1. ABSTRACT

Recently, Internet censorship has been developed to fulfill political and ethical requirements and utilized to restrict people to gain access to specific content online across the countries. This can happen at a local network or the level of ISPs and up in the backbone of the internet itself. Internet censorship mostly was directed by national governments and motivated by politics and served as content filters right after the net was opened up to the general public. In this study, we propose a novel trustworthy framework to monitor a variety of internet censorship deployments that are diverse due to the geolocation of servers. Moreover, this survey included six related papers about Internet censorship measurement in different realm. The first paper provides an open observatory of Network interference, OONI, which aims to collect valuable data for network surveillance, interference, and outright censorship. The second paper proposes Iris, a scalable, ethical measurement system for detecting DNS manipulation. The third paper presents ICLab, a global censorship measurement platform that can measure a wide range of network interference and Internet censorship techniques. The fourth paper introduces Quack, a remote, scalable measurement system that can efficiently detect application-layer interference. The fifth paper proposes FilterMap to identify content filters from responding blockpages. The sixth paper proposes a censorship measurement framework, Disguiser, which enables end-to-end measurement to accurately detect the censorship activities and reveal the censor deployment.

2. LITERATURE REVIEW

2.1 Paper [1]: OONI: Open Observatory of Network Interference

Arturo Filast`o, Jacob Appelbaum. in USENIX Workshop on Free and Open Communications the Internet, 2012

The goal of this paper is to present the OONI framework to collect valuable data by introducing open methodologies, Free Libre Open-Source Software for researchers to conduct their surveillance and censorship measurements. The authors are dedicated to designing a framework to detect network-related tempering and facilitate the discovery of the network topology. A sub-goal of OONI is to raise public awareness on the topic of surveillance and censorship and provide various high-quality data to journalists and scholars to compose stories.

OOIN-probe software enables to detection of the presence of network traffic manipulation and obtaining the type and kind of restricted content from network services. It allows researchers to issue high-speed, large quantities of queries to various notorious protocols, such as HTTP, HTTPS, and DNS. Moreover, the researchers gather interception device information such as the make and model to determine the vendor and product adopted in surveillance and censorship systems.

The research has been initiated through experiments and control groups which were undertaken separately within a test network and actual network. Once the test results of these networks were mismatched that indicated that abnormal network activities have been detected. However, the mismatch of experiment data and control data cannot be categorized as network manipulation, it only shows there has been some kind of tempering that has taken place. This

kind of data is favored by the researchers and other users of the OONI framework, since false positives can be further investigated but false negatives remain in the dark. One awareness has been made through this research that a quantifiable level of risk shall be applied to the OONI users when performing tests on the platform due to political, economic, and legal circumstances. Before running a test on OONI, the users should be fully aware of the high stake they may incur due to the possible testing scenarios.

While drafting this paper, the researchers have given out a prototype of OONI. Thus far, OONI has been developed on top of the Twisted networking framework including test writing, support for locally running tests, and a local reporting mechanism. Moreover, OONI has been designed and allocated connections with OONIProbe which will be implemented in future work.

This paper has provided a broader review of the network surveillance and censorship detection, as well as interference devises utilization, variable data collection, solid and trustworthy framework build-up. However, there are lots of issues remaining to tackle in this study due to the FOCI size constraints. For instance, how much data should be stored pertaining to test results is left unclear. Moreover, there is still short of a method to properly anonymize the user's information and data sources, and neither is a good categorical heuristic for censored domains. In future study, the authors are expecting to address the uncertainty within the threat model. For example, the effort of ISPs in discovering censorship detection tools, and the trade-offs in adjusting it harder to detect the presence of OONIProbe and less reliable test results. In short, the authors will dedicate to implementing specifics for tests that run with OONIProbe and detail censorship detection tests by their framework.

2.2 Paper [2]: Global Measurement of DNS Manipulation

Paul Pearce, Ben Jones, Frank Li, Roya Ensafi, Nick Feamster, Nick Weaver, Vern Paxson. in USENIX Security Symposium, 2017

Censorship measurement remains an open problem, existing empirical measurements are sparse across time, space, and content despite the prevalence. As such, prior studies mainly focus on specific events, one or two countries or particular regions, or can be sparse cross content, namely on censorship of specific types of content such as social media, platforms, search engines, etc. Moreover, the deployed state of the art relies on volunteers which means researchers need someone in that country to perform some measurements or to host the equipment that enables them to perform measurements. As a result, there is a demand for continuous, diverse measurement to understand the scope, scale, and evolution of Internet censorship without user participation.

Recent work has shown nearly two-thirds of all Internet users are subject to some form of government-sponsored Internet censorship. For instance, the notorious Great Firewall of China performs the censorship of content that the government of China deems objectionable in some way. Similarly, a well-known country or large country censors a well-known Internet service that will create headline news and draw the researcher's attention to the problem. Internet censorship is quite large nowadays and there are many different facets for studying in various ways.

As Internet censorship techniques vary, this paper typically focuses on DNS manipulation processing by the governments. The authors want to measure censorship from multiple different countries, multiple different geographic

regions, and also more importantly from multiple vantage points within each of those countries. They also want their work to be longitudinal which means that someone can perform these measurements repeatedly over time with reasonable overhead. Thus, they don't have to go out and recreate an entire new measurement system and explore new vantage points each time. Moreover, they want to ensure that whatever system they build does not require user participation as this will fundamentally limit the scale of the system if it relies on volunteers in every country in every network every time the researchers want to do any type of measurement. Lastly, the measurement should be ethical and ensure no harm to the users or subject users.

Iris is a system to identify DNS manipulation globally and conduct a global measurement study that focuses on identifying the prevalence of censorship worldwide with a particular focus on the heterogeneity of content, countries, and particularly the heterogeneity of censorship within countries. After the authors perform all of the resolutions, they obtain the desired responses which enable them to make some intuitive conceptually graphs.

The first challenge is how to get a number of vantage points all around the globe. To address this challenge is through the use of open DNS resolvers. The use of open DNS resolvers has raised a question of ethics. Prior work is showing that open DNS resolvers are not actual resolvers, if performing measurements against them it will cause harm. In this study, it takes every effort to identify the infrastructure of DNS resolvers. Therefore, these resolvers are not qualified for leveraging to perform the DNS manipulation measurements safely and ethically. The next challenge is how to build a system that can be run repeatedly and how to design and build the system. Lastly, how to identify "wrong" DNS responses which turn out to be problematic in and of themselves and to give a lot of insights

that will touch on. The authors do not rely on consistency and independent verifiability first and foremost before getting into the system.

The authors quantify the generality and effectiveness of each metric to identify unmanipulated responses to assess the metrics with consistency and independent verifiability. This work proposes a method to implement a scalable global measurement for identifying DNS manipulation. To summarize, this paper makes the following contributions: First, Iris is designed, implemented, and deployed according to the Ethics principle. Second, an analysis metric for disambiguating DNS responses natural variation has been built up. Lastly, the authors emphasize the heterogeneity of DNS manipulation which deploys the state of the art of this study.

2.3 Paper [3]: ICLab: A Global, Longitudinal Internet Censorship Measurement Platform

Arian Akhavan Niaki, Shinyoung Cho, Zachary Weinberg, Nguyen Phong Hoang, Abbas Razaghpanah, Nicolas Christin, Phillipa Gill. in IEEE Symposium on Security and Privacy (S&P), 2020

The ultimate goal of this paper is to propose a global, comprehensive, longitudinal censorship measurement platform. Internet censorship detection has been limited within a period of time and in a few countries in the previous studies. Also, they trade-off details for a breadth of coverage. The authors present a novel internet measurement platform, ICLab which is capable of achieving mutual benefits between breadth of coverage and detail of measurement. Moreover, they adopt commercial Virtual Private Network servers (VPNs) as the vantage points to launch censorship measurements, such as DNS manipulation, TCP packet injection, and "block page" redirection. The design of ICLab seeks to minimize false positives and manual effort in the process of validation.

As commercial VPNs serve as vantage points within the Internet measurement, this will tackle the ethnic issue of requesting sensitive content online from politically restricted regions. ICLab engages in monitoring and providing detailed data collection from all levels of the network stacks. Also, ICLab applies volunteer-operated devices (VODs) in a few locations as alternative options for vantage points. Moreover, ICLab has high capabilities to implement new measurements when new censorship techniques come out, update the URLs when needed, and re-analyze old data when applicable. The researchers collect and archive massive observations in detail through ICLab which enables them to compare the blocking techniques deployed in different regions against different types of content. The essential observation of this longitudinal measurement is that censorship consistency has changed concurrently with the political shifts. What's more, unknown block pages and unknown forms of network interference have been detected via ICLab.

ICLab currently is focusing on website accessibility, each time the researchers only test the Alexa top 500, all of the citizen lab's global list of sensitive websites and country specific list. Moreover, the paper displays the packet traces and mimics browser behavior as closely as the manual effort could manage to reveal what happens on the network level and application-level when a website gets censored. Finally, the researchers end up with commercial VPN as vantage points other than volunteers for the consideration of storage capacity and bandit back with consumption. It avoids the practical and ethical problems with remote volunteer work, but it means that they can only access about 60 countries by commercial VPN. There is another critical problem with commercial VPNs that they often lie about the physical locations of their servers.

In this case, Overt censorship means that the government considered a material openly illegal to their citizens thus applying restrictions for access. An end-user tries to access a censored website, she might see an overt block page that states the contract information if she wants to dispute the classification. On the other hand, she might see a completely generic error message which is saying that it tried to connect to the site and equipment and there is no detail and assertion of authority contract information could be reached. On top of that, the site is broken for this moment and it could be called covert censorship. In short, most of the cases in this study proved that the government was using overt censorship for material that was openly illegal in that country and covert censorship for material that was supposed to be allowed.

ICLab plans to expand coverage in "not free" countries, such as Africa and South America in the future. The main difference between ICLab and other platforms is that ICLab is more detail driven than the breadth of detection coverage. While other platforms are intended to obtain as many vantage points as they could, ICLab concentrates on capturing as much detail as possible.

2.4 Paper [4]: Quack: Scalable Remote Measurement of Application-Layer Censorship

Benjamin VanderSloot, Allison McDonald, Will Scott, J. Alex Halderman, and Roya Ensafi. in the Proceedings of the 27th USENIX Security Symposium, 2018

Network filtering detection tools are currently being utilized to detect DNS and IP level interference at a global scale. Yet, there remains a large number of unmonitored types of blocking which is triggered on HTTP and TLS headers to be discovered. Analyzing the blocked keywords in this study, it facilitates building of the application-layer blocking ecosystem and comparing diverse censorship behavior from country to country.

This paper proposes Quack, a remote, scalable censorship measurement framework that ensures high-efficiency application-layer interference detection. Comparatively, the authors engage in a larger magnitude experiment to test for interference across thousands of autonomous systems than in the prior work.

Existing protocols and infrastructure are capable of remotely measuring network interference such as DNS poisoning and blocking between TCP/IP connectivity and remote machines. However, it is still short of a remote method to detect application-layer censorship. The prior work relies on people voluntarily involved in the experiment. Nevertheless, recruiting, maintaining, and coordinating a large set of volunteers is challenging. The design goals of Quack are detecting censored keywords that trigger network interference, minimizing risks for the safety of people in censored countries, ensuring analysis robustness techniques, and performing censorship measurement at scale.

The authors empirically define how long measurements should wait when a blocking event is incurred. Specifically, this allows ordinary servers to recover from stateful DPI disruption. In the experiment, longer timeout displays less likely to set a domain into failure due to a prior sensitive domain having triggered the stateful blocking for the given server. As such, a two-minute delay was identified as a minimum, since the system may need more time to schedule an upcoming test against the disrupted server.

Quack has been used to remotely measure network interference due to application-layer blocking such as DNS poisoning, IP-based blocking, and now application-layer censorship. The experiments regarding disruption technique detection provide insights by producing valuable datasets for political scientists, activists, and other members of the Internet freedom community. This system

also can be used to reveal shifts in application-layer censorship policies. In the future, the ultimate goal is to perform application-layer censorship measurement not merely on echo protocol but also involving other protocols.

2.5 Paper [5]: Measuring the Development of Network Censorship Filters at Global Scale

Ram Sundara Raman, Adrian Stoll, Jakub Daleky, Reethika Ramesh, Will Scottz, Roya Ensafi. in Network and Distributed Systems Security (NDSS) Symposium, 2020

As content filtering techniques have been excessively utilized for Internet censorship, the censorship measurement community lacks a systematic method to monitor the reproduction of these technologies. The need for establishing effective policies calls for a careful and detailed roadmap of the state and the evolution of content filtering censorship. In early works, researchers and policymakers are merely focused on a few types of content filtering technologies that need a heavy workload for manual detection and identification. In this study, the authors present a novel framework, FilterMap, which supports scalably remote monitor content filtering technologies according to their blockpages.

The FilterMap technology first gathers blockpages based on filter development when performing remote, in-network censorship measurement experiments. Then the researchers observed and clustered blockpages to extract signatures for longitudinal tracking. Finally, FilterMap ensures to identify and distinguish each new blockpage, while clustering and clarifying the same blockpages based on the filter deployments. The researchers manually verified each unique blockpage to eliminate false positives. In this research, they launched two large-scale measurements either from a breadth of sensitive test domains triggering as many as content filters or from several months of

longitudinal experiments to reveal the accuracy, scalability, and sustainability of FilterMap.

This study can be evaluated from the following two intervals: Data Collection and Data Analysis. For Data Collection, the measurements for massive domains on all HTTP(S) and Echo servers consume less time by sending requests in parallel. For Data analysis, FilterMap adopts iterative classification which outperforms the other collection methods by identifying 82 blockpages.

This paper has shown the most recent complete view on the deployment of censorship filters which respond with blockpages. In future research, one direction is to focus on other types of filter responses to identify filters such as the certificate returned in HTTPS measurements to further extract signatures and discover filters. FilterMap's analysis techniques have a profound impact on eliminating false positives and reducing noise in the same sort of research. Moreover, FilterMap's measurements inspire circumvention tool developers to develop circumvention strategies based on empirical experience. The longitudinal data collected about filter deployment can help regulate the utilization of filter technology and its illegal proliferation.

2.6 Paper [6]: Understanding the Practices of Global Censorship through Accurate, End-to-End Measurements

Lin Jin, Shuai Hao, Haining Wang, and Chase Cotton. in ACM SIGMETRICS, 2022

Recent years have witnessed steady progress on censorship measurement and censor deployment has been engaged in aligning political criteria and ethnic consideration. In previous works, censorship measurements mostly adopted unscalable manual inspection which identifies false positives and leaves false negatives undetected. In this paper, researchers propose the Disguiser

framework which conducts end-to-end measurements to precisely detect censorship activities. Moreover, this novel framework discloses the censor deployment without user participation.

In this paper, the main innovation is to conduct comprehensive and large-scale measurements on aspects to detect censor behaviors and censorship deployment. First, the authors send requests from vantage points located in different tested regions to control servers established outside the tested regions. Then, the control server will return a static response which is crafted with the purpose to distinguish censorship activities. Once the vantage point receives a response other than the static payload typically, the researchers can identify it was a censorship activity. Second, the authors implement an application traceroute to detect censor deployment along the path. Formerly, each vantage point makes a three-way handshake with the control server and sends the probing requests with incremented Time-to-Live (TTL) values. Then observe ICMP packets as responses to application traceroute progress which reveal the exact censor deployment.

The researchers conduct censorship measurement with all three notable protocols: DNS, HTTP, and HTTPS while other platforms lack the detection capabilities to cover all three protocols. By implementing the heuristics respectively on DNS, HTTP, and HTTPS protocol, they identify 52 false positives out of 254 thousand censorship activities which achieve a 10^{-6} positive rate and zero false negative rate.

In contrast, Disguiser enables large scale end-to-end measurement and receives ground truth on prospective response without censorship intervention which prior censorship studies are short of detection capabilities. Moreover, this

paper is the first to reveal a heuristic mechanism for manual review and has efficiently eliminated false negatives.

Architecture of Disguiser Framework

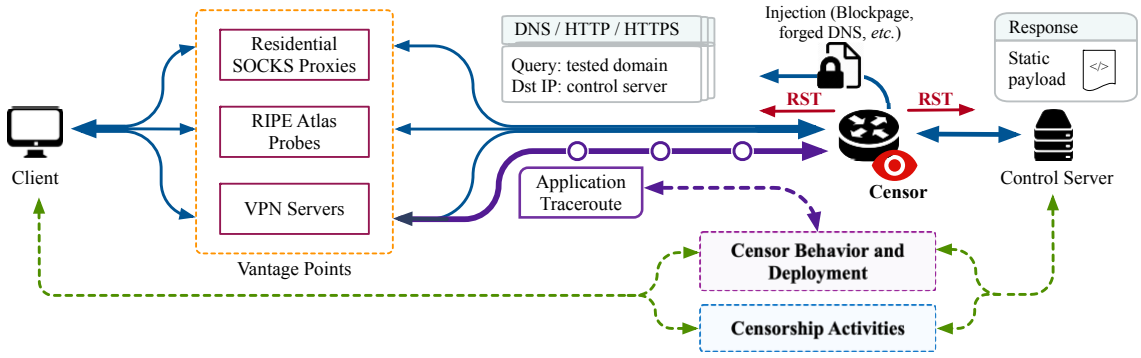

Fig.1. The figure shows the Architecture of Disguiser Framework. [6]

First, the client sends out an HTTP request to the Residential SOCKS Proxies (ProxyRack) from local. Then the Residential SOCKS Proxies reroute the query to a random destination around the world thus has altered the IP address of the query. Second, a query with a Dst IP address headed over to the Control Server and tested domain names embedded in the header was sent out from the Residential SOCKS Proxies. If there are discrepancies in the routing paths within the local vicinity of either endpoint, then a censorship candidate flag is raised. The Censor will adopt the following tactics to block this query to access the contents of tested domains hosted by the Control Server. For instance, blockpage, injected TCP packet or DNS response that reroutes the browser to a server controlled by the censor. That is the Overt censorship. For the Covert censorship, the censor will inject a transparent HTTP proxy (Blockpage), a TCP reset packet, a DNS error or non-reroutable address, or by discarding packets to ultimately stop accessing the destination Server.

Application Traceroute

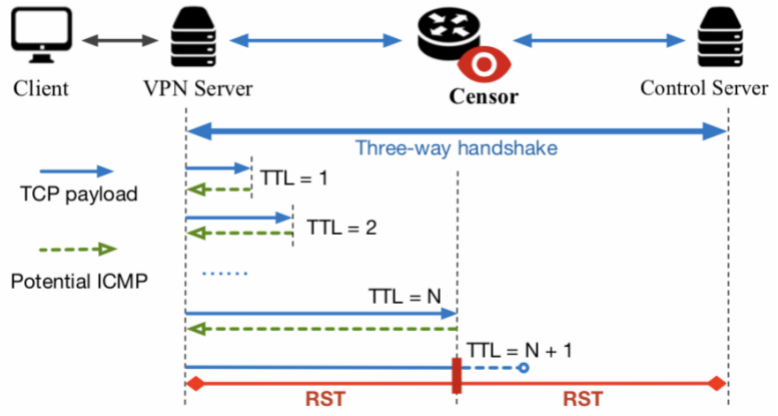

Fig.2. Application Traceroute [6]

The researchers conduct an experiment on exploring censor behavior and deployment by application traceroute. To do so, they sign up for VPN services for obtaining vantage points across the globe. Then, they perform an application traceroute to their control server to identify the censor's location and reveal the censor behaviors. Moreover, the VPN provider should be reliable to their vantage point's geolocation and will not alter the TTL values of packets so the intermediate routers can process the packets properly. The vantage points first connect with the control server by a TCP three-way handshake. Then, it increases the value of TTL of the request that contains sensitive domain names to trigger the censors. If the request reaches the censor on a particular router, it will return an ICMP packet to the client. Once a sign of censorship has been observed or the TTL reaches its limit of 64, The system will stop initiating requests. Note that, the request will reach the control server if there has no censor on the path, the server then relays a static payload back to the client side. As a result, it can be identified as a normal response according to the static payload to distinguish the abnormalities caused by censors.

In this case, the simplest way to tear down the connection is to drop the original packet when there is a censor router on the path. If this happens, the client side will not receive any ICMP packets after certain hops while performing the application traceroute. Then, run a normal traceroute to the control server to locate the IP address of the router that one hop further downstream. Through this method, the researchers observed 10 out of 13 countries deploy in-path censors which simply drop the requests.

3. *Research Plan*: Towards Fine-Grained Censorship Deployment: Demystifying the Impact of Path Diversity on Censorship

3.1 Research Problem

Inspire by the above insights, we develop a novel platform on top of Disguiser to detect censorship deployment variation by sending requests to multiple control servers across the world. We first sign up Proxyrack services for accessing massive vantage points and serves as intermediaries between the client and the control servers. Also, we need to establish multiple control servers in our experiment to reveal the diverse deployment of censors on the path regarding different geolocation of the control servers. Then, the system initiates batches of HTTP requests to our control servers which contain sensitive domain names. As a result, we observe different responses for requesting the same sensitive domain webpage from different control servers. In assumption, the censors should be deployed on all paths for forbidden domains to fulfill the political needs. However, it breaks this assumption based on our observation and brings more insights due to this phenomenon.

3.2 Experiment Methodology and Design

One advantage of the paper [6] that yields better performance than the other censorship measurement studies is that we observe the response from servers on the client side and allowing us to accurately recognize the censorship activities. When the control servers receive the requests from the client side, they will return a crafted static webpage that is different from other webpages. The present system has been built up on top of the Disguiser for scalable remote measurement detecting network interference caused by censorship on HTTP protocol.

For HTTP tests, we archive three types of responses that correspond to particular types of censorship activities. First, we scan our static payload from the response and identify it as being censored if the content text does not match our static payload. This is solid evidence that the webpage has been blocked by censors. Second, the control servers receive our requests successfully and are supposed to relay a static payload back to the client side. However, there is no response given which means the server has responded to a request within the time limit, but the censor has injected a reset package to block the original package return back to the client side. Thirdly, the response shows the server did not answer the request within the time limit, and no content text can be observed.

For each request, we allocate 5 seconds of wait time before we consider there is no response returning from the servers. Yet, we have to retry the queries at most 5 times to ensure the time out is not caused by the network congestion. Before we launch comprehensive experiments in each vantage point, we conduct a proxy-live check to verify the status of the proxy being used.

We use a compiled test list of sensitive domains to measure the censorship activities from country to country, while we set up 7 servers scattered on different continents of the world. The servers we established are located in Virginia (North

America), California (North America), Mumbai (Asia), London (Europe), Sao Paulo (South America), Bahrain (Middle East), and Cape Town (Africa). Based on the small-scale tests we conducted by using VPN, we have observed a phenomenon that the responses we received for acquiring the webpage of a sensitive domain from different servers that geolocation is not in the tested country may vary due to censor deployment along the path. Since the United States has been considered a non-censorship country together with Netherlands and Japan. In prior studies, the researchers set up control servers in these three countries to serve as a comparative group to identify the censorship activities. To this end, we also set up servers on both the east and west coasts of the United States to monitor different kinds of censorship compared to the response from other servers.

REFERENCES

[1] A. Filasto and J. Appelbaum. OONI: Open Observatory of Network Interference. In *USENIX Workshop on Free and Open Communications on the Internet (FOCI)*, 2012.

[2] P. Pearce, B. Jones, F. Li, R. Ensafi, N. Feamster, N. Weaver, and V. Paxson. Global measurement of DNS manipulation. In *USENIX Security Symposium*, 2017.

[3] A. Akhavan Niaki, S. Cho, Z. Weinberg, N. P. Hoang, A. Razaghpanah, N. Christin, and P. Gill. ICLab: A Global, Longitudinal Internet Censorship Measurement Platform. In *IEEE Symposium on Security and Privacy (SP)*, 2020.

[4] B. VanderSloot, A. McDonald, W. Scott, J. A. Halderman, and R. Ensafi. Quack: Scalable remote measurement of application-layer censorship. In *USENIX Security Symposium*, 2018.

[5] R. Sundara Raman, A. Stoll, J. Dalek, A. Sarabi, R. Ramesh, W. Scott, and R. Ensafi. Measuring the deployment of network censorship filters at global scale. In Network *and Distributed System Security Symposium (NDSS)*, 2020.

[6] L. Jin, S. Hao, H. Wang, and C. Cotton. Understanding the Practices of Global Censorship through Accurate, End-to-End Measurements, in *ACM SIGMETRICS*, 2022.