Mapping the Maze: A Study of Internet shutdowns across local and regional geographies in India

Ritik Malik
2018406

BTP report submitted in partial fulfillment of the requirements
for the Degree of B.Tech. in Computer Science & Biosciences
on Dec 17, 2020

BTP Track: Research

BTP Advisors
Dr. Sambuddho Chakravarty
Dr. Aasim Khan

Indraprastha Institute of Information Technology
New Delhi

Student's Declaration

I hereby declare that the work presented in the report entitled **"Mapping the Maze: A Study of Internet shutdowns across local and regional geographies in India"** submitted by me for the partial fulfillment of the requirements for the degree of *Bachelor of Technology* in *Computer Science and Biosciences* at Indraprastha Institute of Information Technology, Delhi, is an authentic record of my work carried out under guidance of **Dr. Sambuddho Chakravarty** and **Dr. Aasim Khan**. Due acknowledgements have been given in the report to all material used. This work has not been submitted anywhere else for the reward of any other degree.

..... **Place & Date: IIIT Delhi, 17th Dec 2020**
Ritik Malik

Certificate

This is to certify that the above statement made by the candidate is correct to the best of my knowledge.

..... **Place & Date: IIIT Delhi, 17th Dec 2020**
Dr. Sambuddho Chakravarty

.....
Dr. Aasim Khan

Abstract

Around the world, powerful governments have frequently used internet shutdowns to curb the freedom of expression and devoice the people, and today, India is no exception. There could be many reasons for such outages, some could be deliberate others may be caused by unintentional reasons, like power failure, device rupture, regional outages, etc. However, we will mainly focus on intentional internet shutdowns. There are currently well-researched methods and are capable of detecting *network outages on a national level scale*, but very few that explicitly talk about the *internet shutdowns* on a more fine-grained *regional and local scale*, and especially beyond the metropolitan geographies in India. Presently, there is no publicly available information that tells us how the internet shutdown is achieved by the ISPs, and whether it was deliberate or perhaps due to something unintentional. Using Big Data, available through other projects that monitor the Internet routinely, we devise a way to understand and predict such a trend. Thus, in this project, we try to address the following questions:

- How does the Indian government implement these shutdowns?
- Is the technique(s) implemented the same across all the ISPs?
- Can we correlate historical shutdowns with some publicly available datasets?
- Is there any correlation between internet shutdowns and outages?
- Can we predict shutdowns in the future after analyzing the current trend?

Acknowledgments

I would like to express my sincere gratitude towards my supervisors Dr. Sambuddho Chakravarty, Dr. Aasim Khan for their guidance and encouragement throughout the project. Further, I would also like to thank Dr. Devashish Gosain for thier constant support and motivation. They motivated me to put in my best and responded to my queries and questions promptly. I will forever be grateful to them for what I learned with their support.

Contents

Chapter 1

Introduction

Internet shutdowns are an absolute restriction placed on the use of internet services due to an order issued by a government body. It may be limited to a specific place and to specific period, time or number of days. It may be limited to mobile internet on smartphones, or the wired broadband that usually connects a desktop - or both at the same time.

The Internet shutdown trend in India is not a new phenomenon. Instead it dates back to 2012 [12]. The reason proposed by government for a shutdown is to control the mob, spreading fake news and violence, but it may have some opposite effect in reality. It creates a lack of medium, suppressing the local voices, which brings social unrest. Not to mention the economic loss caused by it, which accounts for $1.3+ billion for India alone in 2019 [9].

Significantly less work has been done in the field of detecting internet shutdowns at a smaller scale, specifically aiming at India due to the following reasons -

- The internet connectivity is still inferior and distributed unevenly as compared to other countries.
- There is a lack of probes from projects like CAIDA ARK [3], Routeview [19] and RIPE atlas [18], which makes it difficult to track the regional activities
- The majority of the shutdowns are on a microscopic scale, like district wise (mainly due to community violence), making it hard to detect it on the massive global datasets.
- People living in shutdown hotspots face many complications, as their basic amenities are disrupted anytime, and many shutdowns/outages could possibly be treated as normal and may get unreported.

Many governments have adopted Internet shutdowns in the past, like Egypt [11], Libya [13], Iran [1], Sudan [2] and more recently like Belarus [5], and almost all of them have resulted in mass protests and riots against the government, which makes it more tempting to analyze the situation.

1.1 Motivation and Problem Description

In this fast-growing digital era, where almost everyone relies on the internet directly or indirectly, internet shutdowns bring massive disruption to basic amenities in the life of people living in the shutdown hotspots. Especially making it hard for the working class of people like office employees, doctors, drivers, students, etc. This forces us to question the digital freedom of the people.

Different tools have been developed in this field before like CAIDA IODA [4], whose aim is to develop an operational prototype system that monitors the internet, in near-real-time, to identify *macroscopic Internet outages* affecting the edge of the network, i.e., significantly impacting an AS or a *large fraction of a country*. Other projects like RIPE Atlas [18], which is the RIPE NCC's main Internet data collection system. It is a global network of devices, called probes and anchors, that actively measure Internet connectivity. They usually do not talk about internet shutdowns but outages and that too on a country level. Also, there exists other tools like Censys [6], which is used to collects data on hosts and websites through daily port scan of the IPv4 address space with open-source ZMap [23]. More tools and papers discussed in Chapter 2.

To the best of our knowledge, no work has yet been done that detects internet shutdowns on a fine-grained regional scale or talks about the government's techniques.

1.1.1 Problem Description

In this project, we try to figure out how the government across the globe implements these shutdowns and if there exist any publicly available datasets capable of detecting and correlating historical shutdowns with various geopolitical events on a regional scale. We will also observe how internet outages distinguish themselves from internet shutdowns.

Chapter 2

Background and Relevant Work

A plethora of research has been done on Internet outage detection systems, like the Trinocular [24], which uses adaptive probing to detect internet outages at the network edge level. They used ICMP probes according to the Bayesian inference. Their model of the internet was an outage centric model which is populated from long-term observations.

RiskRoute [21], a framework for mitigating Network Outages Threats, it evaluates risk via the concept of bit-risk miles, the geographically scaled outage risk of traffic in a network. The results of their analyses high-light current risks of network infrastructures and how those risks can, in some cases, be significantly mitigated using RiskRoute recommendations.

Detecting Peering Infrastructure Outages in the Wild [22], their methodology relies on the observation that BGP communities, announced with routing updates, are an excellent and yet unexplored source of information, allowing them to pinpoint outage locations with high accuracy.

Analysis of Country-wide Internet Outages Caused by Censorship [20], their primary source of data were BGP interdomain routing control plane data, unsolicited data plane traffic to unassigned address space; active macroscopic traceroute measurements, RIR delegation files, and MaxMind's geolocation database.

Though they all are great in strategy and yield positive results, they do not talk much about *internet shutdowns*, that too on a *regional level scale*. So we aim to find some correlation between the publicly available datasets and the historical shutdown events in India.

Chapter 3

Definitions and Hypothesis

3.1 Definitions

3.1.1 Autonomous System

An autonomous system (AS) is a collection of connected Internet Protocol (IP) routing prefixes (like /24, /18, /16, etc.) under the control of one or more network operators on behalf of a single administrative entity or domain that presents a common, clearly defined routing policy to the internet. AS numbers are assigned in blocks by Internet Assigned Numbers Authority (IANA) [10] to regional Internet registries (RIRs) [17]. The appropriate RIR then assigns ASNs to entities within its designated area from the block assigned by IANA. *E.g.*, BHARTI AIRTEL is AS9498, VODAFONE is AS38266, *etc.*

3.1.2 Border Gateway Protocol

Border Gateway Protocol (BGP) is a standardized exterior gateway protocol designed to exchange routing and reachability information among autonomous systems (AS) on the internet. BGP is classified as a path-vector routing protocol, and it makes routing decisions based on paths, network policies, or rule-sets configured by a network administrator.

BGP used for routing within an autonomous system is called Interior Border Gateway Protocol, Internal BGP (iBGP). In contrast, the Internet application of the protocol is called Exterior Border Gateway Protocol, External BGP (eBGP).

3.2 Hypothesis

Since the overall mechanism of implementation of internet shutdown is currently unknown, we propose two hypotheses that might explain the techniques behind the implementation -

1. The ASes can stop advertising the BGP paths in the radius of internet showdown affected areas. In simple words, they can turn off the router advertising the prefixes so that no traffic can be exchanged beyond that point.
2. They might implement a firewall rule that drops all the outgoing packets, so no one will be able to communicate beyond that point.

We will perform specific tests and measures on historical internet shutdown data to see if they fit any of the two hypotheses. It is equally likely that both the above hypotheses are incorrect and the government uses something completely different. That would be another interesting finding in itself.

Chapter 4

Experimental Setup and Results

4.1 Experimental Setup and Results

Talking of public datasets, we have various sources - CAIDA ARK project [3], IODA [4], Censys [6], RIPE Atlas [18], Routeviews [19]. Our primary focus, for now, is on the publicly available dataset from the "University of Oregon Route Views Archive Project", which provides us with historical BGP dumps from its various nodes across the world. Other datasets are being worked upon and will be included in the next thesis for comparison and cross-validation.

Why and about Routeviews :

- The motivation was to obtain historical BGP information about the global routing system from the perspectives of several different backbones and locations around the internet.
- The Routeviews project has 31 collectors now, spread across the world but mostly concentrated in the US.
- They provide access to historical BGP dumps about the global routing system from their various nodes.
- These dumps are recorded actively every day with the granularity of 2 hours, which sounds very precise for our work, as some shutdowns might only be implemented for a couple of hours due to some *government exam* or maybe some regional outages.

The table 4.1 is an example of general output of a routeviews dump (after formatting).

Prefix	PATH1	PATH2	PATH3	PATH4	PATH5
67.158.52.0/24	37353	37100	6453	9498	135247

Table 4.1: Here 1st column represents the prefix, while other column represents the ASNs

In table 4.1 67.158.52.0/24 is the destination prefix belonging to AS135247, and there exists a unique path from source AS37353 to reach this prefix at AS135247 -

AS37353 -> AS37100 -> AS6453 -> AS9498 -> AS135247

There are millions of such entries with unique paths in a particular timestamp record from an arbitrary node. On average, after searching and combining in the dumps from all nodes, for a particular timestamp, around 100 140 unique paths are observed for a particular prefix.

We created a database using different publicly available data sources, consisting of the list of all prefixes and netblocks for top ISPs in India like Airtel, TATA, Jio, BSNL, Vodafone, Idea, etc., to their respective geolocation data.

The prefixes lists were scrapped from ipinfo.io [15] and CIDR reports [7], and the best effort geolocation was from the Maxmind API [16]. So we got our final "prefix-to-geolocation" database. Now we needed some case studies to verify our hypothesis.

4.1.1 Case Study 1: Rajasthan

Initially, we decided to correlate the dumps with the internet shutdown in Rajasthan in July 2018 [14]. Following are the insights of that shutdown -

- The reason for the shutdown was to prevent cheating in a constable recruitment exam.
- The shutdown was implemented on two consecutive days - 14th and 15th July 2018
- The exam was over four shifts and the shutdown happened for 2 hours in every shift.
- It was implemented within a 5 km radius of every examination center.

So we made a prototype pipeline that performed the following sequential tasks -

- A web scraper to download data from all Routeviews nodes for the full month
- Extract the dumps and perform preprocessing and formatting
- Search specific prefixes in the whole database and dump the raw results in CSVs
- Use the CSVs to make final graphs.

We tried to correlate the dump results with the Rajasthan prefixes obtained from our "prefix-to-geolocation" database. We downloaded data for 30 days of July and 4 timestamps on each day, then we clubbed together all 4 timestamps. Now we need to see the frequency graph from this database to various prefixes belonging to Rajasthan.

If our hypothesis stands correct, the number of such paths to a particular prefix being advertised in the shutdown region should decrease significantly.

We correlated the dumps with Rajasthan prefixes, and we got the output graph as shown in Figure 4.1. Many other graphs were similar to this one.

To verify that this pattern was the repeated across all ISP, we tested it on various ISPs having their prefixes in Rajasthan, which led us to this graph as shown in Figure 4.2. These prefixes were selected randomly from ISPs that belonged to Rajasthan. This graph was obtained for randomly chosen /24 prefixes across all the ISPs from the database and then superimposing them; the results were very encouraging, as they show a familiar pattern.

The initial graphs were a bit vague because we wanted to get a feel of what kind of output we expect in the future to change our scripts accordingly. Another essential thing to mention, that the dumps are in a range greater than 500, contrasting to what we said earlier as 120 to 130. It is because we combined the results of 5 timestamps for a single day.

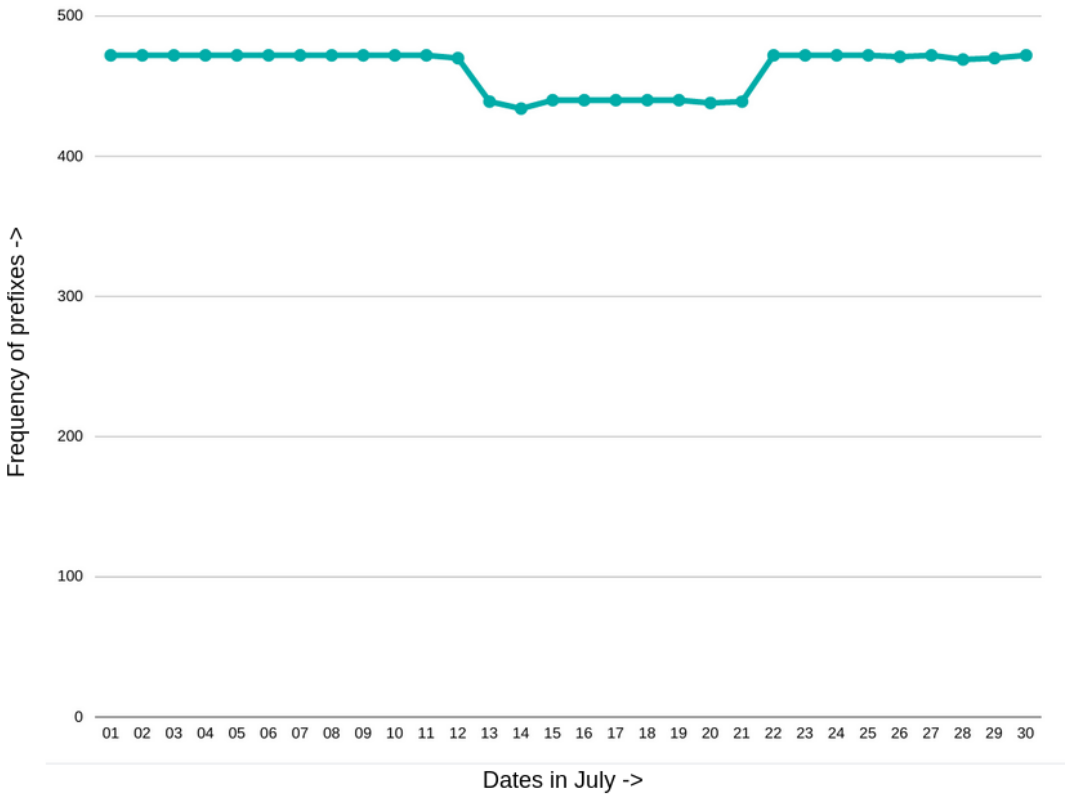

Figure 4.1: Rajasthan #1: X-axis represents the dates in July 2018 While the Y-axis represents the number of unique paths to a particular prefix in Rajasthan's shutdown region

4.1.2 Case Study 2: NRC-CAA Act

This time, we decided to move forward with our hypothesis on a little broader scale. A more recent and infamous event was chosen - *The NRC CAA protest* (Dec 2019) [8], which is known to curb the internet in large parts of India. This time, since the testing was on a broader scale, the old prototype pipeline turned out to be futile to work with such large datasets. It was expected to take a whole month for it to digest the data and return the results.

So a new and more robust pipeline was designed with the following upgrades -

- It uses MongoDB to manage the vast datasets and was feasible to run thousands of queries parallelly compared to the iterative approach earlier.
- It converted the computation of a month into a single day. The overall size of monthly data turned out to be in the range of 400 - 500 GB.
- Learning from the previous case study, we decided to segregate the results based on ISPs and further segregating them based on the number of dips in the advertised prefixes.
- We included regional ISPs too, to get more precise results area wise.
- We also made separate graphs for different timestamps, contradicting what we did last time by clubbing them together. So instead of increasing in the Y-axis, our graphs now stretched their legs on the X-axis.
- In order to get rid of redundant data, we made graphs for only those prefixes which lost more than 40% of their unique paths at least once, in the whole month of Dec 2019

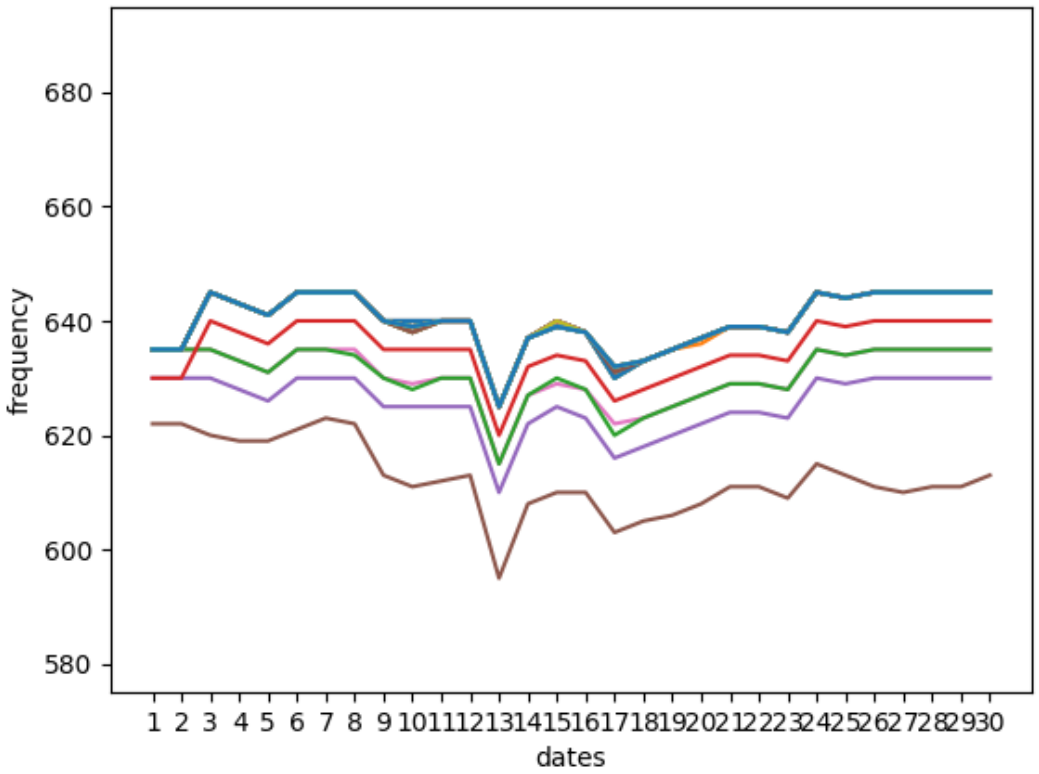

Figure 4.2: Rajasthan #2: X-axis represents the dates in July 2018 While the Y-axis represents the number of unique paths to a particular prefix in Rajasthan's shutdown region

The 6 phases of the new pipeline can be visualized as shown in Figure 4.3

The graph layout was also improved significantly this time:

1. We added additional metadata in their header
2. All the values were normalized according to the highest freq obtained, thus creating a percentile graph
3. Different libraries were used for the graphs this time so tweaking it was much more comfortable and interactive in a web browser
4. The graphs obtained were further divided into two subcategories :
   (a) Those prefixes that starting/stopped advertising themselves completely
   (b) Those prefixes which got a dip greater than 40% and then again rebounded

The graph shown in Figure 4.4 belongs to the *4 (a)* category. The above graph has the following metadata, which is self-explanatory :

AIRTEL_BHARTI_AS9498 -> 23.15.36.0/24 , [max = 140, min = 0]. This graph shows that the following prefix stopped advertising itself at roughly 6 AM UST on 21st Dec 2019. Since we are using a Bokeh plot, the above graph is much more interactive and zoom friendly (for the chaotic X-axis) in a web browser than a typical matplotlib graph.

Coming to the main graphs of interest, the 2nd category, *4 (b)* as shown in Figure 4.5.

ACTFIBRENET_AS131269 -> 106.51.113.0/24 , [max = 125 , min = 0]. The freq shows a hiccup at 1600, 13.12.2019, which coincides with the day after the NRC CAA bill was passed on

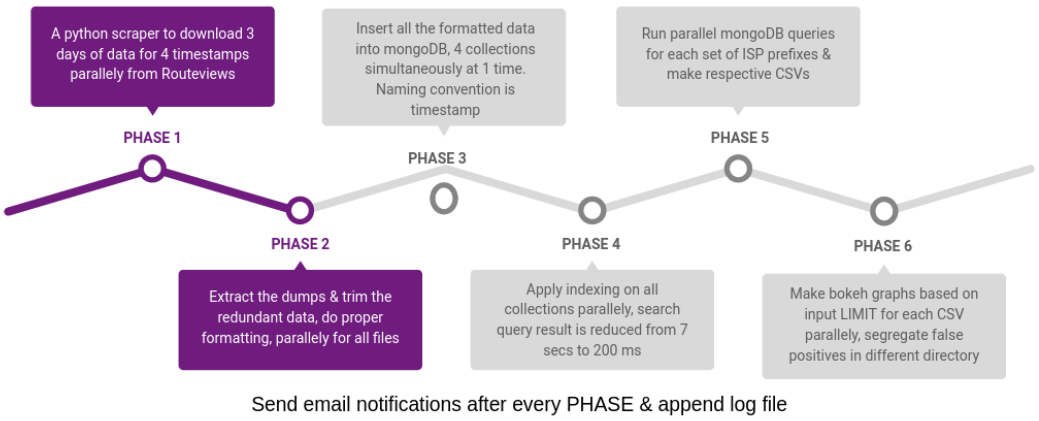

Figure 4.3: Pipeline : The new 6 Phase pipeline requires only starting parameters, it does everything automatically and efficiently

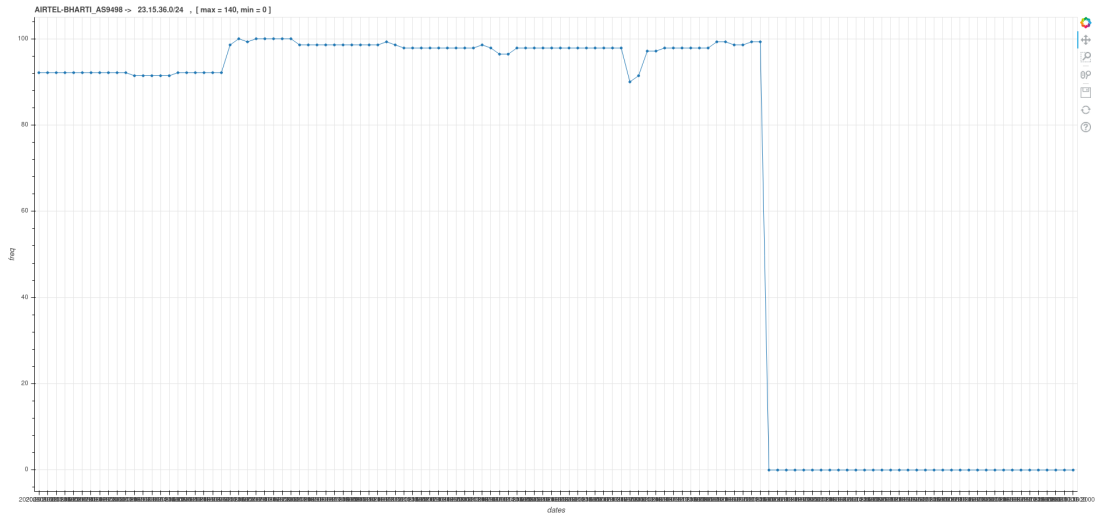

Figure 4.4: Bharti Airtel AS9498: X-axis represents the dates in Dec 2020, while the Y-axis represents the number of unique paths to this particular prefix

11th Dec 2019. Further, we see a sharp dip during which the freq turns to 0. Later it rebounds but remains a little less than it was initially. Many graphs showed the same pattern on different days, regardless of the ISP.

E.g., the following graph from JIO shown in Figure 4.6 gives an excellent correlation. The routes to this particular prefix were nullified for over two days from 18th to 20th Dec.

Many false positives graphs were also obtained like in Figure 4.7. This BSNL graph shows many dips, which appears out of sync with the protests, somewhat too random in nature. The reason for the same is being worked upon.

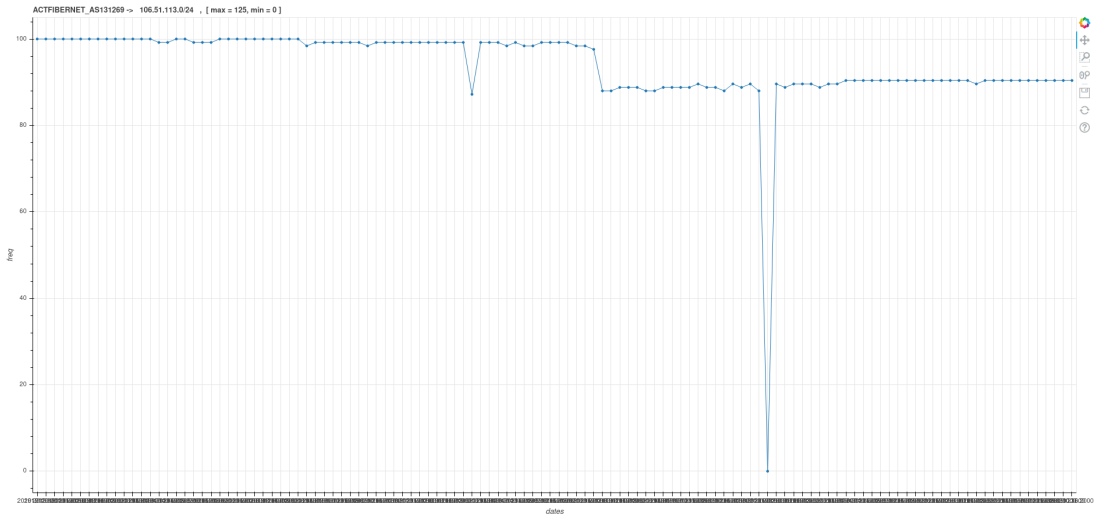

Figure 4.5: ACTFIBRENET AS131269: X-axis represents the dates in Dec 2020, while the Y-axis represents the number of unique paths to this particular prefix

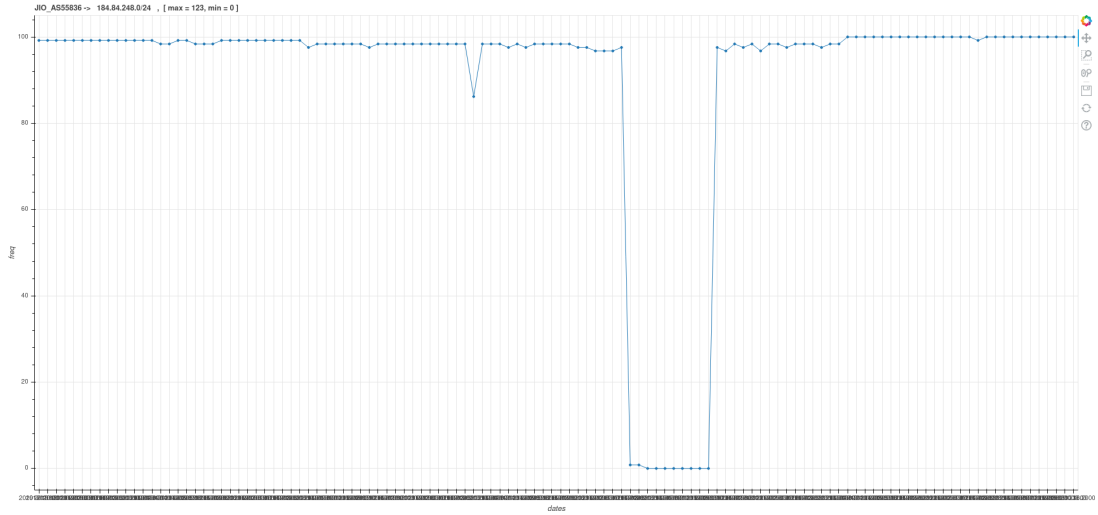

Figure 4.6: JIO AS55836: X-axis represents the dates in Dec 2020, while the Y-axis represents the number of unique paths to this particular prefix

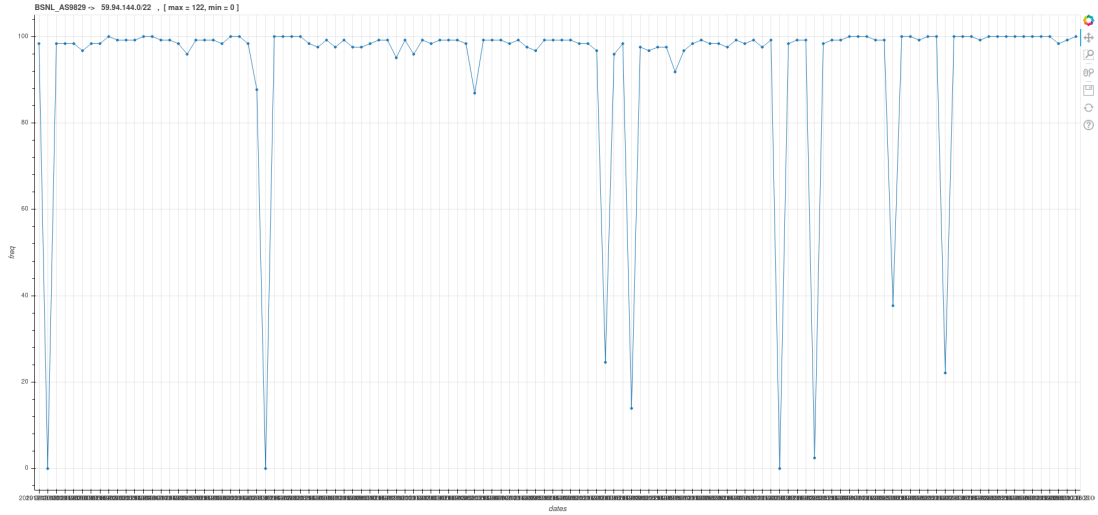

Figure 4.7: BSNL AS9829: X-axis represents the dates in Dec 2020, while the Y-axis represents the number of unique paths to this particular prefix

Chapter 5

Conclusion

Our preliminary investigations indicated that there might exist some relationship between BGP prefix advertisements and internet shutdowns, *i.e.* it is expected that the number of unique BGP paths are decreased for a particular prefix that is advertising its routes in the region of internet shutdowns. However, there are many factors that directly or indirectly affect our results and need to be considered.

Many graphs gave us a strong correlation for our first hypothesis, whereas others seem exactly the opposite. To break the deadlock, we propose several tests to be performed, as mentioned in Chapter 6, to clarify all the above doubts.

Chapter 6

Limitations and future work

6.1 Limitations

There are some limitations that we encountered while working on this project:

- Trusting BGP data is a leap of faith, as BGP itself is a very complex algorithm in which financial relationships have the upper hand against cost and efficiency.
- Since the number of Routeviews probes is limited, it is expected that not all data can be captured.
- If there is some change in a prefix, then according to BGP, it can take hours to get it registered in the global routing tables. Thus a mismatch in time and information can be an expected thing.
- There were many shutdowns/outages which went unreported and thus, it would be difficult to decide whether some dips in the graphs belong to unreported shutdowns or, in actuality, it is a false positive

6.2 Future work

In order to remove all the ambiguity about the results, we plan the following inspections:

- We plan to correlate this data with other months as well, when there was no protest, to prove these graphs' uniqueness.
- We also plan to geolocate the prefixes, which showed a significant dip in the graphs and then verify it with the news.
- In order to increase confidence in our hypothesis, we need to do more of such case studies but this time focussing more on the regional ISPs
- We can correlate for other countries as well, such as Iran 2019 internet blackout, or a more recent one, like in Belarus
- Lastly, we need to correlate the Routeviews data with other publicly available projects, which we mentioned at the beginning, like Censys, CAIDA ARK, IODA and RIPE Atlas.

Bibliography

[1] 2019 internet blackout in iran. https://en.wikipedia.org/wiki/2019_Internet_blackout_in_Iran.

[2] 2019 internet shutdown in sudan. https://globalvoices.org/2020/06/08/internet-shutdowns-in-sudan-the-story-behind-the-numbers-and-statistics/.

[3] Caida ark project. https://www.caida.org/projects/ark/.

[4] Caida ioda project. https://ioda.caida.org/.

[5] Censorship in belarus. https://en.wikipedia.org/wiki/Censorship_in_Belarus.

[6] Censys project. https://www.censys.io/.

[7] Cidr reports. https://www.cidr-report.org/.

[8] Citizenship amendment act protests. https://en.wikipedia.org/wiki/Citizenship_Amendment_Act_protests.

[9] Economic loss due to internet shutdown in india 2019. https://timesofindia.indiatimes.com/business/india-business/india-lost-1-3bn-due-to-4196-hours-of-no-internet/articleshow/73179287.cms.

[10] Internet assigned numbers authority. https://www.iana.org/.

[11] Internet shutdown in egypt 2011. https://en.wikipedia.org/wiki/Internet_in_Egypt#2011_Internet_shutdown.

[12] Internet shutdown in india 2012. https://www.outlookindia.com/website/story/india-news-data-show-internet-shutdown-in-jk-between-2012-17-cost-indian-economy-304-bn/335965.

[13] Internet shutdown in libya 2011. https://en.wikipedia.org/wiki/Internet_censorship_in_the_Arab_Spring#Libya.

[14] Internet shutdown rajasthan, july 2018. http://www.deccanchronicle.com/nation/current-affairs/120718/internet-to-be-shut-during-constable-recruitment-examinations-in-rajas.html.

[15] Ipinfo. https://www.ipinfo.io/.

[16] Maxmind. https://www.maxmind.com/en/home.

[17] regional internet registries. https://www.iana.org/numbers/allocations/.

[18] Ripe atlas project. https://atlas.ripe.net/.

[19] University of oregon route views archive project. http://archive.routeviews.org/.

[20] Dainotti, A., Squarcella, C., Aben, E., Claffy, K. C., Chiesa, M., Russo, M., and Pescapé, A. Analysis of country-wide internet outages caused by censorship. In *Proceedings of the 2011 ACM SIGCOMM Conference on Internet Measurement Conference* (New York, NY, USA, 2011), IMC '11, Association for Computing Machinery, p. 1–18.

[21] Eriksson, B., Durairajan, R., and Barford, P. Riskroute: A framework for mitigating network outage threats. In *Proceedings of the Ninth ACM Conference on Emerging Networking Experiments and Technologies* (New York, NY, USA, 2013), CoNEXT '13, Association for Computing Machinery, p. 405–416.

[22] Giotsas, V., Dietzel, C., Smaragdakis, G., Feldmann, A., Berger, A., and Aben, E. Detecting peering infrastructure outages in the wild. In *Proceedings of the Conference of the ACM Special Interest Group on Data Communication* (New York, NY, USA, 2017), SIGCOMM '17, Association for Computing Machinery, p. 446–459.

[23] Lee, S., Im, S., Shin, S., Roh, B., and Lee, C. Implementation and vulnerability test of stealth port scanning attacks using zmap of censys engine. In *2016 International Conference on Information and Communication Technology Convergence (ICTC)* (2016), pp. 681–683.

[24] Quan, L., Heidemann, J., and Pradkin, Y. Trinocular: Understanding internet reliability through adaptive probing. In *Proceedings of the ACM SIGCOMM 2013 Conference on SIGCOMM* (New York, NY, USA, 2013), SIGCOMM '13, Association for Computing Machinery, p. 255–266.