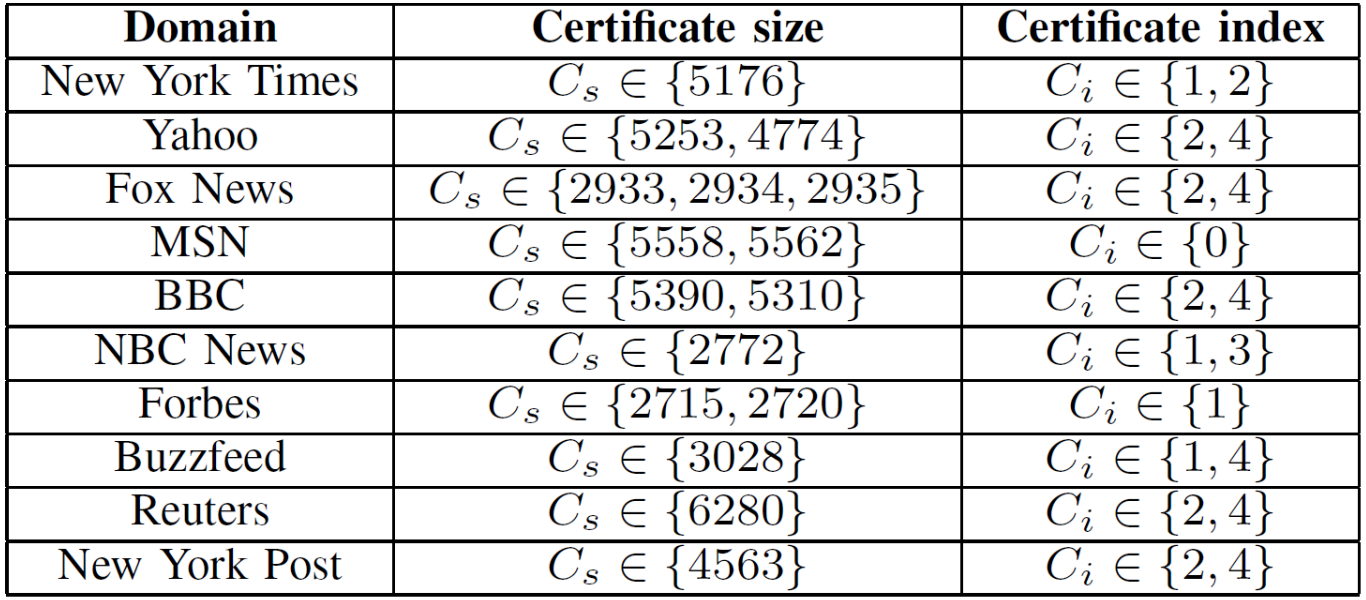
TLS Record

Byte	+0	+1	+2	+3
0	Content type			
1..4	Version		Length	
5..n	Payload			
n..	…			



TLS certificate extraction

- For each repeated connection, the certificate is delivered
  - in similar TLS record index
  - with similar TLS record size



Domain	**Certificate size**	**Certificate index**
New York Times	$C_s \in \{5176\}$	$C_i \in \{1, 2\}$
Yahoo	$C_s \in \{5253, 4774\}$	$C_i \in \{2, 4\}$
Fox News	$C_s \in \{2933, 2934, 2935\}$	$C_i \in \{2, 4\}$
MSN	$C_s \in \{5558, 5562\}$	$C_i \in \{0\}$
BBC	$C_s \in \{5390, 5310\}$	$C_i \in \{2, 4\}$
NBC News	$C_s \in \{2772\}$	$C_i \in \{1, 3\}$
Forbes	$C_s \in \{2715, 2720\}$	$C_i \in \{1\}$
Buzzfeed	$C_s \in \{3028\}$	$C_i \in \{1, 4\}$
Reuters	$C_s \in \{6280\}$	$C_i \in \{2, 4\}$
New York Post	$C_s \in \{4563\}$	$C_i \in \{2, 4\}$

Document size extraction

- Predictable patterns to reconstruct transfer size of main document

- Domain specific reconstruction process

- Sequence based
  - Unbroken TLS records of size $D_i \in D$

- Anchor based
  - Anchor records T_s and T_e

Document size extraction

- Predictable patterns to reconstruct transfer size of main document
- Domain specific reconstruction process
- Sequence based
  - Unbroken TLS records of size $D_i \in D$
- Anchor based
  - Anchor records T_s and T_e

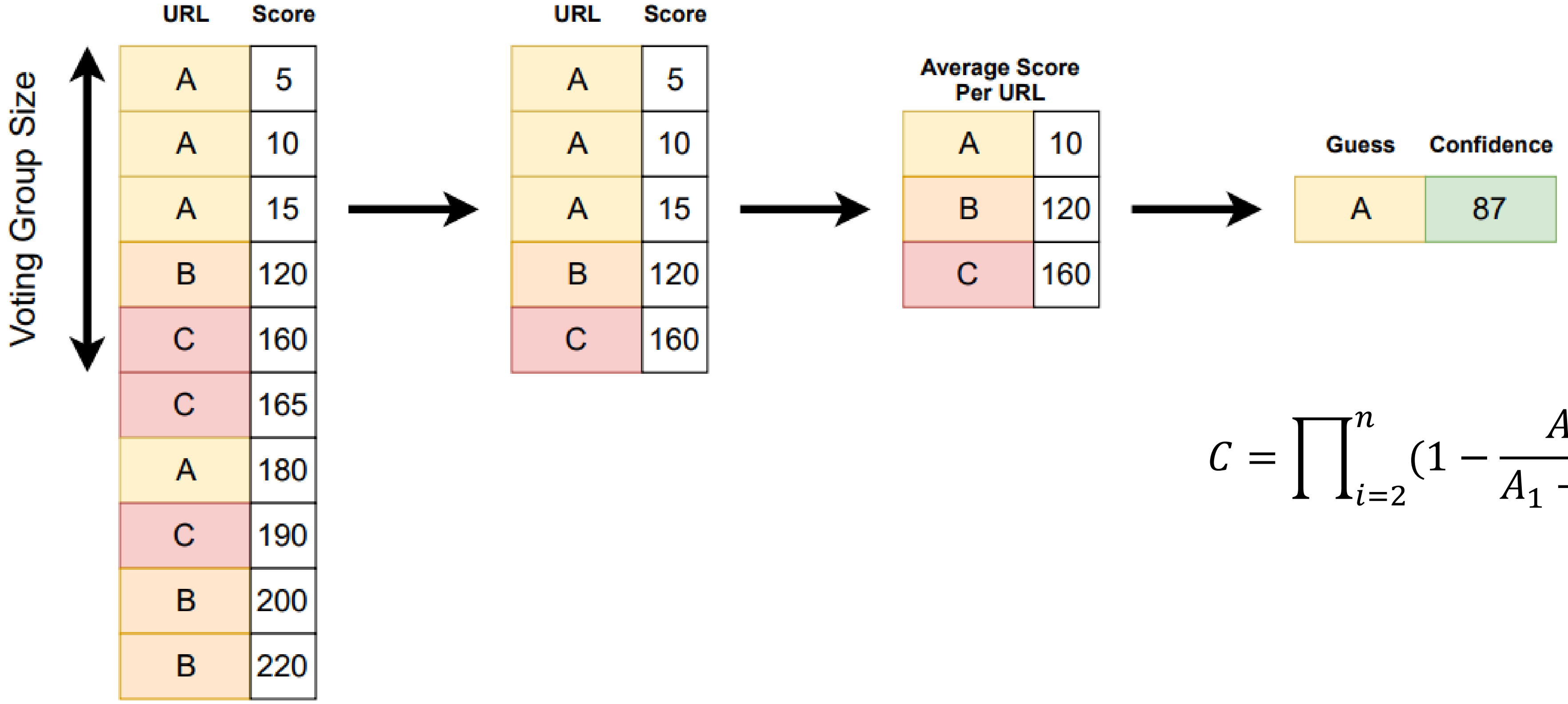
Examples:

New York Times: $D = \{1395, 1055, 202, 40\}$

MSN: $T_s = 33 \quad T_e = 33$

NBC News: $T_s \in \{72, 2907\} \quad T_e \in \{843, \dots, 744\}$

Identification: voting group system



URL	Score
A	5
A	10
A	15
B	120
C	160
C	165
A	180
C	190
B	200
B	220

URL	Score
A	5
A	10
A	15
B	120
C	160

Average Score Per URL	
A	10
B	120
C	160

Guess	Confidence
A	87

$$C = \prod_{i=2}^{n}(1 - \frac{A_1}{A_1 + A_i})$$

Performance testing

- Single-factor experiments

- Data extraction parameters
  - Pages per domain
  - Time window
  - Score deviation

- Identification parameters
  - Voting group size
  - Confidence threshold
  - Score threshold

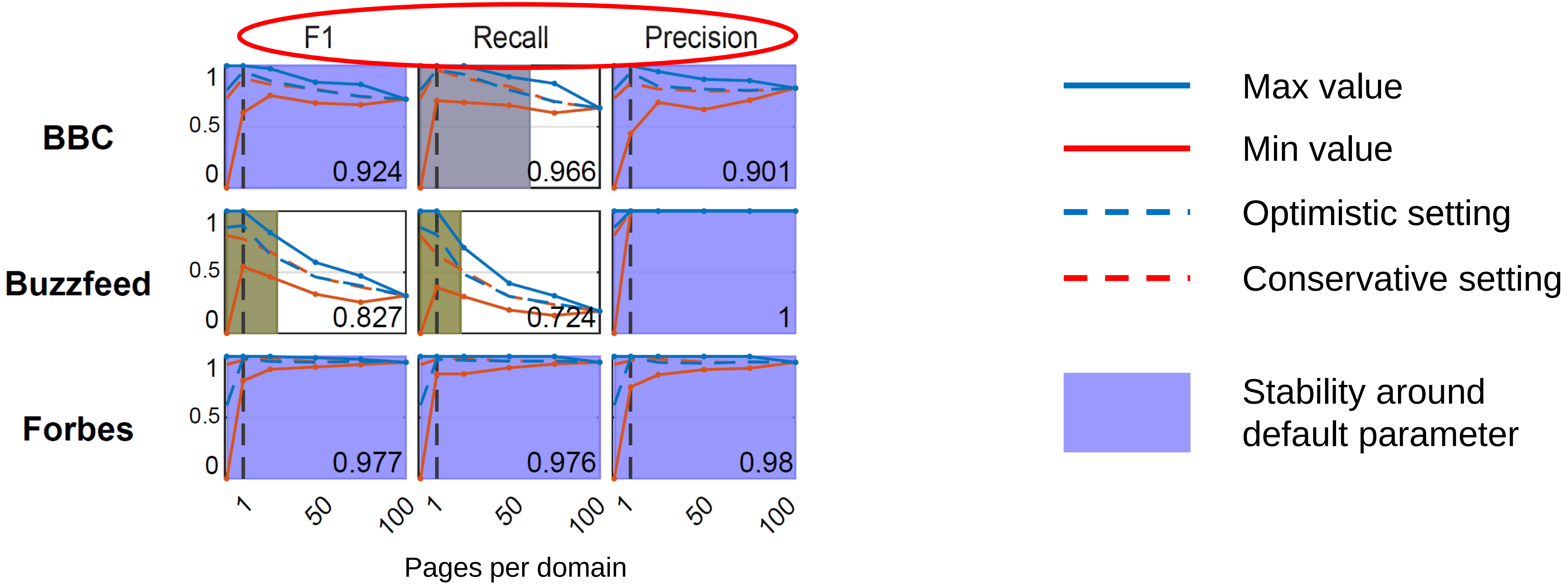
Performance testing

- Single-factor experiments

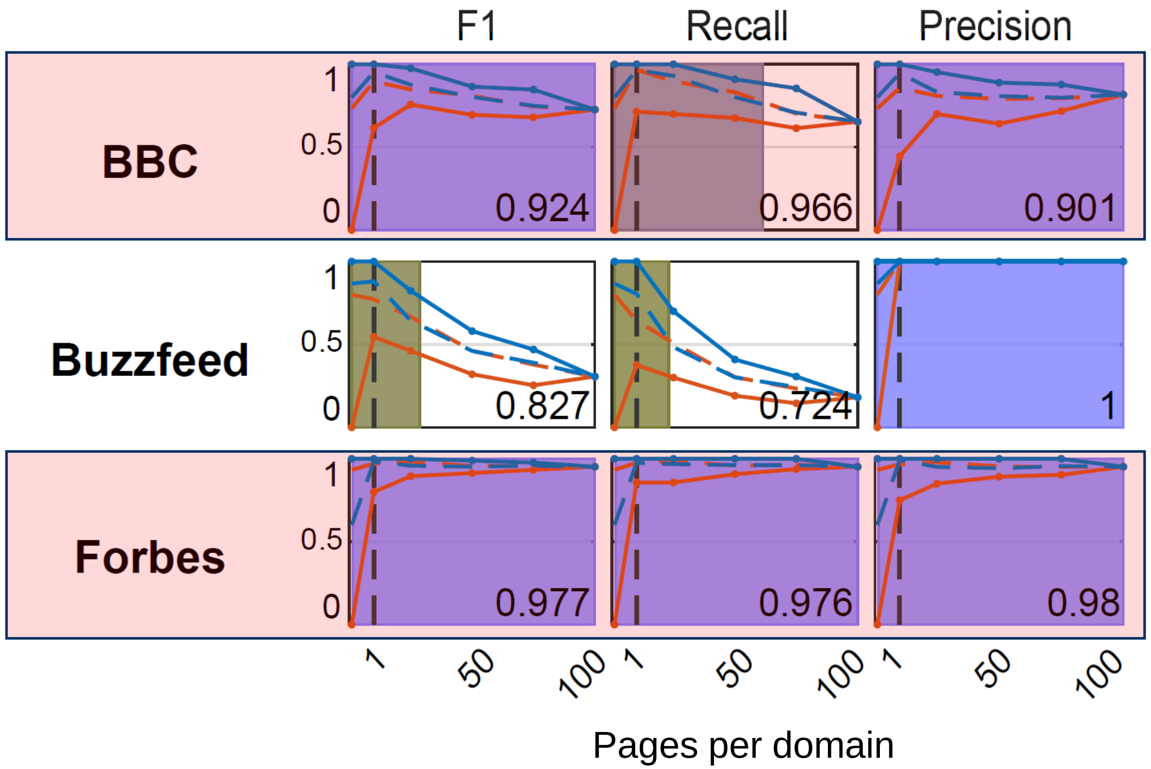
- Data extraction parameters
  - **Pages per domain**
  - Time window
  - Score deviation

- Identification parameters
  - **Voting group size**
  - Confidence threshold
  - Score threshold

Example results: pages per domain



Example results: pages per domain

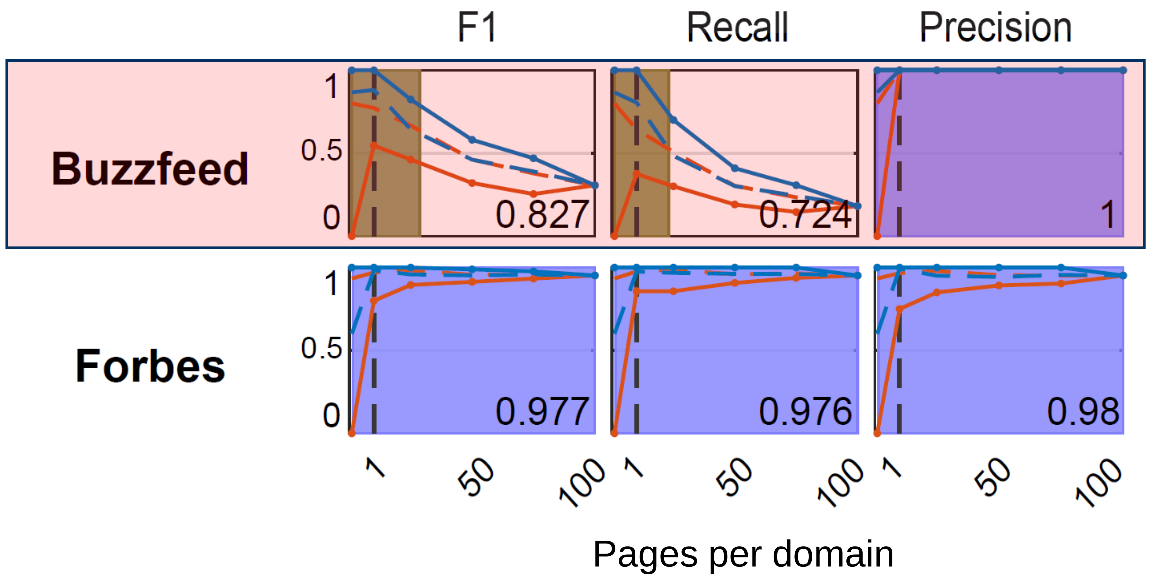


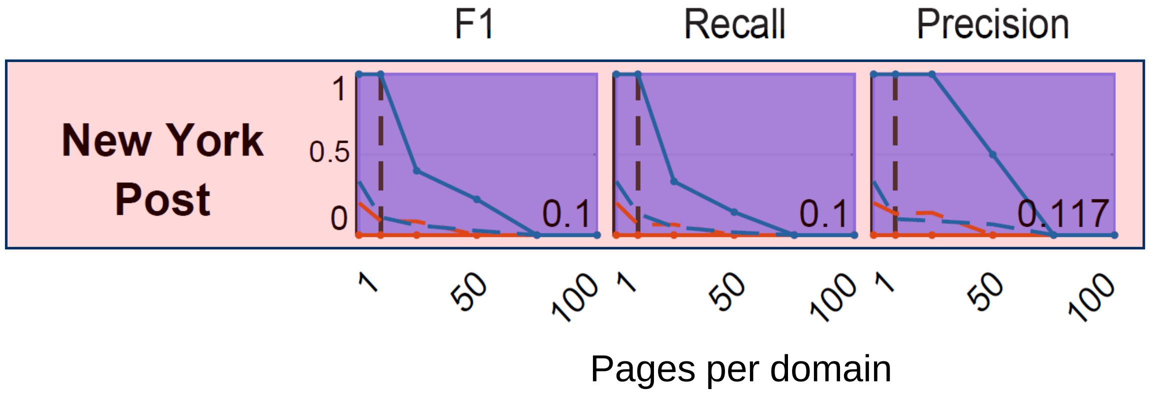
- Attacks performs well
- Only small drops
- High stability

- High metrics
- High stability
- Attack scales well

Example results: pages per domain

- Performance starts well
- Quickly drops
- Precision near 1

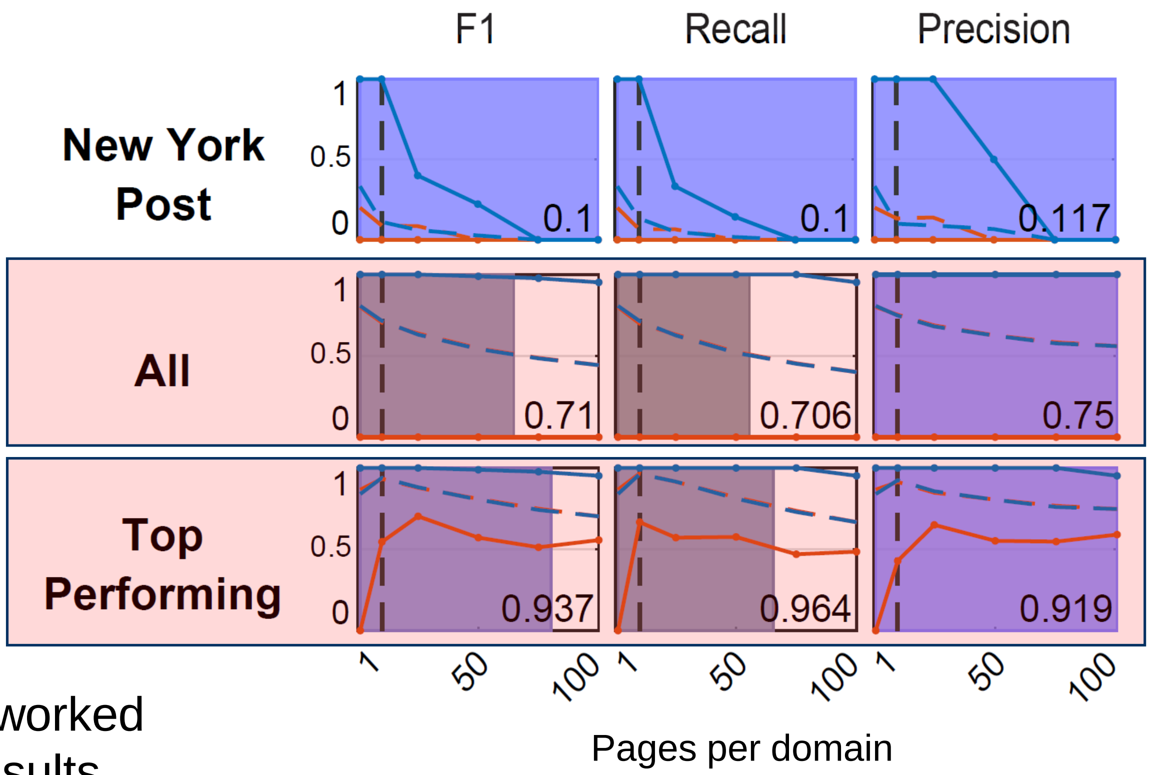




- Poor performance
- No clear TLS record size pattern
- Difficult to extract encrypted sizes

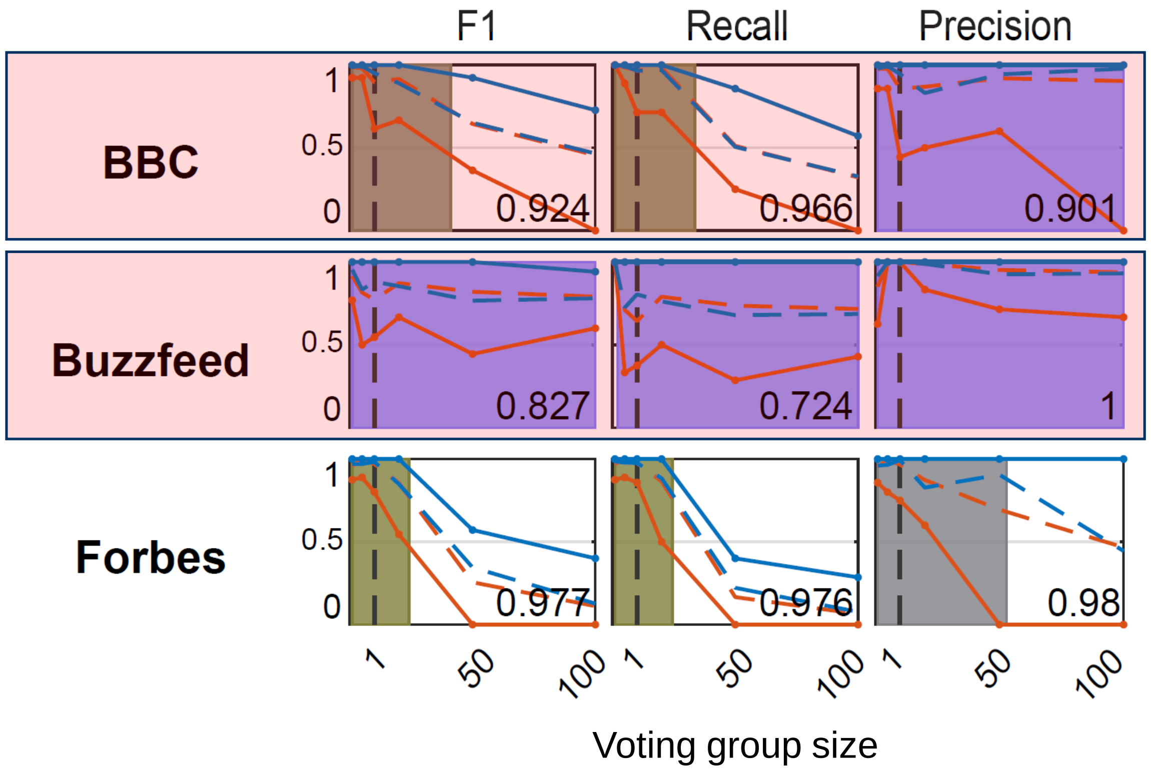
Example results: pages per domain

- In general high results
- Decrease to ~0.5 for all 3 metrics



- For domains where attack worked we see similar but better results

Example results: voting group size

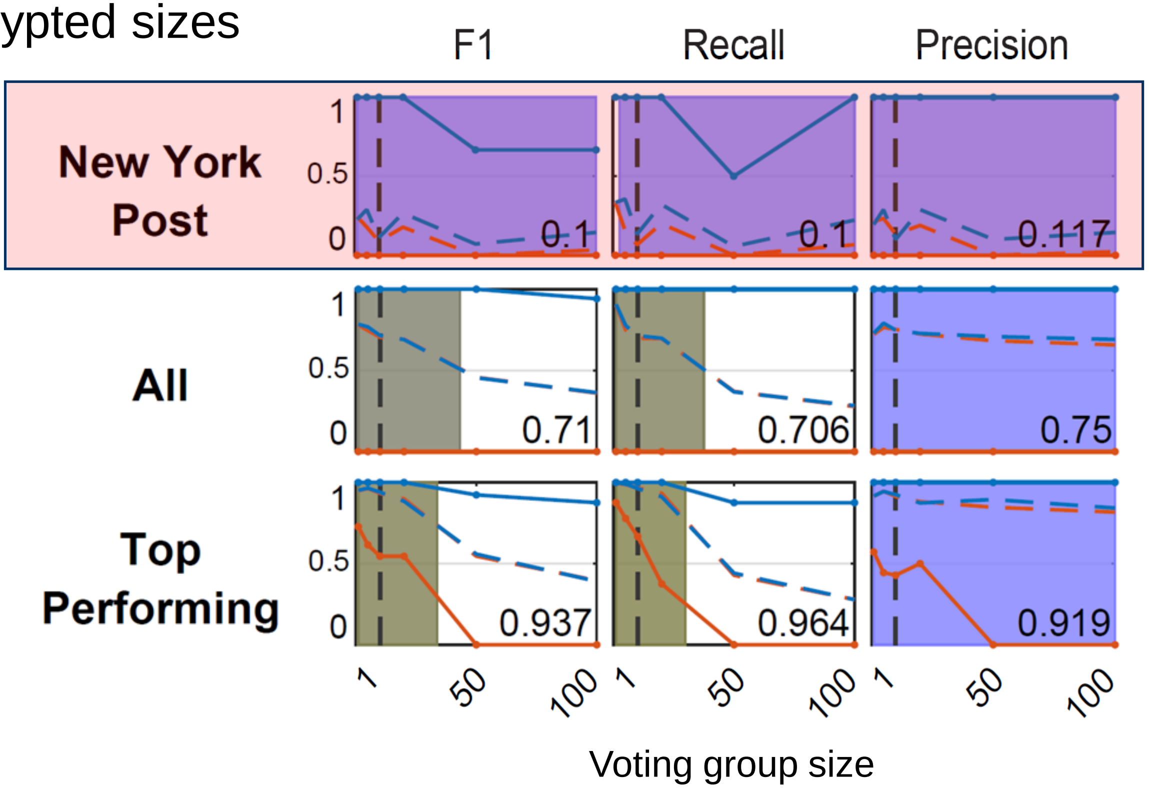


- Tradeoff between F1/recall and precision
- Stability for F1/recall smaller than for precision

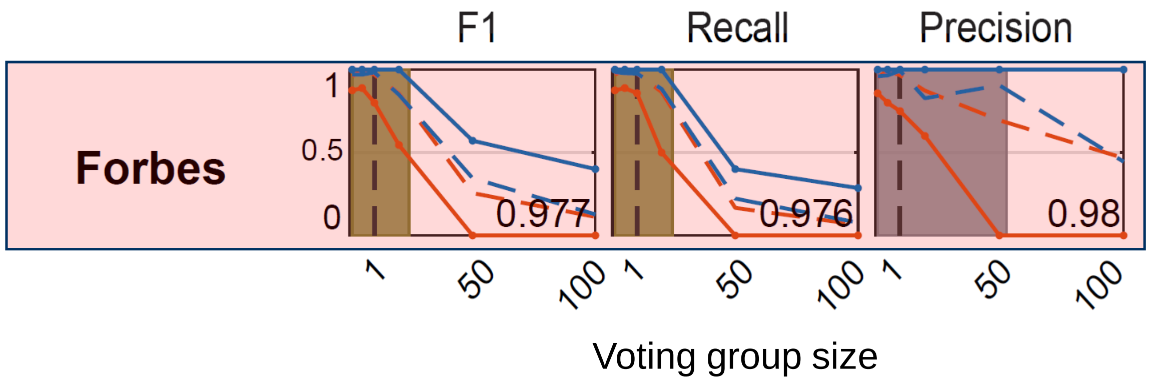
- Stable regardless of voting group size
- High stability for all 3 metrics

Example results: voting group size

- Again, poor performance
- Difficult to extract encrypted sizes



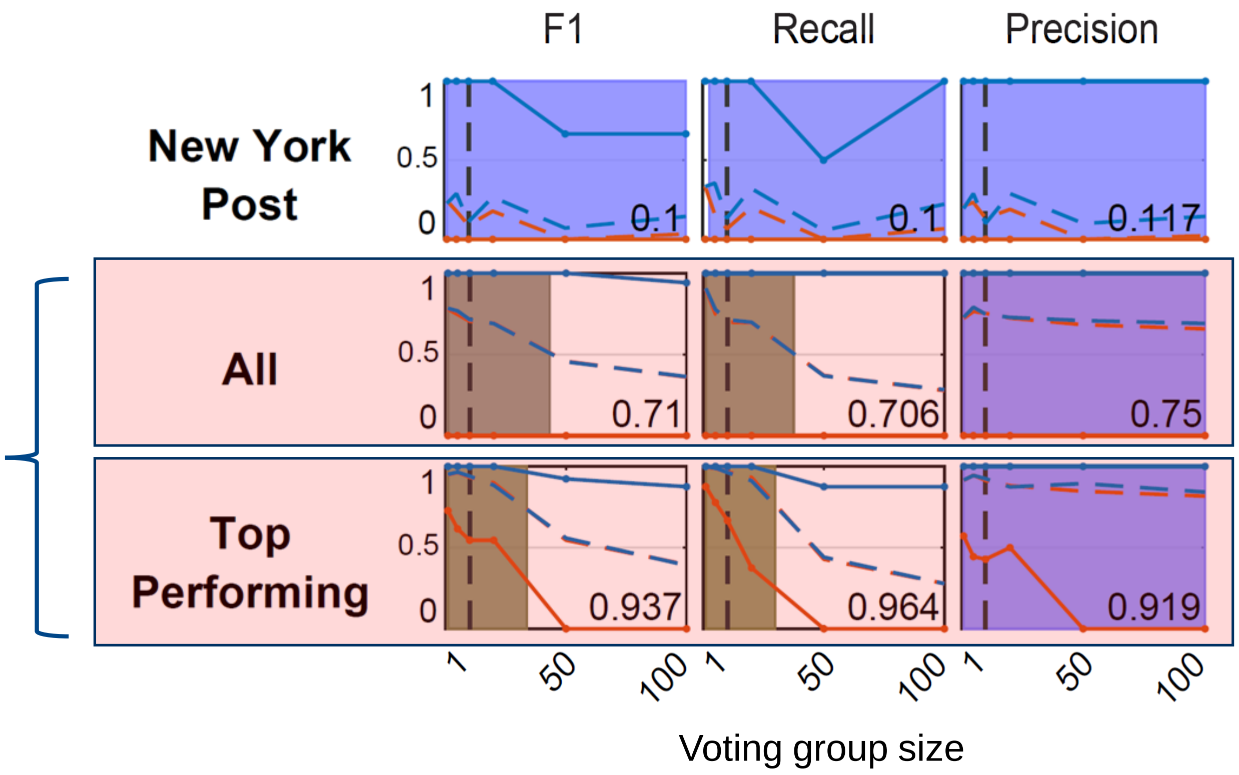
- High performance only with small group size



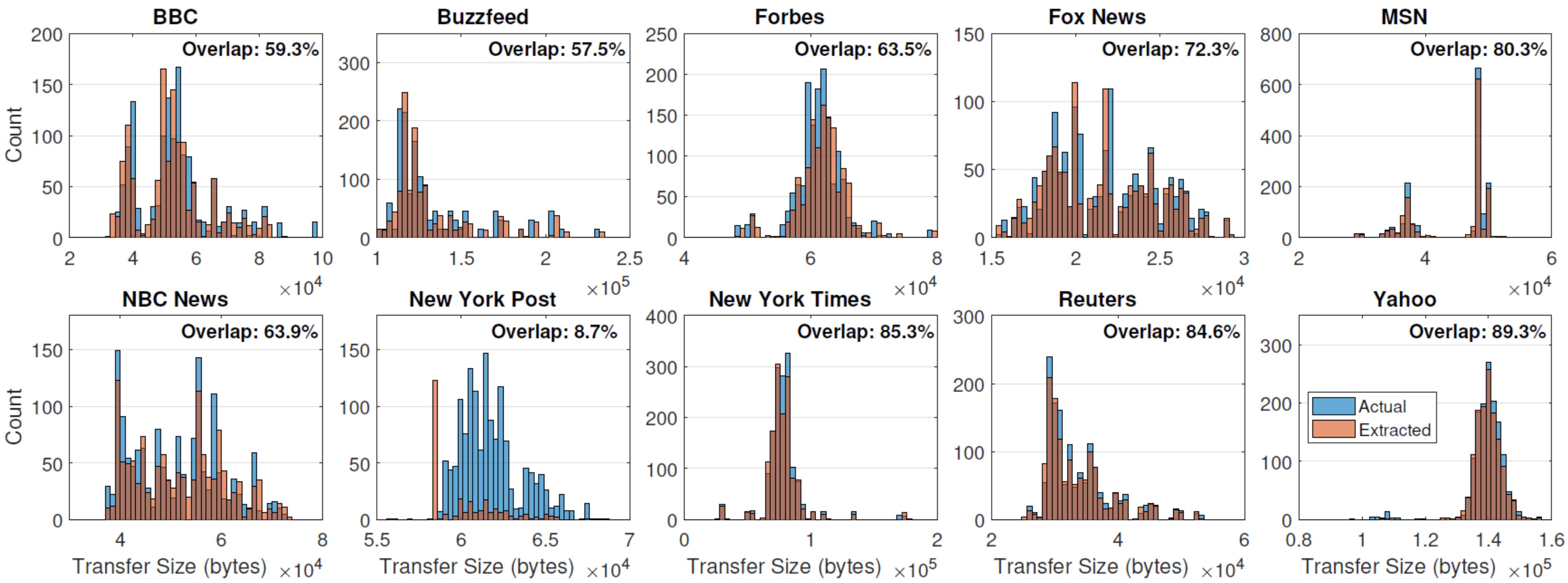
Example results: voting group size

- No significant performance gain when increasing group size

- Size near default value 10 performs well



Transfer size analysis



Discussion: example attack

- High correlation between retweets and reads

- Reads at news websites are heavily skewed
  - Top-10 of links account for 37% of reads/retweets
  - Top-50 for 67%
  - Top-100 for 78%

- News cycle typically changes daily

Discussion: example attack

- Conservative results of precision P_K and recall R_K when fingerprinting the top-K news articles
- Recall R on full set of articles observed is same as R_K
- $P_{LB} = q_K P_K$
  - q_K is fraction of requests to the top-K articles
  - E.g., for a specific domain:

    $q_{10} = 0.37 \quad q_{50} = 0.67$

Discussion: example attack

- Conservative results of precision P_K and recall R_K when fingerprinting the top-K news articles
- Recall R on full set of articles observed is same as R_K
- $P_{LB} = q_K P_K$
  - q_K is fraction of requests to the top-K articles
  - E.g., for a specific domain:

    $q_{10} = 0.37 \quad q_{50} = 0.67$
- $F1_{LB} = \frac{2R_K q_K P_K}{R_K + q_K P_K}$

Discussion: example attack

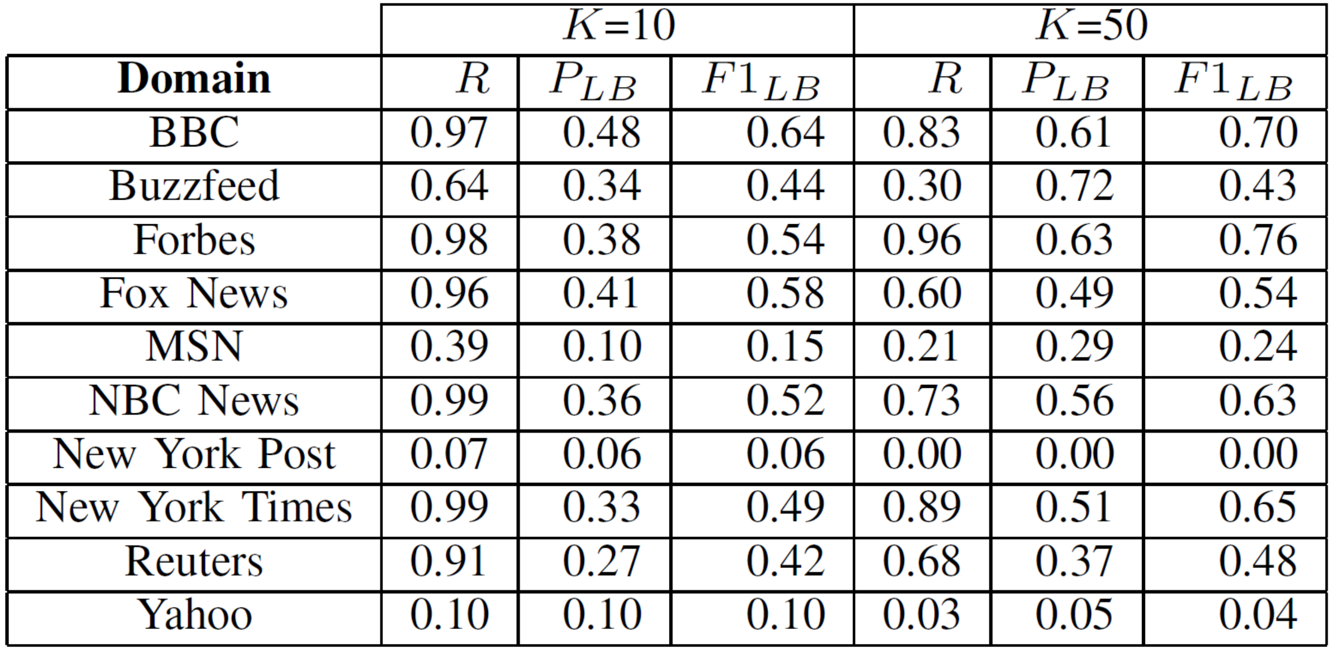
- Conservative results of precision P_K and recall R_K when fingerprinting the top-K news articles

- Recall R on full set of articles observed is same as R_K

- $P_{LB} = q_K P_K$
  - q_K is fraction of requests to the top-K articles
  - E.g., for a specific domain:

    $q_{10} = 0.37 \quad q_{50} = 0.67$

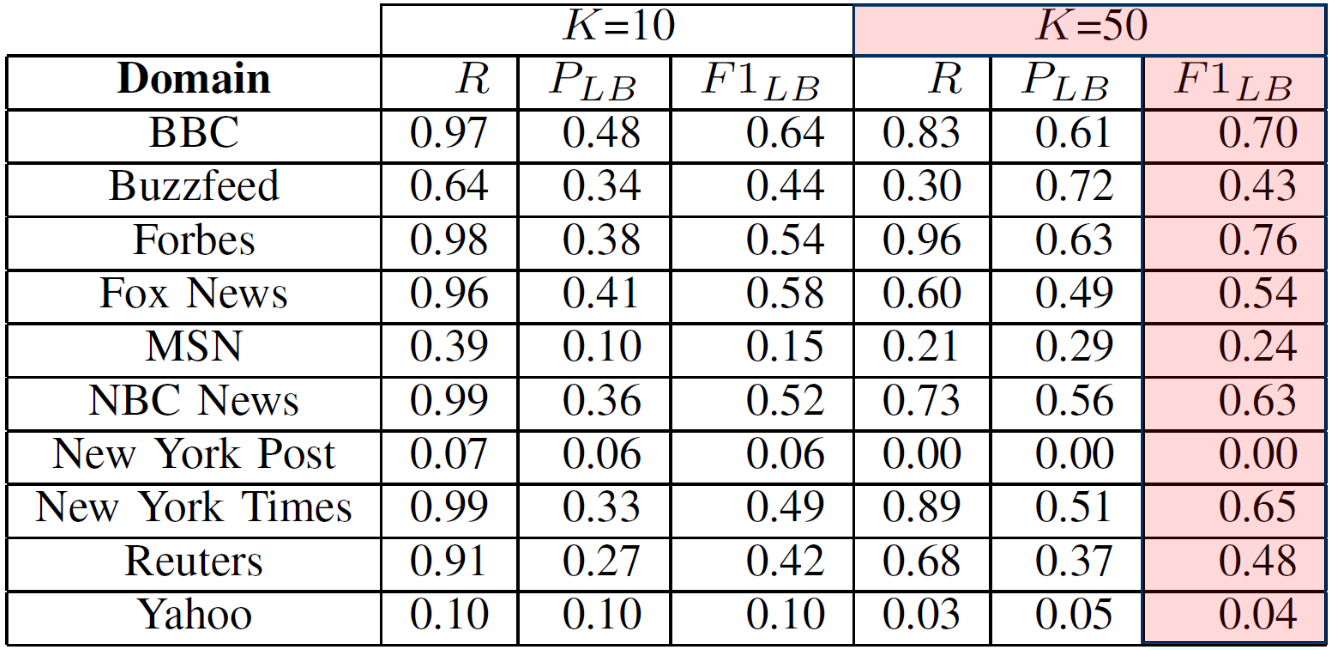
- $F1_{LB} = \frac{2 R_K q_K P_K}{R_K + q_K P_K}$



	K=10			K=50		
Domain	R	P_{LB}	$F1_{LB}$	R	P_{LB}	$F1_{LB}$
BBC	0.97	0.48	0.64	0.83	0.61	0.70
Buzzfeed	0.64	0.34	0.44	0.30	0.72	0.43
Forbes	0.98	0.38	0.54	0.96	0.63	0.76
Fox News	0.96	0.41	0.58	0.60	0.49	0.54
MSN	0.39	0.10	0.15	0.21	0.29	0.24
NBC News	0.99	0.36	0.52	0.73	0.56	0.63
New York Post	0.07	0.06	0.06	0.00	0.00	0.00
New York Times	0.99	0.33	0.49	0.89	0.51	0.65
Reuters	0.91	0.27	0.42	0.68	0.37	0.48
Yahoo	0.10	0.10	0.10	0.03	0.05	0.04

Discussion: example attack

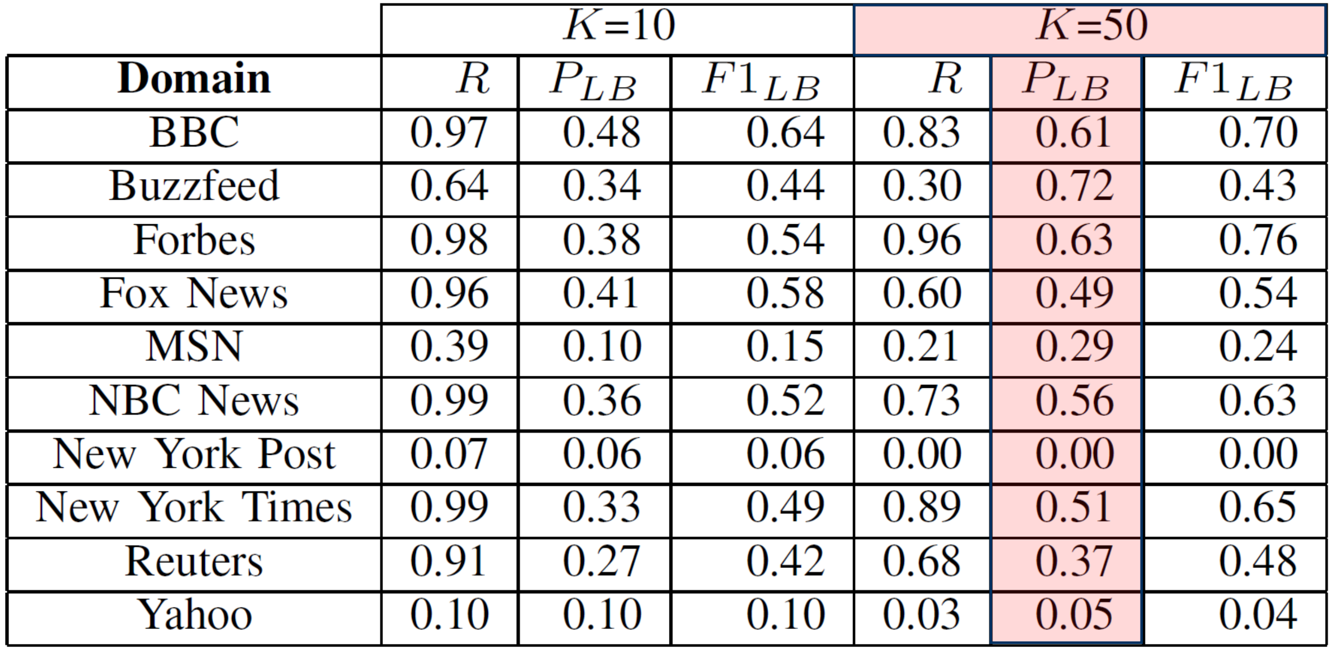
- F1-score > 0.5 for half of domains even with conservative estimates



	K=10			K=50		
Domain	R	P_{LB}	$F1_{LB}$	R	P_{LB}	$F1_{LB}$
BBC	0.97	0.48	0.64	0.83	0.61	0.70
Buzzfeed	0.64	0.34	0.44	0.30	0.72	0.43
Forbes	0.98	0.38	0.54	0.96	0.63	0.76
Fox News	0.96	0.41	0.58	0.60	0.49	0.54
MSN	0.39	0.10	0.15	0.21	0.29	0.24
NBC News	0.99	0.36	0.52	0.73	0.56	0.63
New York Post	0.07	0.06	0.06	0.00	0.00	0.00
New York Times	0.99	0.33	0.49	0.89	0.51	0.65
Reuters	0.91	0.27	0.42	0.68	0.37	0.48
Yahoo	0.10	0.10	0.10	0.03	0.05	0.04

Discussion: example attack

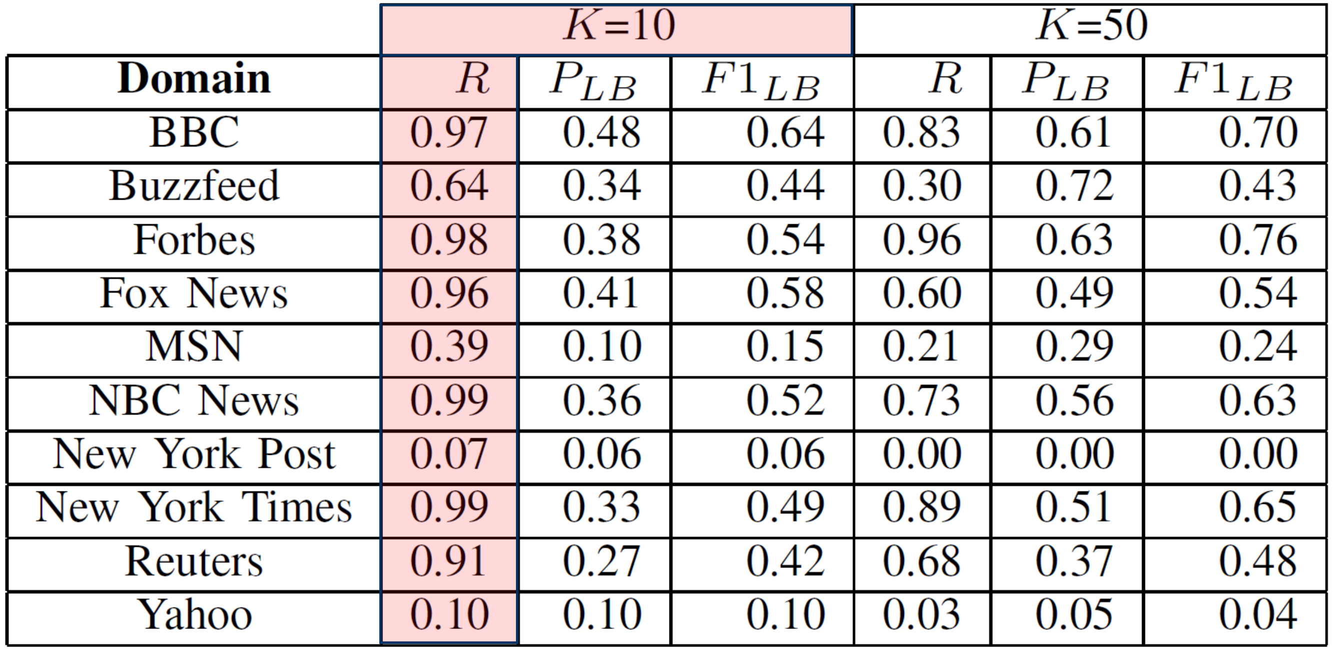
- F1-score > 0.5 for half of domains even with conservative estimates
- Top-50 to increase precision



	K=10			K=50		
Domain	R	P_{LB}	$F1_{LB}$	R	P_{LB}	$F1_{LB}$
BBC	0.97	0.48	0.64	0.83	0.61	0.70
Buzzfeed	0.64	0.34	0.44	0.30	0.72	0.43
Forbes	0.98	0.38	0.54	0.96	0.63	0.76
Fox News	0.96	0.41	0.58	0.60	0.49	0.54
MSN	0.39	0.10	0.15	0.21	0.29	0.24
NBC News	0.99	0.36	0.52	0.73	0.56	0.63
New York Post	0.07	0.06	0.06	0.00	0.00	0.00
New York Times	0.99	0.33	0.49	0.89	0.51	0.65
Reuters	0.91	0.27	0.42	0.68	0.37	0.48
Yahoo	0.10	0.10	0.10	0.03	0.05	0.04

Discussion: example attack

- F1-score > 0.5 for half of domains even with conservative estimates
- Top-50 to increase precision
- Top-10 to increase recall
  - Recall > 0.9 for 6 domains



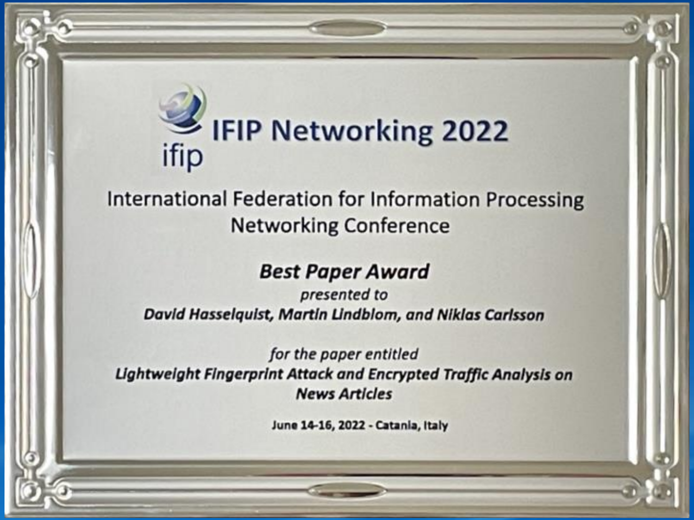
	K=10			K=50		
Domain	R	P_{LB}	$F1_{LB}$	R	P_{LB}	$F1_{LB}$
BBC	0.97	0.48	0.64	0.83	0.61	0.70
Buzzfeed	0.64	0.34	0.44	0.30	0.72	0.43
Forbes	0.98	0.38	0.54	0.96	0.63	0.76
Fox News	0.96	0.41	0.58	0.60	0.49	0.54
MSN	0.39	0.10	0.15	0.21	0.29	0.24
NBC News	0.99	0.36	0.52	0.73	0.56	0.63
New York Post	0.07	0.06	0.06	0.00	0.00	0.00
New York Times	0.99	0.33	0.49	0.89	0.51	0.65
Reuters	0.91	0.27	0.42	0.68	0.37	0.48
Yahoo	0.10	0.10	0.10	0.03	0.05	0.04

Conclusions

- Design and evaluation of lightweight framework
    - Identify individual browsed news articles (internal pages) despite encryption
    - Separate between articles delivered over same infrastructure (e.g., CDN)
- Demonstrate that naive use of HTTPS is not enough to protect users' privacy
    - X.509 certificate size (encrypted with TLS 1.3)
    - Web document size
- Provide insights into why websites are more/less resilient to the attack
- Real-world scenario using Twitter
- Provide insights for websites and users to better protect their privacy

Lightweight Fingerprint Attack and Encrypted Traffic Analysis on News Articles

David Hasselquist
Martin Lindblom
Niklas Carlsson



IFIP Networking 2022
ifip

International Federation for Information Processing
Networking Conference

Best Paper Award
presented to
David Hasselquist, Martin Lindblom, and Niklas Carlsson

for the paper entitled
Lightweight Fingerprint Attack and Encrypted Traffic Analysis on News Articles

June 14-16, 2022 - Catania, Italy

David Hasselquist (david.hasselquist@liu.se)

LINKÖPING UNIVERSITY

WASP | WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

SECTRA