Real time distributed simulation of BGP on the entire current Internet topology and Internet-like topologies

Cyril Trap, *Supervised by Christian Doerr*
TU Delft
Cyber Security Group
2628 CD Delft, The Netherlands
C.H.Trap@student.tudelft.nl

***Abstract*—The Border Gateway Protocol is critical for the correct working of the Internet. When it fails the impact is usually high and therefore failures should be minimized. Unfortunately the configuration of BGP is prone to errors. Besides that, BGP is targeted by attacks of cyber criminals. A simulator capable of running BGP can reduce the number of honest mistakes and successful attacks. In a simulated environment different configurations and optional attacks can be tried out safely. The simulator can also assist with the investigation of real world events by replicating the conditions and triggers that led up to it and providing a detailed view on every aspect. The correct workings of the simulator was verified with real world events testing the simulator from both a macro and micro perspective. The CPU and memory usage during simulation are discussed. Additionally, a quick recap on the working of BGP is provided.**

***Index Terms*—BGP, real time simulation, Internet, BGP-hijack, distributed simulation**

I. Introduction

THE Border Gateway Protocol (BGP) is well known for not being known among Computer Scientists. It operates between Internet Service Providers (ISP) to exchange information about how to reach IP addresses all over the world. Usually this goes well but every now and then parts of the world lose their connection to certain IP ranges. When BGP fails it impacts more than a few users, it has the potential the spread malfunctions over the entire globe. Some of these malfunctions are the result of a honest human error, but hackers managed to exploit the weaknesses of the Border Gateway Protocol as well [1].

These weaknesses are well known in the Internet admin community and it was just a matter of time before they would be exploited. Unfortunately BGP is so well established within the operation of the Internet that is extremely hard to replace although different proposals have been offered by different researchers, as will be discussed in the section Related Work. The incentive to change to these new protocols lacks among network administrators and networking equipment vendors. Since the advantages of these protocols can only be harvested as all network administrators and vendors of networking equipment collectively decided to change to the new protocols. A protocol, after all, needs to be able to run on both sides of a connection in order for it to work properly. If only a few parties decide that they would like to run these new protocols, it is not profitable for the networking equipment vendors to change their products. And on the other side network administrators cannot use the new protocols when their equipment does not support them.

When the actual protocols in use cannot be changed, a simulator could be used to still improve the security of the operation of the Internet. When a network administrator needs to make adjustments to the configuration of the BGP instances ran on the network routers he or she might accidentally make a mistake. If the changes are made directly on the physical network in use these mistakes have a direct impact on the Internet stability and parts of the network might lose its connection to certain IP addresses. With a simulator available the modifications could be first tested for errors within a simulation and the resulting network stability can be observed without impacting the real Internet when it went wrong.

Even more important than preventing network administrators from making honest mistakes a simulator can also be used as a tool to test defences in case of attacks or failures. An existing BGP configuration of a network can be uploaded into the simulator. During the simulation it can be attacked to see if filter rules work as intended and no false routes are inserted into the routing tables used within the network. Also network reachability can be tested from any other viewing points. In theory even counterattacks can be tried out safely within a simulator.

Additionally, a simulator can be used to provide detailed explanations of real world events. After one has gathered information about a certain event of interest the appropriate routers can be configured to trigger the event. Every single BGP instance on the routers of the entire topology can be consulted about routes, filter rules and updates. This provides insights into the chain of events which could explain exactly how things went wrong. The BGP configurations on the routers can be viewed and adjusted during the simulation. This can be useful for example during simulation of BGP hijacks in which certain IP ranges are announced by multiple parties. Usually after attackers have broadcasted a fake announcement, the original and legitimate owner of the IP

C.H. Trap is a Computer Science student at the faculty of Electrical Engineering, Mathematics and Computer Science at the Delft University of Technology in Delft, The Netherlands. Email: C.H.Trap@student.tudelft.nl

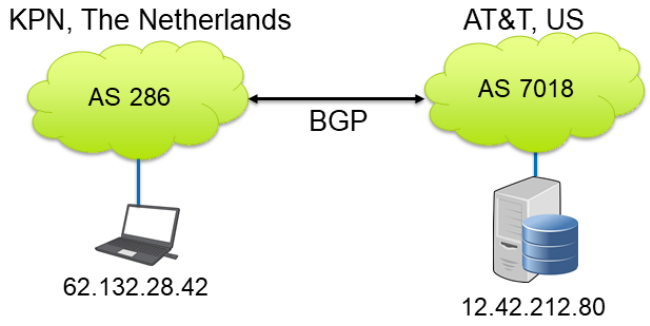

Fig. 1. Where BGP operates within the Internet during connections.

ranges responds with more specific announcements of these IP ranges. Each individual router or entire sections of the simulated network can be disabled to simulate down time of these routers. In the case it is not clear what exactly triggered a specific event different hypotheses can be tested by the simulator to see which lead up to the situation observed in the real world.

This paper will provide a overview of the related work. A quick recap on what the BGP protocol is for those unfamiliar with BGP and those who just need a little refreshment. A reasoning for the need of a simulator is provided. To verify the correct workings of the simulator real world events were re-enacted. A description of the events along with the result of their simulation are given. A comparison between the results and the simulation was made along with a discussion of the differences. To build the current topology of the actual Internet a dataset was constructed. The process of this and a perspective of the scale of the current Internet are provided. The CPU and memory usage of the simulator are visited before ending with a discussion.

II. Related work

To the best of our knowledge this is the first simulation which is able to simulate the actual entire Internet in real time. Related work relied on a partial network or a custom created small network. Correctly configuring BGP entities is hard to do in one go. Therefore researchers suggested a more abstract language to describe the desired properties of a network in an attempt to prevent network administrators from making mistakes. This abstract description will be compiled into BGP configurations for individual routes [2][3][4]. Although a more abstract language might help to make less mistakes it provides no guarantees that no mistakes can be made. The risk of lost connectivity due to erroneous configurations still exists when they are applied to real networks, impacting actual users. The simulator described in this paper can be used to test and verify those configurations without the hazard of shutting down parts of the actual Internet.

It is known that BGP is quite vulnerable to attacks [5] and mistakes. Researchers have published different attacks which exploit various angles [2]. Real life records of incidents prove that these weaknesses are not purely theoretical and are exploited in practice as well [1]. This simulator can be used to re-enact these events and to investigate them in depth.

The small simulations relying on a partial network or a custom created network lack the ability to show or calculate the potential global impact of simulated events [2]. This simulator provides to ability to show how many ASNs, Autonomous System Networks, will be impacted or even compromised by attacks or other events. Therefore the impact on the global Internet stability can be deduced. According to a study performed in 2007 [4] all Internet traffic from more than 50% of all ASNs can be both hijacked and intercepted by a single ASN. The Internet has grown rapidly and has become even more interconnected ever since, so this percentage will be even higher nowadays.

III. What is BGP?

This section provides a quick recap on the working of the Border Gateway Protocol. Readers already familiar with the protocol can skip this section. For a graphical presentation, see figure 1.

When somebody, say Alice in The Netherlands, wants to connect to a server, say Bob's server in the US, BGP is used to look up a high level route. Alice sends to her ISP that she wants to visit Bobs server at 12.42.212.80. Her ISP, say KPN, looks up 12.42.212.80 in its routing tables. This lookup returns a list of ASNs, Autonomous System Networks (networks under a single authority), through which the ISP of Bob can be reached. This enables Alice's traffic to reach Bob's server. When the server sends a response the same mechanism is used only with Alice's IP address.

How are these routing tables filled? BGP can be seen as a sort of interstate highway route map of the internet. It gives you a series of highway numbers (ASNs) to reach to rough location of your destination (the ASN of the receiver) where the local signposting (an internal gateway protocol) takes over to route the traffic precisely to its final destination.

A BGP instance is configured to announce blocks of IP addresses of the computers in its own network or its customers. The notation used for this is called CIDR, Classless Inter-Domain Routing, notation and works as follows: The blocks are denoted by an IP address followed by an forward slash and a number, the number represents the amount of common bits when written in binary form with the given IP address, starting from the left, of the IP addresses of the block, so the larger the trailing number, the smaller the block and vice versa. For example the block 192.168.22.0/24 consists of 256 address of which the first 24 bits are equal to 1100 0000 1010 1000 0001 0110. Or in other words, the block of addresses in the range starting from 192.168.22.0 till 192.168.22.255. These blocks are shared with its neighbors, peers. The neighbors now know that those blocks can be reached via the BGP instance. The neighbors share that information with their own neighbors and so on. This way a chain of neighbors is created called a route, represented as a list of ASNs of the corresponding neighbors.

Each BGP instance has a table with IP blocks and the corresponding routes, called the routing table. As the Internet grows rapidly, more devices are connect to it and more connections between devices are formed. If the growth was limited to just more devices it would not be such a problem since with CIDR notation one can simply use a lower number behind the slash and larger blocks of IP addresses are handled using the same resources as before. Growth in the number of connections between devices, however, means

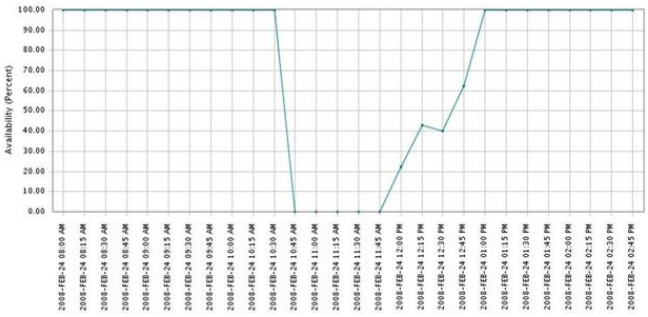

Fig. 2. The availability of YouTube.com during the Pakistan-YouTube 2008 incident as monitored by Keynote Systems, retrieved via cnet.com [6].

that destinations can be reached via more different routes. Which means that when the Internet grows rapidly and thus becomes more interconnected, the routing table grows along rapidly. This is becoming an issues since routers need more and more memory to store these routes.

Multiple routes to a given destination may be present in the routing table. BGP has a set of rules specifying which route to take in such a situation. One of those rules is that the most precise route is chosen, thus the smallest block containing that IP. This rule enables the so-called BGP hijacking, which will be demonstrated in the simulator, but other attacks are also possible. BGP hijacking is the practice of announcing a false route, usually via the criminal, for a more specific block than the original block containing it. This way a portion of the traffic is routed via the criminal, which has a chance to inspect it, remove it and even to modify it, before sending it along to its original destination.

Another rule to decide which route to take in case of multiple equally specific routes is that shorter paths are preferred over longer paths. This influences how large the impact of announcements is when multiple parties are announcing the same equally specific routes.

As a tiebreaker when other rules fail to decide which route to prefer, BGP selects the older already known route instead of the newer one.

Since BGP operates between different ASNs and thus different authorities, it has the unique/interesting aspect (compared to other routing protocols, which usually operate within a single authority) that you cannot always trust the person on the other side. And even if you do trust that person, you have no control on the other side of the connection, the other person might make mistakes.

IV. Pakistan-YouTube 2008

On the 24th of February 2008 YouTube was unreachable for over 2 hours. This started with a state order [7] given by the Pakistan government that YouTube should be blocked within the country. The ISPs reacted by announcing a route to YouTube that was more specific than the routes announced by YouTube itself. Since one of the decision rules of BGP is to select the most specific route when multiple routes to the same destination are present in its routing table, this new more specific route announced by the Pakistan ISPs was chosen. All the traffic intended for YouTube flows through the Pakistan ISPs and they throw it into a blackhole, effectively blocking all communication to YouTube in the country, just as ordered by the state.

Up until this point everything looked normal for the rest of the world.

One of the ISPs, Pakistan Telecom (AS17557), also announced the route to its upstream provider, PCCW (AS3491). PCCW took over this new route and announced it in its turn to its neighbors. After just 15 seconds, 9 ASNs had already taken over this new route. By not long most of the major ISPs worldwide (45 seconds: 47 ASNs. 75 seconds: 93 ASNs. These numbers might not seem as much, but since these are major ISPs they serve together, estimated, over two-thirds of the Internet [8]) had taken over this new route, which ended up in their routing tables as a more specific route to YouTube and thus the preferred route. As a result all traffic to YouTube was send along the new route to Pakistan. This rendered YouTube unreachable for a large part of the world as can be seen in figure 2, the availability of YouTube.com dropped to 0% at 10:45 AM, 24 February 2008 (UTC -8). So this ‘small’ mistake had large consequences for both Pakistan, which had to suddenly handle the world’s YouTube traffic, and the rest of the world, which could not watch its favorite videos on YouTube for over 2 hours.

Halfway, one hour and 20 minutes after the start of the attack, YouTube (AS36561) tried to defend by announcing the hijacked route themselves. This worked partially since about half of the infected ASNs preferred the new legitimate YouTube route 65 seconds after the announcement. Just over 10 minutes later YouTube added 2 additional /25 routes, to make sure all traffic would flow to their servers again, since these routes are more specific than the hijacked prefix.

The attack ended with PCCW (AS3491) withdrawing the hijacked prefix and disconnecting Pakistan Telecom (AS17557). These series of events are derived from dyn.com [8].

V. Propagation of announcements

The 15, 45, 75 seconds times are not surprisingly fast. The default timers of BGP are set to 30 seconds between sending updates to peers. So each 30 seconds routes are propagated to all neighbors. The maximum propagation time can be calculated from the topology of the Internet, which is constructed using the dataset. For each ASN in the dataset a set consisting of the ASN is created. Each time all the peers of the ASNs in the set are added to the set, a counter is incremented until the set consists of all ASNs. The maximum of all counters, 14, multiplied by 30 seconds gives a maximum propagation time of 420 seconds. The average propagation time however will be 15 seconds per router times 9.66 hops equals 144.9 seconds.

One could ask why the Pakistan-YouTube 2008 incident lasted for over 2 hours with such rapid propagation times. This has to do with the fact that ISPs still rely on manual intervention mostly in case of attacks or failures [6]. So only once a human decided to change the BGP configuration of a router the propagation times come into play.

VI. Belarussian traffic diversion

In 2013 Man-In-The-Middle BGP hijacks became a real threat to Internet security with 1500 individual IP blocks

Fig. 3. The normal path and diverted path of Internet traffic during the Belarussian traffic diversion, retrieved via dyn.com [1].

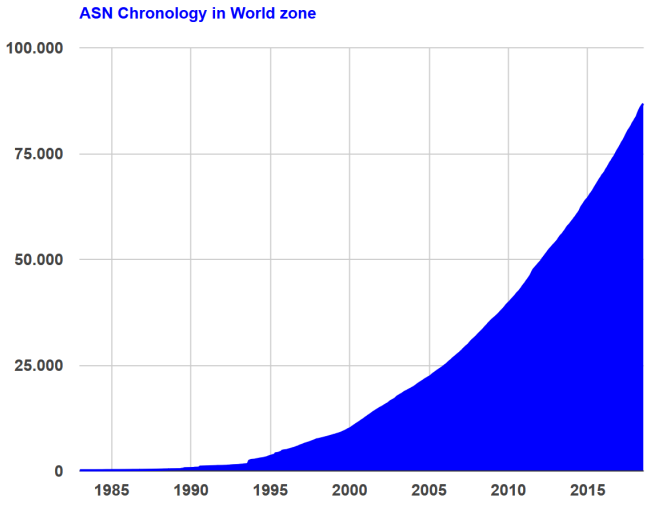

Fig. 4. The number of ASNs is growing faster and faster, retrieved via imtbs-tps.eu [12], modified colors for readability.

hijacked in over 60 incidents [1]. This more sophisticated attack is based on the unintended Pakistan-YouTube 2008 incident, as described in section IV. The victims Internet traffic is diverted via the hacker and then redelivered to the intended recipient, the victim. During this attack a hacker has the change to inspect, remove or even modify the victims Internet traffic. The victim has no way of verifying whether its Internet traffic has been tampered with.

The main difference between the Man-In-The-Middle BGP attack and a simple route hijack, such as the Pakistan-YouTube 2008 incident, is the fact that during the former Internet traffic keeps reaching its intended destination and therefore from the victims perspective everything looks alright. The hacker makes sure that there is at least one outbound path unmodified to deliver the Internet traffic back to the victim. So, after the hacker has inspected or modified the traffic, it is send out via the unmodified outbound path to be delivered to the victim. When the hacker resides on a geographical position somewhere along the natural path from the original source to the victim, they should not notice the increased latency resulting from the interception.

In February 2013, dyn.com [1] daily observed Internet traffic being redirected to a Belarussian ISP GlobalOneBel from various origins around the world. These events lasted for a few minutes to multiple hours and affected major financial institutions, governments and network service providers among others. One recorded traceroute shows traffic from Guadalajara, Mexico destined for Washington DC, US taking a deviation via Minsk, Belarus before actually going to Washington DC, US. It is clear that this Internet traffic should have never crossed an ocean under normal circumstances. See figure 3.

Even when the probably unsuspecting victim decides to verify its connectivity with the Internet by checking his own traceroutes, nothing odd will show up. Only incoming traffic is routed via Minsk, Belarus. While outgoing traffic, and thus the victims own traceroutes, will still follow normal routes.

VII. Dataset

The dataset, used to test the simulator, was constructed by combining information from the following sources: traceroutes from the CAIDA ARK dataset [9], the AS Rank API [10], the BGPview API [11]. All information from the AS Rank API was download. Additionally, a list of ASNs was constructed by downloading all pages from the AS Rank API, this list was used to query the BGPview API. Based on this dataset a topology was build to be ran on the simulator. Mainly the peering and prefix data where used. The raw dataset was too large to be ran on the simulator, so it was filtered. ASNs with no peers were removed. ASNs with exactly one peer, so called stubs, were aggregated into their parent by moving their announced IP blocks to their parent and removing the stub itself from the topology. This aggregation can be done without loss of accuracy since all Internet traffic that was going in and out of the stub had to go through their parent, so from a routing perspective the routes to the parents and the stub are the same routes, except for the extra appended stub. Connections that according to the raw dataset where only going in one direction where removed as well. The filtering resulted in a dataset with 5624 ASNs, which could be used for the simulation.

VIII. Scale

With the Internet continuing to grow faster the number of ASNs is growing along. Currently, as of the first of July 2018, the Internet consists of 86800 ASNs which together announce 3 663 805 752 IPv4 addresses and 15 526 265 935 IPv6 addresses (in /48 blocks). A solution capable of handling these numbers is required for the simulator. And besides the capabilities for the current size of the Internet, with the rapid growth in mind, the solution for the simulator should also be scalable.

IX. Construction of the simulator

The construction of the simulator started off with Mininet [13] to provide connectivity among ASNs and Quagga [14] to provide a BGP implementation. From the Quagga network stack both the BGP daemon and zebra were configured. The zebra daemon was required to provide the interfaces for the TCP connections over which the BGP protocol communicates. The configurations for each router for both daemons were generated dynamically by a script to provide the correct configuration for given topologies.

This setup however was not capable of scaling far enough. In an attempt to solve this, the simulator switched to Mininet Extreme [15], a version of Mininet especially created to be able to run large networks. Unfortunately, it was unable to

run the topology created from the dataset. Since Mininet Extreme was based on a fairly old version of Mininet, 2.1.0, we decided to merge it with the latest version of Mininet. This did not work either, so at last, a custom version of Mininet was written to be used in the simulator. Quagga was modified to fit with this custom Mininet version, but only the launcher was adjusted so it is still compatible.

Instead of relying on the zebra daemon from Quagga to provide the TCP connection which took up much of the memory, Linux bridges were used to connect all BGP daemons via virtual Ethernet interfaces. The memory footprint was reduced further by enabling hosts dynamically and thus eliminating the hosts that did not play any role in the simulation at hand. At this point the simulation fitted in memory. However, the CPU could not cope with the load of all the interrupts raised by the both the daemons and the network interfaces used for the simulation. To solve this issue, a distributed version was made which could be run on multiple servers. The total workload was split up evenly accross the different machines by an algorithm, which will be elaborated upon in section XV. On every server the script injected static routes to the simulated routers on the other machines, so every simulated router could communicate with all other routes. By default BGP sessions are formed between direct neighbors, to enable BGP to make a connection across another interface the next-hop count must be increased. This has to be set via the BGP configuration and is therefor done during the dynamic generation of the configurations by the script. This was required for the distributed version to work across the network connection the multiple servers.

All prefixes to be announced by the BGP routers can be generated as well using the dataset. However, when doing so one does quickly run into memory issues during a simulation since all those prefixes are stored by every single BGP router. Nevertheless, only those prefixes which are relevant to the simulation at hand can be announced. This enables one to still see what happens with the relevant prefixes while the simulation is able to fit in the available memory. Or since the distributed version is run across multiple machines, one can divide the work across even more machines until the simulation with all its prefixes fits in the combined memory. The script contains an option to toggle the generation of the prefixes for all, none or only certain routers.

During the generation of the configuration files for the BGP instances a graph of the given topology is generated to verify the correctness of the topology. This graph is searchable and zoomable to aid in the verification process and for easy reference later on during the simulation. As an example of how such graph looks like, the resulting graph from the Internet topology constructed from the dataset has been included as appendix A.

The simulator comes with a script which can easily generate predefined topologies which allows for quick testing and little setup time. These predefined topologies can be Internet-like or something entirely else, for example abstract shapes, such as a pyramid, straight line or one router connecting multiple straight lines. One can also easily define its own custom topologies by simple listing the connections, thus the edges of the graph. The script can also directly read topologies from simple DOT files, the graph description language [16].

X. Simulation of the Pakistan-YouTube 2008 incident

To verify the correct workings of the simulator from a macro perspective a real world event was re-enacted. The Pakistan-YouTube 2008 incident was chosen since a detailed description of the series of events is available at dyn.com [8].

Both the YouTube router and the Pakistan Telecom were hosting a standard HTTP server with IP 208.65.153.2, serving a static unique oneline page for identification, so cURL [17] could be used to see the result of the address look ups and were exactly traffic is going from the different viewpoints. The IP address 208.65.153.2 was chosen since this is part of all announced blocks during the series of events, so the influence of all actions from the ASNs can be observed. Any other IP address that is part of all IP blocks used during the series of events could have been chosen for the simulation and similar results would follow.

All 4 ASNs on the preferred route from YouTube (AS36040) to Pakistan Telecom (AS17557) were chosen as viewpoints to observe the simulation. These 4 ASNs together provide a nice overview since both endpoints are present and one ASN closer to YouTube is included as well as one ASN closer to Pakistan Telecom. The time line of this simulation can be found in table I.

At Sat Jul 14 03:41:26 CEST 2018 the fake blackhole site was launched. Two minutes later, 03:43:01, the route 208.65.153.0/24 was announced by AS17557. Two minutes later, 03:44:56, the hijacked route 208.65.153.0/24 is also announced by AS36040, as a defence. After 15 seconds the result of this defensive act becomes clear. The defence only works for certain ASNs, only for those in the closer half of the route. Since for routes which are equally specific, shorter routes are preferred. At 03:47:19, due to a manual reboot of AS36040 to load the new configuration, AS24785 falls back shortly to the longer route announced by AS174. Shortly after the reboot of AS36040 is completed with two extra routes, 208.65.153.0/25 and 208.65.153.128/25, AS24785 recovers by preferring the routes to the real YouTube at 03:47:28. Since these new /25 routes are more specific than all other routes so far and BGP prefers the most specific routes, at 03:47:32 things go back to normal.

Since 4 viewpoints do not provide a complete picture of the reachability of YouTube worldwide, the simulation was ran again with more viewpoints. 242 ASNs were randomly selected which all retrieved the website hosted at 208.65.153.2 every second. As can be seen in figure 5 as soon as Pakistan Telecom (AS17557) announces their fake route ASNs start to see the fake blackhole site. After exactly 40 seconds the availability of YouTube.com on the ASNs is decreased to zero. This situation lasts for 1 hour and 20 minutes. YouTube (AS36040) defends itself by announcing the hijack route themselves. This is not effective at all since only 4 out of the in total 242 randomly selected ASNs prefer the /24 route to the genuine YouTube. However, the two additional /25 routes announced by YouTube, 10 minutes later, are effective. Within 72 seconds all of the 242 ASNs can reach the genuine YouTube again.

When one investigates more closely how effective announcing this /24 route from the genuine YouTube precisely was by checking for all 5624 ASNs in the simulation, one

TABLE I
REENACTION OF THE PAKISTAN-YOUTUBE 2008 INCIDENT ON THE SIMULATOR

Timestamp	AS36040	AS24785	AS174	AS17557
Sat Jul 14 03:41:26 CEST 2018	YouTube	YouTube	YouTube	fake blackhole!
⋮	YouTube	YouTube	YouTube	fake blackhole!
Sat Jul 14 03:43:00 CEST 2018	YouTube	YouTube	YouTube	fake blackhole!
Sat Jul 14 03:43:01 CEST 2018	YouTube	YouTube	fake blackhole!	fake blackhole!
⋮	YouTube	YouTube	fake blackhole!	fake blackhole!
Sat Jul 14 03:43:16 CEST 2018	YouTube	YouTube	fake blackhole!	fake blackhole!
Sat Jul 14 03:43:17 CEST 2018	YouTube	fake blackhole!	fake blackhole!	fake blackhole!
Sat Jul 14 03:43:18 CEST 2018	fake blackhole!	fake blackhole!	fake blackhole!	fake blackhole!
⋮	fake blackhole!	fake blackhole!	fake blackhole!	fake blackhole!
Sat Jul 14 03:44:55 CEST 2018	fake blackhole!	fake blackhole!	fake blackhole!	fake blackhole!
Sat Jul 14 03:44:56 CEST 2018	YouTube	fake blackhole!	fake blackhole!	fake blackhole!
⋮	YouTube	fake blackhole!	fake blackhole!	fake blackhole!
Sat Jul 14 03:45:11 CEST 2018	YouTube	fake blackhole!	fake blackhole!	fake blackhole!
Sat Jul 14 03:45:12 CEST 2018	YouTube	YouTube	fake blackhole!	fake blackhole!
⋮	YouTube	YouTube	fake blackhole!	fake blackhole!
Sat Jul 14 03:47:17 CEST 2018	YouTube	YouTube	fake blackhole!	fake blackhole!
Sat Jul 14 03:47:18 CEST 2018	YouTube	fake blackhole!	fake blackhole!	fake blackhole!
⋮	YouTube	fake blackhole!	fake blackhole!	fake blackhole!
Sat Jul 14 03:47:27 CEST 2018	YouTube	fake blackhole!	fake blackhole!	fake blackhole!
Sat Jul 14 03:47:28 CEST 2018	YouTube	YouTube	fake blackhole!	fake blackhole!
Sat Jul 14 03:47:29 CEST 2018	YouTube	YouTube	fake blackhole!	fake blackhole!
Sat Jul 14 03:47:30 CEST 2018	YouTube	YouTube	fake blackhole!	fake blackhole!
Sat Jul 14 03:47:31 CEST 2018	YouTube	YouTube	YouTube	fake blackhole!
Sat Jul 14 03:47:32 CEST 2018	YouTube	YouTube	YouTube	YouTube
⋮	YouTube	YouTube	YouTube	YouTube
Sat Jul 14 03:50:00 CEST 2018	YouTube	YouTube	YouTube	YouTube

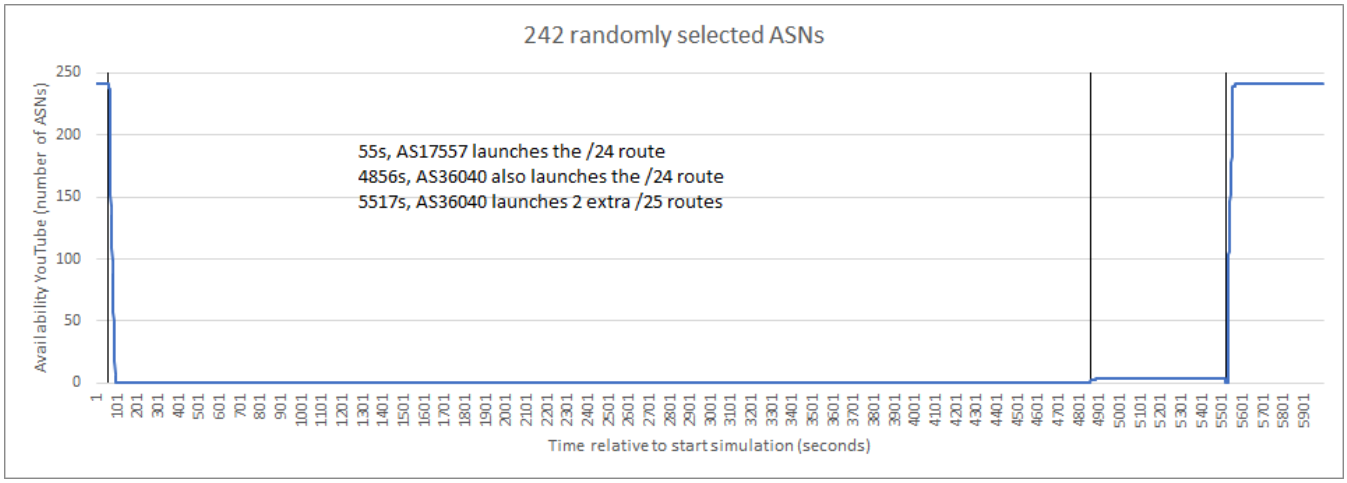

Fig. 5. The availability of YouTube.com on 242 randomly selected ASNs during the simulation of the Pakistan-YouTube 2008 incident.

would draw a similar conclusion. Although the numbers are slightly less worse, they still do not show great effect. Out of the total 5624 simulated ASNs there were 433 ASNs, 7.7%, switched to the genuine /24 route from YouTube (AS36040) while 5191 ASNs, 92.3%, still preferred the fake /24 route announced by Pakistan Telecom (AS17557).

It was however not completely useless of YouTube (AS36040) to announce the hijacked /24 route themselves as well. This is because of filter rules that other ASNs might have in place. They could for example filter out all routes more specific than /24 routes. In that case only announcing the /25 routes would not have solved the problem. Unfortunately, these filter rules or other custom configurations are not publicly known and this lack of data means that they are currently not incorporated into the simulator. However, the simulator does support such configurations.

Pakistan Telecom (AS17557) announced their fake /24 route first and YouTube (AS36040) later. Since one of the decision rules used by BGP is that older already selected routes should be preferred over new ones when the routes

are otherwise equal, the older Pakistan Telecom route has an advantage here. The simulator can also be used to test a hypothetical situation, to show the influence of BGPs preference for older routes. For example, one in which YouTube (AS36040) had already announced the genuine /24 routes as a defence. This greatly impacts the effectiveness of fake announcement by Pakistan Telecom (AS17557). Of all 5624 ASNs the majority with 3580 ASNs, 63.7%, would still prefer the genuine /24 route and 2044 ASNs, 36.3%, switched to the fake route. So in this case significant less ASNs switched compared to the real world event.

XI. Comparison of the Pakistan-YouTube 2008 incident with the simulation

The Internet topology changed quite a bit since 2008, the simulation uses the topology as it was in May 2018. (But if the topology from the 24th of February 2008 would have been available, it could be loaded into the simulator.) Since then Google, the owner of YouTube, moved the YouTube IP blocks to a different ASN: AS36040. So in the simulation the real YouTube IP address block was broadcasted from AS36040. Pakistan Telecom has no longer a peer relationship with PCCW. Since for the verification of the simulator it does not matter which ASN exactly accepts the fake routes from Pakistan Telecom as long as at least one is willing to broadcast them, all ASNs in the simulation accept all announcements and update their routing table accordingly.

The simulation confirms the series of events, with all 4 ASNs on the preferred route from YouTube to Pakistan Telecom providing viewpoints to observe the Internet. Since only the transition between configurations is of interest the intervals between actions by YouTube and Pakistan Telecom were reduced to compact the time required for the simulation. A time line can be seen at table I.

When one compares figure 2 with figure 5 they do not look exactly similar, besides the lower resolution of the data in figure 2. Especially the transition back to full reachability takes significantly longer in 2. This could be explained by the fact that in the simulation no filter rules or other special configurations were present since these are not publicly known. A potential cause for the slower transition could be that some ASNs are filtering routes more specific than /24, thus the /25 routes are not inserted and the transition back might therefore take longer at these ASNs.

XII. Simulation of the Belarussian traffic diversion

As verification of the simulator from a macro perspective the YouTube-Pakistan 2008 incident was re-enacted. To also provide prove of correctness from a micro perspective a different incident was chosen. In 2013 Man-In-The-Middle BGP hijacks became increasingly common, a case in which traffic from Mexico to Washington was redirected via a Belarussian ISP, was ran on the simulator. Some details of the diversion were deducted from dyn.com [1] where possible. The exact mechanism used for the diversion is not known, so the author filled up some gaps between the available observations and built a working mechanism that could have been used.

The key to success in this attack is finding a way to attract Internet traffic destined for the victim while at the same time being able to still send the traffic along for the victim to actually receive it.

To provide a clean outbound path for the Internet traffic to reach its intended recipient a static route was installed to the next hop that the preferred BGP route from the attacker to the victim otherwise would have inserted. To ensure that the rest of this preferred BGP route remains intact a prefix list was used to filter out the fake announcement from the attacker. This fake announcement is required for diverting the Internet traffic to the attacker. The exact same prefix as announced by the ASN connecting the victim was also announced by the attacker. It could not be more specific since in that case all ASNs would eventually prefer this more specific route and this would destroy the clean outbound path. To make sure that as few as possible ASNs would receive the announcement in the first place a prefix filter was used to only inform those ASNs on the preferred path from the origin of the Internet traffic to the attacker. Since BGP spreads around routes to its peers this way at least the fake route appears to be longer, preventing other ASNs from installing it in their routing tables and destroying the clean outbound path. ASNs for which both routes are equally long the decision rule stating that older routes should be preferred protects the clean outbound path from the fake new one.

The Linux `traceroute` command was used to check the route of Internet traffic from its origin to the victim. Under normal conditions it would take this route:

```
traceroute to 63.234.113.110, 30 hops max
 1  pc-gdl1.alestra.net.mx (201.151.31.1)
                                    AS11172
 2  192.168.1.42 (192.168.1.42)     AS3491
 3  192.168.192.64 (192.168.192.64) AS209
 4  63-234-113-110.dia.static.qwest.net
                  (63.234.113.110)  AS209
```

The traffic starts at a server connected via Alestra (AS11172) in Guadalajara, Mexico. PCCW Global (AS3491) transports it to Qwest (AS209). Qwest delivers it to the recipient in Washington DC, US.

With the attack going on the route changes to this:

```
traceroute to 63.234.113.110, 30 hops max
 1  pc-gdl1.alestra.net.mx (201.151.31.1)
                                    AS11172
 2  192.168.1.42 (192.168.1.42)     AS3491
 3  192.168.68.109 (192.168.68.109)
                                    AS20485
 4  192.168.132.146 (192.168.132.146)
                                    AS20940
 5  192.168.192.64 (192.168.192.64) AS209
 6  63-234-113-110.dia.static.qwest.net
                  (63.234.113.110)  AS209
```

The route starts the same but PCCW Global (AS3491) hands over the traffic to TransTeleCom (AS20485) instead of Qwest (AS209). TransTelecom delivers the traffic to the attacker in Minsk, Belarus. The attacker sends the traffic along the clean outbound path via Akamai (AS20940) in order for it to be delivered by Qwest to the victim.

It is unlikely that the victim is aware that its Internet traffic

is being viewed or even modified by somebody in Minsk, Belarus. But if for some reason the victim decides to run its own traceroute as a check, everything will look perfectly normal. This is because the routes carrying content from the victim to their destination have not been tampered with.

A normal outgoing route in the victims traceroute:

```
traceroute to 201.151.31.149, 30 hops max
 1  dca-edge-17.inet.qwest.net
                    (63.234.113.1)  AS209
 2  192.168.1.40 (192.168.1.40)    AS3356
 3  192.168.130.41 (192.168.130.41)
                                    AS11172
 4  pc-gdl2.alestra.net.mx
                (201.151.31.149)  AS11172
```

Qwest (AS209) hands over the Internet traffic to Level3 (AS3356) which transports it to Mexico for delivery by Alestra (AS11172).

XIII. Comparison of the Belarussian traffic diversion with the simulation

The Internet Topology has changed since this event as well. The Internet became more interconnected, resulting in shorter routes between the ASNs of interest. Nevertheless, the attack was still possible.

One could also notice that the IP addresses of the intermediate ASNs are private IPs and are therefore not resolved by DNS. This is because of the fact that to safe resources only those prefixes/addresses of interest are used in the simulation, in this case only the endpoints. This does not change the way BGP and all the routers behave, there are just using other numbers.

Since not all details, and especially the precise mechanism used, are not known, it is possible the the attack happened somewhat different. This does not however invalidate the simulator. The point of this simulation was to verify that the simulator could handle events which would require tiny actions and modifications targeting only very specific routers as well as large events with global impacts as shown by the YouTube-Pakistan 2008 simulation.

XIV. Memory usage of the simulator

During the reconstruction of the Pakistan-YouTube 2008 incident using the Internet topology based on the dataset the simulator used 33.25G of memory for all 5624 ASNs. The larger part of this memory was required for the customized Mininet, using nearly half of it with 14.87G. The BGP daemons come next when considering memory usage with 11.17G. The zebra instances along with the BGP routes divide the rest almost evenly among them with respectively 4.15G and 3.06G. The routes announced during the simulation are 3 unique routes and 1 route was announced twice: 1 times a /22 route, 2 times a /24 route and 2 /25 routes at last. This demonstrates that even a small amount of routes takes up a considerable amount of memory already. This is due to the fact that the routing information is stored on each individual router. Within the simulation 5624 ASNs are present, so the routes are duplicated a lot. One would need a huge amount of RAM to store all routes that are announced in the real world if one would like to use them in

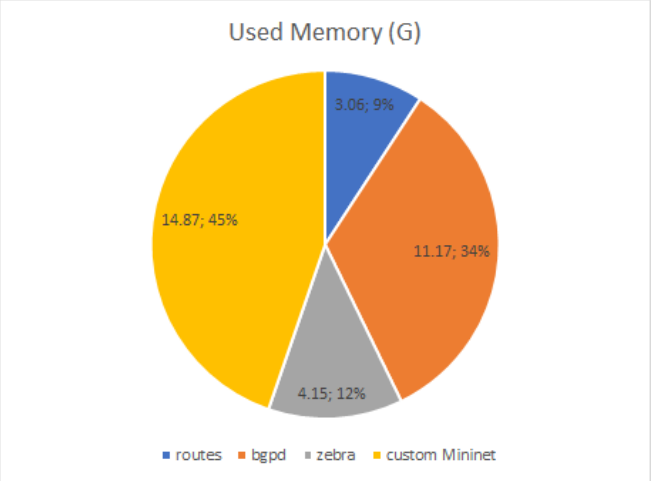

Fig. 6. Memory usage by the simulator for the Internet topology.

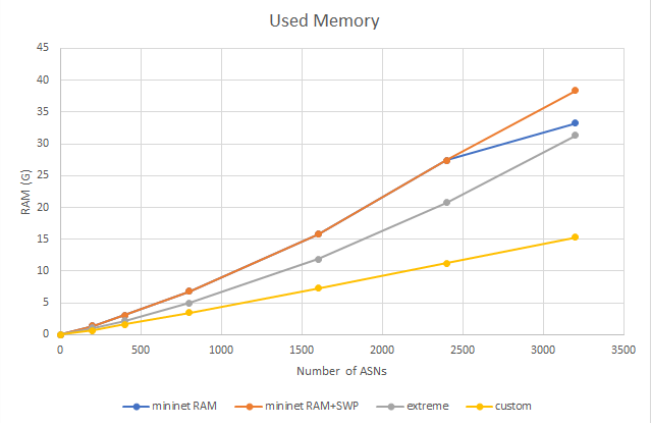

Fig. 7. Memory usage of the custom Mininet in the simulator compared to regular Mininet and Mininet Extreme.

the simulator, therefore it is recommended to only announce those routes which are of interest to the simulation at hand.

As can be seen in figure 7 the custom Mininet version scales considerably better than the original Mininet and Mininet Extreme, enabling the simulation to be ran on the full topology from the dataset, thus the entire Internet. This is one of the features that sets this simulator apart from the rest. With the regular Mininet the computer used for the tests ran out of RAM and used SWAP memory to store the Mininet instances. This was taken into account to prevent the graph from showing an odd discontinuity in the curvature.

XV. CPU usage of the simulator

Each connection between BGP instances has timers to keep it alive and to check for updates. The BGP daemon sets these timers and when they are due the timers send an interrupt service routine request to get processing time.

The simulation for the reconstruction of the Pakistan-YouTube 2008 was first ran on a single machine with two Intel Xeon CPU L5520 processors running at 2.27GHz. When performing the simulation it generated too many interrupts for the CPU to handle, see figure 8. The interrupt service routines are displayed in bright green. From Monday 17:00 onwards they grow from the top of the graph. Since the interrupt service routines were not handled fast

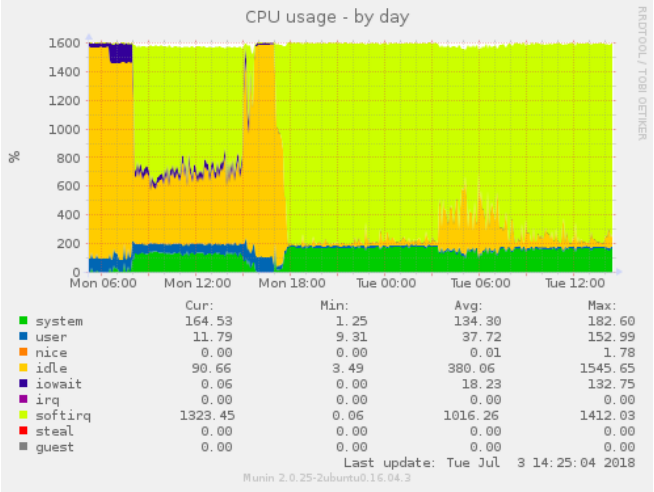

Fig. 8. CPU usage of the custom Mininet in the simulator before splitting load. The simulation started around Monday 17:00. The CPU was fully overwhelmed by the ISR (Interrupt Service Routine) requests by 18:00. The simulation crashed on Tuesday at 03:00.

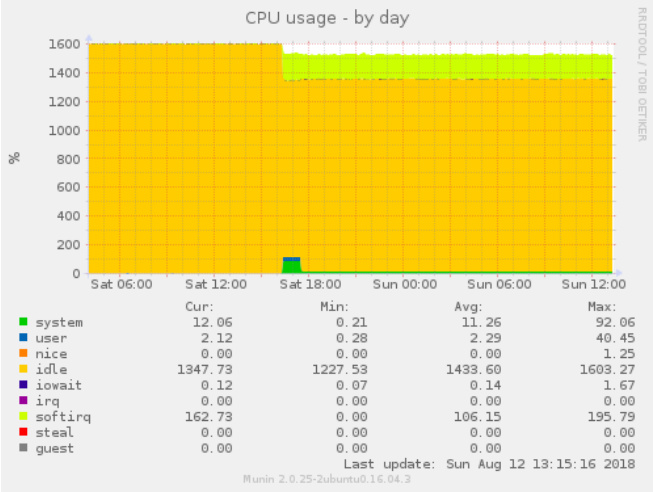

Fig. 9. CPU usage of a single machine after splitting the load across 4 servers. With this reduced load the time spent by the processor handling ISR (Interrupt Service Routine) requests is consistent and no longer growing.

enough, they pilled up until eventually the processor was only serving interrupts and therefore incapable of handling normal processes. Since the topology used for the simulation was already optimized, it was not possible to generate less interrupts by reducing the amount of ASNs simulated.

A somewhat more powerful processor would probably only have postponed the problem since the interrupts overwhelmed the current processor over the course of a few dozen minutes. To solve to problem completely with just a more powerful processor would require a much faster processor which would not be practical any more. So, it was solved instead by combining the processing power of multiple machines by splitting the load across multiple servers. In this case 4 different servers were connected via Ethernet cables and a powerful switch. In combination with forwarding the traffic destined for the other machines via the Linux `ip route add` or the `route add` command. This resulted in a much smaller load per machine as can be seen in figure 9, handling the ISR requests now required a constant amount of processing power.

After the idea to split the work load across different machines the question arose how to split it. A first attempt was made with a very naive and easy approach. The ASNs were split into a lower and upper half depending on their number. This approach showed the potential of splitting, but the lower half of the ASNs was still too heavy to be simulated on one machine. The lower half contained almost all of the larger ASNs, which have significantly more connections. A more even distribution was needed.

As a second attempt equalizing the number of ASNs per machine while minimizing the number of connections between ASNs on different machines was tried. This problem seemed too complex to tackle efficiently since the only option was a brute force algorithm which would do an exhaustive search.

The third and final attempt followed from the idea to see a connection as two endpoints. This way one can just count the number of endpoints per machine and this should give a well balanced distribution, since most of the work done by a machine is determined by the amount of endpoints it has to handle. The number of messages received and send by a machine, and thus the number of interrupt service routines it has to serve, are given by the number of endpoints. A small algorithm was written which started with a random distribution and by shifting ASNs around would try to find an optimum. A distribution was found where the number of endpoints per machine only differed by two. This is an optimum since every ASN in the simulation has at least two connections, as stub ASNs were aggregated into their parents.

After finding an optimal distribution and splitting the ASNs over multiple machines a problem arose. All routers were no longer connected on bridges, which provided direct connections, so ARP can no longer discover the routers on other machines. To overcome this issue the script injected the required ARP entries via the Linux command `arp -s`, after which all routers were able to locate each other again.

The fact that this simulator can run in a distributed way enables it to perform in real time, another feature that makes this simulator unique.

XVI. Discussion

The simulator is capable of running the current entire Internet topology in real time on distributed machines, enabling simulations that were not possible before. It was used to re-enact both the Pakistan-YouTube 2008 incident and the Belarussian traffic diversion. Which together served as a verification for the correct workings of the simulator for both large scale events and events that require detailed in-depth control. Unfortunately the simulation encountered some deviations from the real world since ISPs might use custom configurations with for example filter rules applied which are not publicly known. ISPs usually treat their ASN configurations as some kind of corporate secrets. However if those configurations would be publicly known the simulator would be able to use them. Therefore it is a limitation on the availability of data and not of the simulator. Some researchers have proposed ways to overcome this problem partially, but these proposals will never be entirely accurate [18][19].

The simulator works best in a distributed fashion, unless smaller topologies than the current Internet are used in which case a single machine could be used as well. Since the Internet is growing larger and larger with an increasing pace the possibility to run the simulation distributed across multiple machines ensures compatibility with future Internet topologies for the foreseeable future.

With this simulation Internet administrators can prevent both honest mistakes and deliberate attacks. The simulator provides them a tool to test new configurations with the hazard of real network failures and existing configurations against attacks and failures. This increases the Internet security and stability.

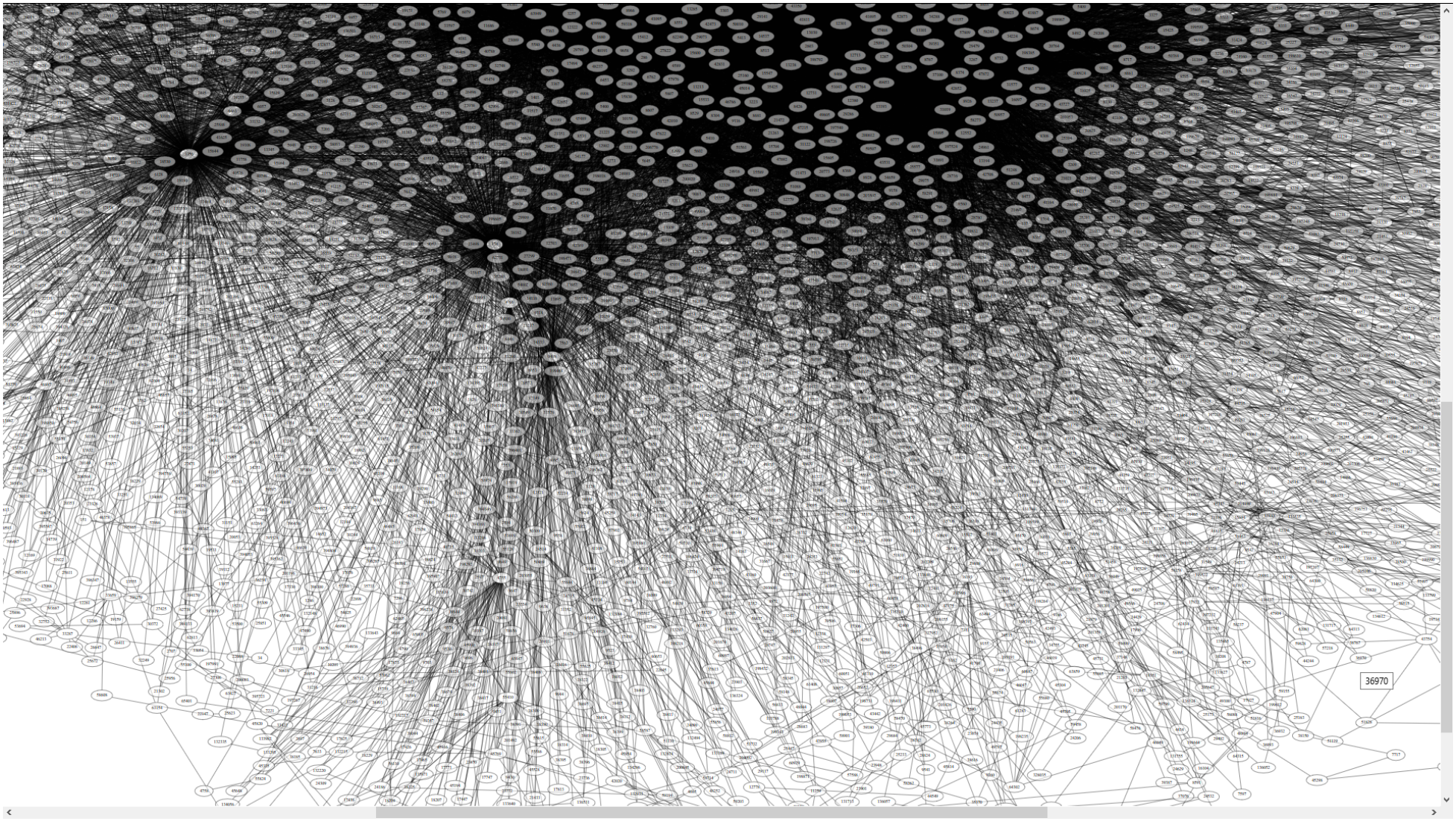

Fig. 10. Part of the graph used in the simulation. The graph is exported as .svg file to make it zoomable and searchable. When the cursor is hovered over an edge or node, it displays respectively the ASN number of the node or the ASN numbers of the endpoints. In the topleft corner an ASN, AS3356, with many peers is visible which nicely shows up as a concentration of edges around the node. But also ASNs with just 2 peers are visible, see the bottom left, AS58608. On the bottom right the mouse hovers over node AS36970 and its ASN number is shown.

APPENDIX A
RENDER OF THE GRAPH USED IN THE SIMULATION

References

[1] Jim Cowie. *The New Threat: Targeted Internet Traffic Misdirection*. Nov. 2013. URL: https://dyn.com/blog/mitm-internet-hijacking/.

[2] Jintae Kim et al. "A BGP Attack Against Traffic Engineering". In: *Proceedings of the 2004 Winter Simulation Conference, 2004.* (2004). DOI: 10.1109/wsc.2004.1371332.

[3] Palanivel A. Kodeswaran et al. "A declarative approach for secure and robust routing". In: *Proceedings of the 3rd ACM workshop on Assurable and usable security configuration - SafeConfig 10* (Oct. 2010). DOI: 10.1145/1866898.1866906.

[4] Hitesh Ballani, Paul Francis, and Xinyang Zhang. "A study of prefix hijacking and interception in the internet". In: *Proceedings of the 2007 conference on Applications, technologies, architectures, and protocols for computer communications - SIGCOMM 07* (Aug. 2007). DOI: 10.1145/1282380.1282411.

[5] Pavlos Sermpezis et al. "A Survey among Network Operators on BGP Prefix Hijacking". In: *ACM SIGCOMM Computer Communication Review* 48.1 (2018), pp. 64–69. DOI: 10.1145/3211852.3211862.

[6] Declan McCullagh. *How Pakistan knocked YouTube offline (and how to make sure it never happens again)*. Feb. 2008. URL: https://www.cnet.com/news/how-pakistan-knocked-youtube-offline-and-how-to-make-sure-it-never-happens-again/.

[7] Government Of Pakistan, Pakistan Telecommunication Authority, Zonal Office Peshawar, 2008. URL: https://web.archive.org/web/20130131083222/http://www.renesys.com/blog/pakistan_blocking_order.pdf.

[8] *Pakistan hijacks YouTube*. Feb. 2008. URL: https://dyn.com/blog/pakistan-hijacks-youtube-1/.

[9] Mar. 2018. URL: http://data.caida.org/datasets/topology/ark/.

[10] *API Help*. URL: http://as-rank.caida.org/api/v1.

[11] Apiary. *BGPView API - Apiary*. URL: https://bgpview.docs.apiary.io/.

[12] Patrick Maigron. *Regional Internet Registries Statistics*. URL: http://www-public.imtbs-tsp.eu/~maigron/RIR_Stats/RIR_Delegations/World/ASN-ByNb.html.

[13] Mininet Project. *Mininet*. URL: http://mininet.org/.

[14] Released under GNU General Public License. *Quagga Routing Software Suite*. URL: https://www.quagga.net/index.html.

[15] The Mininet Extreme Team and The Mininet Project. *Mininet Extreme*. URL: https://github.com/sk2/mininet-extreme.

[16] Open Source. *DOT (graph description language)*. URL: http://www.graphviz.org/documentation/.

[17] Daniel Stenberg. *cURL*. URL: https://curl.haxx.se.

[18] Phillipa Gill, Michael Schapira, and Sharon Goldberg. "A survey of interdomain routing policies". In: *ACM SIGCOMM Computer Communication Review* 44.1 (2013), pp. 28–34. DOI: 10.1145/2567561.2567566.

[19] Ruwaifa Anwar et al. "Investigating Interdomain Routing Policies in the Wild". In: *Proceedings of the 2015 ACM Conference on Internet Measurement Conference - IMC 15* (Oct. 2015). DOI: 10.1145/2815675.2815712.