The Evolution of National AS Chokepoints and Their Connection to Internet Freedom

June 28, 2018

1 Abstract

Paths on the Autonomous Systems (AS) graph of the Internet derived from the Border Gateway Protocol (BGP) can be used by researchers to understand the dynamics of Internet topology and to interpret how that topology may enable nations to enact censorship, surveillance, or other Internet control measures. Unfortunately datasets of these paths are not generally made publicly available, and paths collected from measurements such as traceroute or BGP probes tend to be incomplete. Simulation frameworks have been used to generate AS paths based on inferred AS relationships. We introduce TOOL_NAME , a suite of open source, cross platform, and efficient tools for monitoring national chokepoints on the AS graph. We introduce chokepoint potential as an important measure of a nation's ability to control Internet traffic, either through censorship or surveillance. Previous research endeavors in this direction have only identified chokepoints in single snapshots. We apply our path simulation and routing tree datasets to view snapshots of the Internet over multiple years in order to introduce a new technique to investigate the evolution of the complex and dynamic AS level Internet topology. Through this approach we can more carefully evaluate the state of the Internet than was previously possible. As an application of TOOL_NAME we compare Freedom House's Freedom On The Net (FOTN) score for Internet freedom and with our chokepoint potential measure in order to interpret the relationship between AS-level topology and actual censorship activity, providing an illustration of new ways to monitor which governments can easily control the flow of information in their nations and whether they are acting on that potential.

2 Introduction

The Autonomous Systems layer of the Internet has grown and changed dramatically over its history. In the early 2000s, there were around 10,000 ASes connected to the global AS graph. Today there are over 60,000 of these ASes. The locations and relationships of these ASes determine how many chokepoints of AS-level routing exist and what strength these chokepoints have in regards to paths intercepted. This rapid expansion of the AS graph suggests a dynamic system with behavior that is important for understanding things like path robustness, AS hierarchies, and the potential for governments to control information as it flows in and out of their borders. It is certain that structrural qualities of national AS graphs have evolved differently from nation to nation, whether that be due to economic decision making, infrastructural necessities, or efforts to build a powerful censorship and surveillance network.

The Internet has been used as a tool for the citizens of authoritarian nations to voice opinions, organize revolutionary movements, and connect with other nations' governments and citizens to seek aid. Much of the organization, revolutionary momentum, and broadcasting of happenings during the Arab Spring can be attributed to social media communications through the Internet [11]. Because of this potential, national governments may take interest in maximizing their ability to control the flow of information on the Internet.

Internet censorship and surveillance have become profound social and technical issues facing the world. The portion of the global population that uses the Internet has increased substantially in recent years. According to the International Telecommunication Union, the number of individual users of the Internet has increased from 1.024 million in 2005 to 3.578 million in 2017 [3]. Of the Internet population, a majority of them are subject to some lack of Internet freedom. The 2017 Freedom on the Net report from freedom house reports that 64% of Internet users belong to a nation with Internet that is not free or partly free [2]. It is evident that Internet control is a powerful force for governments to wield. Whether or not AS-level topology supports these efforts is a research question pivotal to an understanding of the dynamics of Internet censorship and surveillance.

Every nation has a different layout and count of ASes. The censorship and surveillance strategies of nations also differ. For instance, China both conducts keyword filtering in border ASes and in internal provincial nodes [15]. Another nation that conducts extensive Internet censorship, Iran, has been found to route all of its Internet traffic through a centralized facility [6]. Having views into the AS topology of a nation, then, will both help researchers identify nations that could easily conduct censorship and also provide possible insight into what kind of censorship is likely being con-

ducted. The measure of chokepoint potential, defined later in this paper, is an effective way to capture the important properties of border ASes related to these capabilities.

The primary contributions of our work are as follows: **1.** We introduce the measure of chokepoint potential and motivate its value as way to interpret the capability for a nation to enact censorship of its Internet traffic. **2.** We provide an overview of the evolution of national AS-level chokepoints over time, showing the evolution of the Internet allowing for a comparision between nations. **3.** We develop a new tool, TOOL_NAME , that provides the capability for efficiently evaluating chokepoints for a given state of the Internet. Finally, **4.**, we show that our chokepoint measure has a significant relationship to Internet freedom as measured by a qualitative source.

3 Background

The AS-layer of the Internet is the highest level of organization of the Internet. ASes generally represent ISP networks, large university networks, or government entities. Interdomain routing between these ASes is governed by BGP, and individual ASes may have local preferences for how they wish to send packets along, but they cannot govern the paths selected by other ASes. AS topology researchers often use measurement tools to create a picture of the global Internet. These measurement tools, for instance BGP Updates from BGPStream [**?**] or empirical paths from traceroute [**?**], suffer from incompleteness and innacuracies [13].

AS-level studies are further complicated by the lack of ground truth data for AS relationships and BGP paths. For AS relationships, inference is often used based on economic considerations to classify the relationships of ASes. This was first done by Gao [8] by maximizing the occurances of certain economic rules on the AS graph by choosing a particular set or relationships. This technique was only evaluated against a single ISPs set of true relationships, however. As part of the CAIDA project, this inference technique was extended, leading to the CAIDA AS-relationship dataset [12], [1]. We choose these relationships for our purposes, and they are the current research standard.

Finding BGP paths is more of a challenge. Packet based simulations, wherein BGP is simulated directly, are computationally infeasible for the scales relevant to a global study. An accurate, but realistic, simulation technique must be chosen to provide useful research potential in this regard. The BGPSim algorithm [9] is one such simulation technique that is suitable for this study's purposes. BGPSim takes a set of AS relationships as input, such as those provided by the CAIDA dataset [1], and returns routing trees based off of these relationships. These paths are found via a modified breadth-first seach (BFS) algorithm. The BFS adds edges to routing trees first according to local preference (LP), then shortest path (SP), and finally tiebreak (TB). A resulting routing tree contains all the equally reasonable paths (according to economic concerns) that exist between source ASes (within the routing tree) to destination ASes (the root of the routing tree). We find this technique suitable for our purposes, so we extend BGPSim (in ways explained further in this paper) as part of TOOL_NAME .

4 Chokepoint Potential

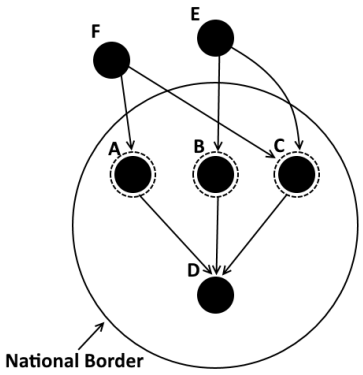

Figure 1: Chokepoint potential example. ASes A,B, and C are all border nodes. AS D is an internal node. ASes E and F are both external nodes. The out-to-in chokepoint potentials of A,B, and C are 0.25, 0.25, and 0.5 respectively.

In order to identify AS chokepoints and compare nations, we need a measure that can be calculated from the many paths between ASes. First, we decided to use a measure that is evaluated only on border ASes, or ASes that lie one hop from an AS belonging to another nation. We argue that this is intuitive because border ASes are the first opportunity for nations to censor incoming traffic and the last opportunity for nations to block access to foreign websites from within their borders. Additionally, while internal chokepoints may intercept many paths, those paths are required to have entered through a border AS (in the case of out-to-in paths) or exit through a border AS (for in-to-out paths). This simplification has the added benefit of making calculations more efficient.

We define the chokepoint potential of a border node to be the ratio of paths intercepted by that border node to the number of paths intercepted by all other border nodes. Consider a border node b, belonging to country c. If P_c is the set of paths in to or out of country c and B_p is the set of border nodes within path p, then the chokepoint potential of b, CP_b is defined formally in equation 1. This measure captures the relative strength exhibited by a border nodes in regards to what portion of paths it intercepts. This is calculated seperately for in-to-out paths (those starting from a source in the country in question) and out-to-in paths.

$$CP_b = \frac{|\{p \in P_c \text{ s.t. } \{b\} \subseteq B_p\}|}{\sum_{p \in P_c} |B_p|} \quad (1)$$

Given a set of BGP paths, this measure is an intuitive way to compare individual ASes and different nations. With this measure, the chokepoint potential sum over all border nodes for a country is 1.0, as in, all of the border nodes collectively control the flow of information over the nation's border. To compare one nation to another, we can inspect how many border nodes minimum are required to obtain a certain chokepoint potential. The more border nodes required, the more difficult it would be for that nation to perform censorship or surveillance. This is a clear way to differentiate nations based on the topology of their ASes.

5 TOOL_NAME

TOOL_NAME is a new tool that calculates the chokepoint potential for every border AS given a set of AS relationships. TOOL_NAME first takes the entire AS graph, as in from the CAIDA dataset [1], and uses the principles of BGPSim [9] to generate a set of routing trees. These routing trees, as well as a set of AS country codes (for identifying which nation an AS belongs to) are used to determine the chokepoint potentials for every border node. In our experiments, we used the country codes returned from Team Cymru's IP to ASN whois service [4].

In order to calculate the routing trees, we use an extended version of the BGPSim algorithm developed by Gil et. al in [9]. In our work we addressed the following limitations of BGPSim: (1) BGPSim returns a set of ASes for each path it considers but not the order in which they are visited; (2) Once routing trees are determined, they cannot be accessed later without recalculation; (3) BGPSim relies on the outdated parallelization framework DryadLinq for C#. To address these issues, we use a Python implementation we dub BGPSimPy. BGPSimPy returns ordered paths from its routing trees, saves routing trees to disk after calculation, and is parallelized with MPI via the mpi4py library. These improvements have the added benefit of yielding a cross platform routing tree algorithm that is ready to work on most hardware.

Once BGPSimPy generates the routing trees, they can be processed to determine chokepoint potentials. This is done by iterating over every path between each AS-pair. Because we use the same random tie-break as BGPSim, there is only 1 path between each AS-pair considered, even if multiple exist in the routing tree. Once a path is determined, it is traversed. For each node visited, the number of paths intercepted by that node is incremented. This is done for both in-to-out paths and out-to-in paths, so only one traversal is necessary per path. Additionally, the number of paths of each type that belong to each nation is tallied, as this makes up the denominator in equation 1.

6 Experimental Setup

There is ongoing research interest into ASes that intercept large portions of Internet paths [5], [10]. Some questions remain unanswered however. For instance, to what extent does the current state of the Internet support national governments' attempts to censor Internet traffic? How does this vary from nation to nation? Is the Internet developing more powerful chokepoints, or becoming more evenly accessible?

To probe these questions we first used our chokepoint evaluation technique to investigate the chokepoint potential of all nations for the current Internet as well as the change in chokepoint potentials over time. We looked at multiple snapshots of AS relationships from the CAIDA dataset (2012-2018). For each timestamp, we generated routing trees based on these relationships. Then we calculated the chokepoint potential for every border node per snapshot. As a result we can investigate how countries have changed overtime in their capability to enact censorship and surveillance. This is an attempt to understand what topological trends have developed historically. With this test we can compare nations, and see which ones have overtime increased their capability to control the flow of information across their borders.

If chokepoint potential can be leveraged to determine if a nation can easily implement censorship, it stands to reason that their might already be a negative relationship between the chokepoint potential of a nation and its Internet freedom. If a significant relationship were to be found it would strengthen chokepoint potential as a measure of a nation's censorship capability and it would increase the value in monitoring chokepoint potentials across the globe.

We tested whether a significant relationship exists between our measure of national chokepoints and a qualitative evaluation of Internet Freedom. For Internet Freedom we used the Freedom House's Freedom On The Net report [2]. FOTN scores quantify the level of Internet freedom in countries. Each country receives a numerical score from 0 (the most free) to 100 (the least free). To evaluate our measure, we first must define a way to rank each nation in regards to our chokepoint potential measure. Acharya et al. [5] chose to determine how many ASes were needed (globally) to intercept 90% of paths as a measure of the AS hierarchy. We choose a similar measure. Because we are looking at national comparisons, however, we record how many border ASes are required to incercept 90% of in-to-out or out-to-in paths for each nation. We compared this number with each nation's (of those recorded by Freedom House) freedom on the net score.

7 Results

7.1 Nations Over Time

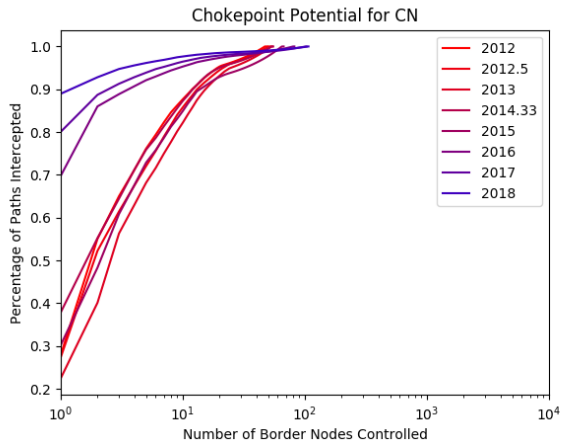

Figure 2: The evolution of China's in-to-out Chokepoint Potential

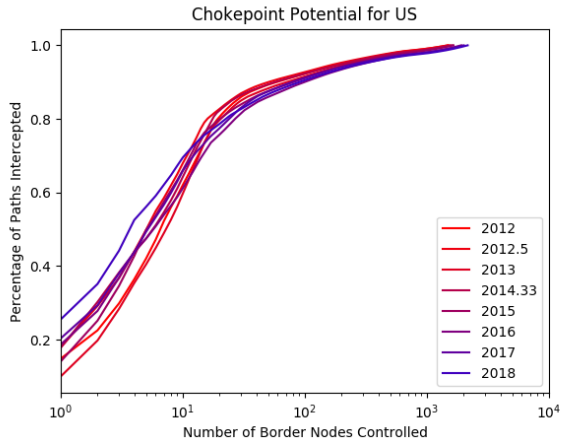

Figure 3: The evolution of the US's Chokepoint Potential

To inspect the changes over time for each nation we arrange all the border nodes belonging to a nation into a list. The list of border ASes is reverse sorted so that the first AS has the highest chokepoint potential. The sum of all these nodes' chokepoint potential is 1.0. Then we step through the list, and record the cumulative chokepoint potential of all the ASes seen so far. If we plot the number of ASes controled vs the cumulative chokepoint potential for that number of ASes, we can see how many ASes are required for a certain nation to control different percentages of paths. We repeat this process for each snapshot to highlight changes over time.

For instance, consider figures 2 and 3. The x-axis (log-scale, so that countries with large differences in AS counts can be compared more easily) is the number of border nodes controlled, and the y-axis is the ratio of in-to-out paths intercepted. Each line represents a different snapshot, with the more red lines being farther in the past and more blue lines being more recent.

The United States and China are shown here to exhibit their dramatic differences. First, it is worth pointing out that the United States has many more ASes than China, hence it's line extends further to the right in these plots. We also see that China, in all cases, can control a much larger portion of its paths with much fewer ASes than the United States. This result is entirely expected. Of more interest is the trends that can be seen over time. China clearly has developed an AS topology such that very few ASes can intercept nearly all BGP paths. The US has evolved in a different way. While it has become somewhat easier for the US to intercept most of its paths, it has become more difficult to intercept around 70% of its paths and higher. This suggest an expansion of ASes on the interior and more connections, as well as a strengthening of the very top ASes.

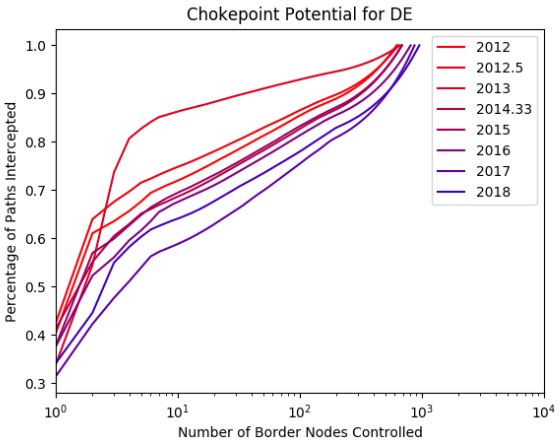

Figure 4: The evolution of Germany's in-to-out Chokepoint Potential

Not all nations have evolved to a state where it easier to control paths. One example is Germany, as shown in figure 4. For Germany, a fairly constant trend shows that it has become more difficult to intercept BGP paths. Unlike the other examples, any amount of German border ASes intercept less paths in more recent tests.

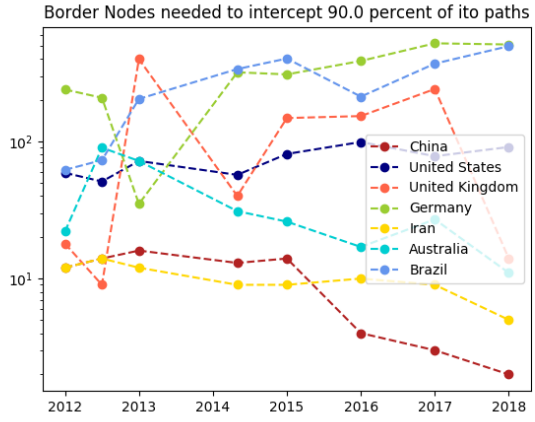

Figure 5: Number of ASes to intercept 90% of ito paths for multiple nations over time

In order to compare multiple nations, we have plot-

ted the number of ASes needed to intercept 90% of in-to-out paths for each timestamp recorded. These results are detailed in figure 5.

7.2 Internet Freedom

In our test of the relationship between Internet Freedom and chokepoint potential, we plotted the Freedom on the Net score of each nation vs the number of border ASes that that nation needed to in order to intercept 90% of in-to-out paths. Additionally, we evaluated the relationship with an Ordinarly Least Squares (OLS) fit, and found that the relationship was statistically significant, with a p-value ≤ 0.002 for 2017. The relationship held for other timestamps as well. The results for 2017 are shown in figure 6

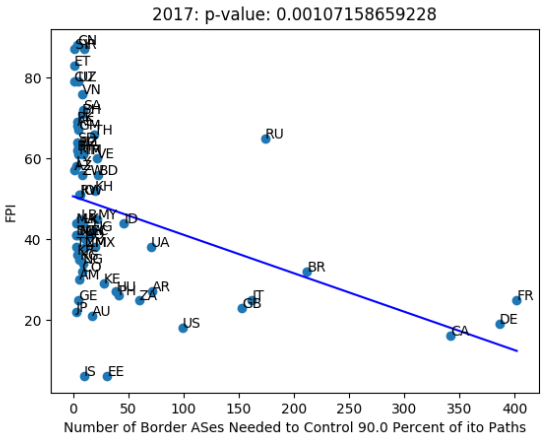

Figure 6: Number of ASes to intercept 90% of ito paths vs Freedom On The Net Score (2017)

For each timestamp we see what appears to be two general modes of behavior. Nations that are not free or partly free tend to requrie few ASes to intercept 90% of their paths. Free nations on the otherhand require varying degrees of large numbers of ASes to control the same portion of their paths. There are interesting outliers for both situations, however. Countries like Estonia (EE) or Iceland (IS) are very free but require few nodes to control most of their paths. The reason for these outliers is likely that their overall AS counts are very low. On the other hand, Russia (RU) is a very interesting outlier in that it is found to be not free by FOTN, but requires a large number of ASes to control most of its paths. This suggests that the censorship efforts in Russia might be of types that do not require AS level chokepoints.

8 Related Work

Previous studies have used BGP path models to find ASes that intercept a high fraction of paths. One such project in [5] identified that 90% of paths on the Internet could be intercepted with only 30 or so ASes. The researchers in [5] generated paths by first using only paths to top websites as defined by the Alexa top websites project, and then appending additional edges to those paths from the full AS graph according to the principles defined by Gao [8]. Many of the ASes that were found to intercept a large number of paths were found to be within nations that conduct censorship. As an extension of these results, another paper [10] revealed that ASes that intercept many paths could be utilized for decoy routing. This approach for identifying AS level chokepoints has several potential pitfalls that are remedied by the approach taken in this paper. First, the authors in [5] didn't consider that many of the paths found from the Alexa top websites would have destinations in nations that censor Internet traffic, particularly China due to its large Internet population. This artificially inflates the chokepoint nature of Chinese ASes. As an alternative, in this paper we consider paths from every source-destination AS pair, and we make a distinction between in-to-out paths and out-to-in paths. Secondly, chokepoints have previously only been identified at a single snapshot of the Internet. This makes it difficult to discuss the evolution of Internet topology, and whether or not chokepoints are anomalous or common. Finally, the aforementioned approach for identifying chokepoints does not allow a clear comparison between different nations. It can be said that one nation controls a large portion of Internet paths, but not how easily traffic directed through that country could be intercepted on a national level. For this, some aggregate measure across all of the border ASes within a country must be considered, as it is in this paper.

In [15], Xu et al. investigated the AS level topology of China to identify where keyword filtering occurred. They found that the most effective ASes with which to deploy keyword filtering devices are those in the backbone of the Chinese AS topology. A relevant contribution of [15] is that, while most filtering occurs in border ASes, some filtering is controlled by non-central provincial ASes. China has a diverse strategy for Internet censorship, both targeting chokepoints and the Chinese provincial network. The potential for various forms of censorship in regards to various AS level topologies motivates the question: Is centralized censorship or decentralized censorship more common? Instead of directly identifying censorship devices on the AS graph, we instead quantify the chokepoint potential of ASes on the national level, and then compare that with qualitative Internet freedom measures and empirical censorship events.

We are not the first to investigate the relationship between Internet freedom and AS-level topology. Similar techniques have been used to classify nations according to the connectivity of their ASes [14]. This has only been done for a single moment in time, however, making the results limited in terms of stability and predictability. Additionaly, previous work has not used country level chokepoints as the link between Internet topology and censorship or surveillance practices. The work in [14] chose to relate Internet topology to the Freedom of the Press measure from freedom house instead of the Freedom On The Net score. They chose to do this to include more nations. Additionally, they

didn't include the United States and Russia in their experiments because they were outliers in regards to their topologies. Through our approach we hope to extend this previous work by finding an interesting measure for understanding the dynamics of all nations, as well as targeting our results more specifically to Internet freedom by using the Freedom On The Net score as our measure for Internet freedom. Through our techniques, we provide a simple measure that not only sheds light on the relationship between topology and Internet freedom, but reveals currently free nations that could easily implement censorship if their governments decided to.

9 Discussion

9.1 Routing Trees Dataset

In the hope to further AS topology research, we have open sourced the routing tree datasets generated in this study. While we generated routing trees with an efficient algorithm, it still requires considerable time to calculate them, particulary for multiple timestamps. Additionally, each set of routing trees takes up on the order of 50GB of disk space. By releasing these datasets we hope that researchers looking for a particular set of routing trees will find working with these simpler than recalculating them. This also provides an alternative to research projects that might otherwise use measurement tools to estimate BGP paths.

9.2 Future Work

While linking chokepoint potential to FOTN scores is a substantial contribution, it still stands that FOTN can serve only as a proxy for censorship. This is useful for large trends and general understanding, but a more fine-grained approach could be used to interpret the direct connection between actual censorship events and the shifts in the AS graph. The Open Observatory of Network Interference, or OONI, [7] provides Internet users around the world with a tool called the ooniprobe. The ooniprobe lets users run a suite of tests to identify censorship anamolies of various types, and the results are recorded in the large OONI database. Matching up changes in OONI measurements, such as increased censorship campaigning in an authoritarian nation, with shifts in chokepoint potential would be a major step in understanding the interplay of censorship and AS-level chokepoints. This process involves designing a way to classify censorship events and chokepoint potential changes, and as such lies beyond the scope of this study.

References

[1] Caida as relationships. http://data.caida.org/datasets/as-relationships/.

[2] Freedom on the net 2017. https://freedomhouse.org/sites/default/files/FOTN_2017_Final.pdf.

[3] International telecommunication union internet statistics. https://www.itu.int/en/ITU-D/Statistics/Pages/stat/default.aspx.

[4] Ip-asn mapping tool from team cymru. https://www.team-cymru.com/IP-ASN-mapping.html.

[5] HB Acharya, Sambuddho Chakravarty, and Devashish Gosain. Few throats to choke: On the current structure of the internet. In *Local Computer Networks (LCN), 2017 IEEE 42nd Conference on*, pages 339–346. IEEE, 2017.

[6] Simurgh Aryan, Homa Aryan, and J Alex Halderman. Internet censorship in iran: A first look. In *FOCI*, 2013.

[7] Arturo Filasto and Jacob Appelbaum. Ooni: Open observatory of network interference. In *FOCI*, 2012.

[8] Lixin Gao. On inferring autonomous system relationships in the internet. *IEEE/ACM Transactions on networking*, 9(6):733–745, 2001.

[9] Phillipa Gill, Michael Schapira, and Sharon Goldberg. Modeling on quicksand: dealing with the scarcity of ground truth in interdomain routing data. *ACM SIGCOMM Computer Communication Review*, 42(1):40–46, 2012.

[10] Devashish Gosain, Anshika Agarwal, Sambuddho Chakravarty, and HB Acharya. The devil's in the details: Placing decoy routers in the internet. In *Proceedings of the 33rd Annual Computer Security Applications Conference*, pages 577–589. ACM, 2017.

[11] Philip N Howard, Aiden Duffy, Deen Freelon, Muzammil M Hussain, Will Mari, and Marwa Maziad. Opening closed regimes: what was the role of social media during the arab spring? 2011.

[12] Matthew Luckie, Bradley Huffaker, Amogh Dhamdhere, Vasileios Giotsas, et al. As relationships, customer cones, and validation. In *Proceedings of the 2013 conference on Internet measurement conference*, pages 243–256. ACM, 2013.

[13] Zhuoqing Morley Mao, Jennifer Rexford, Jia Wang, and Randy H Katz. Towards an accurate as-level traceroute tool. In *Proceedings of the 2003 conference on Applications, technologies, architectures, and protocols for computer communications*, pages 365–378. ACM, 2003.

[14] Rachee Singh, Hyungjoon Koo, Najmehalsadat Miramirkhani, Fahimeh Mirhaj, Phillipa Gill, and Leman Akoglu. The politics of routing: Investigating the relationship between AS connectivity and internet freedom. In *6th USENIX Workshop on Free and Open Communications on the Internet (FOCI 16)*, Austin, TX, 2016. USENIX Associa-

[15] Xueyang Xu, Z Morley Mao, and J Alex Halderman. Internet censorship in china: Where does the filtering occur? In *International Conference on Passive and Active Network Measurement*, pages 133–142. Springer, 2011.