

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

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

NDSS 2019, 25 February 2019

Security researchers rely on top websites rankings

*“We perform a comprehensive analysis on **Alexa’s Top 1 Million** websites”*

*“We collected the benign pages from the **Alexa top 20K** websites”*

*“The list of websites we chose for our evaluation comes from the **Alexa Top Sites** service, the source widely used in prior research on Tor”*

 **Scott Helme** ✓
@Scott_Helme

Hey @AlexaInternet have you stopped providing your Top 1 Million Sites list?
s3.amazonaws.com/alexa-static/t...

10:38 AM - 19 Nov 2016

 **Alexa Support**
@Alexa_Support Follow

Replying to @Scott_Helme

@n0x00 @LewisArdern @adamcaudill @dongjiuju @TimmehWimmy Yes, the top 1m sites file has been retired.

6:40 PM - 21 Nov 2016

 **Scott Helme** ✓
@Scott_Helme Following

This is a shame and a real blow to my research. The data costs \$2,500 to get from the API, not something I can afford!
@AlexaInternet

10:50 AM - 19 Nov 2016

 **Martin Schmiedecker**
@Fr333k Follow

Replying to @jw_sec @Alexa_Support @Scott_Helme

feedback: this move actively hinders progress in computer science ... well done!

8:06 PM - 21 Nov 2016

 **Adam Caudill** ✓
@adamcaudill Follow

Replying to @Scott_Helme @AlexaInternet

Wow, that's a blow to a lot of security researchers - quite a loss to the community.

3:26 PM - 19 Nov 2016

 **isaac**
@_wirepair Follow

well shit

Alexa Support @Alexa_Support
Replying to @Scott_Helme

@n0x00 @LewisArdern @adamcaudill @dongjiuju @TimmehWimmy Yes, the top 1m sites file has been retired.

5:00 AM - 22 Nov 2016



Alexa Support

@Alexa_Support

Follow



Replying to [@Alexa_Support](#) [@n0x00](#) and 4 others

The file is back for now. We'll post an update before it changes again.

10:06 PM - 22 Nov 2016

Browser vendors make security decisions based on top websites rankings

Mozilla Security Blog



Delaying Further Symantec TLS Certificate Distrust



Wayne Thayer

“While the situation has been improving steadily, our latest data shows **well over 1% of the top 1-million websites** are still using a Symantec certificate that will be distrusted.”

<https://blog.mozilla.org/security/2018/10/10/delaying-further-symantec-tls-certificate-distrust/>

We studied four free, large and daily updated top websites rankings



Cisco Umbrella

Quantcast

How do these rankings **affect** research?

Can malicious actors **abuse** the rankings?

Can we **improve**?

Inherent properties

→ affect

Large-scale manipulation

→ abuse

A new ranking: Tranco

→ improve

Inherent properties

→ affect

Large-scale manipulation

→ abuse

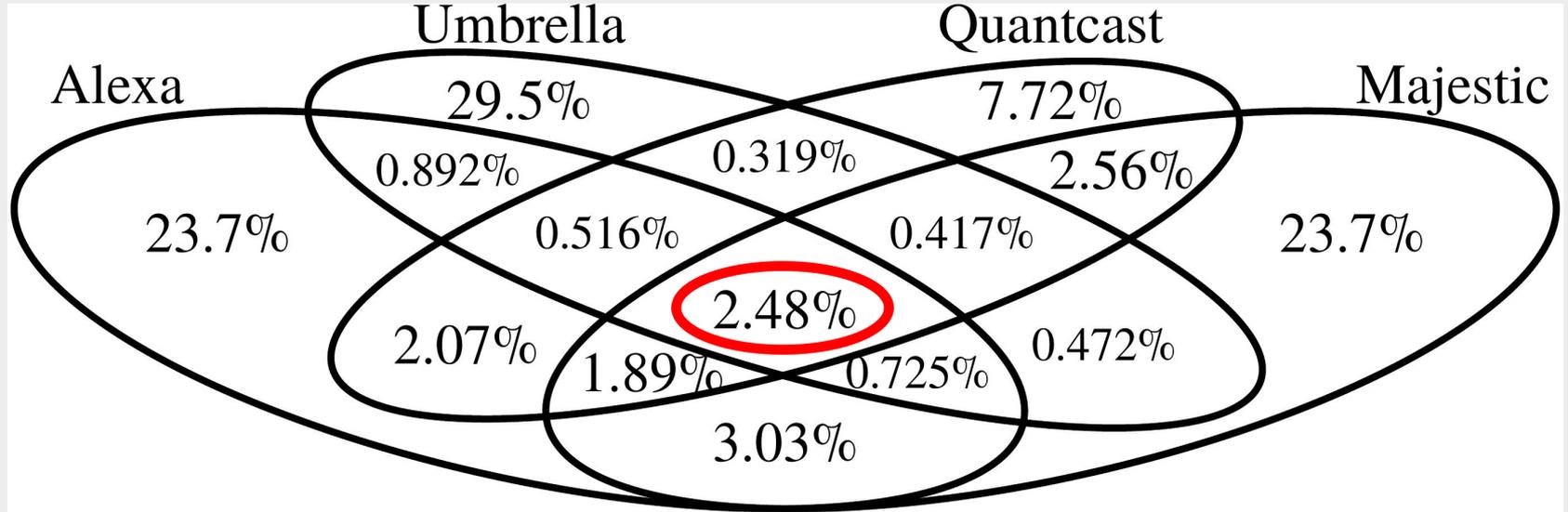
A new ranking: Tranco

→ improve

Inherent properties can skew conclusions of studies

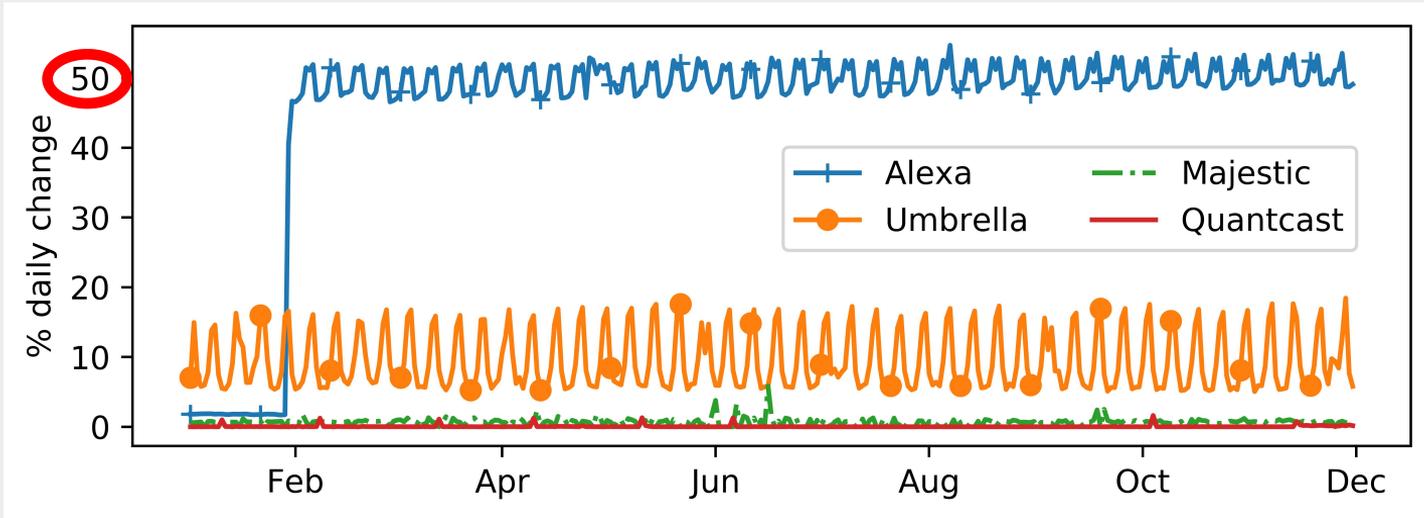
Inherent properties can skew conclusions of studies

- › Low agreement



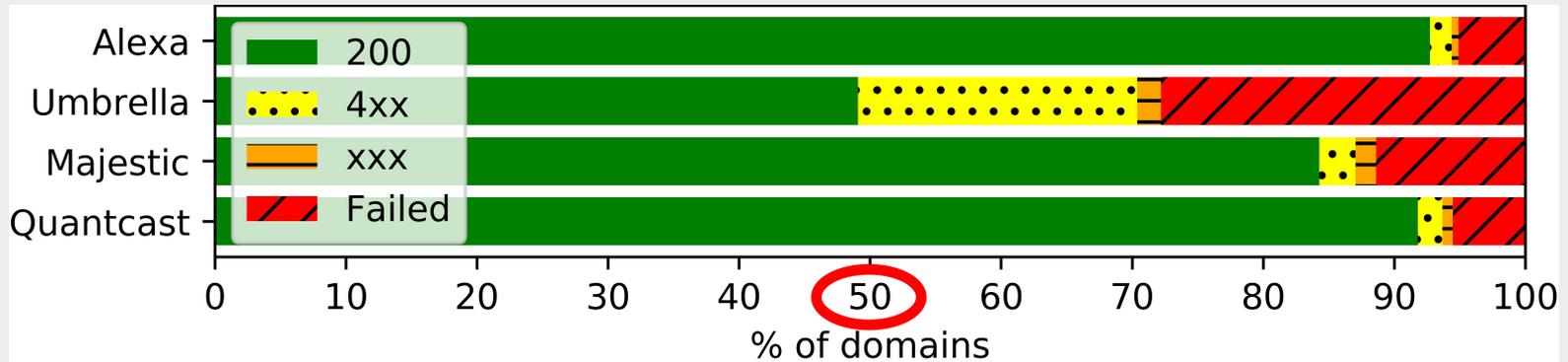
Inherent properties can skew conclusions of studies

- › Low agreement
- › Varying stability



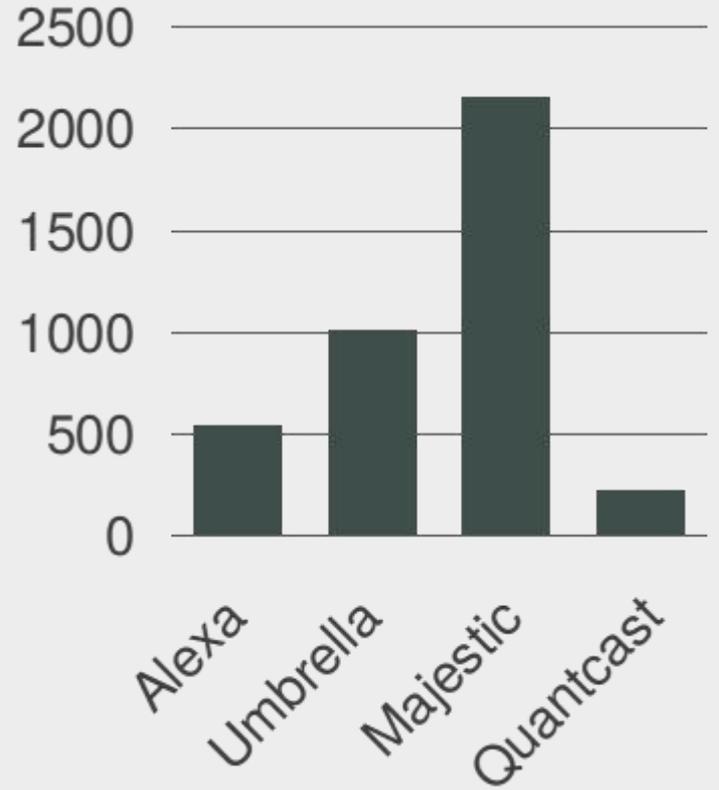
Inherent properties can skew conclusions of studies

- › Low agreement
- › Varying stability
- › Unresponsive sites



Inherent properties can skew conclusions of studies

- › Low agreement
- › Varying stability
- › Unresponsive sites
- › **Malicious sites**



Inherent properties can skew conclusions of studies

- › Low agreement
- › Varying stability
- › Unresponsive sites
- › Malicious sites

Inherent properties of rankings impact the **validity** and **reproducibility** of research

Inherent properties

→ affect

Large-scale manipulation → abuse

A new ranking: Tranco

→ improve

Malicious actors have incentives to manipulate rankings

incentive to manipulate

achieved by promoting

whitelisting malicious domains

own domains

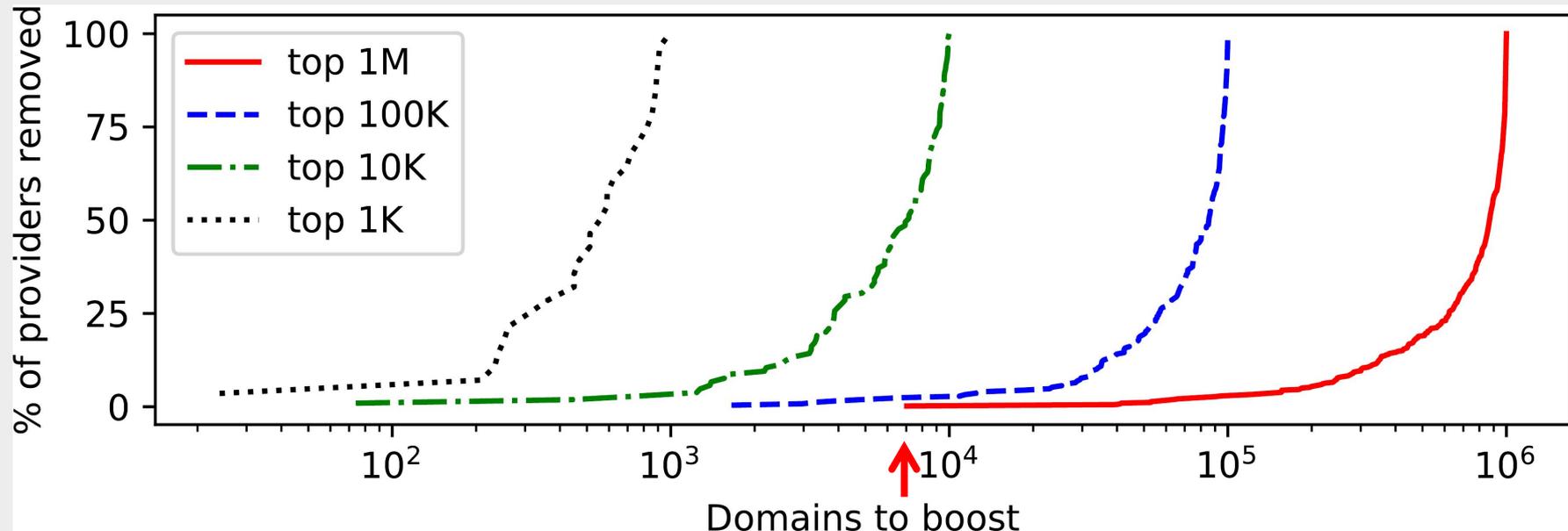
hiding malicious practices

other domains

changing prevalence of issue

'good'/'bad' domains

With large-scale manipulation of rankings, fingerprinting providers can remain undetected

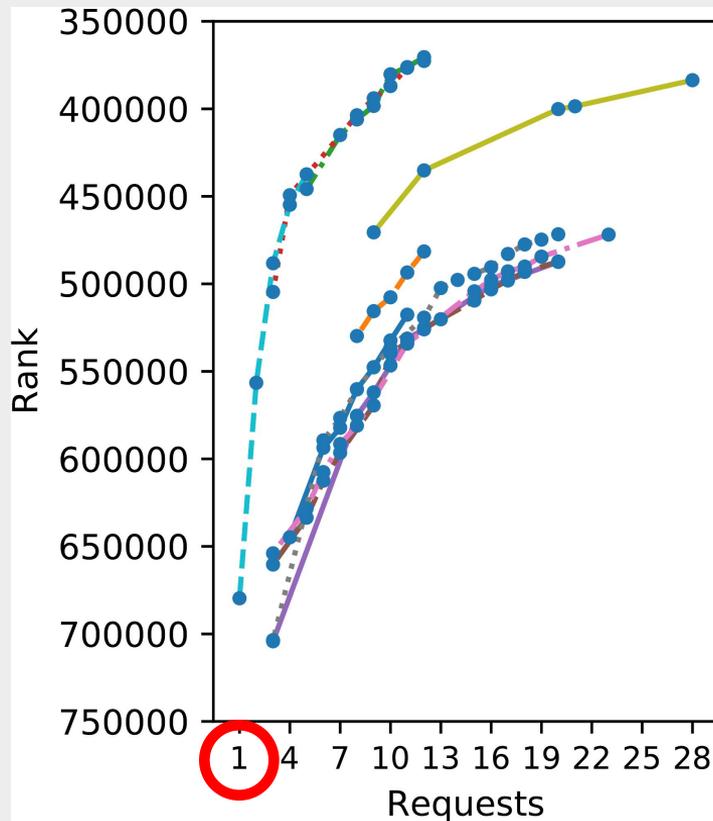


Simple, low-cost techniques make this manipulation possible on a **large scale**

Simple, low-cost techniques make this manipulation possible on a large scale

- › Alexa: browser extension

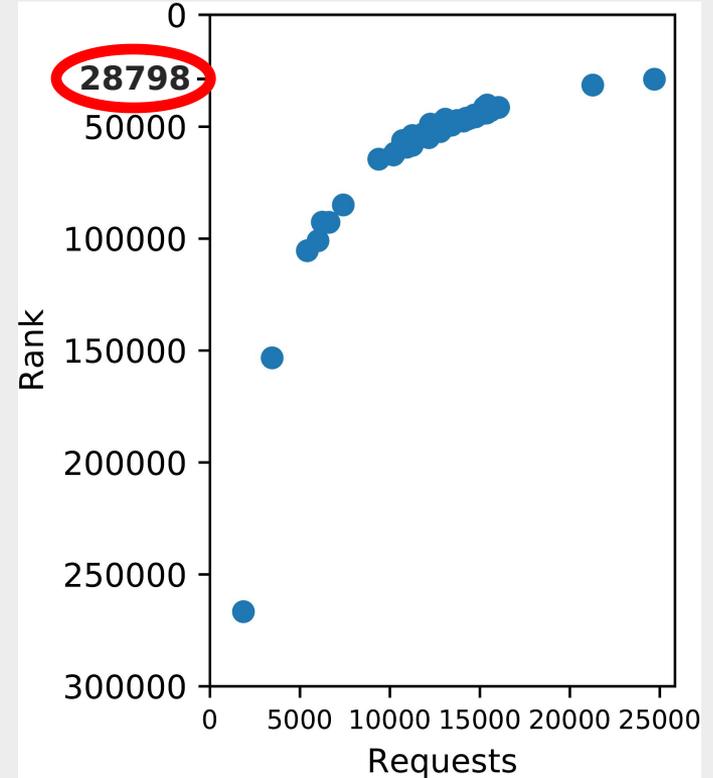
A single request
is sufficient to get
into the top million



Simple, low-cost techniques make this manipulation possible on a large scale

- › Alexa: analytics script

A malicious actor
can easily reach
a **very good rank**



Simple, low-cost techniques make this manipulation possible on a large scale

		Monetary	Effort	Time
Alexa	Extension	none	medium	low
	Analytics script	medium	medium	high
Umbrella	Cloud providers	low	medium	low
Majestic	Backlinks	high	high	high
	Reflected URLs	none	high	medium
Quantcast	Analytics script	low	medium	high

Simple, low-cost techniques make this manipulation possible on a large scale

		Monetary	Effort	Time
Alexa	Extension	none	medium	low
	Analytics script	medium	medium	high
Umbrella	Cloud providers	low	medium	low
Majestic	Backlinks	high	high	high
	Reflected URLs	none	high	medium
Quantcast	Analytics script	low	medium	high

Malicious actors may want to **manipulate** rankings, and such manipulation is feasible at a **large scale**

Inherent properties

→ affect

Large-scale manipulation

→ abuse

A new ranking: Tranco → improve

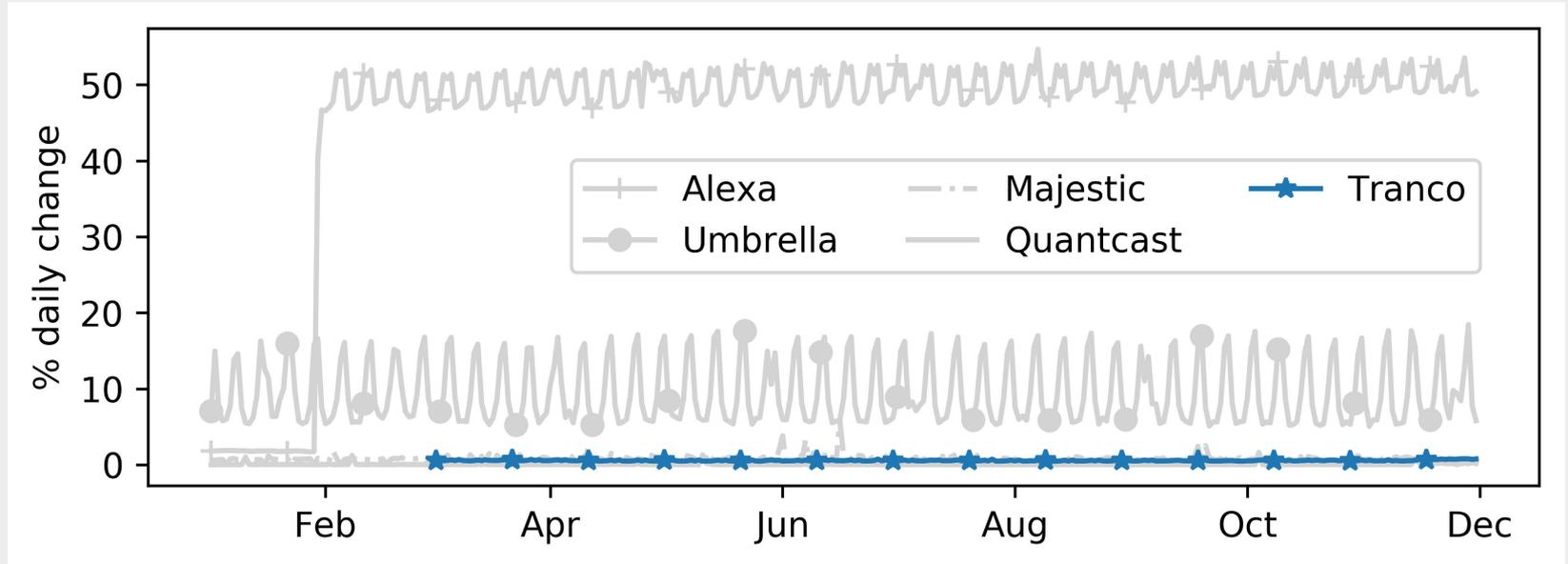
Tranco: an improved approach to top sites rankings

- › Aggregate existing rankings intelligently
- › Default settings: all providers, 30 days
- › Customizable: tailor to purpose of study
 - ›› Other combinations of providers/days
 - ›› Filters on specific services
 - ›› Remove unresponsive/malicious sites

Tranco improves on properties important for research

Tranco improves on properties important for research

› Stability



Tranco improves on properties important for research

- › Stability
- › Reproducibility

Information on the list with ID R2L9

[Download list](#)

Composition

This list combines the lists provided by **Alexa**, **Umbrella**, **Majestic**, **Quantcast** from 2019-01-06 to 2019-02-04 (**30 days**). [Read more](#) on the methods used to compose each of these lists to understand each list's properties and potential shortcomings.

These lists were combined using the **Dowdall rule** (the first domain gets 1 point, the second 1/2 points, ..., the last 1/N points and unranked domains 0 points). This method roughly reflects the observation of **Zipf's law** and the "long-tail effect" in the distribution of website popularity.

For each list, all domains were used.

The following filters were applied to the domains:

- Only pay-level domains were retained.

Of the combined and filtered list, the 1000000 first domains were used.

The list was first generated on 2019-02-04.

Tranco improves on properties important for research

- › Stability
- › Reproducibility
- › **Manipulation**

Tranco improves on properties important for research

- › Stability
- › Reproducibility
- › Manipulation

We provide Tranco, an **improved** ranking that is more suitable for **research** and is hardened against **manipulation**

We demonstrate how these rankings can affect **research results**

We uncover how attackers can **abuse** rankings to **influence** research results

We provide Tranco, an **improved** ranking to **strengthen** security research

Download the Tranco ranking:

<https://tranco-list.eu/>

Get the source code:

<https://github.com/DistriNet/tranco-list>

 DistriNet

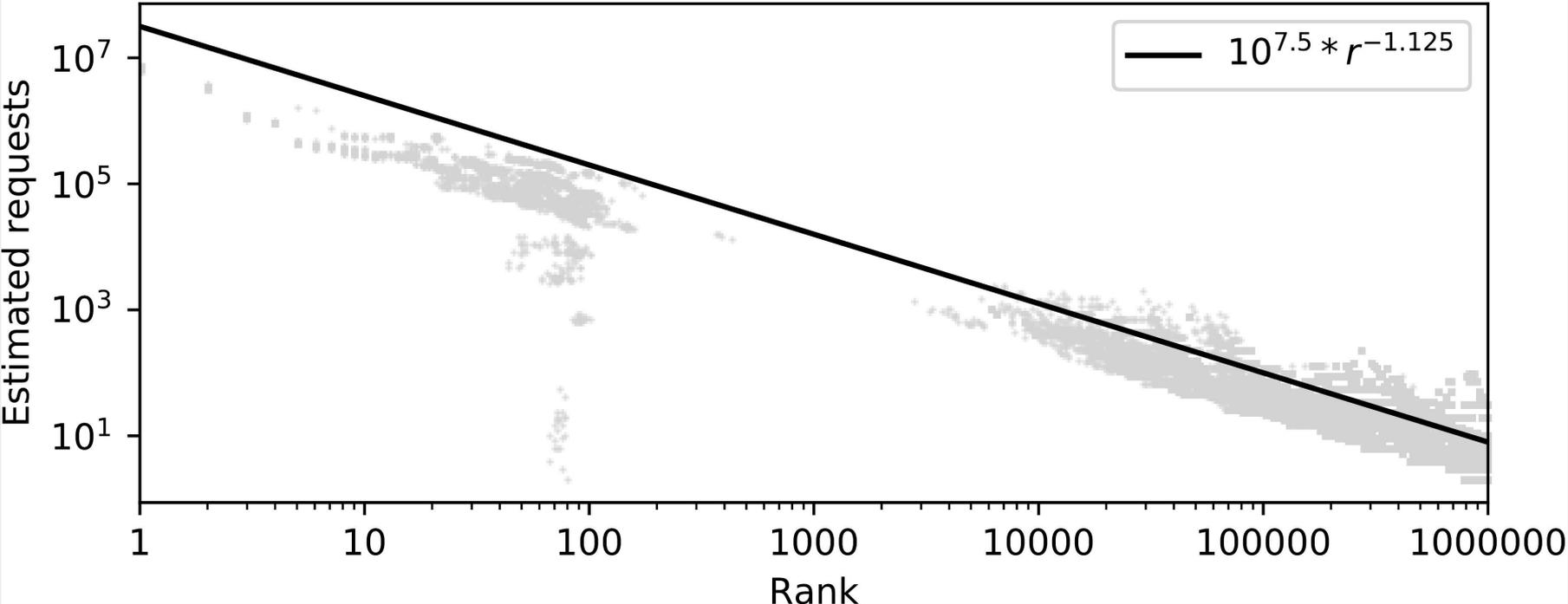
Thank you!

`victor.lepochat@cs.kuleuven.be`

References

1. Konoth, R.K., Vineti, E., Moonsamy, V., Lindorfer, M., Kruegel, C., Bos, H., and Vigna, G., “MineSweeper: An In-depth Look into Drive-by Cryptocurrency Mining and Its Defense,” in Proc. CCS, 2018, pp. 1714-1730. DOI: 10.1145/3243734.3243858
2. Kharraz, A., Robertson, W., and Kirda, E., “Surveylance: Automatically Detecting Online Survey Scams,” in Proc. SP, 2018, pp. 70-86. DOI: 10.1109/SP.2018.00044
3. Rimmer, V., Preuveneers, D., Juarez, M., Van Goethem, T., and Joosen, W., “Automated website fingerprinting through deep learning,” in Proc. NDSS, 2018. DOI: 10.14722/ndss.2018.23105
4. Scheitle, Q., Hohlfeld, O., Gamba, J., Jelten, J., Zimmermann, T., Strowes, S.D., & Vallina-Rodriguez, N., “A Long Way to the Top: Significance, Structure, and Stability of Internet Top Lists,” in Proc. IMC, 2018, pp. 478-493. DOI: 10.1145/3278532.3278574
5. G. Acar, C. Eubank, S. Englehardt, M. Juarez, A. Narayanan, and C. Diaz, “The web never forgets: Persistent tracking mechanisms in the wild,” in Proc. CCS, 2014, pp. 674–689. DOI: 10.1145/2660267.2660347
6. S. Englehardt and A. Narayanan, “Online tracking: A 1-million-site measurement and analysis,” in Proc. CCS, 2016, pp. 1388–1401. DOI: 10.1145/2976749.2978313
7. J. Fraenkel and B. Grofman, “The Borda count and its real-world alternatives: Comparing scoring rules in Nauru and Slovenia,” *Australian Journal of Political Science*, vol. 49, no. 2, pp. 186–205, 2014.

Estimated number of forged requests



Limitations

- › What if one list goes down?
 - › Still works with 3 other lists
 - › Change is permanently recorded and mentioned on list page
- › Completely resilient to manipulation?
 - › No, we rely on manipulable sources, but the required effort is higher
- › How permanent is the link?
 - › We are looking into more permanent archival (OSF)