

Sound data sets and methods for web security research

Victor Le Pochat

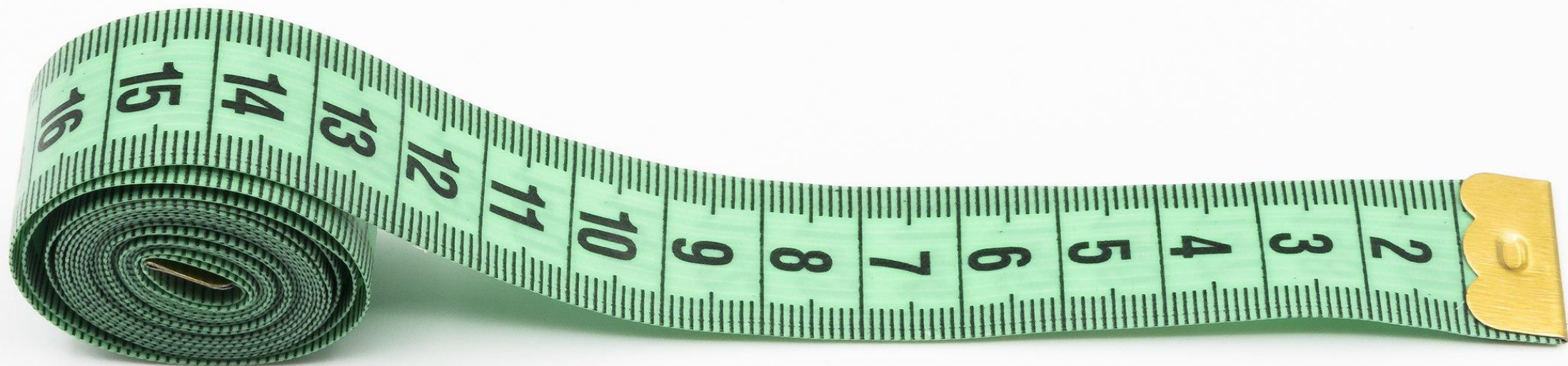
make the web
more **secure**







sound
research



**Empirical
real-world
measurements
are crucial**


```
graph LR; Design --> Develop; Develop --> Select;
```

Design

Develop

Select

Design

Develop

Select

Collect

Analyze

Interpret

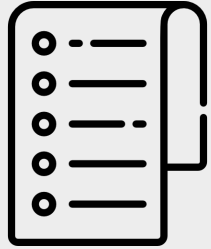


accurate comprehensive
representative transparent

Highlight the *importance* and *challenges*
for **sound data sets and methods**

to *understand* and *improve*
how to approach **web security research**

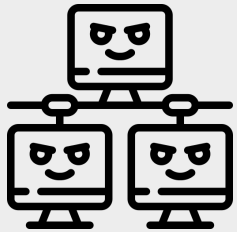
Case studies on **sound *data sets and methods***



1. Top sites rankings

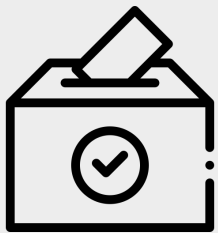
→ open *data sets* to improve *methods*

Automated decision-making systems



2. Botnet domains

→ adapting *methods* for missing *data sets*



3. Political ads

→ auditing *methods* with novel *data sets*

sound data sets



1

Analyzing and improving top sites rankings

Victor Le Pochat, Tom Van Goethem, Samaneh Tajalizadehkhoob, Maciej Korczyński, Wouter Joosen.
Tranco: A Research-Oriented Top Sites Ranking Hardened Against Manipulation. NDSS 2019

What is a **top sites** ranking?

- › Ranking of most popular websites / domain names

1, google.com

6, netflix.com

2, youtube.com

7, akamaiedge.net

3, facebook.com

8, epicgames.com

4, a-msedge.net

9, twitter.com

5, microsoft.com

10, instagram.com

- › **Essential data source**

Select

- › Potential **impact** on measurements and findings

Commercial rankings proved **unsuitable**

- › **Opaque** methods: *unclear data sources, processing, ...*
- › **Undesirable** properties: *dissimilar, volatile, mismatch with expectations*
- › **Manipulable**: *easy to insert domains at scale*
- › **Ill-suited** for reproducibility: *no reference or archive*
- › **Difficult** to discover: *undocumented resources, uncertain availability, ...*

In 2019,
we started a new
research-oriented
top sites ranking:
Tranco

Tranco improves on **research-oriented** properties

- › **Opaque** methods → **transparent** construction method
- › **Undesirable** properties → more **suitable** properties
- › **Manipulable** → **hardened** against manipulation
- › **Ill-suited** for reproducibility → **emphasizes** reproducibility
- › **Difficult** to discover → **easy to access**

Tranco

A Research-Oriented Top Sites Ranking Hardened Against Manipulation

By [Victor Le Pochat](#), [Tom Van Goethem](#), [Samaneh Tajalizadehkhoob](#), [Maciej Korczyński](#) and [Wouter Joosen](#)

[Download the latest Tranco list](#)

We are tracking the apparent lack of updates to the Alexa ranking since 1 February 2023. We are in the process of deciding whether and how to maintain Alexa in Tranco. We are also working on adding the Chrome User Experience Report and Cloudflare Radar rankings to Tranco.

Researchers in web security or Internet measurements often use rankings of popular websites. However, [in our paper](#) we showed that these rankings disagree on which domains are most popular, can change significantly on a daily basis and can be manipulated (by malicious actors).

As the research community still benefits from regularly updated lists of popular domains, we provide **Tranco**, a new ranking that improves upon the shortcomings of current lists. We also emphasize the reproducibility of these rankings and the studies using them by providing permanent citable references.

Access the Tranco ranking

We advise to use the **latest standard Tranco list**:
one million domains obtained by averaging all four rankings over the past 30 days.

[Retrieve latest list](#)

A new research-oriented top sites ranking: *Tranco*

- › Academia and industry recognize the **impact** of Tranco
 - › Considered *standard* data set
 - › Increasingly used in *papers* (>400 citations)
 - › Used by Mozilla, Cloudflare, EFF, Internet Society, Brave, ...

CYBERSECURITY ARTIFACTS COMPETITION
IMPACTFUL DATASET AWARD

Tranco: A Research-Oriented Top Sites Ranking
Hardened Against Manipulation

*Victor Le Pochat, Tom Van Goethem, Samaneh Tajalizadehkhoob,
Maciej Korczyński, Wouter Joosen*

*“The papers [...] used [...] Tranco [...],
which shows that works that aim
to provide **best practices**
have a **positive impact**
on our community”*

sound data sets

sound
methods

Auditing automated decision-making systems









2

Classifying DGA domains for the Avalanche botnet takedown

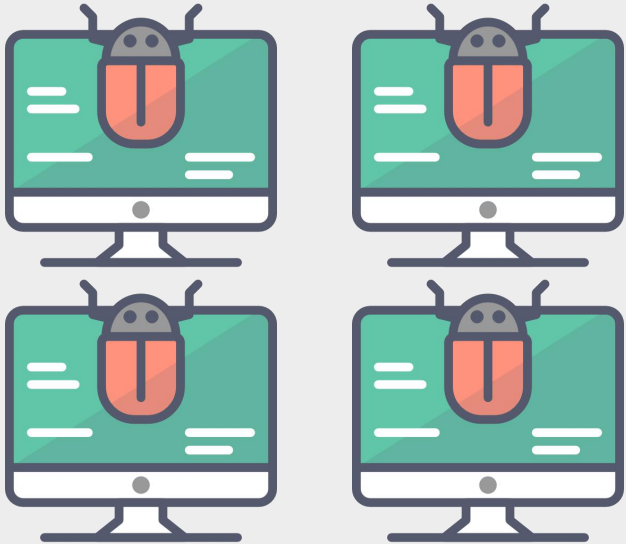
Victor Le Pochat, Tim Van hamme, Sourena Maroofi, Tom Van Goethem, Davy Preuveneers, Andrzej Duda, Wouter Joosen, Maciej Korczyński. *A Practical Approach for Taking Down Avalanche Botnets Under Real-World Constraints*. NDSS 2020

“the world’s largest and most sophisticated cybercriminal syndicate law enforcement has encountered”

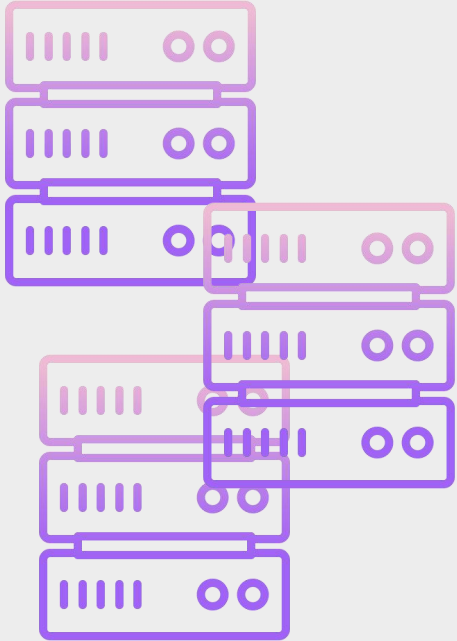
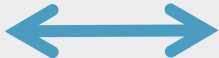
[Wai17]

Avalanche operated an advanced infrastructure

Infected client



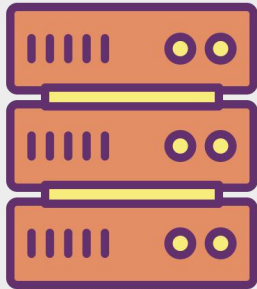
Infected hosts
serving as entrypoints



Layered network
of proxy servers

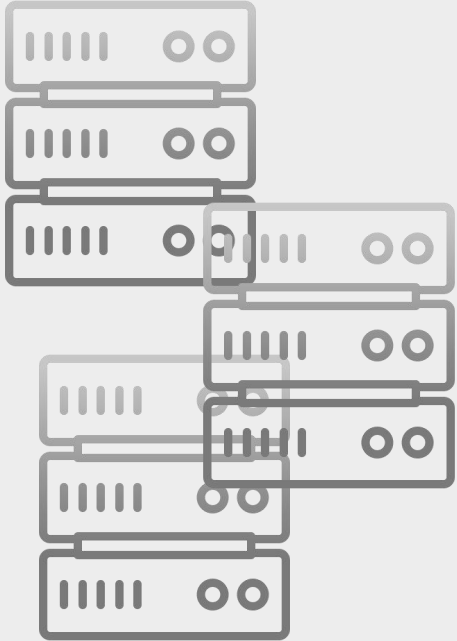
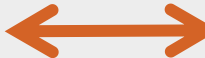
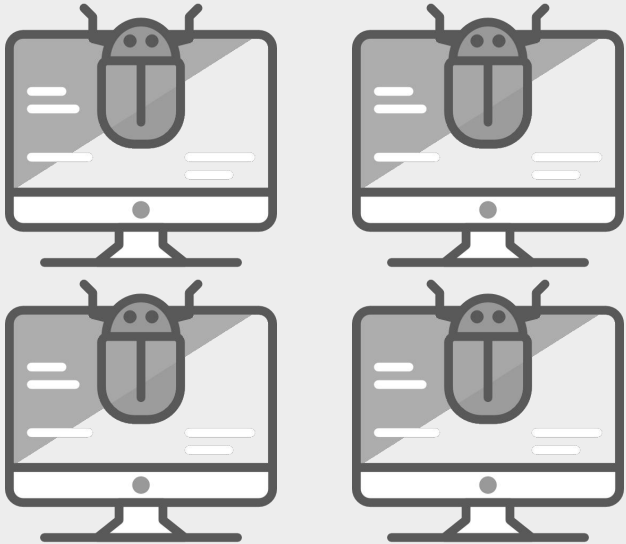


Core C&C
server

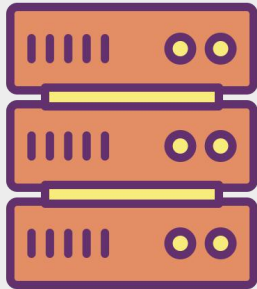


Avalanche operated an advanced infrastructure

Infected client



Core C&C server



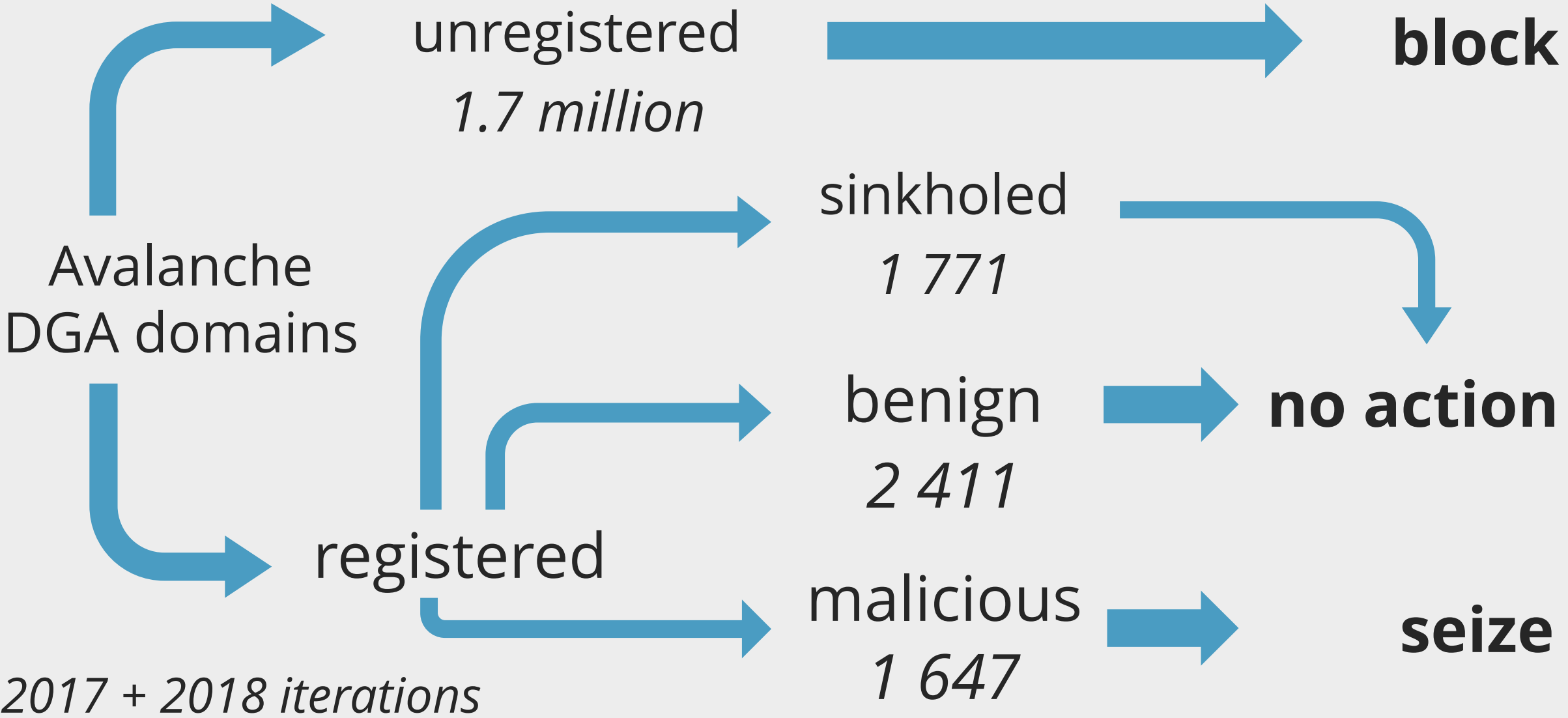
Domain Generation Algorithms

0a85rcbe2wb5n5f.com

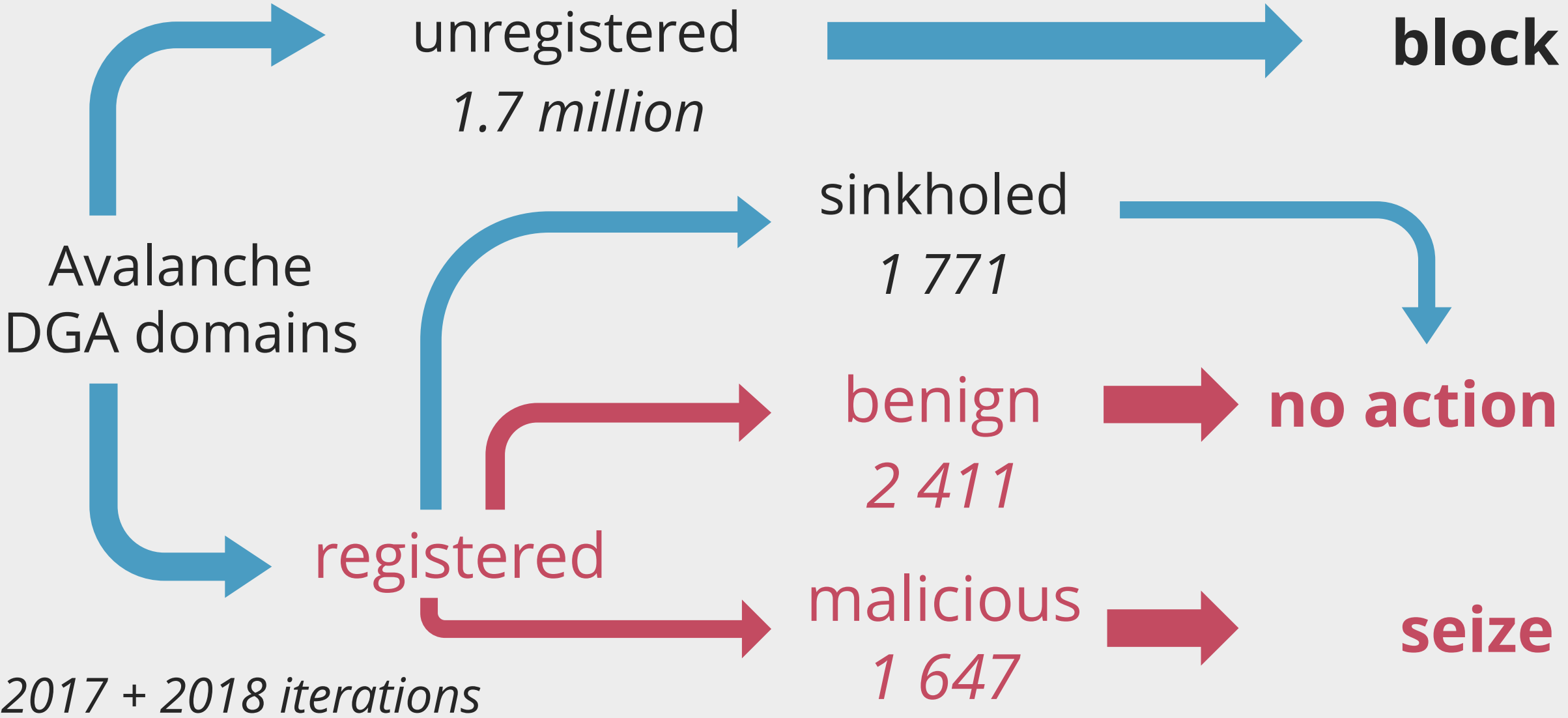
researchmadness.com

arbres.com

Law enforcement has to classify registered domains



Law enforcement has to classify registered domains



We design a hybrid model
for a real-world use case
of classifying DGA domains

We evaluate our model on the *Avalanche* takedown

- › Design an **automated** approach to reduce extensive **manual classification effort**
 - › Classify benign vs. malicious registered DGA domains
- › and assist in making **accurate** decisions
 - › Take down a *benign* domain: service interruption
 - › *Not* take down a *malicious* domain: botnet can respawn
- › contributing to **real-world** law enforcement operations
 - › Constrained in (ground-truth) data, indicators, approach

Requirements and data affect our **methods**

- › Strict **accuracy requirements**
 - **Hybrid** model: *Human in the loop*
 - **Trade-off** between accuracy (97.6%) and effort saved (66.2%)
- › Some domains have **missing data**
 - Classify **all** domains using **ensemble** model
- › **Data availability** affects **performance**
 - Partially covered by **redundancy** between data sets



3

An Audit of Facebook's Political Ad Policy Enforcement

Victor Le Pochat, Laura Edelson, Tom Van Goethem, Wouter Joosen, Damon McCoy, Tobias Lauinger. *An Audit of Facebook's Political Ad Policy Enforcement*. USENIX Security 2022



N-VA

Sponsored · Paid for by Nieuw-Vlaamse Alliantie

ID: 646422706820328



Vlaamse begroting just hits different.

-
-

#politiekememes #politiekehumor #meme #humor #bdw



Vooruit

Sponsored · Paid for by Vooruit

ID: 171946665667153



Kom jij ook mee denken over ons onderwijs? Afspraak op 22/04 in Sint-Niklaas. Schrijf je hier in. Tot dan! 🙌



Online political advertising is **powerful**,
but has a risk of **abuse**

**TikTok and Facebook fail to detect
election disinformation in the US,
while YouTube succeeds**

*Facebook Says It Won't Back Down
From Allowing Lies in Political Ads*

**Facebook 'influence operations' run by Russia and China were
shut down by Meta**

Online political advertising is **powerful**, but has a risk of **abuse**

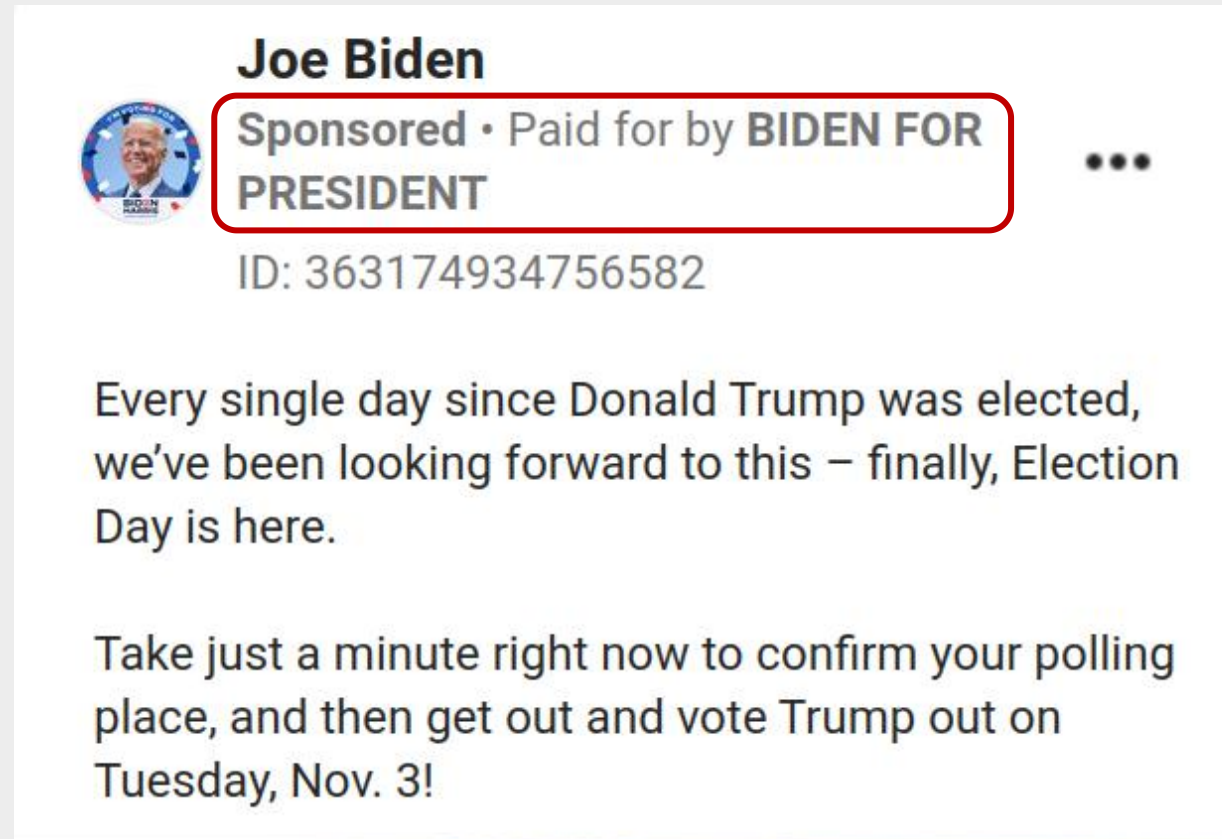
- › Lack of up-to-date legislation
- › **Oversight** falls to major advertising platforms
 - Google, Twitter, Facebook (Meta)
 - › developing their own **policies**
 - › implementing their own **enforcement**

Facebook's **political ad policy** enforcement

› “Ads about **social issues, elections and politics**”

(“political ads”)

›› Additional requirements:
identity, verification,
labeling



The screenshot shows a Facebook ad for Joe Biden. At the top left is a circular profile picture of Joe Biden with the text 'I AM VOTING FOR BIDEN FOR PRESIDENT'. To the right of the profile picture, the name 'Joe Biden' is displayed. Below the name, a red-bordered box contains the text 'Sponsored • Paid for by BIDEN FOR PRESIDENT'. To the right of this box are three dots indicating a menu. Below the red box, the ID 'ID: 363174934756582' is shown. The main text of the ad reads: 'Every single day since Donald Trump was elected, we've been looking forward to this – finally, Election Day is here.' Below this, it says: 'Take just a minute right now to confirm your polling place, and then get out and vote Trump out on Tuesday, Nov. 3!' The bottom of the screenshot shows a dark banner with the text 'TAKE THE FIRST STEP TO' and a partial view of an elderly person's head on the right side.

Joe Biden

Sponsored • Paid for by **BIDEN FOR PRESIDENT**

ID: 363174934756582

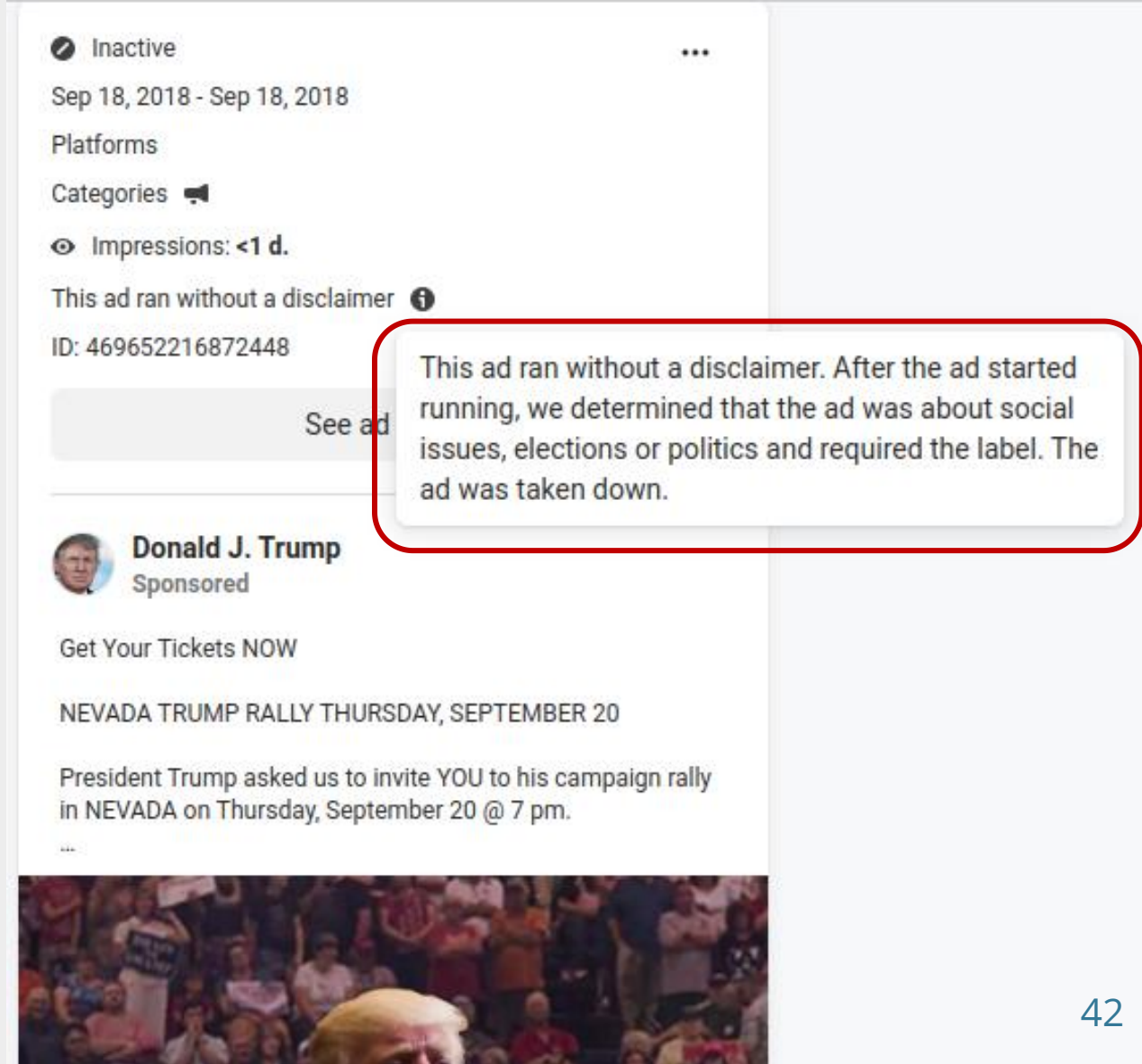
Every single day since Donald Trump was elected, we've been looking forward to this – finally, Election Day is here.

Take just a minute right now to confirm your polling place, and then get out and vote Trump out on Tuesday, Nov. 3!

TAKE THE
FIRST STEP TO

Facebook's **political ad policy enforcement**

- › **Catch undeclared ads** that are political
 - ›› Largely automated
 - ›› Retroactively label and disable the ad



The screenshot shows a Facebook ad interface. At the top, it indicates the ad is 'Inactive' and ran from 'Sep 18, 2018 - Sep 18, 2018'. Below this, it lists 'Platforms', 'Categories', and 'Impressions: <1 d.'. A notification bubble with a red border states: 'This ad ran without a disclaimer. After the ad started running, we determined that the ad was about social issues, elections or politics and required the label. The ad was taken down.' The ad itself is from 'Donald J. Trump' and is a 'Sponsored' post. The text of the ad reads: 'Get Your Tickets NOW', 'NEVADA TRUMP RALLY THURSDAY, SEPTEMBER 20', and 'President Trump asked us to invite YOU to his campaign rally in NEVADA on Thursday, September 20 @ 7 pm.' At the bottom, there is a partial view of a crowd of people at a rally.

<https://www.facebook.com/business/m/election-integrity>

<https://www.facebook.com/business/help/167836590566506>

We audit
Facebook's enforcement
of its political ad policy

We needed more **data** than Facebook makes readily available

- › Facebook's transparency tool: the *Ad Library*
- › *API*: archive of all (known) political ads
 - › **Insufficient**: ad is unavailable
 - 1) while not yet caught or 2) if never caught
- › *Web portal*: all currently active ads for all pages
 - › **Reverse-engineering** internal API; ads **disappear** once inactive
 - › **Custom-built tool** to archive these active ads

We **audit** ads deemed political by Facebook or us

Detected as political by Facebook

Not detected by Facebook

40,191 * False positive (subsection 5.2)	32,487 * True positive (subsection 5.1)	116,963 § False negative (subsection 6.3)
---	--	--

Not political

Actually political

Precision: 0.45

Recall: 0.22

F₁ score: 0.29

* Across all advertisers worldwide; estimate based on 55% FP rate in U.S.

§ Across political advertisers worldwide.

False positives in enforcement

Detected as political by Facebook		Not detected by Facebook
40,191 * False positive (subsection 5.2)	32,487 * True positive (subsection 5.1)	116,963 § False negative (subsection 6.3)
Not political	Actually political	
<i>Precision: 0.45</i>	<i>Recall: 0.22</i>	<i>F₁ score: 0.29</i>

* Across all advertisers worldwide; estimate based on 55% FP rate in U.S.

§ Across political advertisers worldwide.

False positives in enforcement

Detected as political by Facebook

Not detected by Facebook

40,191 *

False positive
(subsection 5.2)

32,487 *

True positive
(subsection 5.1)

116,963 §

False negative
(subsection 6.3)

Not political

Precision: 0

* Across all advertise
§ Ac

55% of *detected* ads
are ***incorrectly*** detected
(~40,191 ads)

Example falsely flagged ads

 **Friendly Ford - Las Vegas**
Sponsored
ID: 3631264633564503

Shopping for a new car? Check out our new models and competitive pricing.

FRIENDLY FORD Nevada's Only 19 Time **FRIENDLY SAVES**



Text Us @ 702-904-9820
www.FriendlyFordLV.com
600 N. Decatur Blvd - Las Vegas, NV 89107

FRIENDLY SAVES YES!

2020 Ford F-150 XLT | For only \$43,345

FRIENDLY FORD



Text Us @ 702-904-9820
www.FriendlyFordLV.com
600 N. Decatur Blvd - Las Vegas, NV 89107

2020 Ford F-250 | For only \$66,800

 **The News Break - News Break**
Gesponsord
ID: 3230741887020937

US News: Silky Peanut Butter Pie Recipe. Install news app trusted by millions to stay informed of latest US local news!

Never Miss Any Trending NEWS

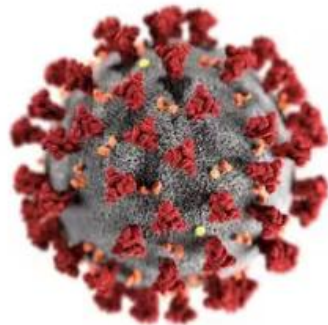


WWW.NEWSBREAKAPP.COM
Breaking news from US!
Connecting to the iTunes Store.

Nu installeren

 **SmartNews : Local Breaking News**
Gesponsord
ID: 272146773868279

11 new cases of COVID-19 confirmed in Madison County; 3,031 total cases. Install SmartNews for free to read more.



[HTTPS://APPS.APPLE.COM/US/APP/SMARTNEWS-LOCAL-BREA...](https://apps.apple.com/us/app/smartnews-local-breaking-news/id1467738682)
Local Madison County News
SmartNews is the award-winning news app downloaded by 40+ million readers in 100+ countries! SmartNews analyzes...

False negatives in enforcement

Detected as political by Facebook

Not detected by Facebook

40,191 *

False positive
(subsection 5.2)

32,487 *

True positive
(subsection 5.1)

116,963 §

False negative
(subsection 6.3)

Not political

Actually political

Precision: 0.45

Recall: 0.22

F₁ score: 0.29

* Across all advertisers worldwide; estimate based on 55% FP rate in U.S.

§ Across political advertisers worldwide.

False negatives in enforcement

Detected as political by Facebook

Not detected by Facebook

40,191 *

32,487 *

116,963 §

False positive
(subsection 5.2)

True positive
(subsection 5.1)

False negative
(subsection 6.3)

116,963 undeclared ads
from political actors
go **undetected**

political

F_1 score: 0.29

on 55% FP rate in U.S.
dwide.

Example missed ads

 **Vlaams Belang**
Sponsored
ID: 969996336828420

Geen moskee in onze gemeente. LIKE als je het eens bent!



Vlaams Belang
Political Party, Organization
604,813 people like this

[Like Page](#)

 **MR - Mouvement Réformateur**
Sponsored
ID: 3294645283949556

Inscrivez-vous au Libre Débat et abordons ensemble les thématiques clés de l'avenir du pays ! Évènement 100% en ligne, accessible à tous !



 **Elizabeth Warren**
Sponsored
ID: 812272636232502

Gear up and show the world that you are fighting for big, structural change! Our limited time sale ends at MIDNIGHT tonight. Enter code SHOPFROMHOME20 at checkout for 20% off your order.



 **Progressive Turnout Project**
Sponsored
ID: 2783976068510291

100% of proceeds support our work.



 **Don Young**
Sponsored
ID: 589568758425595



Limitations in enforcement

1. Approach:

automated solutions do not *learn*

obvious signals of political intent

2. Consistency: unequal performance globally

3. Transparency:

bad enforcement → bad transparency

- better transparency
- better accountability
- better enforcement
- **better security**

sound
methods

Conclusion

The importance of **data sets** and **methods**

- › *Sound research* needs **suitable** data sets and methods:

Case study: domain rankings

- ›› **Tranco**: increasing transparency, reproducibility
- ›› Research is increasing and **maturing**

The importance of **data sets** and **methods**

- › *Sound research* needs **available** data sets and methods

Case study: automated decision-making systems

- ›› Developing/evaluating/auditing needs **data access**
- ›› Access to crucial data is **worsening**

Sound data sets and methods

enable more **scientifically
grounded and inclusive,**

and ultimately **better**
web security research

make the web
more **secure**

Sound data sets and methods for web security research

Victor Le Pochat

References

- › [Sch18] Quirin Scheitle, Oliver Hohlfeld, Julien Gamba, Jonas Jelten, Torsten Zimmermann, Stephen D. Strowes, and Narseo Vallina-Rodriguez. 2018. A Long Way to the Top: Significance, Structure, and Stability of Internet Top Lists. In 2018 Internet Measurement Conference (IMC '18), 478–493. doi: 10.1145/3278532.3278574.
- › [LeP19] Victor Le Pochat, Tom Van Goethem, Samaneh Tajalizadehkhoob, Maciej Korczyński, and Wouter Joosen. 2019. Tranco: A Research-Oriented Top Sites Ranking Hardened Against Manipulation. In 26th Annual Network and Distributed System Security Symposium (NDSS '19). doi: 10.14722/ndss.2019.23386.
- › [LeP19b] Victor Le Pochat, Tom Van Goethem, and Wouter Joosen. 2019. Evaluating the Long-term Effects of Parameters on the Characteristics of the Tranco Top Sites Ranking. In 12th USENIX Workshop on Cyber Security Experimentation and Test (CSET '19). <https://www.usenix.org/conference/cset19/presentation/lepochat>.
- › [Dem20] Nurullah Demir, Matteo Große-Kampmann, Tobias Urban, Christian Wressnegger, Thorsten Holz, and Norbert Pohlmann. 2022. Reproducibility and Replicability of Web Measurement Studies. In ACM Web Conference 2022 (WWW '22), 533–544. doi: 10.1145/3485447.3512214.
- › [Rut22] K. Ruth, D. Kumar, B. Wang, L. Valenta, and Z. Durumeric. “Toppling Top Lists: Evaluating the Accuracy of Popular Website Lists”. In: 22nd ACM Internet Measurement Conference. IMC '22. doi: 10.1145/3517745.3561444.

References

- › [Jau20] Julian Jaurisch. Defining Online Political Advertising. How Difficulties in Delineating Paid Political Communication Can Be Addressed. Stiftung Neue Verantwortung, Nov. 2020. URL: https://www.stiftung-nv.de/sites/default/files/snv_definingpoliticalads.pdf.
- › [Lee19] Paddy Leerssen, Jef Ausloos, Brahim Zarouali, Natali Helberger, and Claes H. de Vreese. “Platform ad archives: promises and pitfalls”. In: Internet Policy Review 8.4 (Oct. 2019). DOI: 10.14763/2019.4.1421.
- › [Ede20] Laura Edelson, Tobias Lauinger, and Damon McCoy. “A Security Analysis of the Facebook Ad Library”. In: 2020 IEEE Symposium on Security and Privacy. SP '20. 2020, pp. 661–678. DOI: 10.1109/SP40000.2020.00084.
- › [Mat22] J. Nathan Matias, Austin Hounsel, and Nick Feamster. SoftwareSupported Audits of Decision-Making Systems: Testing Google and Facebook’s Political Advertising Policies. In: 25th ACM Conference on Computer-Supported Cooperative Work and Social Computing. CSCW '22. 2022.
- › [Sos21] Vera Sosnovik and Oana Goga. “Understanding the Complexity of Detecting Political Ads”. In: The Web Conference 2021. WWW '21. 2021, pp. 2002–2013. DOI: 10.1145/3442381.3450049.
- › [Ede21] Laura Edelson, Jason Chuang, Erika Franklin Fowler, Michael M. Franz, and Travis Ridout. A Standard for Universal Digital Ad Transparency. 2021. Knight First Amendment Institute. <https://knightcolumbia.org/content/a-standard-for-universal-digital-ad-transparency>