# Transparency into the Causes of Website Inaccessibility

Sadia Afroz
ICSI and UC Berkeley

Huilin Chen
UC Berkeley

Mobin Javed
LUMS

Marc Juarez
KU Leuven and ICSI

Vern Paxson
ICSI and UC Berkeley

Shoaib Asif Qazi
LUMS

Shaarif Sajid
LUMS

Michael Carl Tschantz
ICSI

Suppose you run a test to find that a website will load in the US but not in Pakistan. If the website is politically sensitive, it is not unreasonable to suspect censorship, but numerous other possibilities exist, and some cases lack transparency into which possibility actually occurred. Here, we discuss our efforts to document why some websites do not properly load in some countries.

In recent work [?], we looked at cases where websites were transparent about their reasons for blocking. We loaded webpages from the US and Pakistan, and when a page would only properly load in just one of the locations, we examined error codes and block pages, if any, to see whether they provided a reason. We found websites explicitly blocking users by country, which could be misinterpreted as middlebox censorship by measurement studies that do not check the reasons for failures. In particular, we found websites explicitly blocking to avoid compliance with the EU's GDPR. We also found websites explicitly using the country-wide and IP blocking features provided by the CDN Cloudflare, apparently for security reasons.

In each of these cases, we relied upon the website to provide transparency into it, as opposed to a middlebox censor, was doing the blocking. We can also determine whether a block is caused by the website or a middlebox in the absence of such transparency. In this work [?], we use traceroutes to determine whether a block is likely occurring at the server or somewhere before the server on the path to it from the client. Using this approach we have identified websites that are likely using country-based blocks.

In ongoing work, we are now developing methods for detecting other factors that may cause clients in different locations to experience differences with a website. We have developed a method of isolating some of these factors using SOCKS proxies. For instance, to measure the impact of placing measurement probes exclusively in academic networks, as opposed to diversifying the types of networks, we use pairs of SOCKS proxies: one in an academic network and another in a personal residence. The advantage of using a SOCKS proxy is that we can pin down the client to focus on just what varies between the two networks. Our preliminary results show significant differences in reachability between different types of networks.

Additionally, we are studying the implications that differential treatment by webservers of different locations – whether it is due to censorship or not – may have for censorship measurements and on privacy. We found cases in which different redirections at the server-side led to errors in our crawls, despite using exactly the same client. These are consistent client-side errors that might be naively confused with censorship events. We also found cases in which some locations were consistently not redirected to HTTPS, whereas they always were in the US. In particular, up to 40 sites in our dataset were never redirected to HTTPS in China but always were in the US. Even though, these do not necessarily represent cases of censorship, these observations have implications for the security and privacy of users of those sites and make them more vulnerable to application-layer censorship. Censorship studies that only look for content-blocking might miss pages that were successfully loaded but had their contents modified.

## References

[1] Afroz, S., Javed, M., Paxson, V., Qazi, S. A., Sajid, S., and Tschantz, M. C. A bestiary of blocking: The motivations and modes behind website unavailability. *ArXiv 1806.00459* (June 2018). To appear at FOCI 2018.

[2] Afroz, S., Tschantz, M. C., Sajid, S., Qazi, S. A., Javed, M., and Paxson, V. Exploring server-side blocking of regions. *ArXiv 1805.11606* (May 2018).