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

Victor Le Pochat, Tom Van Goethem, Samaneh Tajalizadehkhooob, 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

Repeating @Scott_Helme

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

6:40 PM - 21 Nov 2016

Scott Helme
@Scott_Helme

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

Repeating @jwוכה @Alexa_Support @Scott_Helme

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

8:06 PM - 21 Nov 2016

Adam Caudill
@adamcaudill

Repeating @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

Alexa Support
@Alexa_Support

Repeating @Scott_Helme

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

5:00 AM - 22 Nov 2016

isaac
@_wireair

well sh*t
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

Oct 10 2018

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
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

![Graph showing % daily change from February to December for different data sources: Alexa, Majestic, Umbrella, and Quantcast. The graph fluctuates significantly throughout the year, with peaks and troughs.](image)
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 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

<table>
<thead>
<tr>
<th>Incentive to manipulate</th>
<th>Achieved by promoting</th>
</tr>
</thead>
<tbody>
<tr>
<td>whitelisting malicious domains</td>
<td>own domains</td>
</tr>
<tr>
<td>hiding malicious practices</td>
<td>other domains</td>
</tr>
<tr>
<td>changing prevalence of issue</td>
<td>'good'/'bad' domains</td>
</tr>
</tbody>
</table>
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

<table>
<thead>
<tr>
<th>Service</th>
<th>Technique</th>
<th>Monetary</th>
<th>Effort</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alexa</td>
<td>Extension</td>
<td>none</td>
<td>medium</td>
<td>low</td>
</tr>
<tr>
<td></td>
<td>Analytics script</td>
<td>medium</td>
<td>medium</td>
<td>high</td>
</tr>
<tr>
<td>Umbrella</td>
<td>Cloud providers</td>
<td>low</td>
<td>medium</td>
<td>low</td>
</tr>
<tr>
<td>Majestic</td>
<td>Backlinks</td>
<td>high</td>
<td>high</td>
<td>high</td>
</tr>
<tr>
<td></td>
<td>Reflected URLs</td>
<td>none</td>
<td>high</td>
<td>medium</td>
</tr>
<tr>
<td>Quantcast</td>
<td>Analytics script</td>
<td>low</td>
<td>medium</td>
<td>high</td>
</tr>
</tbody>
</table>
Simple, low-cost techniques make this manipulation possible on a large scale

<table>
<thead>
<tr>
<th></th>
<th>Monetary</th>
<th>Effort</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alexa</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Extension</td>
<td>none</td>
<td>medium</td>
<td>low</td>
</tr>
<tr>
<td>Analytics script</td>
<td>medium</td>
<td>medium</td>
<td>high</td>
</tr>
<tr>
<td>Umbrella</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cloud providers</td>
<td>low</td>
<td>medium</td>
<td>low</td>
</tr>
<tr>
<td>Majestic</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Backlinks</td>
<td>high</td>
<td>high</td>
<td>high</td>
</tr>
<tr>
<td>Reflected URLs</td>
<td>none</td>
<td>high</td>
<td>medium</td>
</tr>
<tr>
<td>Quantcast</td>
<td>Analytics script</td>
<td>low</td>
<td>medium</td>
</tr>
</tbody>
</table>

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

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
Thank you!

victor.lepochat@cs.kuleuven.be
References


Estimated number of forged requests

\[ \text{Estimated requests} = 10^{7.5} \times r^{-1.125} \]
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)