BTP Presentation

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

Ritik Malik
2018406

●●●

BTP Track : Research

BTP Advisors
Dr. Sambuddho Chakravarty
Dr. Aasim Khan

Internet Shutdown

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.

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.

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, Routeview and RIPE atlas, 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.

Motivation

the overall mechanism of implementation of internet shutdown is currently unknown,

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.

here 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

Problem statement

– 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?

Definitions :

- **Autonomous System :** A collection of hundreds or thousands of routers under some network operator, these AS are assigned a number by the Internet Authority called ASN

- **Border Gateway Protocol :** Border Gateway Protocol (BGP) is a gateway protocol designed to exchange routing and reachability information among autonomous systems (AS) on the internet. In simple words, it tells us the path which we should follow, to reach from one router to another

  BGP used for routing within an autonomous system is called Interior Border Gateway Protocol, Internal BGP (iBGP).

  the Internet application of the protocol is called Exterior Border Gateway Protocol, External BGP (eBGP).

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

Experimental Setup and Results

Talking of public datasets, we have various sources -

CAIDA ARK project, IODA, Censys, RIPE Atlas, Routeviews. Our primary focus, for now, is on the publicly available dataset from the **”University of Oregon Route Views Archive Project”,**

Other datasets are being worked upon and will be included in the next thesis for comparison and cross-validation.

Why and about Routeviews :

- It captures Historical BGP dumps about the global routing system from its various nodes across the world
- The Routeviews project has 31 collectors now, spread across the world but mostly concentrated in the US
- 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.

Experimental Setup and Results

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

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 and CIDR reports, and the best effort geolocation was from the Maxmind API.
So we got our final “prefix-to-geolocation”database.
Now we needed some case studies to verify our hypothesis.

Case Study 1 : Rajasthan

Initially, we decided to correlate the dumps with the internet shutdown in Rajasthan in July 2018.

Following are the insights of that shut down -

- The shutdown was imposed on two consecutive days - 14th and 15th July 2018
- The reason for the shutdown was to prevent cheating in a constable recruitment exam
- 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 -

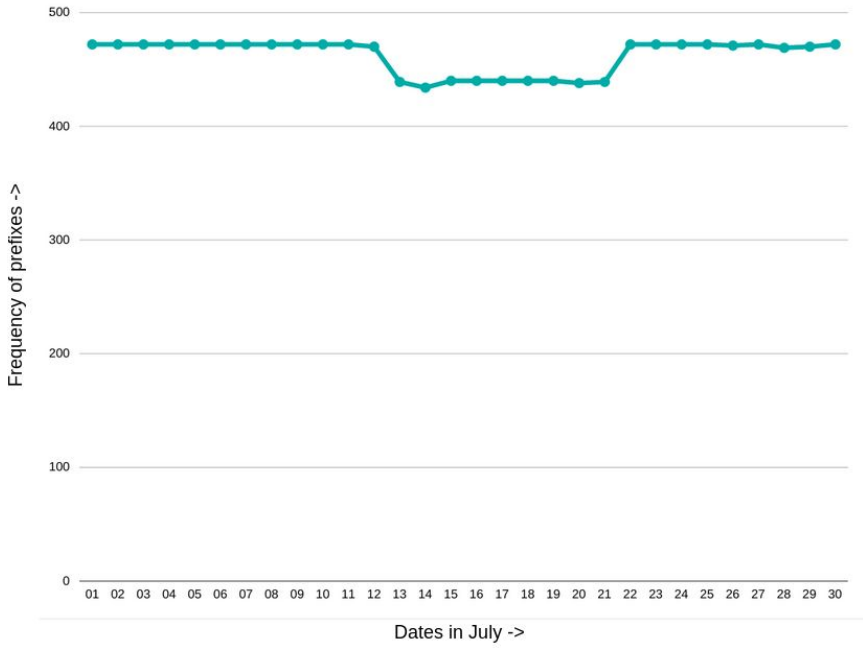
- A web scraper to download data from all Routeviews nodes for the full month
- Extract the dumps and perform preprocessing and formatting
- Search specific prefixes in the whole database and dump the raw results in CSVs
- Use the CSVs to make final graphs.

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 ->

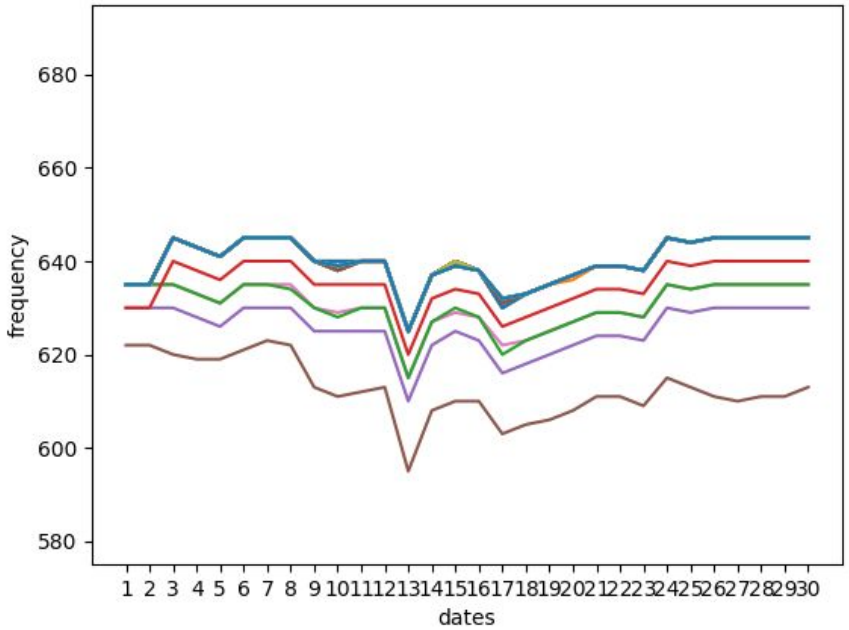


To verify that this pattern was the repeated across all ISP, we tested it on various ISPs having their prefixes in Rajasthan, which led us to this graph as shown in Figure.

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. Itis because we combined the results of 5 timestamps for a single day.



Case Study 2 : NRC - CAA Act

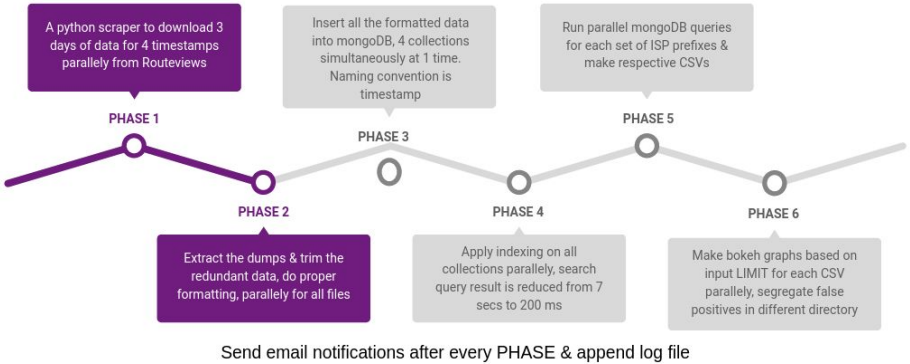
This time, we decided to move forward with our hypothesis on a little broader scale. A morerecent and infamous event was chosen -The NRC CAA protest(Dec 2019), which is known to curb the internet in large parts of India. This time, since the testing was on a broader scale, the old prototype pipeline 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 parallely 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

The 6 phases of the new pipeline can be visualized as shown -

The graph layout was also improved significantly this time:

1. We added an 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



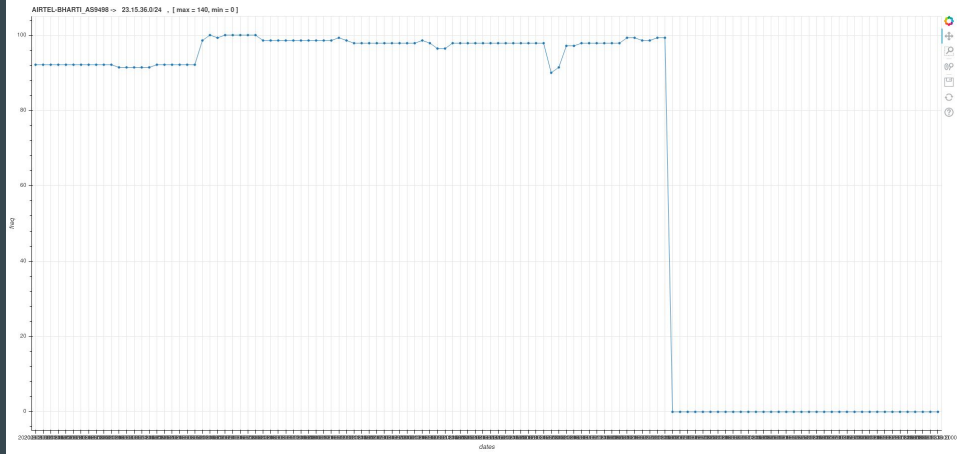
The graph shown in Figure belongs to the 4(a) category.

The 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.



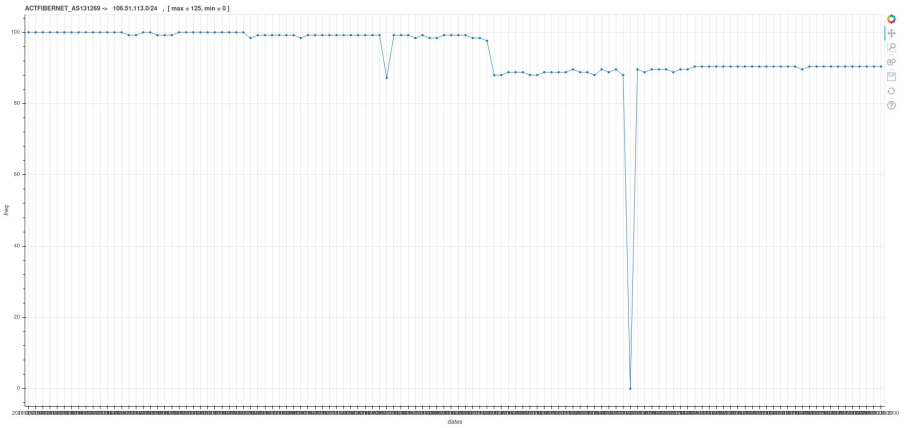
Coming to the main graphs of interest, the 2nd category, 4(b) as shown in top Figure -

ACTFIBRENET_AS131269 ->106.51.113.0/24 , [max = 125 , min = 0].

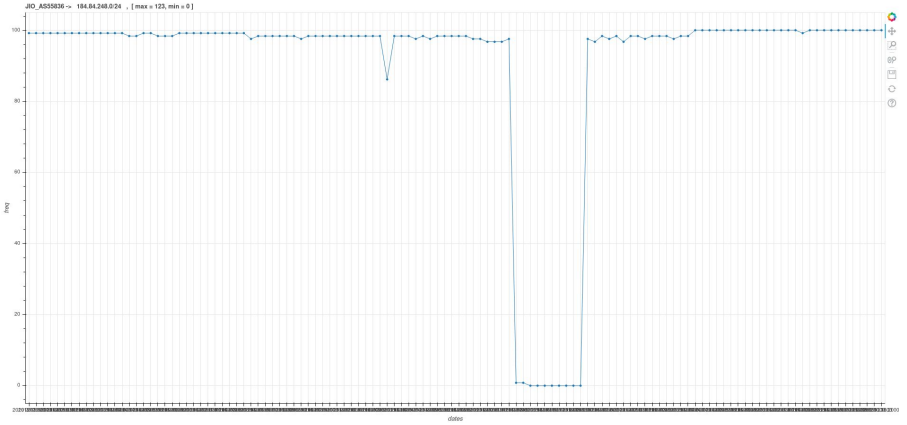
The freq shows a hiccup at 1600, 13.12.2019, which coincides with the day after the NRC CAA bill was passed on 11th Dec 2019. Further, we see a sharp dip during which the freq turns to 0. Later it rebounds but remains a little less than it was initially.

Many graphs showed the same pattern on different days, regardless of the ISP.

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



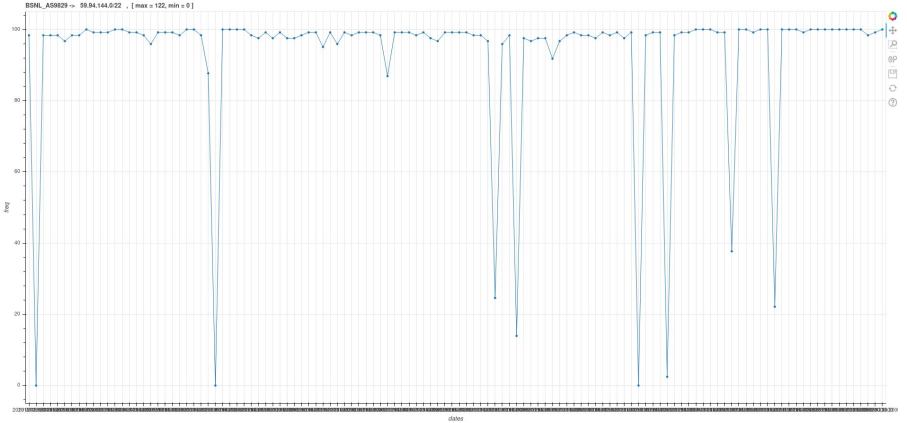
ACTFIBRENET_AS131269



JIO_AS55836

Many false positives graphs were also obtained like in this Figure.

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.



BSNL_AS9829

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. Those will be discussed in limitations in a while.

Further, we will propose several tests to be performed

Limitations and future work

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 expected
- 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

Limitations and future work

Future Work

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.

Thank You