Title

Identifying Traffic Anomalies Interfering with IBR Based Outage Detection

Author

Gupta, Ojas

Publication Date

2018

UNIVERSITY OF CALIFORNIA SAN DIEGO

Identifying Traffic Anomalies Interfering with IBR Based Outage Detection

A thesis submitted in partial satisfaction of the
requirements for the degree
Master of Science

in

Computer Science

by

Ojas Gupta

Committee in charge:

Kimberly C. Claffy, Chair
Geoff Voelker, Co-Chair
Alberto Dainotti
Aaron Schulman

2018

The thesis of Ojas Gupta is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

____

____

____
Co-Chair

____
Chair

University of California San Diego

2018

DEDICATION

I dedicate my thesis to my friends and family. My friends, for keeping me company while I was working on my project. My family - parents and grandparents, for their constant support and guidance at every stage of my life.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I wish to thank my committee members who were more than generous with their expertise and precious time. I thank my advisor Alberto Dainotti for his invaluable guidance both as my principal investigator as well as my mentor. I thank Alistair King for continually providing his valuable knowledge and patiently answering all my questions. I thank them for the flexibility and freedom they allowed me in picking which projects to work on. Although there is still plenty of room for improvement, I am a much better presenter, writer, researcher, and programmer due to their detailed and honest feedback. I thank my committee chair Professor Kimberly C. Claffy for providing me the opportunity to work in CAIDA. I thank Professor Geoff Voelker and Professor Aaron Schulman for agreeing to be part of my M.S. thesis committee.

ABSTRACT OF THE THESIS

Identifying Traffic Anomalies Interfering with IBR Based Outage Detection

by

Ojas Gupta

Master of Science in Computer Science

University of California San Diego, 2018

Professor Kimberly C. Claffy, Chair
Professor Geoff Voelker, Co-Chair

Internet Background Radiation (IBR) is network pollution, composed of all kinds of Internet traffic, namely backscatter data (due to spoofing), botnet scans and other traffic pollution (due to misconfigurations in the networking devices). For many years, IBR collected at routed, but unused address spaces, known as Network Telescope aka Darknet, has been used in multiple network research applications such as identifying malicious Internet activities, detecting Internet outages, etc. The Internet Outage Detection Analysis (IODA) is a research project of UCSD which uses a predictable signal that can be extracted from IBR to monitor the Internet for macroscopic Internet connectivity outages. Due to the varied composition of IBR, extracting consistent ”normal” IBR traffic is difficult, but at the same time is required to detect outages accurately. In this report, I investigate and analyze IBR collected at the UC San Diego Network Telescope (UCSD-NT), with the goal of developing a better understanding of events distorting the coherent nature of the IBR signal and subsequently devising approaches to detect and remove traffic triggered by these events. These distortions in the IBR signal can be caused by many events, such as Distributed Denial of Service (DDoS) attacks, botnet scans, Domain Name Service (DNS) poisoning, etc. I investigate these events to study their causes by various means of statistical and experimental tools. In my analysis of three years of IBR data, I detect many short-term events which distort a stable signal extracted from IBR (used by IODA), by generating traffic bursts. Also, I identify and analyze a large-scale event: SYNRENT (coined from SYN-BitTorrent), which caused an increase in the TCP-SYN traffic reaching the UCSD-NT for over two years. I present a broad characterization of SYNRENT in terms of source IPs, countries, Autonomous Systems (AS), operating systems, etc. to better understand the phenomenon. This thesis also identifies possible causes of SYNRENT such as the Distributed Hash Table (DHT) poisoning of BitTorrent traffic in the Great Firewall of China. To mitigate such distorting events and bursts, I present a software solution to detect and filter them in real-time: the primary idea is to use the history of previously observed unique source IPs on destination ports and IPs, to determine the occurrence of a burst. In addition to proposing and implementing the solution, I provide substantial evidence to prove that it successfully detects and removes the components that distort the signal from IODA preferred IBR.

Chapter 1

Introduction

1.1 Internet Background Radiation

For more than a decade, Internet measurement research has benefited from the use of unsolicited one-way traffic also known as Internet Background Radiation (IBR)[1][2]. IBR is network pollution, generated by a vast number of computers all around the globe. The traffic generated from these computers, is often unknown to their original users. The resulting traffic aggregate has proven to be a useful source of data for observing network characteristics that are not revealed by other types of data.

Scientists study and analyze IBR collected at unused but routed network blocks. The passive traffic monitoring system built on these network blocks are commonly known as Network Telescope aka Darknet. The communication flow on these network blocks is always unidirectional, i.e., the network telescope does not respond to traffic it receives. Currently, University of California San Diego (UCSD) operates one of the world's largest Network Telescopes[3] (UCSD-NT) which observes 99% of a /8 IPv4 address block, i.e., 1/256 of the whole IPv4 address space (Figure 1.1). The volume of IBR collected at this large telescope is also enormous. Currently, UCSD-NT captures around 1,200 GB of compressed IBR per day (approximately 50 GB per hour).

Internet Background Radiation (IBR) is composed of all kinds of Internet traffic. There are three primary causes of IBR traffic: (i) backscatter data from spoofed denial-of-service (DoS) attacks, (ii) botnet scans, and (iii) bugs and misconfigurations in networked devices. A DoS attack tries to flood a victim with traffic from randomly spoofed packets. An attack of high magnitude can reduce or crash its victim's ability to serve users. The response packets (e.g., SYN-ACK TCP packets in reply to SYNs from the attacker) are sent back to the spoofed IP addresses, resulting in backscatter data, which is collected at the telescope. Scanning of the address space in search of vulnerable victims, caused by automated bots or manual initiation, is another fundamental component forming IBR. Misconfiguration and bugs in networking systems (firmware or software) can also induce IBR traffic, e.g., by setting a wrong IP address on DNS or setting wrong byte order. This misconfiguration can assign incorrect network addresses to a device, generating unsolicited traffic.

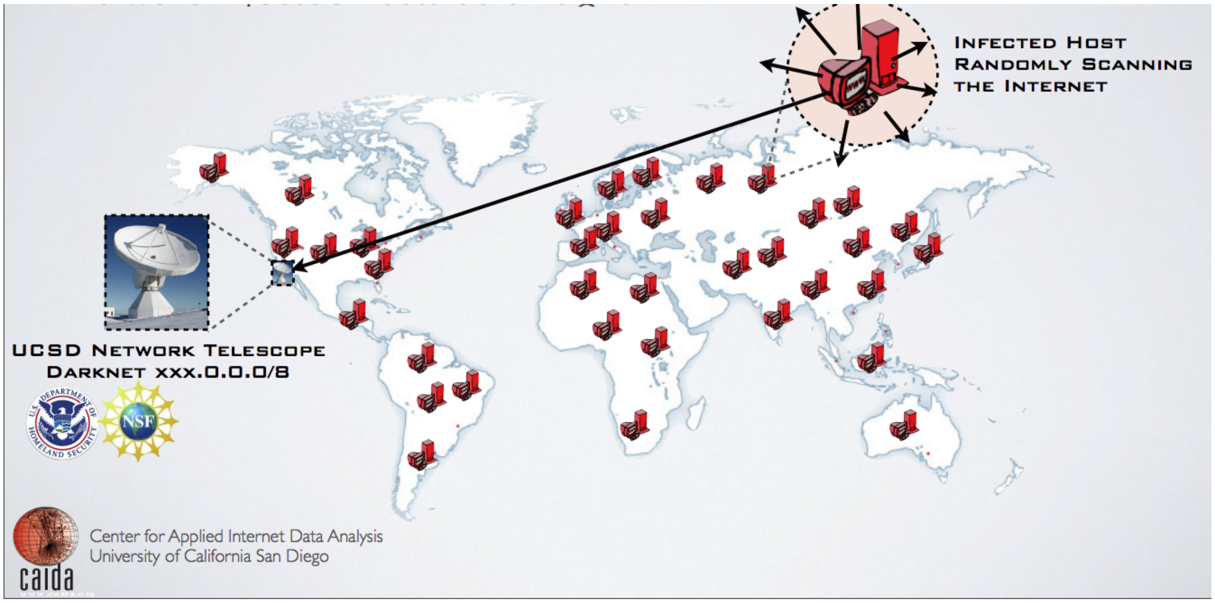


Figure 1.1: Illustration of Internet background radiation reaching UCSD Network Telescope from all around the world. IBR traffic is composed of primarily backscatter data, botnet scans, and traffic generated by bugs and misconfigurations in networked devices.[4]

Internet background radiation has been extensively used in many opportunistic Internet measurement studies, ranging from discovering malicious Internet activities such as identifying botnet scans[5][6] to performing opportunistic network measurements[7]. Many researchers have used a signal extracted from IBR that exhibits a regular pattern, to detect Internet connectivity disruptions in the world[8][9][10] (discussed in the next section).

1.2 Internet Outage Detection and Analysis

Internet Outage Detection and Analysis (IODA) is a large Internet research project by the Center for Applied Internet Data Analysis (CAIDA)[11], which monitors the Internet to identify macroscopic Internet outages. The IODA system (Figure 1.2) processes and analyzes measurements from three sources: i) Global Internet routing (Border Gateway Protocol data), ii) Internet Background Radiation (IBR collected at UCSD Network Telescope) and iii) Active probing. IODA inference system combines information from these three data sources, establishes the possibility of an outage and generates alerts. This activity is monitored and visualized by a time series visualization tool known as Charthouse[12][13]. Since many years, IODA has been successful in identifying and detecting numerous Internet outages all around the globe (Figure 1.3).

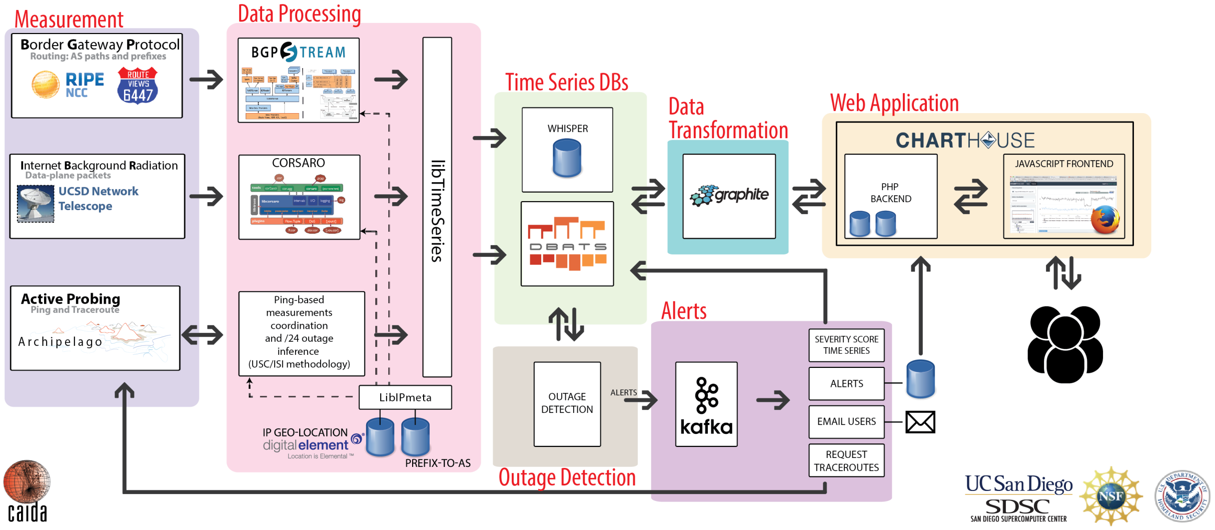
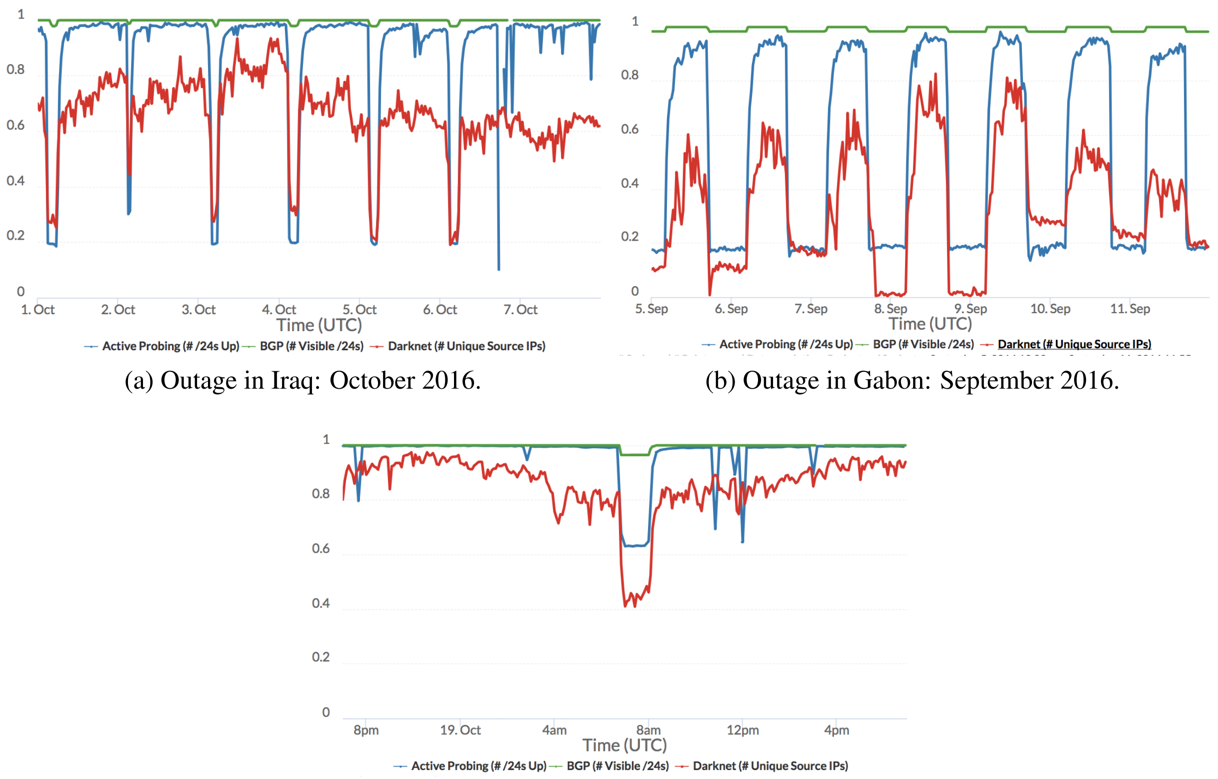


Figure 1.2: A high level view of IODA architecture.[14]

Internet background radiation is one of the primary sources of data in IODA for detecting Internet outages. It is established in previous studies[18], that IBR consists of network pollution sent from all over the world. Sources sending IBR represent a significant amount of total address space in terms of IPs, /24 blocks, autonomous systems, regions, and countries. This composition of IBR makes it an essential and pervasive source of information in IODA. Casado et al. were the first to explore opportunistic measurement opportunities from IBR in [7]. Following their work in [8] and [9], Dainotti et al. extracted signals that revealed new dynamics of long-term connectivity disruptions. They used the fact that IP addresses from disconnected networks stop sending IBR traffic. Mapping these addresses to their geolocation enables an estimation of the geographic impact of an event on communications infrastructure.



(a) Outage in Iraq: October 2016.

(b) Outage in Gabon: September 2016.

(c) Outage in Mexico: October 19, 2016

Figure 1.3: Example outages detected by IODA in 2016. Blue represents Active Probing, green represents BGP and red represents IBR collected at the UCSD-NT. a) For the first week of October 2016, IODA detected daily outages in Iraq that affected Iraqi Internet Connectivity. b) In September 2016, IODA detected Outages in Gabon, which were imposed amid political clashes. c) On 19th October 2016, IODA detected an outage that affected Mexican Internet connectivity for over an hour. [15][16][17]

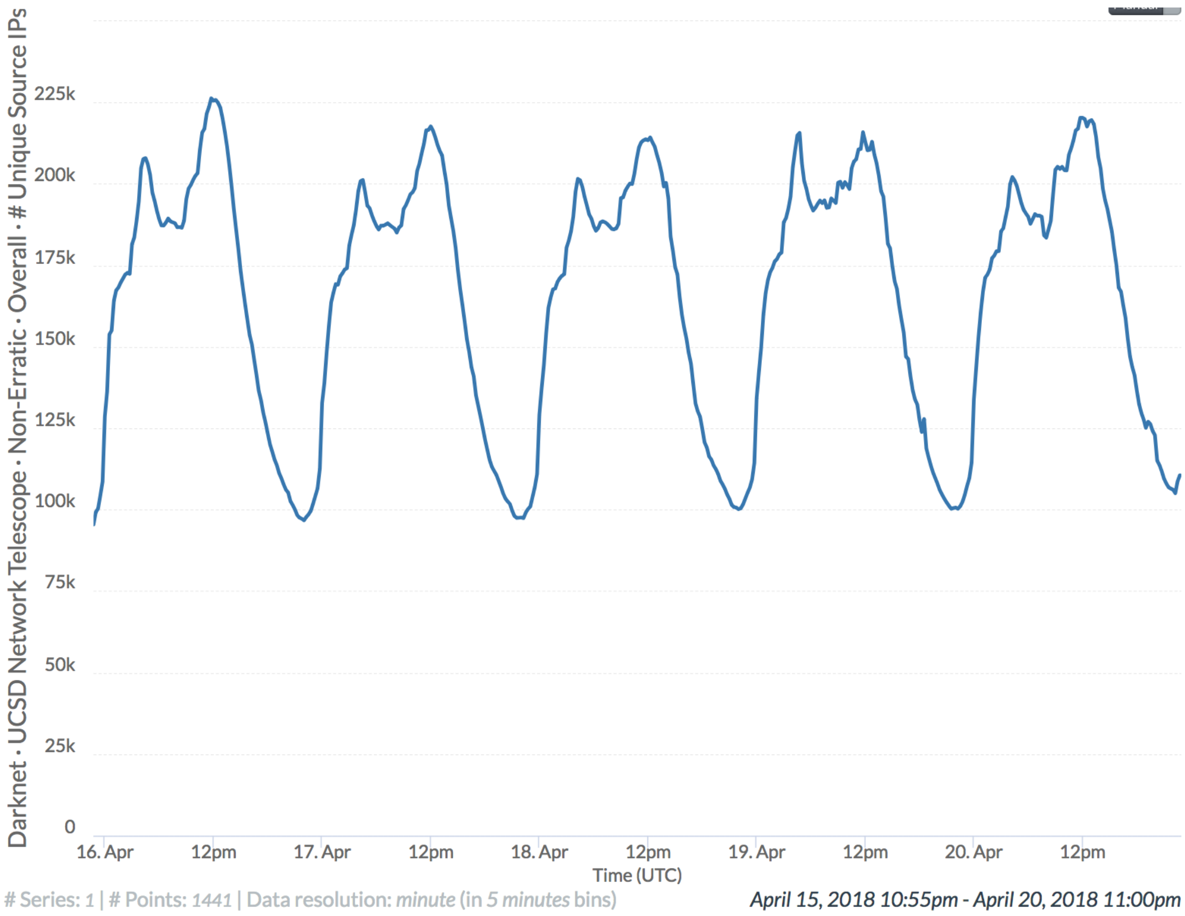


Figure 1.4: Diurnal pattern observed in IBR collected at UCSD Network Telescope to illustrate predictable IBR behavior. The cause of this pattern is because during work hours machines are active and generate more traffic than during non-work hours.

The IODA project detects an outage using a stable signal extracted from IBR in the following two steps: firstly, it geolocates all the source IPs in the extracted signal to obtain the traffic composition regarding countries, ASes, and regions, and secondly, it checks for sudden changes (drops or deviations) in the packet rate from a particular area, suggesting outages. If a drop in the traffic originating from a region or country is encountered, IODA combines this alert with the results from BGP and active probing, to predict an outage.

Ideally, the signal extracted from IBR to be used in IODA should be strong (statistically significant), stable (low noise), and globally pervasive (seen in most networks). However, due to its nature, IBR is not stable and does not always have a consistent pattern. Generally, IBR follows a diurnal pattern, since more machines are active during work hours (see Figure 1.4). This regular pattern in IBR is disrupted by bursts of traffic, resulting in disturbances (random spikes and drops) in IBR signal. These are typically caused by DoS attacks, scans, etc. Since IODA uses traffic drops to detect outages, with the presence of these irregularities, it becomes difficult for IODA to infer outages from IBR.

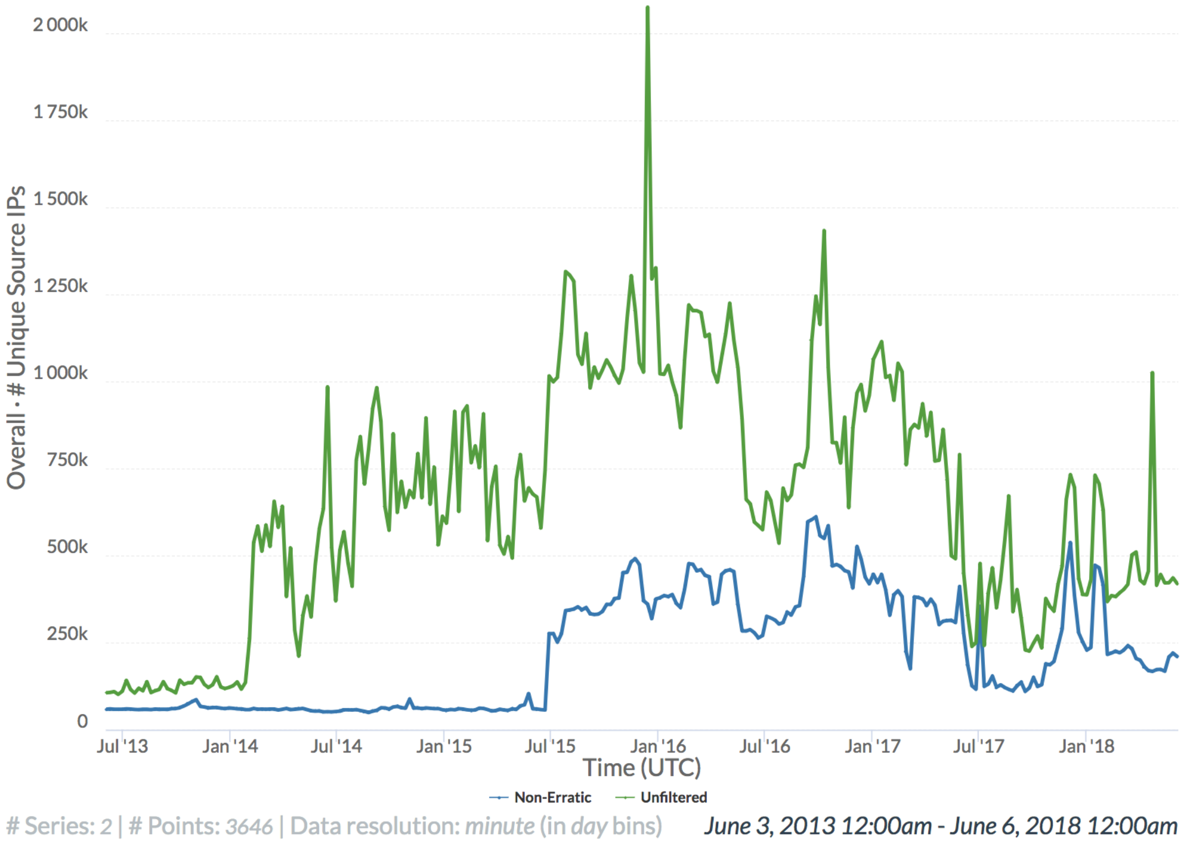


Figure 1.5: Time series graph of Internet Background Radiation collected at UCSD for previous five years. Unique source IPs are used to illustrate IBR traffic. Before July 2015, IBR signal was consistent and stable for IODA, as it was continuously monitored and filtered for events disrupting IBR signal (for e.g. spoofed traffic, conflicker scanning, potential backscatter traffic, BitTorrent traffic collected on UDP ports). After July 2015, IBR has included unusual traffic from many events (BitTorrent traffic on TCP, MIRAI botnet) which distort its predictable signal.

Figure 1.5 portrays the time series graph of the extracted IBR signal (in terms of unique source IPs) on UCSD-Network Telescope used by IODA for the last five years. The graph shows that before July 2015, IBR signal is consistent and stable for IODA, as it was continuously monitored and filtered for events disrupting IBR signal (for e.g. spoofed traffic, conflicker scanning, potential backscatter traffic, BitTorrent traffic collected on UDP ports). After July 2015, IBR has witnessed unusual traffic from many events (BitTorrent traffic on TCP-SYNRENT, MIRAI botnet) that contaminated its consistent signal, thus reducing the accuracy of outage detection. There is an immediate need to filter these irregularities for IODA to detect outages accurately.

1.3 Objective

Previously, Internet background radiation had been filtered and sanitized[19][8], to stabilize its signal for IODA, but this was done only till June 2015. Since then a significant amount of evolved traffic has contaminated the stable signal extracted from IBR (refer Figure 1.5). The existing filters fail to remove this emerged erratic traffic due to its unprecedented traffic composition. The primary objective of my thesis is to investigate and analyze IBR collected at UC San Diego Network Telescope, with the goal of developing a better understanding of these events distorting the coherent nature of IODA preferred IBR signal and subsequently devising approaches to detect and remove these events.

1.4 Organization of the Thesis

The remainder of this thesis is organized as follows: Chapter 2 presents an overview of the previous works done in studying and sanitizing IBR along with the experimental tools. Chapter 3 provides empirical details about the events identified and explains the characterization of the events. Chapter 4 covers the proposed solutions, relevant experiments, and results. Chapter 5 concludes the thesis, mentioning ideas for future work.

Chapter 2

Background and Related Work

Internet background radiation (IBR), the unsolicited one-way Internet traffic collected at unassigned address space (network telescope), has been used in Internet measurement and network analysis studies for many years. IBR reflects fundamentally nonproductive traffic, either malicious (flooding backscatter, scans for vulnerabilities, worms, DDoS attacks) or benign (misconfiguration at different layers of the protocol stack). Due to this varied composition of IBR, defining coherent ”normal” IBR traffic is difficult. To use IBR in certain types of opportunistic network analysis, such as outage detection, sanitizing IBR by filtering out disturbances is essential. The primary objective of my thesis is to investigate these disturbances in IBR and devise strategies to detect and filter them.

Section 2.1 covers Literature Review regarding Internet Background Radiation in the following manner: Subsection 2.1.1 discusses the significant components of IBR, followed by Subsection 2.1.2 mentioning previous studies and work done in identifying malicious Internet activities using IBR. Subsection 2.1.3 describes the previous work done in establishing the importance of IBR in Internet outage detection. Subsection 2.1.4 discusses previous work in sanitizing IBR to make the IBR signal consistent for IODA. The last section 2.2 highlights the main experimental tools I used in the analysis of IBR.

2.1 Literature Review

2.1.1 Composition of Internet Background Radiation

There has been much work done in analyzing Internet Background Radiation. Benson et al. performed a major characterization of IBR traffic observed at two large network telescopes [18]. They extracted various IBR components, using the datasets from UC San Diego Network Telescope[3] and Merit Network Telescope[20]. UCSD-NT and Merit-NT cover 99% and 67% of different /8 IPv4 address blocks respectively, thus collecting a massive amount of IBR. In her work, Benson focuses towards answering questions like ”Who sends IBR?”, ”What is IBR made of?” etc. They analyzed many aggregations (IP addresses, /24 IP blocks, prefixes, ASes, and countries), and observed a significant amount of sources compared to the total address space announced in BGP. Along with the sources, they studied the composition of IBR. They accomplished this task by characterizing IBR along two basic dimensions: transport layer protocol and application. Their work has shown that TCP (due to scanning and backscatter) and UDP (due to Peer-to-Peer traffic and misconfigurations in the networking devices) were the major contributors of IBR in 2012 and 2013. Their study concluded that IBR has the potential to represent an Internet-wide view concerning the address space it covers as well as its overall traffic composition. Due to these salient attributes, IBR has played a significant role in many Internet research studies.

2.1.2 IBR in detecting malicious Internet activities

In [3], using the UCSD Network Telescope, Dainotti et al. identified and studied a stealth botnet scan of VoIP-related SIP servers, carried out in 2011. They presented multiple approaches and techniques to determine the scanning behavior of botnet. They used a Hilbert-curve map[21] to project the 16 million IP addresses in the telescope. Using the Hilbert curve, they identified the reverse byte ordering sequence in the destination IPs in packets from the botnet. To further investigate the characteristics of the botnet, they inferred machine composition using p0f tool[22], distribution of IP addresses at country-level and AS-level, etc. Thus, IBR along with these sophisticated analysis methods [6] provides us a means to detect and fight against large-scale malicious activities. Since 2011, many unprecedented events have occurred that need to be studied and analyzed. In my work, I investigate and examine some of these suspicious events by using the approaches and tools mentioned in the above study.

2.1.3 IBR in Internet Outages Detection

In addition to identifying malicious Internet activities, IBR has also been used to analyze macroscopic Internet events that are unrelated to malware. One significant use of IBR has been done in analyzing country-wide Internet connectivity disruption, caused by censorship due to political events in two countries: Egypt and Libya[8]. In this work, Dainotti et al. studied Internet outages using multiple sources of large-scale Internet data such as BGP routing control plane, IBR, active probing measurements, etc. They used a geolocation database to determine the IP addresses of the given countries in IBR. They observed many plummets in the number of packets captured per second by the UCSD-NT from the source IPs of Egypt and Libya. These sudden drops in the Internet traffic from the above countries matched the withdrawals and re-announcements of BGP routes of the countries' networks. Their work was one of the first instances where IBR was used to study Internet outages and acted as a stepping stone for the IODA project. The main idea proposed in their work is to study IBR to observe and detect changes in the traffic from the affected regions to determine an outage.

Another notable work in detecting Internet outages is done in [9], where Dainotti et al. studied the impact of geophysical events (like earthquakes) on the Internet connectivity. They proposed generating and using a new metric of local IBR activity based on the unique number of source IP addresses per hour. Subsequently, to detect Internet outages at a finer granularity, Benson et al. analyzed IBR to gain insight into AS-level outages in [10]. These studies prove that a consistent IBR signal is a helpful source of information in detecting Internet outages. My work in this thesis is fundamental in preserving the consistency in erratic IBR signal used by IODA. For my analysis, I use their proposed metric of unique source IP addresses. Use of this metric as opposed to packet rate makes the data consistent and immune to bursts of traffic from a few hosts.

2.1.4 Sanitizing IBR

To use IBR, we must understand that unpredictable events and decisions influence the reliability of IBR as a source. IBR contains bursts of traffic, spoofed IP addresses, backscatter data, botnet scans, etc., making the data inconsistent with irregular spikes and drops. To provide consistency in IBR traffic, Dainotti et al. identified and filtered spoofed traffic from IBR, while estimating Internet address space utilization in [19]. Presence of spoofed traffic can significantly distort results by implying that fake addresses are active. To mitigate the effects of spoofing on IBR measurements, they identified large portions of spoofed traffic by using various heuristics. They also built signatures for their filters by identifying suspicious traffic components. They manually isolated and analyzed the suspicious traffic, and further defined filters to remove it.

In [8], Dainotti et al. also devised various IBR cleansing techniques to make it usable for outages detection. To make the IBR signal consistent, they identified and filtered Conficker scanning traffic by selecting TCP SYN packets with destination port 445 and packet size 48 bytes. They also removed potential backscatter traffic flows by filtering TCP SYN-ACK or RST, or ICMP echo replies (which are captured by telescope when its IPs are spoofed in DoS attacks). Their work provides an exhaustive study of detecting and filtering noise from regular IBR and thus acts as an inspiration for my thesis. Earlier, the events were being identified and filtered by manual driven methods. However, it is important to note that these previous works have sanitized IBR before 2015 and since then IBR has become extremely noisy and incoherent. The existing filters fail to detect and remove the new evolved erratic traffic. My work entails a semi-automatic approach to detect and remove this erratic traffic caused by these events.

Benson further sanitized IBR, by studying the pattern of BitTorrent traffic in her Ph.D.

dissertation[1]. She discovered that BitTorrent traffic collected on UDP ports accounted for a sizable portion of IBR. Since it has unpredictable and bursty temporal behavior, it is difficult to leverage BitTorrent traffic for opportunistic inferences. She investigated the origin of the massive BitTorrent traffic. She also provided evidence that this traffic is a part of intentional or unintentional index poisoning attacks on the Distributed Hash Table (DHT). DHT index poisoning is the pollution of the DHT with erroneous entries. Her work also explains the strategies used in extracting precise signatures from the UDP payload of BitTorrent traffic and creating appropriate filters.

Given the advantages of Internet Background Radiation outlined in previous sections; it can be concluded that IBR is an essential building block for comprehensive monitoring, network analysis, and detection of macroscopic Internet events. However, irregularity in IBR signal, caused by erratic events is hampering its usage in Internet outage detection by the IODA. Thus, it is crucial to sanitize the Internet Background Radiation and make it ideal for the IODA project.

2.2 Experimental Tools

Analyzing and extracting crucial information from IBR is a challenging task, because of its large size. UCSD-NT collects approximately 50 GB of compressed IBR traffic every hour. To process this massive amount of IBR data, I use a wide range of tools and software in my thesis.

2.2.1 Corsaro

Corsaro[23] is an open source software suite for performing large-scale analysis of trace data. It is specifically designed to be used with IBR captured by network telescopes. Corsaro allows quick analysis of trace data on a per-packet basis and provides a mechanism for aggregating results based on customizable time intervals. The trace analysis logic is separated into a set of plugins, that can be added or removed as per user's requirements. In addition to the Core Plugins that are distributed with Corsaro, the plugin framework makes the creation of new plugins as simple as possible.

The Corsaro distribution also includes several other supporting tools for fundamental analysis of Corsaro output data. In this thesis, these tools are used in processing the flowtuple data. Moreover, the proposed solutions are developed as plugins in Corsaro.

2.2.2 Flowtuples

Examining network traffic from pcap files is a time and effort-intensive task. In most of the studies, only header fields are required to understand the traffic instead of the complete packet. An alternative approach to analyzing and parsing a humongous pcap file is to extract the important parts of the header for quick analysis. To extract the important details from the pcap data, Corsaro creates flowtuples. Flowtuple only includes the eight tuple flow information namely: i) source IP, ii) destination IP, iii) source port, iv) destination port, v) protocol id, vi) TCP-flags, vii) time to live and viii) packet length. Converting pcap data to flowtuple format makes it easy and accessible for researchers in their studies.

2.2.3 Charthouse

Charthouse[12] is a time series visualization tool created by CAIDA to monitor and visualize Internet background radiation captured by UCSD Network Telescope. This tool is capable of exploring real-time and historical time-series data. Charthouse provides an interactive platform to build custom visualizations using specialized post-processing functions. In my thesis, I have used Charthouse to analyze IBR traffic and also to generate relevant graphs.

Chapter 3

Analysis of Irregular events

Previous chapters show that Internet background radiation collected at the UCSD-NT is an invaluable source of information for opportunistic network analysis. IBR demonstrates typical traits which make it an ideal data source in detecting macroscopic Internet connectivity outages. The IODA[24] project uses a stable signal extracted from IBR to detect outages. It starts its detection by geolocating source IPs in IBR to determine the traffic contribution of countries and regions. Internet connectivity disruption in a region results in less IBR traffic generation (drop in packet rate). In the event of a substantial packet drop from a region, IODA sends an alert regarding the possibility of an Internet outage, which is further confirmed by the results of BGP and active probing.

Ideally, the signal extracted from IBR to be used in IODA should be strong (statistically significant), stable (low noise), and globally pervasive (seen in most networks). IBR has been previously sanitized by filters created by manual driven methods[19][8][1] before 2015. These filters effectively remove erratic data (irregular traffic) from a predictable pattern of IBR. Figure 3.1 portrays non-erratic signal extracted from IBR by deploying filters. However, with time, new events have emerged with intermittent erratic behavior. Figure 1.5 shows that the last three years of IBR traffic collected at the UCSD-NT is erratic with random spikes and drops. Erratic traffic

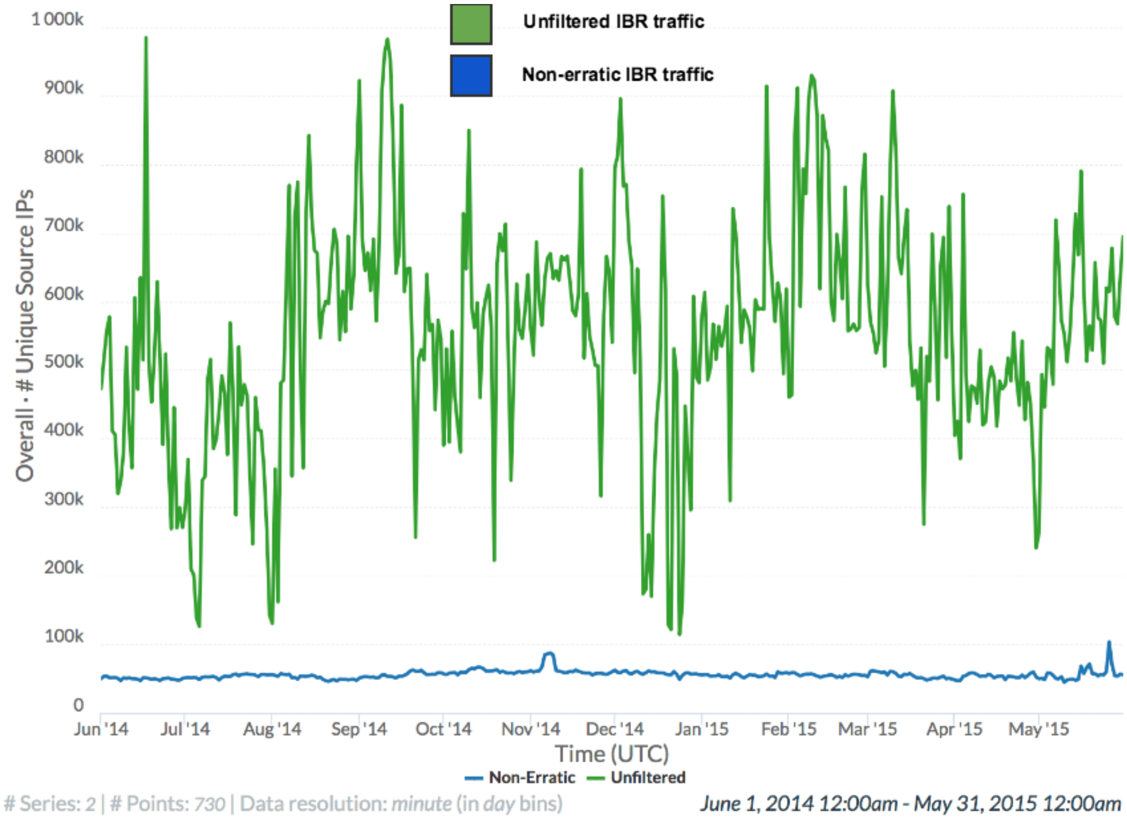


Figure 3.1: Time series graph of unfiltered IBR traffic and non-erratic IBR traffic collected at the UCSD-NT. Both traffics are represented in terms of unique source IPs. Non-erratic IBR traffic is obtained by deploying filters mentioned in [19][8][1].

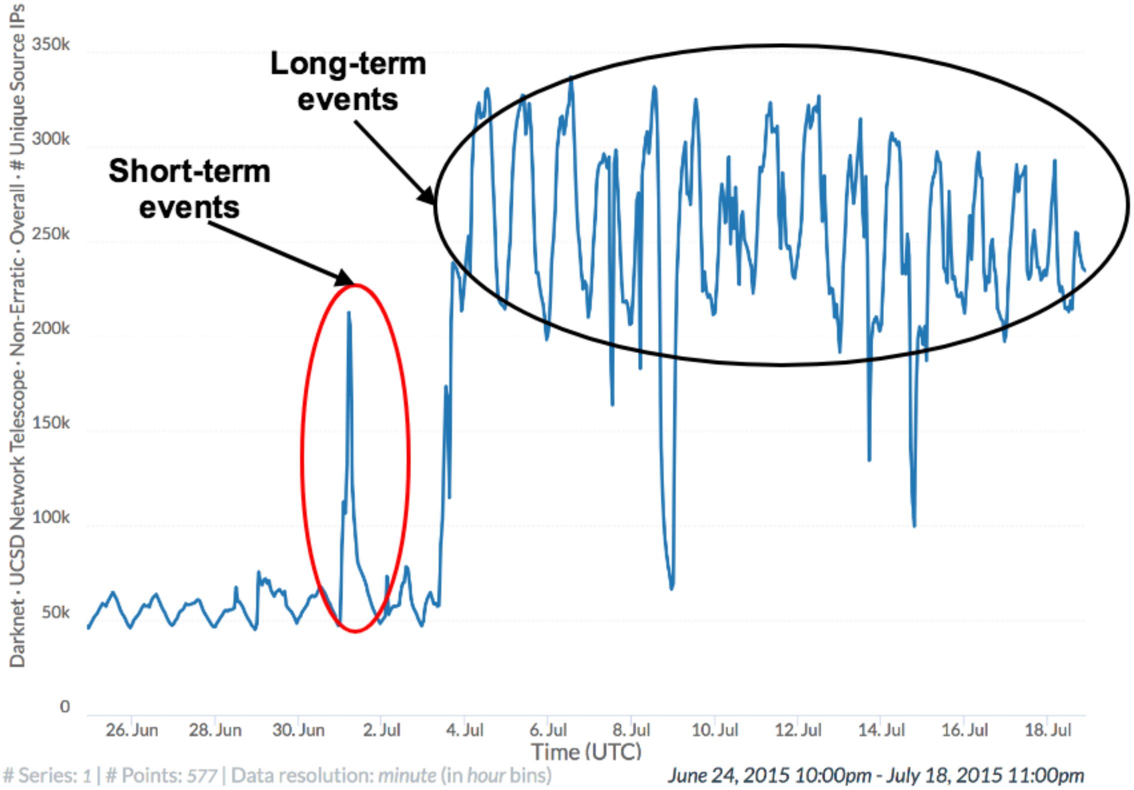


Figure 3.2: Illustration of short-term and long-term events. Traffic bursts with duration of smaller than one day are called short-term events. Traffic bursts with duration of more than one day are called long-term events.

in IBR interferes with IODA's outage inference system. Therefore, to make IODA's inference system effective, new events generating erratic traffic need to be analyzed and filtered.

The first step towards sanitizing Internet background radiation is to understand the traffic responsible for the erratic behavior in IBR. Understanding the traffic corresponds to identifying the commonly targeted ports, IPs, protocols, etc. After identifying the components of the erratic traffic, signatures can be extracted and used to create appropriate filters. Figure 1.5 depicts irregular bursts of traffic observed after July 2015; thus, I started my analysis by studying the traffic of July 2015.

In my analysis, I encounter many erratic events in IBR. These events are identified and classified into two major categories based on their duration (Figure 3.2): i) short-term events (exists for less than one day) and ii) long-term events (continues for more than one day). The rest of the chapter contains a detailed analysis of both types of events. The basis of the categorization is the difference in their respective nature and duration of traffic. As the name suggests, short-term events are non-persistent and must be removed. Coarser filters can be used in removing them, without much overhead and traffic loss. On the other hand, removing long-term events is not always desirable. In some scenarios, long-term events generate persistent and continuous traffic, maintaining the regular behavior of IBR signal. Thus, long-term events should be first analyzed and if need be, removed.

In my thesis, I use two types of data: flowtuple and pcap. First, I inspect IBR traffic using flowtuple data to gain insights about its traffic composition, followed by investigating pcap data, by parsing payload of network packets.

3.1 Short term events: Traffic bursts

Short-term events are non-persistent network glitches which result in generating or redirecting bursts of Internet traffic. These events contribute toward an intermittent and erratic signal in IBR traffic. Since these short-term events last for a small duration of time (less than one day), they can cause substantial packet rate variations in IBR. These traffic bursts can be caused either by malicious (backscatter, botnet scans) or benign (misconfiguration, bugs, erroneous entries in DNS) activities.

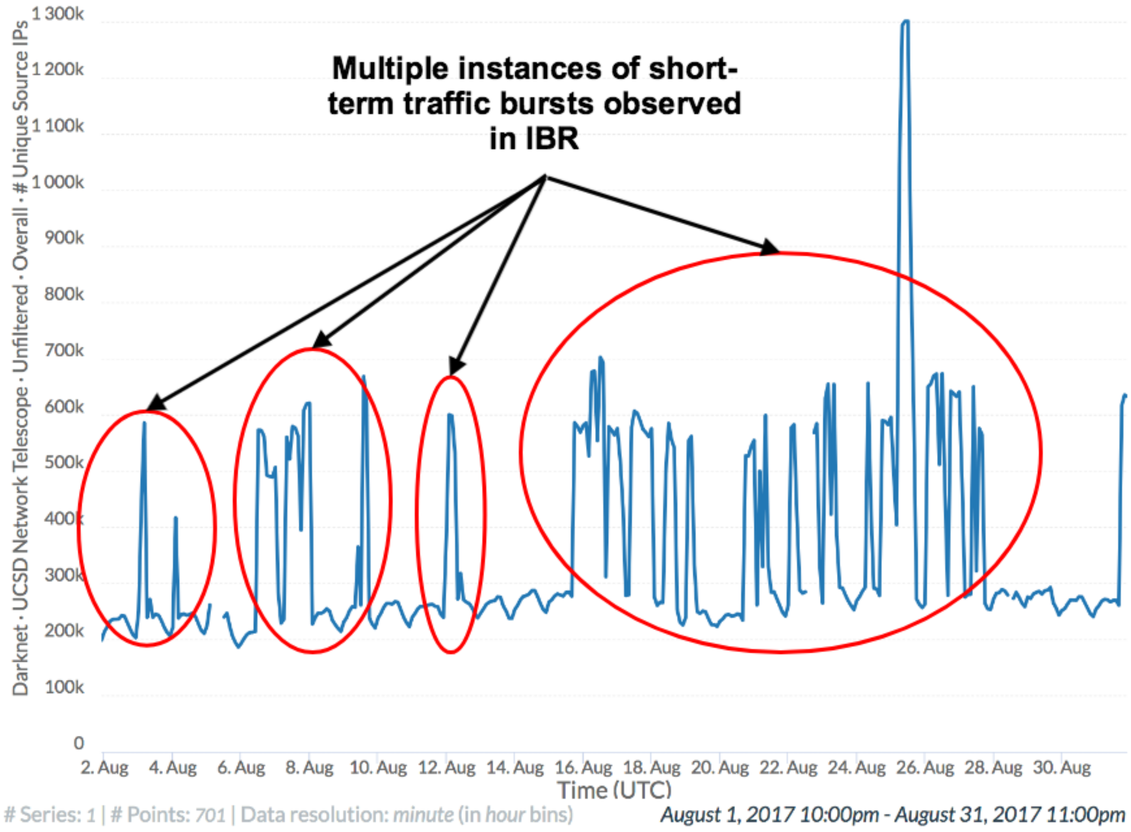
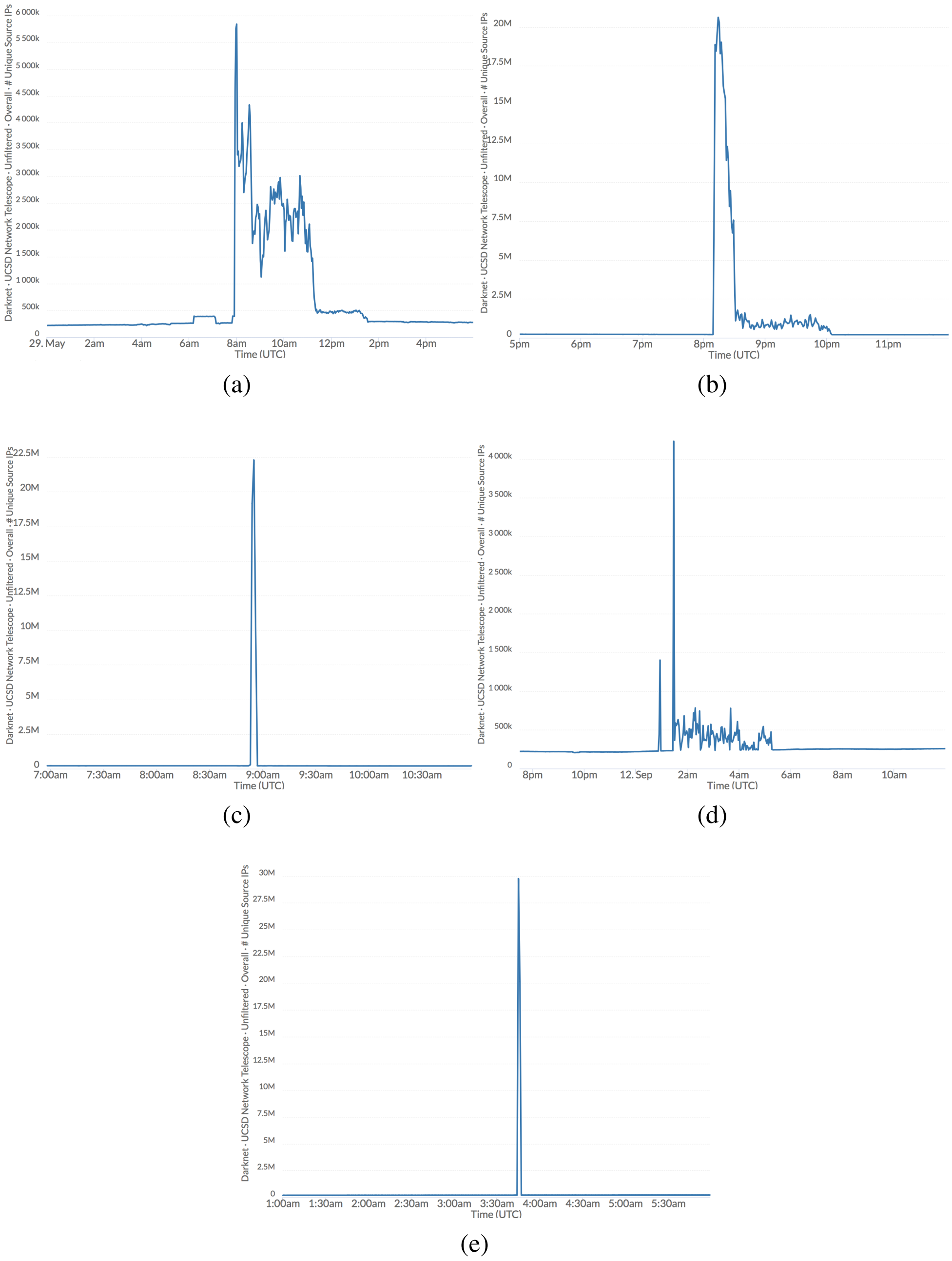


Figure 3.3: Multiple instances of traffic bursts generated by short-term events, observed in IBR traffic (in terms of unique source IPs). The graph shows the IBR traffic captured by the UCSD-NT for August, 2017.

Presence of such intermittent bursts of traffic in IBR affects the utility of IBR in inferring opportunistic network measurements. In my analysis of the last three years of IBR traffic captured by the UCSD-NT, I observe many short-term events. Figure 3.3 portrays multiple instances of traffic bursts generated by short-term events in IBR collected at the UCSD-NT. These traffic bursts interfere with the detection of Internet connectivity outages. Due to the nature of these packet rate variations, traffic bursts can overwhelm the predictable signal of IBR, generated from countries with few IP addresses. Since IODA's heuristics check for drops and deviations in the predictable signal of IBR, bursts can significantly affect the effectiveness of its use for IODA inferences.



(a)

(b)

(c)

(d)

(e)

Figure 3.4: Example of different types of individual bursts observed in IBR. a) Traffic burst observed on a single destination port 5673 on 29 May 2015. b) Traffic burst observed on 4 July 2017 on destination IP X.217.86.34 and destination ports: 27015-28014. c) Traffic burst observed on 26 July 2017 on destination IP X.18.32.14 and destination ports: 27015-28014. d) Traffic burst observed on 12 Sept 2017 on destination IP X.217.31.103 and destination port 34001. e) Traffic burst observed on 7 Oct 2017 on destination IP X.33.13.233. The figures indicate that that short-term traffic bursts usually target a few set of destination IPs and destination ports.

The first step towards removing these erratic bursts is to study and characterize a common signature and pattern present in individual bursts. To study traffic composition of these short-term events, I extracted the traffic generated by them. I studied half a dozen of randomly selected short-term traffic bursts from IBR. On analyzing few of these individual events, I observe that short-term traffic bursts usually target a few set of destination IPs and destination ports. Figures 3.4 depict individual traffic bursts observed in IBR signal, affecting specific destination ports and IPs. These set of destination IPs and ports can be used as a criterion in creating filters to remove short-term bursts[19][1].

3.2 Long term events: SYNRENT

After the analysis of short-term events (traffic bursts), I investigate long-term events which have a lifespan of more than one day. Figure 1.5 depicts that non-erratic IBR signal is consistent till 4th July 2015 due to previously deployed filters. However, 4th July 2015 onwards, UCSD-NT started capturing an enormous amount of erratic traffic, which was not filtered by the existing filters. The average number of unique source IPs soared from ∼50,000 to ∼350,000 per day. This rise in the number of unique source IPs in overall IBR signal is the result of an increase in TCP component of IBR. I have termed this event ”SYNRENT” after its composition of TCP-SYN packets and its relevance to BitTorrent.

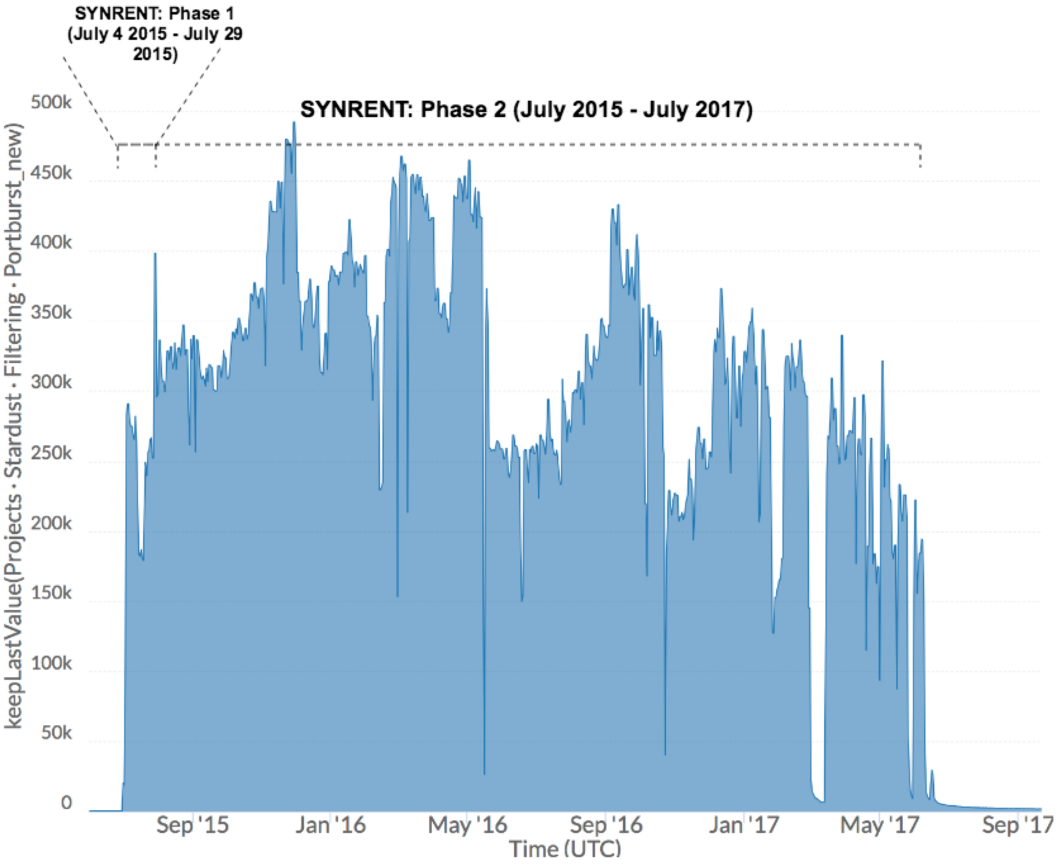


Figure 3.5: Time series graph of SYNRENT traffic in terms of unique source IPs collected at the UCSD-NT. It categorizes the SYNRENT traffic into two phases depending upon the nature of the traffic, in terms of affected destination IPs and destination ports. The first phase started on 4th July 2015 till 29th July 2015. The second phase started on 29th July 2015 and continued till June 2017.

UCSD-NT observes SYNRENT in two phases as per its extracted signature. The major differences in both of these phases are the destination ports and address spaces targeted. Figure 3.5 illustrates both the phases of SYNRENT traffic. Phase one started on 4th July 2015 and continued till 29th July 2015. Phase two started on 29th July 2015 and continued for two years till June 2017. The following subsections entail the detailed analysis and traffic composition of both the phases of SYNRENT.

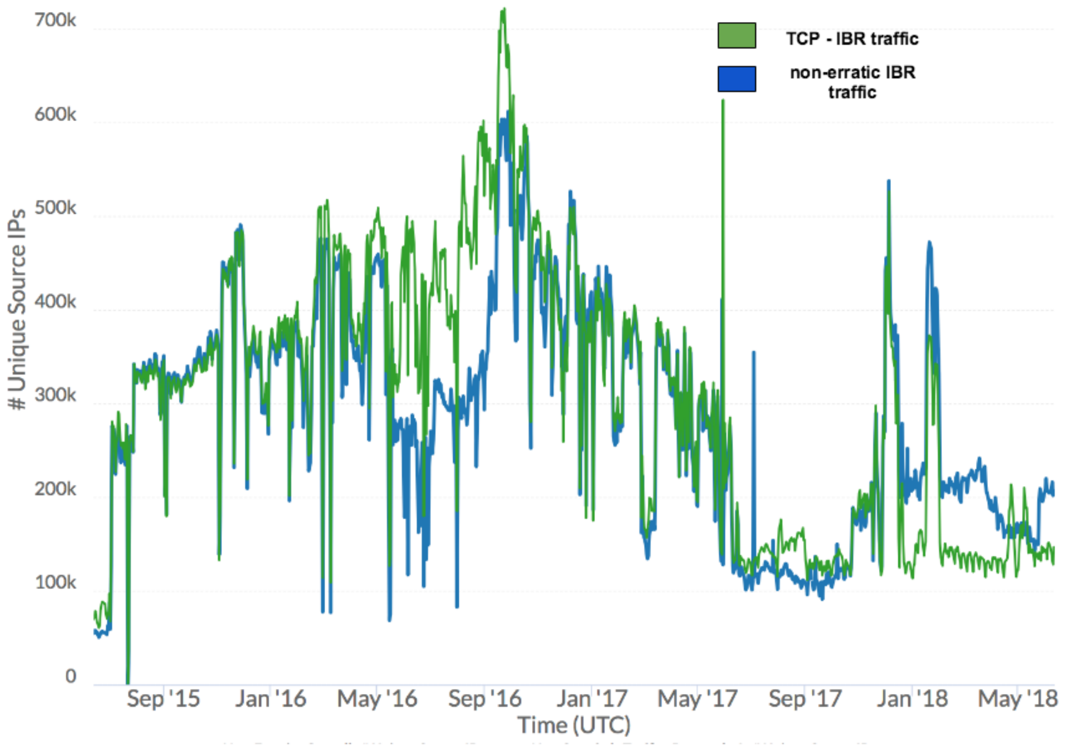


Figure 3.6: Time series graph of non-erratic IBR and TCP traffic collected at the UCSD-NT for previous three years. The collinearity in the graph shows that TCP traffic is primarily responsible for the rise in non-erratic IBR traffic.

3.2.1 SYNRENT: Phase 1

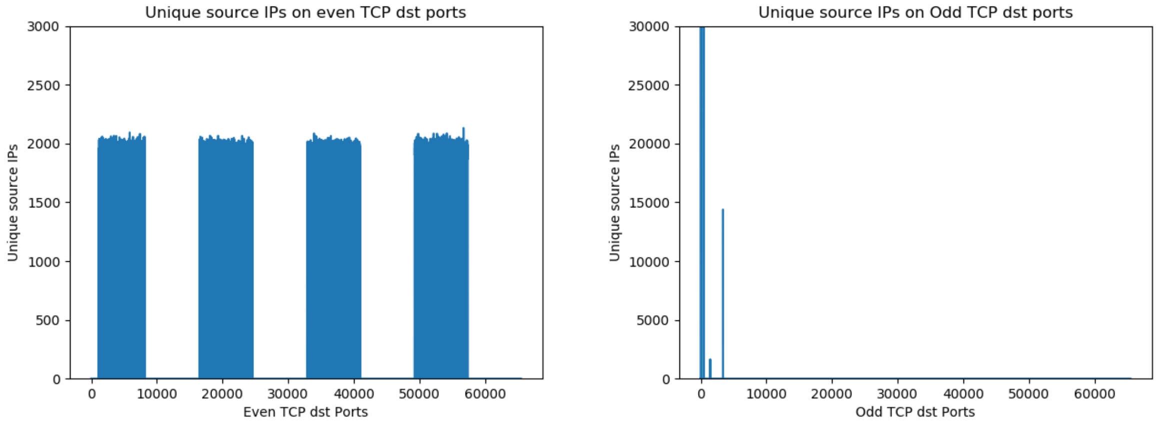
On 4th July 2015, UCSD-NT observed an unusual increase in non-erratic IBR traffic. Previously deployed filters failed to remove the recently observed erratic traffic. The pattern formed by non-erratic signal of IBR resembled the overall behavior of TCP traffic captured by the UCSD-NT. The collinearity in Figure 3.6 demonstrates that TCP traffic is primarily responsible for the rise in non-erratic IBR traffic. To begin my analysis, I studied the traffic composition corresponding to TCP traffic. I extracted the flowtuple data of two days (3rd and 4th July 2015). Flowtuple data contains compressed eight-tuple information of packet headers (see chapter 2).

Analysis of Destination Ports

To gain insights into SYNRENT traffic, I studied the TCP component of IBR in terms of destination ports. In my analysis, I observe that certain TCP destination ports receive more traffic than the others. Figure 3.7 illustrates the distribution of SYNRENT traffic in terms of unique source IPs collected on even and odd-numbered destination ports. The graphs show that SYNRENT traffic targets a specific set of even-numbered destination ports (Figure 3.7a). On the other hand, odd-numbered ports show no sign of any irregular traffic (Figure 3.7b). Traffic is observed on destination ports with the following bit pattern: destination_port & 0x2081 = 0x0080.

Table 3.1: Bit-pattern formed by the SYNRENT (phase one) targeted destination ports.

Bytes of destination port	Bit-pattern
1st byte	_ _ 0 _ _ _ _ _
2nd byte	1 _ _ _ _ _ _ 0



(a) Even-numbered TCP destination ports.

(b) Odd-numbered TCP destination ports.

Figure 3.7: Unique source IPs on even and odd TCP destination ports. The graphs show that SYNRENT traffic targets a specific set of even-numbered destination ports. Odd-numbered ports show no sign of any irregular traffic. Traffic observed on destination ports form the bit pattern: destination_port & 0x2081 = 0x0080.

After studying the distribution of SYNRENT traffic on destination ports, I observe that traffic collected on these TCP ports consists of TCP-SYN packets with TCP_FLAGS = 0x02. This obtained information (TCP protocol, TCP-SYN flags and destination ports) is combined to create a signature representing SYNRENT. I use this derived signature as a criterion to extract SYNRENT traffic for two days (48 hours). After analyzing the extracted traffic, I observe that the bit-pattern of destination ports remains constant till 29th July 2015.

Analysis of Destination IPs

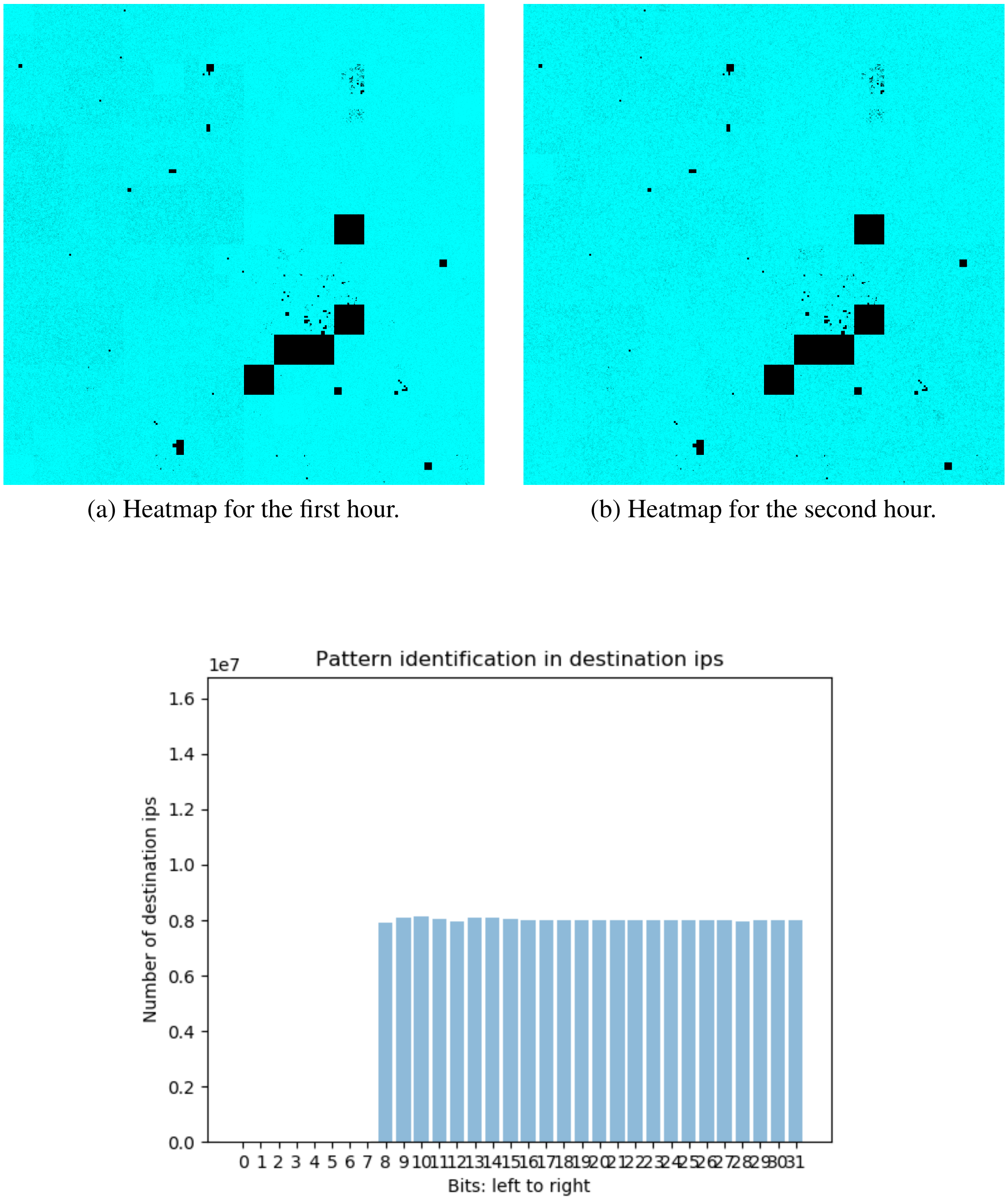
To further understand SYNRENT, I investigate its traffic in terms of destination IPs. This analysis provides a better understanding of the effect of SYNRENT on IBR traffic collected at the UCSD-NT. The Hilbert graph (Figure 3.8) highlights the address space of the UCSD-NT targeted by SYNRENT. The graph portrays that phase one of SYNRENT traffic is observed on the entire /8 address space of the UCSD-NT. The dark spots present in the graph represent the routed address blocks in the UCSD-NT. The sudden rise in the TCP component of IBR on specific destination ports raises concerns regarding the intent of the traffic. SYNRENT profoundly resonates with the classic TCP-SYN flood, attacking the entire UCSD-NT's address space. Subsection 3.2.3 discusses the possible causes of SYNRENT.

3.2.2 SYNRENT: Phase 2

The second phase of SYNRENT started on 29th July 2015 and continued for almost two years until June 2017. This phase of SYNRENT observes more traffic than the previous phase. Figure 1.5 depicts the increase in non-erratic IBR signal towards the end of July 2015. This new expanded traffic is similar to SYNRENT phase one traffic regarding its protocol and TCP-SYN flag.

Analysis of Destination Ports

Phase two traffic targets a different but specific set of destination ports, rather than the bit-pattern formed by phase one traffic. Phase two concentrates on a specific set of ports (19332-19624). Figure 3.9 depicts the SYNRENT traffic in terms of unique source IPs on destination ports. It shows that 175,000 unique source IPs are observed every hour on each of these TCP ports.

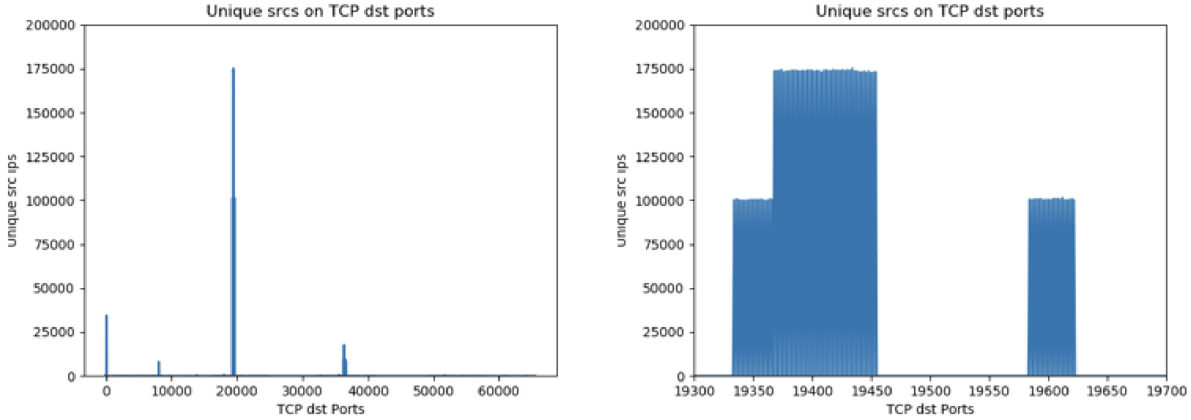


(a) Heatmap for the first hour.

(b) Heatmap for the second hour.

(c) Destination IP address graph representing the set bits in SYNRENT Phase 1.

Figure 3.8: Heatmap of destination IPs generated by the Hilbert curve using the SYNRENT Phase one traffic. The graphs (a and b) highlight the address space of the UCSD-NT targeted by SYNRENT for two different time periods. They show that phase one of SYNRENT traffic is observed on the entire /8 address space of the UCSD-NT for the entire duration till 29 June. c) Bit representation of destination IPs in the form of set bits.



(a) TCP Even Destination Ports.

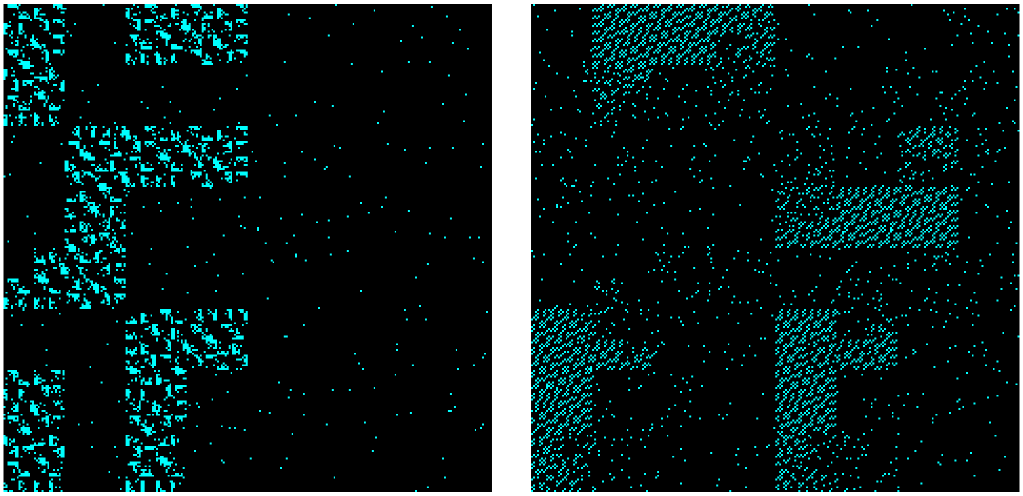
(b) TCP Even Destination Ports between 19300-19700.

Figure 3.9: Unique source IPs on even TCP Destination ports.

After obtaining the signature of phase two traffic (TCP protocol, TCP-SYN flags and destination ports), I extracted the TCP component of IBR for six days between 21st September 2015 and 26th September 2015. The data from these six days (144 hours) provides a holistic view of the entire SYNRENT phenomenon. I have used this data in all my further analysis and studies.

Analysis of Destination IPs

Hilbert curves (Figure 3.10) represent the address space of the UCSD-NT targeted by phase two of SYNRENT traffic. Instead of covering the entire address space, it forms a specific bit-pattern in destination IPs. Figure 3.11 and Table 3.2 portray the bit-pattern observed in the UCSD-NT. UCSD-NT IP addresses (X.B.C.D) satisfying: B & 0x88 = 0x00, C & 0x38 = 0x08 or 0x10 and D & 0x09 = 0x01, observes SYNRENT traffic. To validate the consistency in bit-pattern of destination IP, I plotted the Hilbert curves of the UCSD-NT address space for 144 hours. I observe that the pattern formed by SYNRENT in targeting specific destination IPs remains the same. The pattern followed in targeting the destination IPs and destination ports is unusual. A plausible explanation for this abnormal pattern can be a bug in the pseudo-random number generator (PRNG) used by an Internet application that determines which addresses and ports to target.



(a) Representing slash eight address space.

(b) Representing slash sixteen address space.

Figure 3.10: Heatmap of destination IPs generated by Hilbert curve in SYNRENT Phase 2. a) Each pixel represents /16 address block in the UCSD-NTs address space. b) Each pixel represents /24 address block in the UCSD-NT's address space.

Table 3.2: Bit-pattern formed by the SYNRENT (phase two) targeted destination IPs.

Bytes of destination IP (IPv4)	Bit-pattern
1st byte	x.x.x.x.x.x.x.x
2nd byte	0._._._.0._._._
3rd byte	_._.0.1.0._._._ _._.0.0.1._._._
4th byte	_._._._.0._._.1

Analysis of source IPs

Another important task in understanding the traffic is to examine the source IPs that generate SYNRENT. One of the main concerns is to see if the same set of source IPs is visible every hour. Figure 3.12 addresses the question by depicting the presence of source IPs and /24

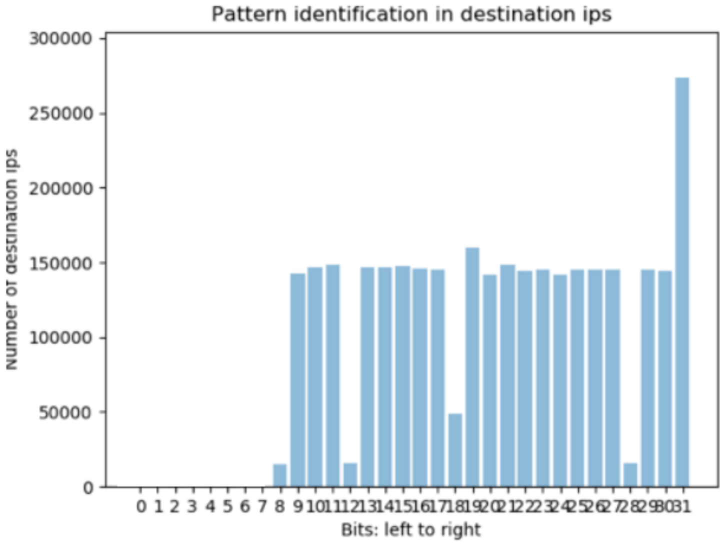


Figure 3.11: Destination IP address graph representing the set bits in SYNRENT Phase 2.

address blocks per hour for 144 hours. The first figure portrays that most of the source IPs are non-persistent, i.e., they are observed only for a short duration (1-2 hours), while the second figure shows that unlike source IPs, a few /24 blocks are observed every hour in the test dataset.

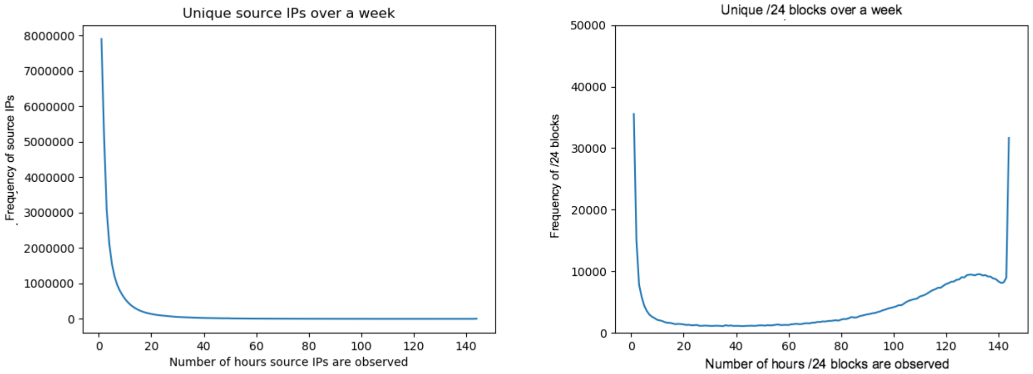


Figure 3.12: Graph represents the frequency (in terms of hours) of source IPs and /24 address blocks for six days (144 hours). The first figure portrays that most of the source IPs are non-persistent, i.e., they are observed only for a short duration (1-2 hours), while the second figure shows that unlike source IPs, a few /24 blocks are observed every hour in the test dataset.

Figure 3.13 helps in understanding the arrival of source IPs generating SYNRENT traffic.

It represents the arrival of new source IPs as a cumulative percentage over the period of 144 hours. The graph indicates that new source IPs are introduced almost every minute. An implementation bug or a misconfiguration in the networked device can redirect the traffic from a region resulting in the arrival of new source IPs every hour.

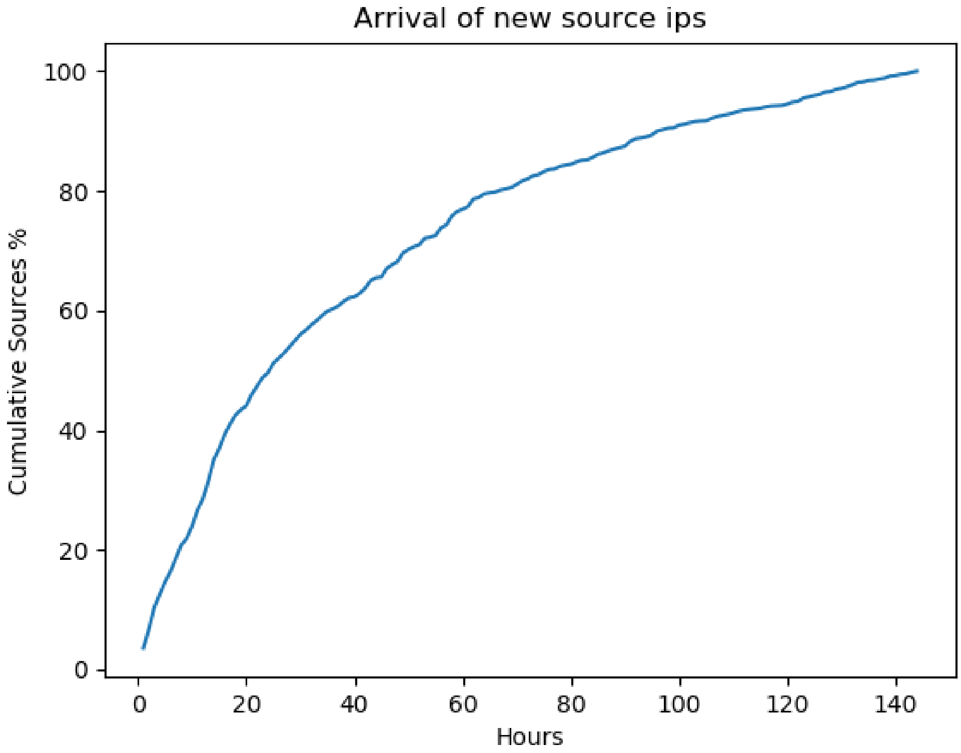


Figure 3.13: Arrival of new source IPs as cumulative percentage for 144 hours.

Analysis of source IPs in terms of operating systems

For understanding and characterizing the source IPs and traffic representing the phenomenon, I use the p0f tool [22] to extract machines/operating systems accountable for the received packets. The p0f tool is provided with the trace data corresponding to the SYNRENT traffic. It categorizes the packets according to different operating systems by identifying typical fingerprints in them. Table 3.3 lists all the OSes encountered while analyzing the SYNRENT

Table 3.3: Operating systems responsible for the SYNRENT traffic extracted by Michal Zalewski's p0f tool[22].

Operating System	Unique source IPs in SYNRENT traffic	Unique source IPs in Control traffic
Windows 7 or 8	45,623,881	11,183,037
Windows XP	12,542,438	29,570,919
Linux 3.1-3.10	8,192,539	5,460,600
Linux 2.2.x-3.x	4,887,910	9,460,195
Windows NT kernel 5.x	3,534,271	712,619
Unknown operating system	1,049,203	127,959,106
Windows NT kernel	486,855	1,841,275
Linux 3.x	327,632	2,063,906
Linux 2.2.x-3.x (no timestamps)	262,333	8,180,656
Linux 2.6.x	245,706	7,900,819
Linux 2.4.x	213,037	76,702,741

traffic and the control traffic (not SYNRENT). 82.7% and 16.4% of SYNRENT traffic are generated from Windows users and Linux users respectively. On the other hand, 15.5% and 39.3% of control traffic (not SYNRENT) is generated from Windows and Linux users respectively. These stats highlight that the participation of traffic from Windows users rose from merely 15.5% in the control traffic to 82.7% in the SYNRENT traffic. It shows that the Windows operating system is the primary source of SYNRENT traffic.

Analysis of source IPs in terms of countries and regions

Now that the operating systems behind the event have been identified, the next step is to determine the pervasive nature of the SYNRENT traffic. It is important to note whether the traffic is originating from all over the world or some specific region or country. To identify the origin of the SYNRENT traffic on a broader granularity of countries, I use an IP geolocation database to map every source IP to their respective countries. Table 3.4 illustrates the origin of SYNRENT traffic (in terms of unique source IPs per country/region) from all over the world. The table demonstrates that China is the most significant source, contributing around 99.76%

of traffic. Other Chinese speaking regions like Taiwan and Hong Kong, also contribute 0.07% towards the total SYNRENT traffic.

Table 3.4: Table representing the origin of SYNRENT traffic in the world. The data is obtained by geolocating source IPs. Out of 28.43M unique source IPs, 28.36M (99.77%) IPs originate from China.

Country/Regions	Number of unique source IPs	Percentage contribution in total traffic (28,431,740 IPs)
China	28,364,187	99.76%
Taiwan	15,134	0.05%
Russia	7,840	0.02%
USA	7,417	0.02%
Japan	6,164	0.02%
Hong Kong	5,836	0.02%
South Korea	2,901	0.01%
Malaysia	1,692	<0.01%
Singapore	1,673	<0.01%
Ukraine	1,573	<0.01%
Canada	1,290	<0.01%
Great Britain	1,061	<0.01%
Spain	954	<0.01%
Australia	927	<0.01%
Others(43 Regions)	13,091	<0.04%

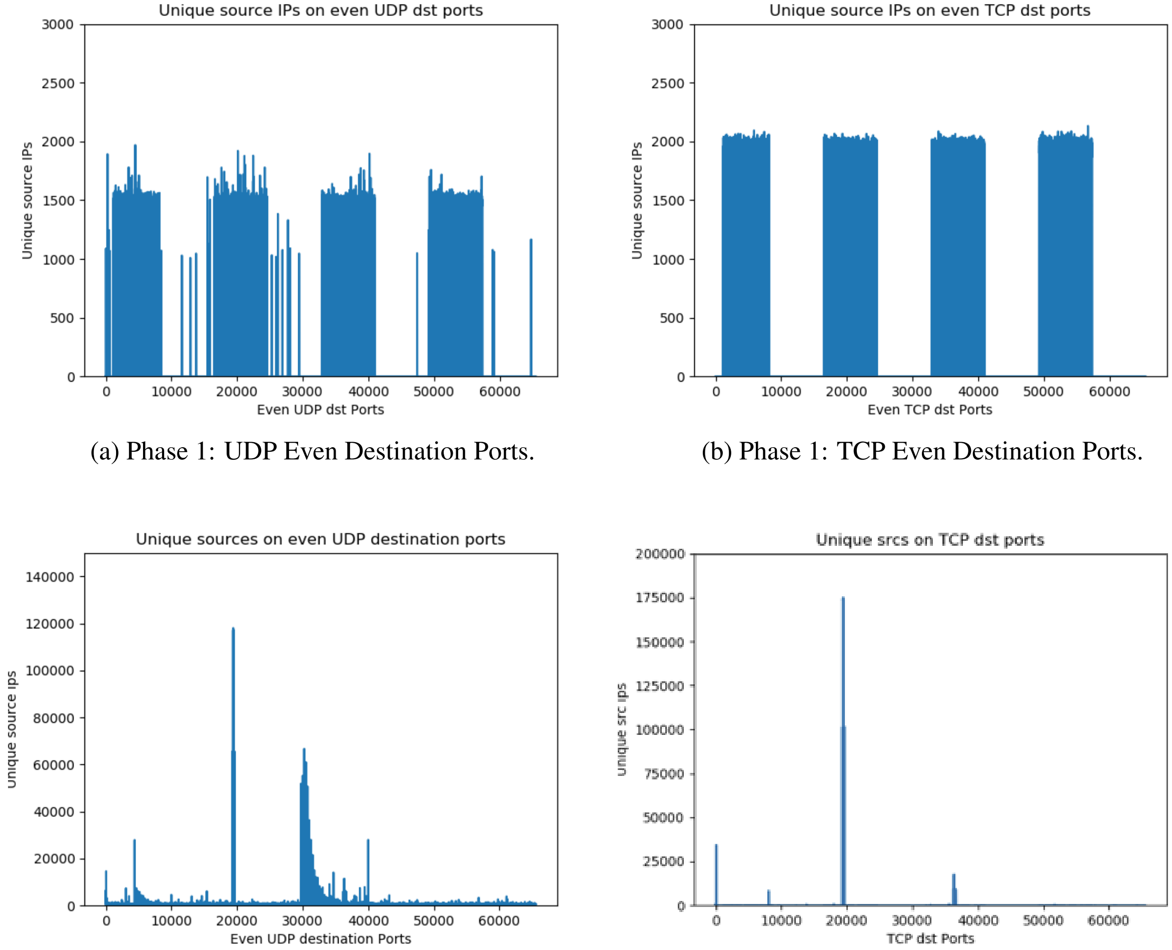
Analysis of source IPs in terms of autonomous systems

Studying the origin of the traffic at a finer granularity of autonomous systems provides significant insight into its origin. I extracted the autonomous systems of the source IPs to study the traffic pattern. It turns out that as many as 4,280 unique ASes are responsible for 28.43M unique source IPs. This distribution implies that a multitude of ASes cause and generate SYNRENT traffic. A plausible explanation behind this significant traffic from China could be the presence of a common malfunctioning component which affects multiple networks and autonomous systems.

For example, a bug in the implementation of Great Firewall of China can result in the redirection of Internet traffic of a significant magnitude.

3.2.3 Potential causes of SYNRENT

At first, the SYNRENT event is only observed and analyzed in TCP traffic, but surprisingly, even UDP traffic demonstrates similar characteristics. To investigate the association of SYNRENT with UDP traffic, I plotted UDP traffic on specific destination ports. Figure 3.14 depicts the traffic on UDP destination ports covering both phases of SYNRENT. After observing the traffic distribution on both TCP and UDP destination ports, I conclude that traffic on both the protocols follows an identical pattern. To further validate the hypothesis that traces of SYNRENT are visible in both protocols, I extracted the common source IPs that send traffic on selected TCP and UDP destination ports. Figure 3.15 represents the common source IPs sending both TCP and UDP traffic in the same hour bin.



(a) Phase 1: UDP Even Destination Ports.

(b) Phase 1: TCP Even Destination Ports.

(c) Phase 2: UDP Even Destination Ports.

(d) Phase 2: TCP Even Destination Ports.

Figure 3.14: Unique source IPs on even TCP and UDP destination ports during both the phases of SYNRENT.

On an average around 7.36M source IPs are common to TCP (47%) and UDP (93%) traffic per hour. These stats provide evidence that SYNRENT traffic on both of these protocols is tightly coupled. UDP packets contain payloads, which can be further parsed to study the nature of the traffic. On parsing the pcap data for UDP packets, for common source IPs on specific ports, I observe that UDP packets contain get_peers, ping queries of DHT protocol[25]. This discovery opens compelling doors for its connection with the BitTorrent traffic.

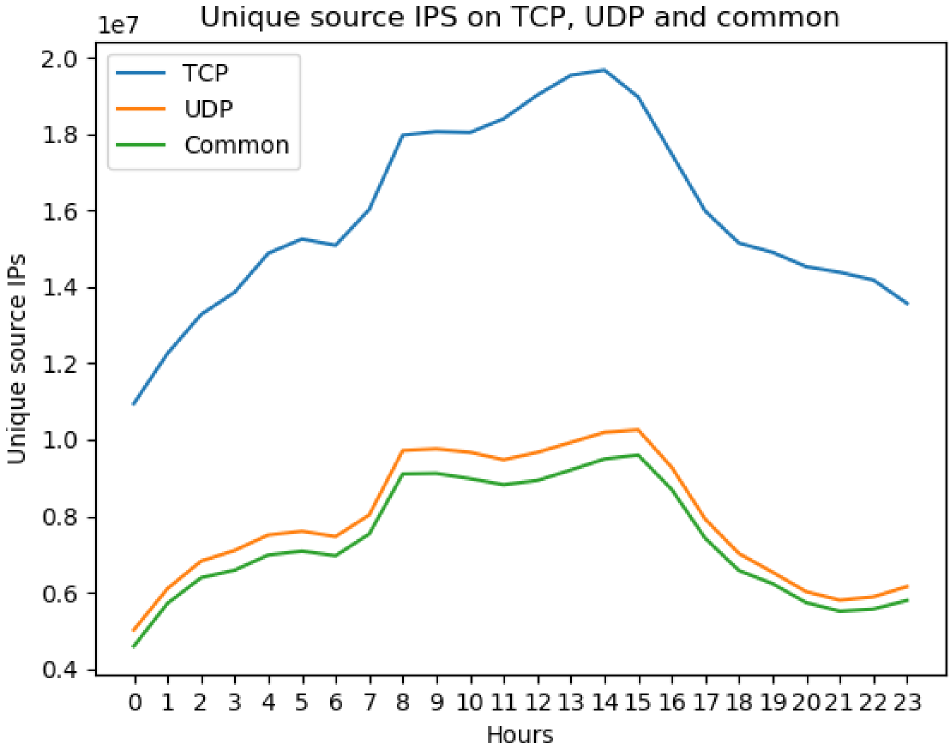


Figure 3.15: Common source IPs observed in TCP and UDP components of SYNRENT for 24 hours. The graph depicts that both TCP and UDP follow an identical pattern in source IPs, tightly coupling both TCP and UDP formats of SYNRENT.

Benson in her dissertation [26] studied the vast BitTorrent traffic captured by the UCSD-NT in the datasets of 2012 and 2013. BitTorrent traffic was in the form of Distributed Hash Table (DHT) queries (get_peers, find_node and ping) in UDP packets. She analyzed the source of these UDP packets and found the majority of the traffic originating from China. A potential reason for this sizable portion of BitTorrent traffic in IBR can be the presence of erroneous entries in DHT. These entries could have led many BitTorrent clients to search for trackers and seeders in the UCSD-NT address space. These incorrect entries may have further traversed amongst a large group of BitTorrent clients, resulting in the rise of BitTorrent traffic in IBR.

BitTorrent uses the DHT protocol on UDP to share metadata and location of the file in peers. On the other hand, TCP is used to share the contents of the file between seeders and clients[25]. Before July 2015, BitTorrent traffic in IBR was observed only on UDP ports in the form of DHT queries[26]. Erroneous entries in the DHT might have caused BitTorrent clients from Chinese speaking regions to misconstrue UCSD-NT address space as DHT nodes. This might have resulted in the massive inflow of UDP traffic. In July 2015, behavior in the BitTorrent traffic changed substantially. Its possible explanation could be that instead of sharing metadata, clients started TCP connections with DHT nodes. One of the causes of these TCP connections could be mistaking DHT nodes as seeders, or sending announce messages via TCP[27]. Often BitTorrent clients check the HTTP error codes provided by the seeders/servers to check for 404 error code (File Not Found) and consequently stop sending file requests[28]. Since UCSD-NT supports only one-way traffic, clients never receive 404 error response messages. BitTorrent clients never receive this message and keep on retrying TCP connections[29]. This activity might also be the reason behind the enormous rise in TCP-SYN traffic in IBR.

In 2008, the Chinese Government upgraded its censorship system by restricting access to all prominent peer to peer file sharing platforms like BitTorrent, PirateBay, etc[30][31]. Previously, users accessing these websites were redirected to the Chinese search giant Baidu[31], until 2015, when multiple network security reports and blogs [32][33] claimed the presence of DoS attacks from BitTorrent clients from China. Several websites across the world were affected by a large number of TCP-SYN packets[34]. The source IPs responsible for these attacks originated from China. SYNRNET could be either caused by an intentional (user exploiting China's censorship system and redirecting the traffic to target random address spaces) or unintentional (DNS bug in the Great Firewall of China or erroneous entry in the DHT redirecting requests for Chinese files and videos) activities.

Chapter 4

Proposed solution

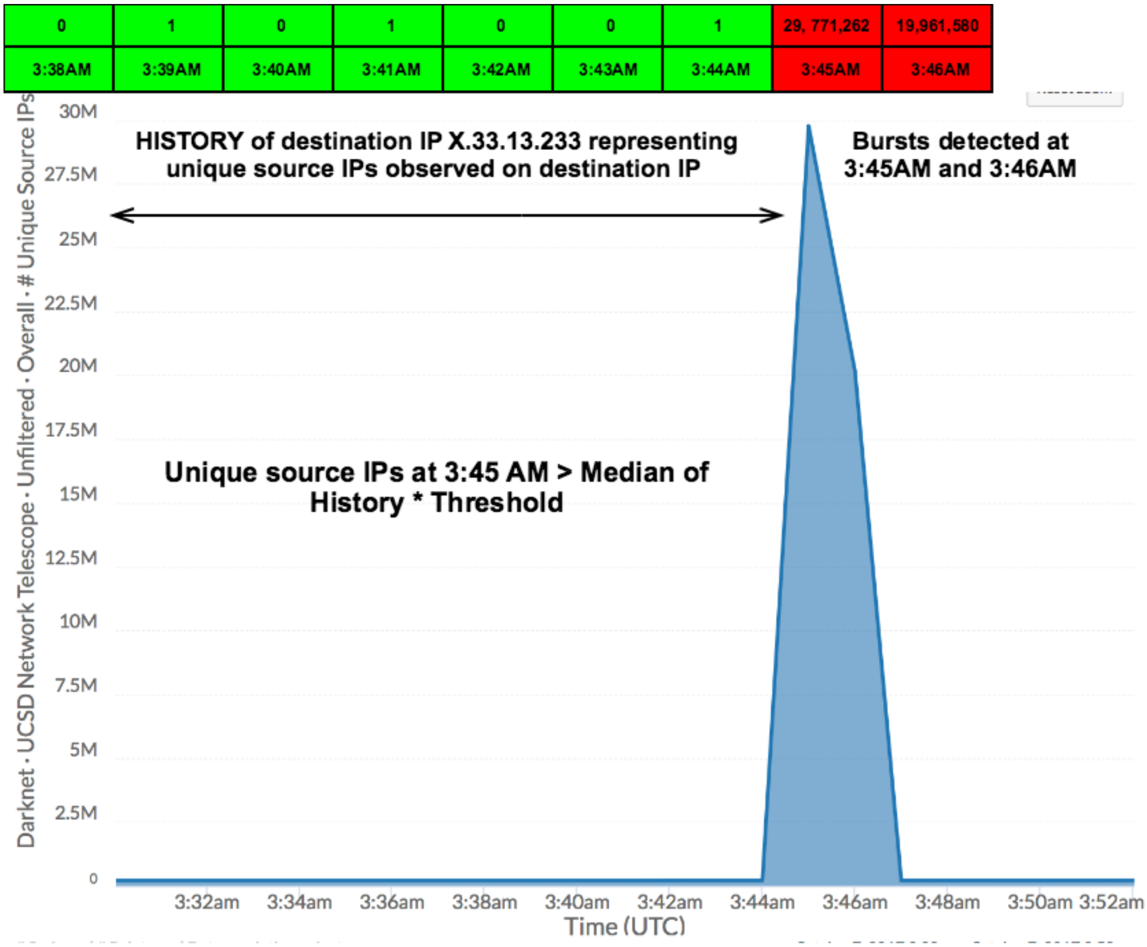
Internet background radiation is a complex mixture of many Internet phenomena. These phenomena generate enormous bursts of erratic traffic. Ideally, IBR signal required by the IODA project needs to be consistent and stable with predictable traffic behavior. Presence of erratic traffic can paralyze its ability to detect outages. Thus, sanitizing IBR is an essential task in detecting outages effectively. Previously, IBR was sanitized by manually identifying signatures of the erratic traffic and creating appropriate filters to remove them. However, with time significant amount of evolved traffic has contaminated the stable signal extracted from IBR (refer Figure 1.5). The existing filters still work but are insufficient because of the emergence of new unprecedented anomalous traffic. This traffic distort the stable signal extracted from IBR, by introducing intermittent erratic traffic. There is an immediate need to come up with a robust solution that does not only detect and identify erratic traffic but also processes large chunks of data in real-time with minimal human intervention.

In previous chapters, I analyzed and studied the events responsible for the erratic traffic in IBR. Analysis of both short term and long term events provide relevant insight into the characteristics of the erratic traffic. Since I use the metric of unique source IPs per minute, its usage excludes the effect of traffic caused by bursts of packets from a few hosts. In my analysis, I have observed that erratic traffic (both short-term and long-term) observed after July 2015 is different from pre-July 2015 erratic traffic. Most events target a small set of destination IPs and ports. Based on this knowledge, I have devised a software solution to detect most of these events that culminate in traffic bursts in IBR.

This chapter is organized as follows: Section 4.1 describes the proposed solution with the detailed architecture, Section 4.2 reports the results obtained after applying the solution. Section 4.3 sheds light on interesting findings while using the solution.

4.1 Solution architecture

The proposed solution is structured in two steps. Step one is detecting erratic traffic and identifying the destination IPs and destination ports affected by this sudden rise in traffic rate. Once relevant ports and IPs are identified, step two uses these detected ports and IPs in creating appropriate filters to remove erratic traffic from IBR. The software solution is developed as a plugin in Corsaro[35], which is a large-scale trace data analysis tool (see Chapter 2).



0	1	0	1	0	0	1	29, 771,262	19,961,580
3:38AM	3:39AM	3:40AM	3:41AM	3:42AM	3:43AM	3:44AM	3:45AM	3:46AM

Figure 4.1: Detection of a burst using history of previously observed unique source IPs sending to specific destination IPs. At 3:45 AM, destination IP X.33.13.233 observed 29.7M unique source IPs. This value is greater than the median value of the stored history times the threshold index. The user can configure the value of the threshold. For this experiment, the value of the threshold used is 500.

The method of detecting erratic traffic has two main components: i) Hostburst detection, and ii) Portburst detection. Both of these components inspect the IBR traffic rate, in terms of unique source IPs observed per minute. As the name suggests, Hostburst detects erratic traffic on destination IPs while Portburst detects erratic traffic on destination ports. The critical aspect of these approaches is keeping history in one-minute bins (e.g., 30 minutes oh history) to store previously observed unique source IPs on all 16M destination IPs (UCSD-NT) and 65,535 destination ports. Size of history is configurable by users.

For every new minute, the current traffic is checked against the median value of history. If the number of unique source IPs in the current minute overshoots the median value of the history, a traffic burst is detected. Hostburst and Portburst report specific destination IPs and ports, which observe erratic traffic. These detected ports and IPs are further used in creating filters. Overshooting criteria is configured by a user-defined threshold value, which determines the difference in the current traffic and the traffic observed in the history. Figure 4.1 illustrates an example of using history in detecting erratic traffic.

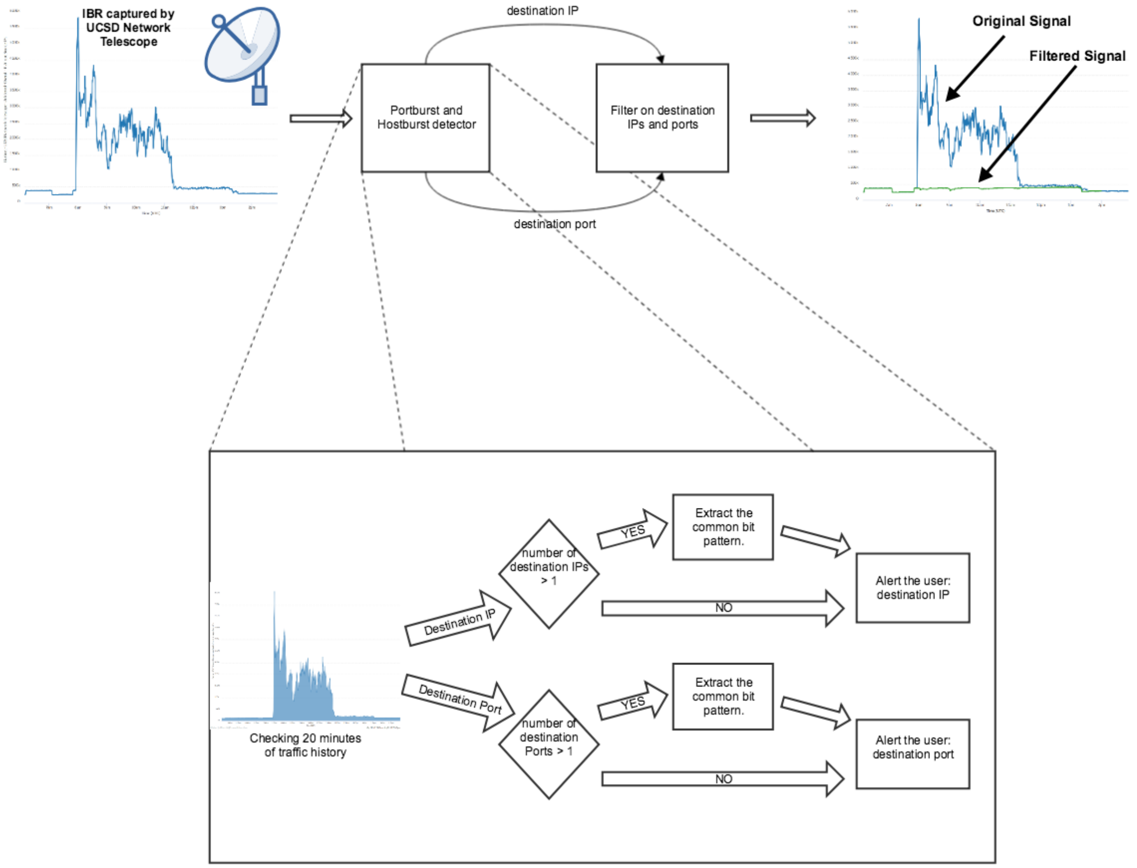

IBR captured by UCSD Network Telescope
destination IP
Portburst and Hostburst detector
Filter on destination IPs and ports
destination port
Original Signal
Filtered Signal
Checking 20 minutes of traffic history
Destination IP
Destination Port
number of destination IPs > 1
number of destination Ports > 1
YES
YES
NO
NO
Extract the common bit pattern.
Extract the common bit pattern.
Alert the user: destination IP
Alert the user: destination port

Figure 4.2: Architecture of history based Hostburst and Portburst software solution.

Figure 4.2 demonstrates the overall architecture of the proposed solution. When Hostburst and Portburst detect erratic traffic, they extract the corresponding destination ports and IPs associated with the traffic. Previous chapters show that instead of targeting a single destination port and IP, many events (e.g., SYNRENT, short-term bursts) target a set of ports and IPs, forming a specific bit-pattern. If the number of detected ports and IPs are greater than one, Hostburst and Portburst are capable of identifying the bit-pattern and reporting it to the user.

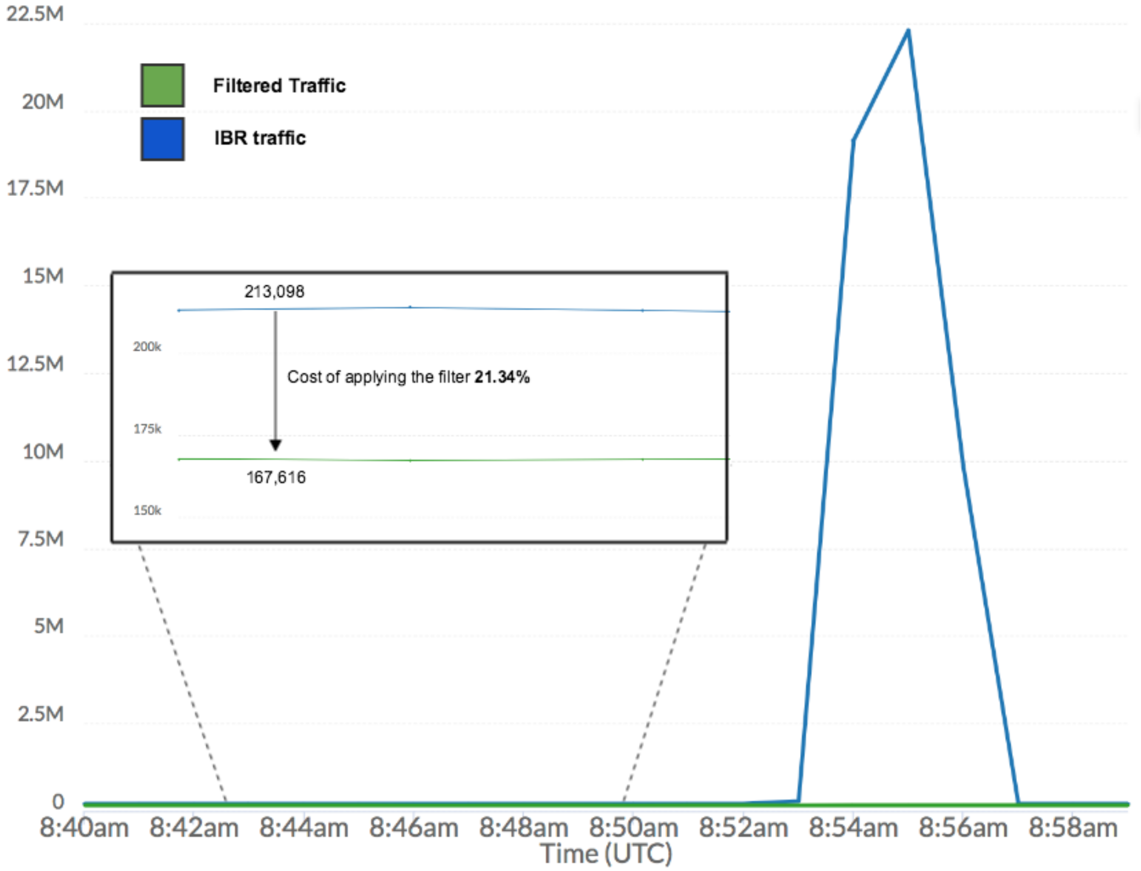


Figure 4.3: Example of the cost incurred in applying the filter. Filtering IBR traffic on IP xxx.18.32.14.21 observes 34% reduction in the number of source IPs generating regular, predictable traffic.

It is, however, important to note the limitations of the proposed solution. Since the deployment of filters removes the entire traffic from selected ports and IPs, the filters also remove the regular and stable traffic present on those destination ports and IPs. This loss of stable traffic causes a downward shift in the non-erratic IBR signal (Figure 4.3). It creates a trade-off between using the filter and discarding it. This cost is referred to as the filter cost associated with applying the filter.

4.2 Results

Along with proposing the solution to the problem of identifying erratic traffic, my thesis also provides results to justify its accuracy. To validate the correctness, I have selected multiple instances of erratic traffic identified by Hostburst and Portburst. To test my solution, I ran it for a period of 10 months from January to October in 2015 and on specific periods in 2017.

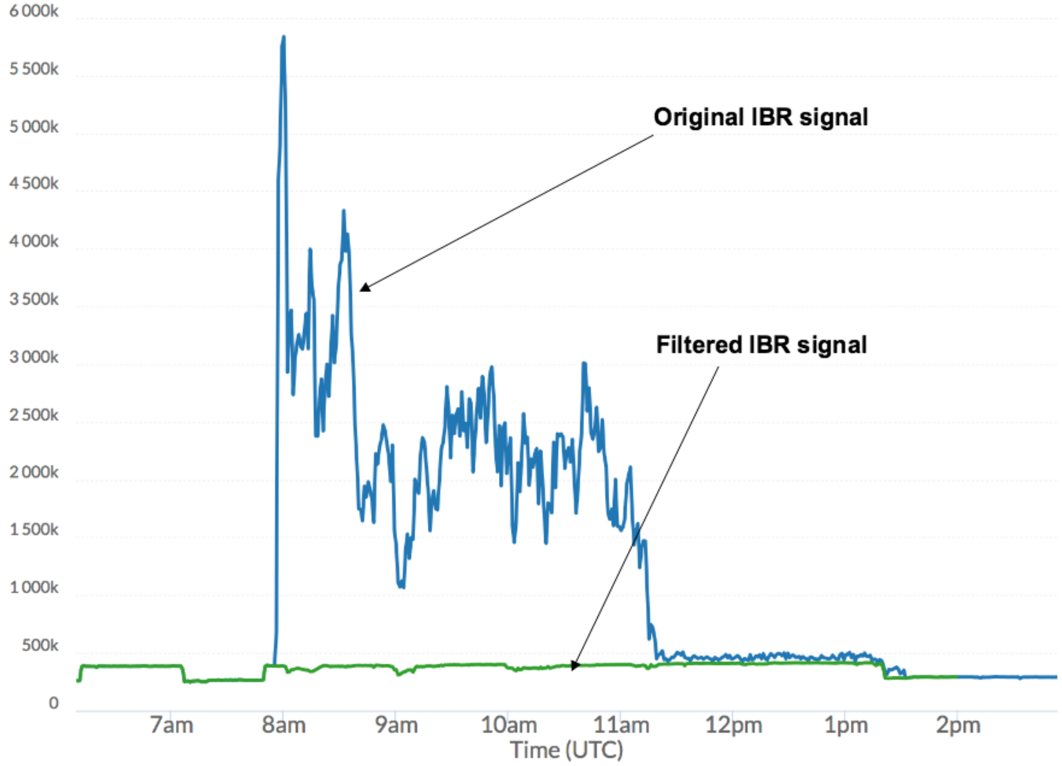


Figure 4.4: Detecting and filtering traffic burst on 29th May 2017 on destination port 5673.

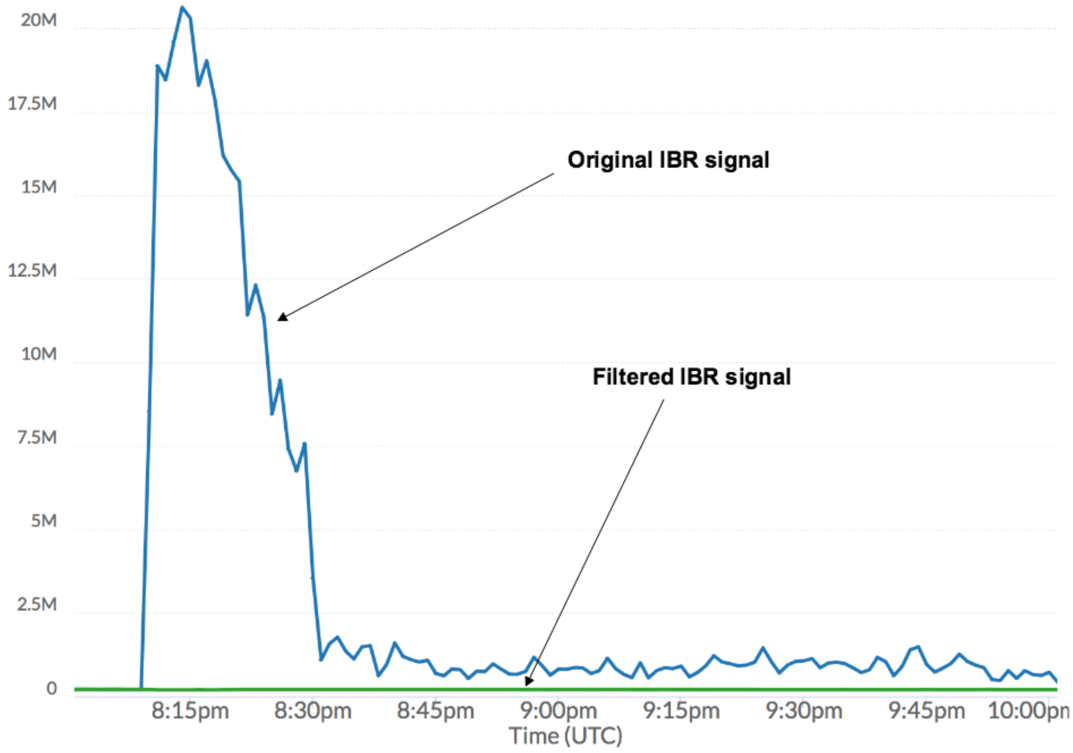


Figure 4.5: Detecting and filtering traffic burst on 4th July 2017 on destination IP X.217.86.34. A range of ports 27015-28014 received the traffic with bit pattern dst_port & 0x3C80 = 0x6900.

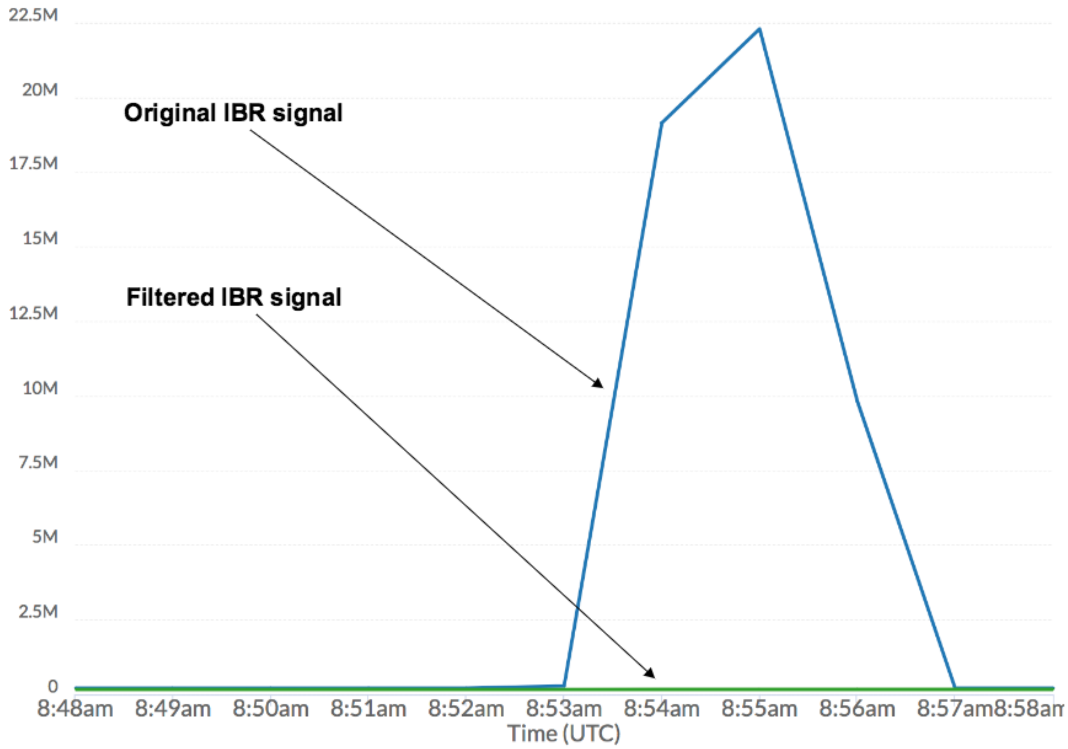


Figure 4.6: Detecting and filtering traffic burst on 26th July 2017 on destination IP X.18.32.143. A range of ports 27015-28014 received the traffic with bit pattern dst_port & 0x3C80 = 0x6900.

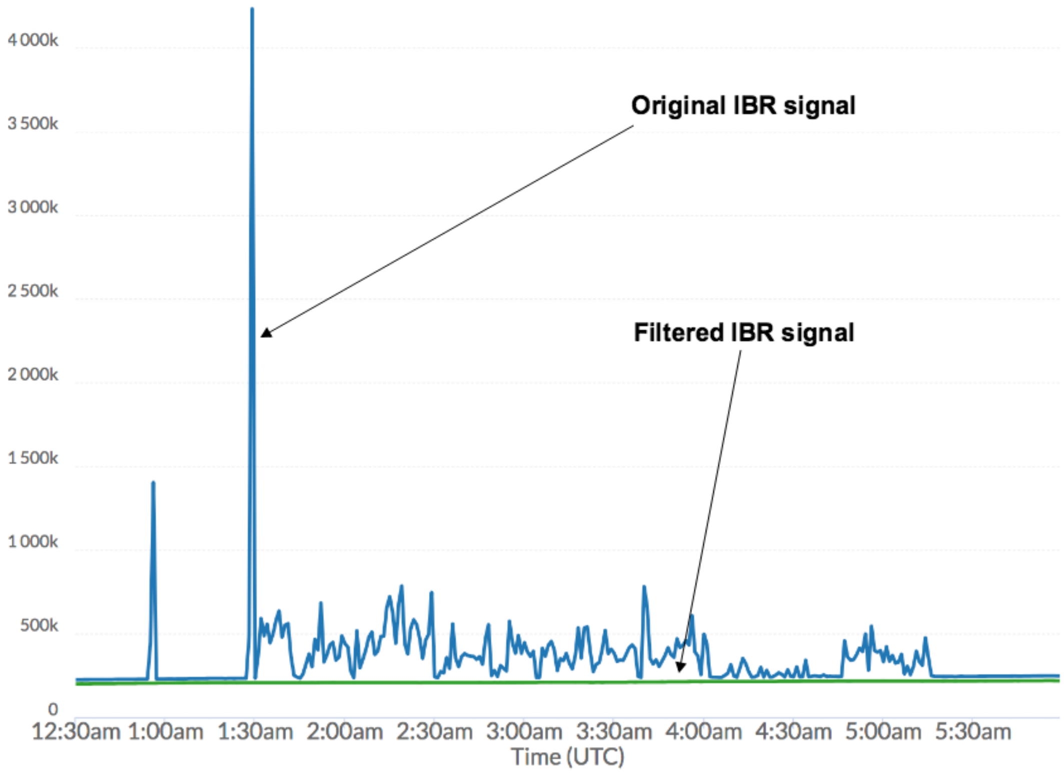


Figure 4.7: Detecting and filtering traffic burst on 12th Sept 2017 on destination IP X.217.31.103 and destination port 34001.

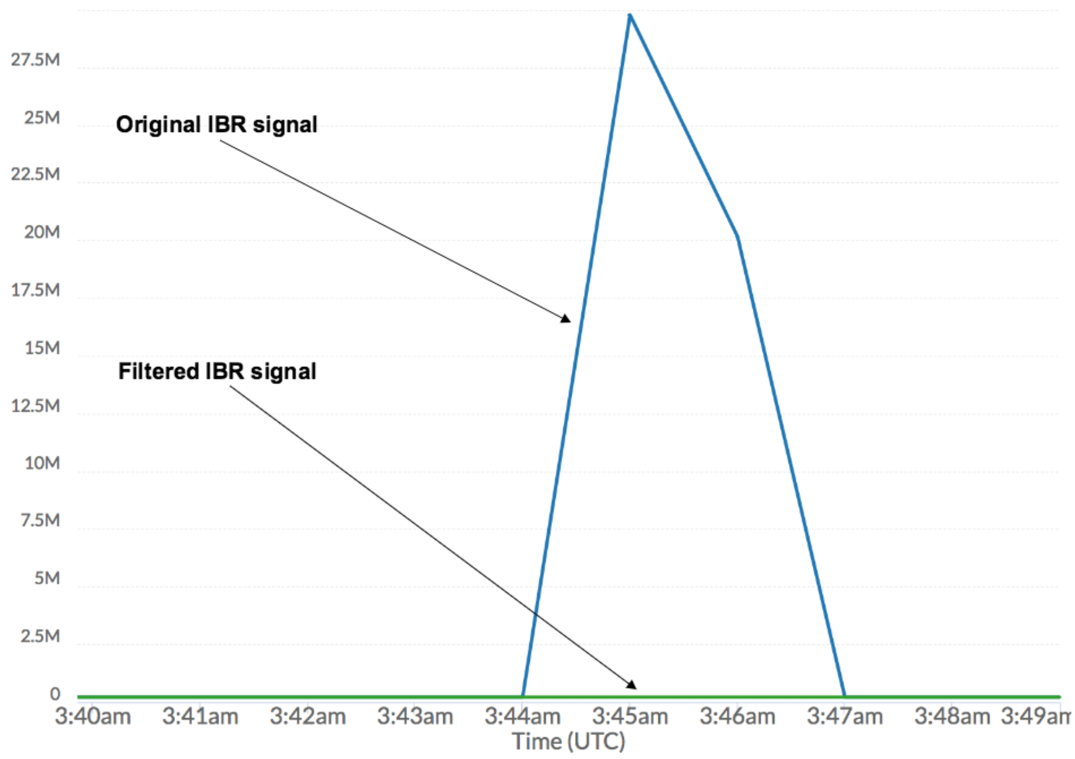


Figure 4.8: Detecting and filtering traffic burst on 7th Oct 2017 on destination IP X.33.13.233.

Figures 4.4-4.8 prove that the proposed solution does an efficient work in identifying and removing erratic traffic from IBR signal. This sanitization of IBR signal makes the extracted signal stable and ideal for effective outage detection. This validates that Hostburst and Portburst are capable of detecting and filtering short-term as well as long-term erratic traffic events. Removing long-term events is not always desirable. In some scenarios, they generate persistent and continuous traffic, maintaining the regular behavior of IBR signal. Figure 3.2 depicts a scenario, where the magnitude of IBR traffic increases and unlike other traffic bursts, doesn't decrease in time. Due to their persistent nature, long-term events cannot be referred to as erratic events without the complete analysis of their duration. However, my solution uses the stored history of observed source IPs to determine the persistent nature of events and consequently alerts the user. The user can further decide whether to filter the observed long-term event.

4.3 Interesting Findings

Long-term event SYNRENT observed in IBR captured by the UCSD-NT, doesn't explain the erratic IBR traffic observed in two periods: September 2016 and Nov 2017 - Feb 2018. To further detect these events, I used Hostburst and Portburst to analyze IBR data in those periods. The obtained results reveal the presence of two interesting events in IBR: 1) MIRAI botnet, and 2) Phase three of SYNRENT.

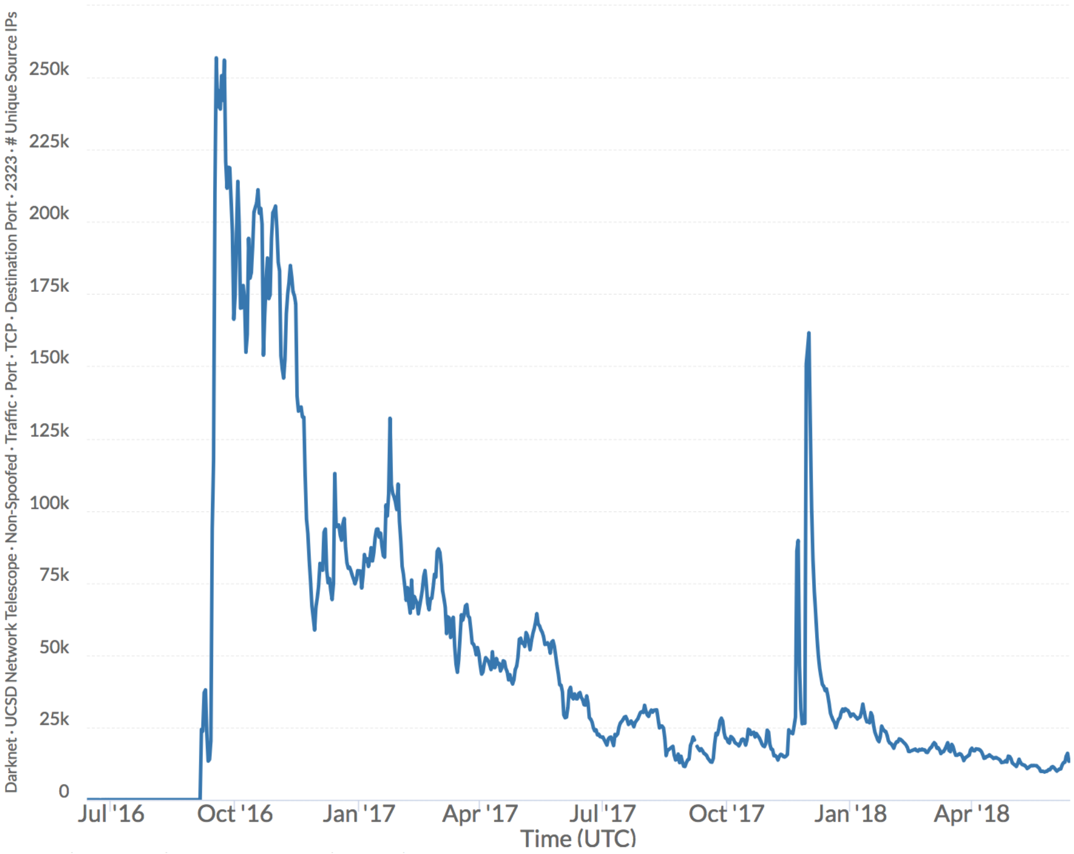


Figure 4.9: Mirai botnet traffic observed on port 2323 in September 2016

On running Portburst in the time frame of May 2016-Sept 2016, I detected persistent traffic in IBR. Traffic suddenly increased on certain TCP destination ports 23 and 2323. Figure 1.5 portrays an increase in the IBR traffic observed on ports 2323 in Sept 2016. Ports 23 and 2323 are used for Telnet. On further analysis, I found out that the event responsible for the traffic is

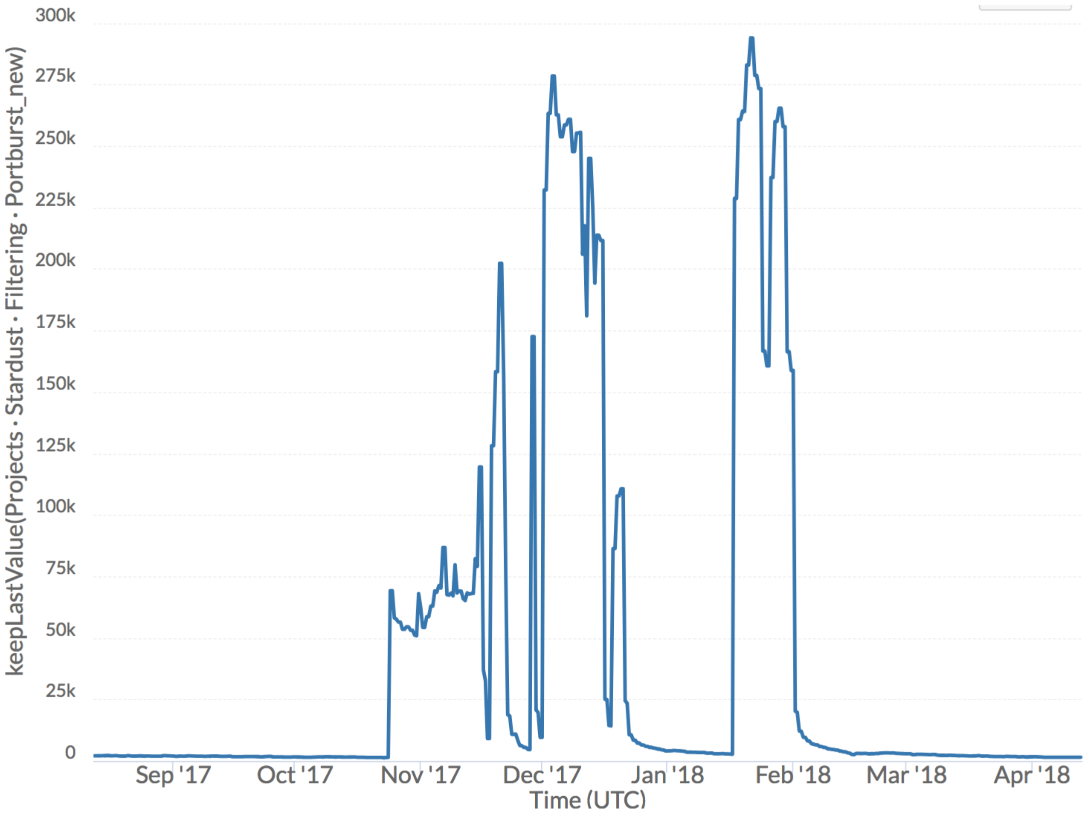


Figure 4.10: SYNRENT traffic observed in Nov 2017 - Feb 2018.

caused by the infamous ”Mirai” Botnet[36]. The Mirai botnet caused one of the biggest DDoS attacks by leveraging compromised Internet-of-Things (IoT) devices. It expanded its botnet by targeting networked devices running Linux like home routers, IP cameras, etc.

In previous sections, I concluded that events responsible for generating SYNRENT traffic stopped in June 2017. However, after analyzing the IBR traffic of Nov 2017 - Feb 2018, Hostburst and Portburst reveal the presence of a large amount of TCP and UDP traffic on destination ports 19332-19622. On further investigation, I observe that this new detected traffic resembles SYNRENT (phase two) in terms of its signature (range of destination ports, destination IPs, TCP-SYN flags, BitTorrent, etc.). Figure 4.10 shows the traces of SYNRENT traffic present during Nov 2017 till Feb 2018.

Chapter 5

Conclusion and Future Work

Internet background radiation is network pollution, generated by a vast number of computers, often unknown to their original users. IBR collected at unused but routed network blocks, has been used in Internet measurement and network analysis studies for many years. The passive traffic monitoring system built on these network blocks are commonly known as Network Telescope aka Darknet. UCSD's Internet Outage Detection and Analysis (IODA) is a research project which uses a signal extracted from IBR (that exhibits a regular and non-erratic pattern), to monitor the Internet for macroscopic Internet connectivity outages. IBR reflects fundamentally nonproductive traffic, either malicious (backscatter, botnet scans, worms, DoS attacks) or benign (misconfiguration or bugs). Ideally, the signal extracted from IBR to be used in IODA should be strong (statistically significant), stable (low noise), and globally pervasive (seen in most networks), but due to the varied composition of IBR, extracting stable ”normal” IBR traffic is difficult. Previously, Internet background radiation had been filtered and sanitized to stabilize its signal for IODA, but this was done only till June 2015. After July 2015, significant new erratic traffic components have been observed, disrupting the regular non-erratic signal extracted from IBR. These events result in generating traffic bursts which interfere with IODA's inference system.

In my thesis, I have investigated and analyzed three years of IBR traffic (2015-2018) collected at the UCSD-NT. During my analysis, I detected many short-term events resulting in traffic bursts. Along with short-term events, I analyzed a long-term event SYNRENT (TCP-SYN - BitTorrent) which disrupted the regular pattern in non-erratic signal extracted from IBR. SYNRENT affected the TCP component of IBR traffic by sending a large magnitude of TCP-SYN requests, for over two years. I presented a broad characterization of SYNRENT in terms of source IPs, countries, autonomous systems, operating systems, etc. to better explain the phenomenon. This thesis also identified the potential cause of SYNRENT as the Distributed Hash Table (DHT) poisoning of BitTorrent traffic (erroneous entry in DHT) originating from China. To mitigate such distorting events and bursts in IBR, I presented a software solution (Hostburst and Portburst) to detect and filter this erratic traffic in IBR in real-time. The primary idea is to use the history of previously observed unique source IPs on destination IPs and ports, to determine the occurrence of a burst/irregular event. In addition to proposing and implementing the solution, I provided substantial evidence that proves the effectiveness of my solution in detecting and filtering irregular/erratic events.

In the current version, Hostburst and Portburst follow a coarser approach in filtering out irregular traffic. In the event of burst detection, entire traffic on detected destination ports and IPs is removed, resulting in traffic loss. This scenario can be improved by creating a less-coarse solution by using more information like protocol IDs, flags, packet size, etc., in creating the filters. It will result in less overhead in terms of traffic loss from the removed ports and IPs. By observing the recent trends in IBR, it is fair to conclude that unprecedented events like SYNRENT, MIRAI, etc., will continue to disrupt the extracted non-erratic IBR signal with more intensity and sophistication. Even though Hostburst and Portburst will identify most of the erratic events in IBR, manual investigation of IBR traffic from time to time will ensure its effectiveness and accuracy.

Bibliography

[1] "Internet Background Radiation.," *https://en.wikipedia.org/wiki/Internet_background_noise*.

[2] R. Pang, V. Yegneswaran, P. Barford, V. Paxson, and L. Peterson, "Characteristics of internet background radiation," in *Proceedings of the 4th ACM SIGCOMM Conference on Internet Measurement*, IMC '04, (New York, NY, USA), pp. 27–40, ACM, 2004.

[3] A. Dainotti, R. Amman, E. Aben, and K. Claffy, "UCSD Network Telescope," *http://www.caida.org/data/passive/network_telescope.xml*, 2010.

[4] "Illustration of IBR reaching UCSD Network Telescope.," *http://www.caida.org/publications/presentations/2017/ioda_sdsc/ioda_sdsc.pdf*.

[5] A. Dainotti, A. King, K. Claffy, F. Papale, and A. Pescap, "Analysis of a "/0" Stealth Scan from a Botnet," *IEEE/ACM Transactions on Networking*, vol. 23, pp. 341–354, Apr 2015.

[6] A. Dainotti, A. King, and K. Claffy, "Analysis of Internet-wide Probing using Darknets," in *Building Analysis Datasets and Gathering Experience Returns for Security (BADGERS)*, Oct 2012.

[7] M. Casado, T. Garfinkel, W. Cui, V. Paxon, and S. Savage, "Opportunistic Measurement: Spurious Network Events as a Light in the Darkness," in *ACM Fourth Workshop on Hot Topics in Networks (HotNets-IV)*, 2015.

[8] A. Dainotti, C. Squarcella, E. Aben, K. Claffy, M. Chiesa, M. Russo, and A. Pescap, "Analysis of Country-wide Internet Outages Caused by Censorship," in *Internet Measurement Conference (IMC)*, pp. 1–18, Nov 2011.

[9] A. Dainotti, R. Amman, E. Aben, and K. Claffy, "Extracting benefit from harm: using malware pollution to analyze the impact of political and geophysical events on the Internet," *ACM SIGCOMM Computer Communication Review (CCR)*, vol. 42, pp. 31–39, Jan 2012.

[10] K. Benson, A. Dainotti, k. claffy, and E. Aben, "Gaining Insight into AS-level Outages through Analysis of Internet Background Radiation," in *Traffic Monitoring and Analysis Workshop (TMA)*, Apr 2013.

[11] "Center for Applied Internet Data Analysis (CAIDA).," *https://www.caida.org/home/*.

[12] "Charthouse.," *http://www.caida.org/interactive/*.

[13] "IODA Dashboard.," *https://ioda.caida.org/ioda/dashboard*.

[14] "A high level view of IODA architecture.," *http://www.caida.org/projects/ioda/*.

[15] "Internet connectivity outage in Iraq.," *https://ioda.caida.org/ioda/dashboard/inspect #from=1475280000&until=1475884800&view=inspect&entity=country/IQ*.

[16] "Internet connectivity outage in Gabon.," *https://ioda.caida.org/ioda/dashboard/inspect #view=inspect&entity=country/GA&from=1473033600&until=1473638400*.

[17] "Internet connectivity outage in Mexico.," *https://ioda.caida.org/ioda/dashboard/inspect #view=inspect&entity=country/MX&from=1476817200&until=1476903600*.

[18] K. Benson, A. Dainotti, k. claffy, A. Snoeren, and M. Kallitsis, "Leveraging Internet Background Radiation for Opportunistic Network Analysis," in *Internet Measurement Conference (IMC)*, Oct 2015.

[19] A. Dainotti, K. Benson, A. King, k. claffy, M. Kallitsis, E. Glatz, and X. Dimitropoulos, "Estimating Internet address space usage through passive measurements," *ACM SIGCOMM Computer Communication Review (CCR)*, vol. 44, pp. 42–49, Jan 2014.

[20] "Merit Network, Inc. Merit Darknet IPv4.," *http://software.merit.edu/darknet/*.

[21] M. R., "xkcd: Map of the Internet.," *http://xkcd.com/195/*, 2006.

[22] Z. M., "p0f v3," *http://lcamtuf.coredump.cx/p0f3//*, 2012.

[23] A. King, "Corsaro.," *http://www.caida.org/tools/measurement/corsaro/docs/*, 2012.

[24] "Internet Outage Detection Analysis.," *https://ioda.caida.org/*.

[25] A. Loewenstern and A. Norberg, "DHT Protocol.," *http://www.bittorrent.org/beps/bep_0005.html*, 2008.

[26] K. Benson, *Leveraging Internet Background Radiation for Opportunistic Network Analysis*. PhD thesis, 2016. Copyright - Database copyright ProQuest LLC; ProQuest does not claim copyright in the individual underlying works; Last updated - 2016-10-25.

[27] "BitTorrent message formats: Announce.," *http://www.bittorrent.org/beps/bep_0003.html*.

[28] "Article by BitTorrent on DDoS attacks.," *https://engineering.bittorrent.com/2015/01/29/a-note-on-the-ddos-attacks/*.

[29] "TCP Timeout and Retransmission.," *http://www.pcvr.nl/tcpip/tcp_time.htm*.

[30] "China banning torrents.," *https://readwrite.com/2009/12/10/torrent-china-government/*.

[31] "China Hijacks Popular BitTorrent Sites.," *https://torrentfreak.com/china-hijacks-popular-bittorrent-sites-081108/.*

[32] T. Fox-Brewster *https://www.forbes.com/sites/thomasbrewster/2015/01/26/china-great-firewall-causing-ddos-attacks/#19e28d5c6a47*, 2015.

[33] Ernesto *https://torrentfreak.com/zombie-pirate-bay-tracker-fuels-chinese-ddos-attacks-150124/*, 2015.

[34] *http://blog.devops.co.il/post/108740168304/torrent-ddos-attack*, 2015.

[35] A. King and A. Dainotti, "Corsaro.," *www.caida.org/tools/measurement/ corsaro/*, 2018.

[36] "MIRAI botnet: Wikipedia.," *https://en.wikipedia.org/wiki/Mirai_(malware).*