FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Anomaly Detection in Enterprisal Networks

André Moreira

Mestrado Integrado em Engenharia Electrotécnica e de Computadores

Supervisor: Ricardo S. Morla (Professor)

Co-supervisor: Jaime S. Cardoso (Professor)

February 2011

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Technological advances in communications have sustained the quick development of the Internet over the past years to the point where it became well carved into people's culture. From a simple chat to complex business transactions, the truth is that people rely on it for pretty much everything with unquestionable benefits. However, as the dependency grows so does the demand for higher quality services. An important contribution to satisfy such quality levels is to ensure the security and reliability of the Internet's underlying protocols.

One of the key protocols behind this global scale network is the Border Gateway Protocol (BGP). As the current standard of inter-domain routing protocols, BGP is responsible for connecting together the distinct networks that form the Internet. Hence, its correct operation is fundamental to achieve global communication. Unfortunately, unexpected behavior occurs in several layers of the Internet and BGP is not an exception. Recently, Butler et al. [1] presented a survey on BGP security issues and solutions. Typical communication security problems like denial-of-service attacks, attacks against message integrity and attacks against message confidentiality are described. On the other hand, Elmokashfi2010 et al. [2] carried out an analysis of BGP churn evolution. In this analysis, the reasons behind some anomalies in the BGP churn time series were investigated, leading to interesting results. Most of the update bursts detected were caused by duplicate announcements. Specifically, 40% of route announcements were considered redundant and not needed for correct protocol behavior. Additionally, some sustained high levels of activity were found out to be caused by misconfigurations. Such anomalies represent a severe and unnecessary overload to the routers.

From malicious attacks to misconfigurations, several kinds of pathologies can compromise the protocol responsible for the Internet's connectivity. To make things worse, BGP doesn't define mechanisms to ensure security or performance, leaving those to be addressed by some other means. Independently of their nature, these events can be traced by searching for anomalies in BGP routing information dynamics. However, due to the enormous volume of BGP data, automatic methods are mandatory to help pinpoint such anomalies. Therefore, the purpose of this work is to study passive anomaly detection techniques that do not require modifications to the Internet's infrastructure and its underlying protocols.

Chapter 2

State of the Art

2.1 Network

2.1.1 Autonomous Systems

An Autonomous System (AS) is a group of IP networks run by one or more network operators with a single and clearly defined routing policy [3]. Each AS is identified by a globally unique number referred to as autonomous system number (ASN). An ASN is a 16-bit number which allows for a maximum of 65536 assignments. However, not every number in that range is available for public use (see Table 2.1). Due to exhaustion problems 32-bit ASNs were introduced in 2007 but the same principles apply.

Range	Description
0	Reserved (may be used to identify non-routed networks)
1-58367	Designated for public use
58368-64495	Reserved
64496-64511	Reserved (for use in documentation and sample code)
64512-65534	Designated for private use
65535	Reserved

Table 2.1: 16-bit ASNs Distribution (February 2011)

Furthermore, depending on their connectivity and operating policy ASes can be classified as one of the following types:

- Stub AS – connects only to one AS.
- Transit AS – connects to more than one AS allowing traffic to flow between those neighboring ASes through itself. A good example of a transit AS is an ISP which allows traffic to flow between its customers and the rest of the internet.
- Multihomed AS – connects to more than one AS. However, unlike a transit AS it does not allow traffic to flow between its adjacent ASes through itself. One objective is to increase

connection reliability, by assuring the availability of a backup AS. This technique is called multihoming.

2.1.2 Internet Topology

The internet is a global distributed network composed by many smaller interconnected networks. Each of these networks is majorly comprised of end-systems called hosts and intermediate-systems called routers. Keeping all these different networks connected at all times is crucial to achieve the purpose of global communication. This worldwide structure can be depicted as a set of nodes with links between them. For purposes of this work, the interest goes to an autonomous system level conceptual topology, where nodes are autonomous systems and links between them represent some kind of relationship.

2.1.3 Internet Addressing

The internet protocol (IP) is the communications protocol of choice for relaying messages across networks. It uses blocks of data called datagrams to carry the desired information along with the source and destination internet protocol addresses (IP addresses). These addresses uniquely identify a system within a network using the internet protocol. Currently there are two types of IP addresses in active use: IP version 4 (IPv4) and IP version 6 (IPv6). IPv4 was initially deployed on 1 January 1983 and is still the most commonly used version [1] [4]. IP addresses are 32-bit numbers often expressed as 4 octets in "dotted decimal" notation (e.g. 192.10.0.5). Regardless of the version, all IP addresses are made up of two parts: network number and host number.

Upon the introduction of the internet protocol [5], the addressing space was divided in 3 network classes as shown in Table 2.2. Two additional classes currently exist that encompass special purpose addresses. Eventually this originated the term classful addressing.

Class	Most Significant Bits	Network/Host	#Networks	#Hosts
A	0	n.h.h.h	128	16,777,216
B	10	n.n.h.h	16,384	65,536
C	110	n.n.n.h	2,097,152	256

Table 2.2: Main IPv4 Network Classes

However, quick growth of the internet and inefficiency of this approach caused concerns over the long-term scaling properties of the class A/B/C system [6]. This led to the introduction of the Classless Inter-Domain Routing (CIDR) which presented the concept of classless addressing. As the name suggests the idea of classes became deprecated. This alternative consists on the use hierarchical blocks of IP addresses called prefixes where the size of the network number within the address is explicitly indicated. As a result, according to CIDR notation, a prefix is represented as an IP address followed by “/” and the number of most significant bits that correspond to the

[1] For this reason, IP address will refer from now on to IPv4 address

network address. Some examples including how legacy classes can be represented using classless notation are shown in Table 2.3.

Prefix	Comment
n.0.0.0/8	Legacy class A
n.n.0.0/16	Legacy class B
n.n.x.0/20	Network with size between legacy A and B classes
n.n.n.0/24	Legacy class C

Table 2.3: CIDR Notation

This new addressing methodology has two major advantages:

- Flexibility – liberty of choice on the network size allows for a more rational use of the address space.
- Aggregation – the hierarchical structure allows the aggregation of many smaller prefixes into one larger prefix hence reducing the size of routing tables on the internet.

Classful addressing isn't completely dead as it is still used for example in local networks. Also it is so embedded in people's culture that classful terminology is used even when the other applies. Nevertheless, at internet scale and principally for routing purposes, classless addressing is preferred for the advantages mentioned above.

2.1.4 Routing

Single knowledge about a packet's destination address is not enough to deliver it. Information about existing paths is also necessary. The selection of a path for transmission is called routing [5] and is accomplished by intermediate-systems called routers.

Every router stores a routing table that correlates final destinations with next hop addresses. However, most only hold partial routing information meaning they don't know the routes to all destinations. In such cases, packets with unknown destinations are successively forwarded to a default router called default gateway, until at some point it arrives to a router which has reachability information to the desired location and effectively indicates the path. Either way the packet will travel hop-by-hop through several networks until it reaches its destination. It is therefore fundamental that routing tables are kept up-to-date, under the risk of occurring unreachability phenomena.

Manually entering routing information is unthinkable over the size of routing tables and constant network topology variations. A more dynamical approach is used to quickly propagate changes by using routing protocols. The job of a routing protocol is to disseminate routing information between routers so that they keep a correct view of the network topology at all times. Depending on its operation mode, protocols fall in one of the following categories:

- Distance-vector routing protocol – each node stores information about where to send packets for a certain destination and how many hops are necessary to get there. Best path is selected as the one with minimum number of hops.

- Link-state routing protocol – each node stores a connectivity map of the network and the path is calculated with a Shortest Path First algorithm taking into account not only the number of hops but also its cost (e.g. bandwidth of the link).

- Path-vector protocol – each node stores information about where to send packets for a certain destination and the path (consisting in a set of nodes) necessary to get there.

Obviously, storing and sharing information about every single connection is impossible at such large scale. For that reason, routing information in one network is filtered according to a routing policy that defines which routes are to be actually learned and which ones are to be further propagated. Another key factor is the AS organization of the Internet (see Section 2.1.2). Since by definition all networks within an AS present the same routing policy, all routing information belonging to them can be treated and delivered to other ASes as whole (aggregated), providing reachability information while hiding the inner details. Despite that, routers inside an AS eventually need to exchange routing information between them so that they all have a consistent view of the outside. Conveniently, the concepts of interior gateway protocol (IGP) and exterior gateway protocol (EGP) exist. An IGP is a routing protocol used to exchange routing information among routers within a single autonomous system. Conversely, an EGP is a routing protocol used to exchange routing information among autonomous systems within the Internet. Several IGPs can be used within an AS. However, there's only one EGP currently in use: the Border Gateway Protocol (BGP), specifically BGP-4 [3]. As the de facto inter-domain routing protocol, BGP is responsible for holding the disparate parts of the Internet together.

2.1.5 BGP

As the de facto standard of inter-domain routing protocols, BGP is responsible for keeping the disparate parts of the Internet together. This protocol can interconnect any set of autonomous systems using an arbitrary topology. To achieve that, it suffices that each AS designates one or more of its routers to speak on its behalf with other ASes using BGP. Any router that implements BGP is referred to as BGP speaker and its primary function is to exchange network reachability information with other BGP systems. Because BGP stores the AS path it is considered a path-vector routing protocol.

BGP uses TCP on port 179 for communication between neighbors. When BGP neighbors establish a TCP session, they start exchanging BGP information in the form of messages. Each message starts with a fixed size header with the structure shown in Table 2.4.

Following the header are the message contents whose structure depends on the message type. The existing types of messages are Open, Update, Notification and Keepalive.

Marker	Length	Type	...
16 bytes	2 bytes	1 byte	

Table 2.4: BGP message header structure

Among these, update messages are of particular interest since those are the ones used to exchange routing information. Specifically, they are used to announce feasible routes and/or withdraw unfeasible routes. The structure of the update message is shown in Table 2.5.

...	WR length	Withdrawn Routes	PA length	Path Attributes	NLRI
	2 bytes	Variable	2 bytes	Variable	Variable

Table 2.5: BGP update message structure

In the context of BGP, a route is defined as a unit of information that pairs a set of destinations with the attributes of a path to those destinations [3]. The “Withdrawn Routes” and “Network Layer Reachability Information” (NLRI) fields carry a list of prefixes associated with routes to be removed and routes being announced respectively. Several prefixes can be announced in the same message as long as they share the same path attributes. Path information associated with the destinations in the NLRI field is included in the “Path Attributes” field. The main attributes are:

- Origin
- AS path
- Next hop
- Multi Exit Discriminator
- Local Preference

Path attributes are used to select the best route when several routes to the same destination network exist. All pertinent information is recorded in the Routing Information Base (RIB) of the BGP speaker.

2.2 Anomaly Detection

2.2.1 General Aspects

Anomaly detection refers to the problem of finding patterns in data that do not conform to expected behavior. Such abnormal patterns are usually referred to as anomalies or outliers.

Input data to these techniques consists on several data instances, each one being described by a set of features. If one data instance is represented by only one feature it is considered univariate, otherwise it is considered multivariate. Each of these features can be of different numeric or symbolic types and influence the type of techniques that can be used.

Another important aspect associated with a data instance is the data label. The data label specifies whether a specific instance of data is normal or anomalous. Obtaining accurate labeled data for every kind of behavior is difficult and expensive. Moreover, since anomalies represent a small fraction of the whole behavior and are more unpredictable, it is more difficult to get labeled data for anomalous behavior than for normal behavior. Depending on the availability of data labels, different modes of operation can be employed:

- Supervised anomaly detection
- Semi-supervised
- Unsupervised anomaly detection

Regarding the output of anomaly detection techniques, those are usually represented by labels, indicating whether a test instance was considered normal or anomalous, or scores that indicate the degree to which a test instance is considered anomalous.

There are many fields of application for these techniques, each one with its data characteristics and performance requirements that make some techniques more suitable than others.

2.2.2 Classification Based Techniques

Classification based techniques make use of a model known as classifier capable of distinguishing between normal and anomalous classes. The classifier is built during the training phase using labeled data. In multi-class settings data instances can belong to multiple normal classes, while in one-class settings only one normal class exists. Finally, the learned classifier is used in the testing phase to determine if a given test instance is normal or anomalous. The way classifiers are learned and represented depends on the technique used. The main possibilities are:

- Neural Networks-Based
- Bayesian Networks-Based
- Support Vector Machines-Based
- Rule-Based

Consequently, classification based techniques can be used to distinguish between several classes. The testing phase is fast since test instances need only to be matched to the models. However a comprehensive set of labeled training data is necessary. Also, the output is usually in the form of labels which can be or not a disadvantage.

2.2.3 Nearest Neighbor Based Techniques

Nearest neighbor techniques require a measure of distance between two data instances to be defined. Also, they assume that normal data instances occur in dense neighborhoods and anomalies occur far from these. One approach used to detect anomalies is using the $k^{(}th)$ nearest neighbor. Techniques using this approach calculate the distances from the test instance to the $k^{(}th)$ nearest neighbors, which is proportional to the anomaly score. Another way is to use relative density. In this case, the density around the test instance is calculated and inversely proportional to the anomaly score. In both cases thresholds can be applied to obtain a labeled output.

Nearest neighbor techniques are inherently unsupervised, requiring no labeled data or models to characterize it. Also, they are easy to adapt, sufficing that an appropriate distance measure in defined for the data. On the other hand, if the assumption about neighborhood of normal and anomalous classes stated above doesn't hold, false alarms and/or undetected anomalies can occur. The computational complexity is another shortcoming, since algorithms involve calculating the distances between each test instance and the other data instances.

2.2.4 Clustering Based Techniques

Clustering is used to group similar data instances into clusters. Depending on the technique, different assumptions about the displacement of normal and anomalous data instances exist. Clustering based techniques can operate in unsupervised mode and have fast testing phases since test instances need only to be compared against a small fixed number of clusters. Also, adaptation of the techniques requires only the adaption of the clustering algorithm. The computational complexity of the clustering algorithm is a shortcoming of these techniques.

2.2.5 Statistical Techniques

Statistical anomaly detection techniques are based on the principle that normal data instances are located in high probability regions of a stochastic model while anomalous data instances are located in the low probability regions. A statistical model is defined to fit the available data and statistical inference is used to determine if a given test instance belongs to the model or not with some degree of probability. Both parametric and non-parametric techniques exist. Parametric techniques like Gaussian model-based, regression model-based and mixture model-based techniques, assume knowledge of the data distribution. On the other hand, non-parametric techniques like histogram-based and kernel function-based do not assume such knowledge.

These techniques highly depend on the accuracy of the model chosen to represent the data. If an accurate model is used the results are statistically justifiable. If the model used is also robust to anomalies, these techniques can operate in an unsupervised setting.

2.2.6 Informational Theoretic Techniques

Information theoretic based techniques apply information theoretic measures like entropy to the data. It is assumed that anomalies generate irregularities in the information content of this data.

These techniques can operate in an unsupervised setting and do not make assumptions about the statistical distribution of the data. However, these kind of techniques highly depend on the information theoretic measure used, which can be difficult to choose.

2.3 Related Work

There are many security and performance issues concerning BGP. Recently, Butler et al. [1] presented a survey on BGP security issues and solutions, while Elmokashfi et al. [2] presented some interesting results regarding anomalies in the BGP churn time-series, showing for example that most of the update bursts detected in their work were caused by duplicate announcements. As the only inter-domain routing protocol available for use, BGP must be able to overcome these problems, although it doesn't define mechanisms to address them. Therefore, locating and justifying these issues has been an extensive area of research and many techniques have been proposed for that purpose.

The objective of this chapter is to present passive anomaly detection techniques which represent the state of the art on BGP anomaly detection. Because of the different characteristics of each problem, a separation is made between BGP security anomaly detection and BGP dynamics anomaly detection which is much more general problem.

2.3.1 BGP Security Anomaly Detection

Several techniques have been proposed to address BGP security issues [1]. One of the most straightforward type of BGP attack is prefix hijacking which is analyzed in more detail by [7]. Based on implementation aspects, security solutions can be grouped in different categories. Some solutions like [8] make use of cryptographic techniques in order to eliminate security breaches in the protocol thus ensuring secureness of BGP. Although they have a preemptive action their adoption is expensive. Butler et al. [1] stated that "there is a resistance in the operations community to using any sort of cryptography in networks, largely due to the costs imposed". Other solutions like [9, 10, 11] do not make use of cryptography but require changes to the routing protocol, router software, router configuration or network operations. For example, Goodell et al. [11] proposed a new protocol that would work in concert with BGP, helping Autonomous Systems detect and mitigate accidentally or maliciously introduced faulty routing information. Finally, solutions based on anomaly detection techniques exist that require only passive monitoring of data. Because they are easier to deploy they are more attractive despite suffering from false alarms. Solutions like [12] use both control plane data and forwarding plane data to detect security anomalies. However, because BGP routing information is publicly available in many resources, special interest lies in solutions that are based only on control plane data. Lad et al. [13] present PHAS, a light weight real-time monitoring system aimed at prefix hijacking detection which can be easily deployed due to its simplicity. PHAS monitors BGP routing data in search for changes in BGP prefix origins. These changes correspond to potential threats and are emailed to registered prefix owners who can easily identify real hijack alerts and filter out normal origin changes. In an evaluation procedure known prefix hijacking events were detected by PHAS. Kruegel et. al [14] propose a readily deployable topology-based solution. The algorithms build a model of the autonomous system connectivity from BGP routing information and whois database. Route advertisements are checked to make sure that they are consistent with the network topology, preventing routers from accepting invalid routes. Visualization-based tools also exist to detect MOAS events [15].

2.3.2 BGP Dynamics Anomaly Detection

Rather than inspection of the contents of BGP update messages, BGP dynamics anomaly detection techniques usually perform a coarser analysis of the data. A common approach is to count the number of BGP updates in a defined time interval forming a BGP data volume time series. BGP update data can be retrieved from a single vantage point or multiple vantage points and can be analyzed under different perspectives (e.g. one prefix along several views or vice-versa). Furthermore, depending on the efficiency of the algorithms used some techniques can be used in online configuration while others must be used offline.

Zhang et al. [16] presents two approaches, signature-based detection and statistics-based detection, to search for anomalous BGP routing dynamics. Raw input data is retrieved from one single collector and filtered for processing. The purpose of the combination of these techniques is to overcome the weakness of each other. However, these tools need extensive processing on the content of the collected routing data, and are therefore more appropriate for off-line post-processing. Teoh et al. [17] extended this work by combining the techniques in [16] with visual data mining techniques in a near-real-time anomaly detection system.

Some works showed that BGP traffic exhibits self-similarity [18]. This property makes wavelet preferred over other frequency analysis tools like FFT when frequency analysis is to be performed, as it performs multi-scale analysis of the data. Works like [18, 19, 20] make use of this. Praskash et al. [18] proposed a tool for automated analysis of BGP updates called BGP-lens. The input consists of raw BGP data retrieved from one monitor transformed into a time-series with the number of updates received by a router every b seconds (called bin size). The analysis comprises both temporal and frequency analysis. In temporal analysis, “clothesline” phenomena may be observed in the log-linear plot if an appropriate bin size is chosen. On the other hand, frequency analysis allows detection of spikes and prolonged spikes. Due to the self-similarity of the signal, Discrete Wavelet Transform (DWT) is used, specifically Haar wavelets. In the resulting scalogram, two types of “tornados” patterns may be observed corresponding to spikes in the time domain.

Despite all algorithms are linear on the number of updates this tool still needs some work to be deployed as a non-stop monitoring tool. Mai et al. [19] proposed a framework for BGP anomaly detection with wavelet named BAlet. The tool is aimed at detection of groups of BGP updates that are likely triggered by the same events. It achieves network-wide event detection based on BGP updates received at a single vantage point. The procedure takes a 2-dimensional matrix of localized update counts as input. Once again, self-similarity of the BGP update volume time-series favors the use of wavelet. Wavelet analysis is performed on each row of the matrix followed by a two-dimensional clustering, which allows localization of anomalies in both spatial and temporal dimensions. Due to extensive computational effort required over the data, mainly due to the cluster analysis, it is currently used offline. J. Zhang et al. [20] carried out a similar work proposing an instance-learning framework to identify anomalies based on deviations from the normal BGP-update dynamics for a given destination prefix and across prefixes. In their scheme, BGP update behaviors are represented by a feature vector mapped into a multidimensional space. Wavelet transform, specifically Haar wavelets, is employed to extract update dynamic features such as burst duration and inter-burst intervals. The mapped feature vectors are then clustered into normal and abnormal groups using k-means clustering.

Chapter 3

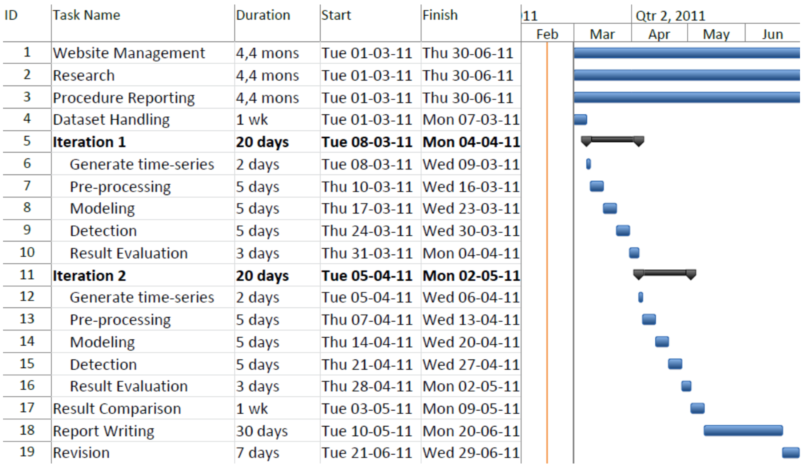
Work

3.1 Problem Definition

Several factors can affect the stability of Border Gateway Protocol. By analyzing publicly available BGP data, specifically BGP update logs, anomalies can be detected. Raw data can be treated to generate different time-series, specifically across time, views and prefixes. The objective of this work is to apply different models and automatic anomaly detection algorithms to different BGP time-series. Anomalies like BGP update bursts or spikes are the most basic examples. In the end, it is expected a comparison of the different models and algorithms employed by analyzing the results.

3.2 Work Plan

The following Gantt chart (Figure 3.1) depicts a schedule of the tasks planned in order to accomplish the previous defined objectives.



ID	Task Name	Duration	Start	Finish
1	Website Management	4,4 mons	Tue 01-03-11	Thu 30-06-11
2	Research	4,4 mons	Tue 01-03-11	Thu 30-06-11
3	Procedure Reporting	4,4 mons	Tue 01-03-11	Thu 30-06-11
4	Dataset Handling	1 wk	Tue 01-03-11	Mon 07-03-11
5	**Iteration 1**	**20 days**	**Tue 08-03-11**	**Mon 04-04-11**
6	Generate time-series	2 days	Tue 08-03-11	Wed 09-03-11
7	Pre-processing	5 days	Thu 10-03-11	Wed 16-03-11
8	Modeling	5 days	Thu 17-03-11	Wed 23-03-11
9	Detection	5 days	Thu 24-03-11	Wed 30-03-11
10	Result Evaluation	3 days	Thu 31-03-11	Mon 04-04-11
11	**Iteration 2**	**20 days**	**Tue 05-04-11**	**Mon 02-05-11**
12	Generate time-series	2 days	Tue 05-04-11	Wed 06-04-11
13	Pre-processing	5 days	Thu 07-04-11	Wed 13-04-11
14	Modeling	5 days	Thu 14-04-11	Wed 20-04-11
15	Detection	5 days	Thu 21-04-11	Wed 27-04-11
16	Result Evaluation	3 days	Thu 28-04-11	Mon 02-05-11
17	Result Comparison	1 wk	Tue 03-05-11	Mon 09-05-11
18	Report Writing	30 days	Tue 10-05-11	Mon 20-06-11
19	Revision	7 days	Tue 21-06-11	Wed 29-06-11

Figure 3.1: Gantt Chart

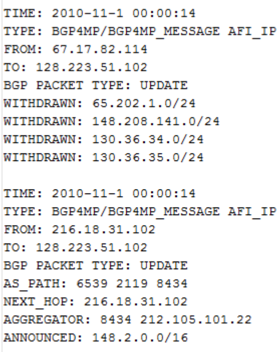
The three upper tasks are continuous tasks that are to be performed in parallel with the main work. Regarding, the main iterations are predicted. The first one is intended two accomplish a basic functional version of the framework. Successive iterations will be used to develop enhanced versions of the framework in the previous iterations. This repetitive iteration structure will be carried out following a spiral development model.

Regarding the tools, the following separation is made:

- Data and processing tools
  - Route Views project: resource providing BGP routing data and BGP routing table dumps. The dataset to be used will be retrieved from here.
  - Perl Script: script used to convert BGP MRT dumps to ASCII format
  - Matlab: software used to implement and run the anomaly detection algorithms
- Documentation management tools
  - LaTex: document processor.
  - BibTex: format used to describe and process lists of references, mostly in conjunction with LaTeX documents.
  - Jabref: used to manage references in the BibTex format

3.3 Initial Work

Some experiences with the dataset were already carried out. These consisted on downloading, converting and inspecting BGP route data from the Route Views Project. The downloaded data corresponds to the BGP messages received by a BGP monitor in the London Internet Exchange Point (LINX) during the month of November 2010. It comes in compressed packets of 15 minute MRT dumps. By converting MRT files to ASCII, messages like in Figure 3.2 could be observed.



```
TIME: 2010-11-1 00:00:14
TYPE: BGP4MP/BGP4MP_MESSAGE AFI_IP
FROM: 67.17.82.114
TO: 128.223.51.102
BGP PACKET TYPE: UPDATE
WITHDRAWN: 65.202.1.0/24
WITHDRAWN: 148.208.141.0/24
WITHDRAWN: 130.36.34.0/24
WITHDRAWN: 130.36.35.0/24

TIME: 2010-11-1 00:00:14
TYPE: BGP4MP/BGP4MP_MESSAGE AFI_IP
FROM: 216.18.31.102
TO: 128.223.51.102
BGP PACKET TYPE: UPDATE
AS_PATH: 6539 2119 8434
NEXT_HOP: 216.18.31.102
AGGREGATOR: 8434 212.105.101.22
ANNOUNCED: 148.2.0.0/16
```

Figure 3.2: Update messages in ASCII format

Due to the huge size of a single 15 minute dump when uncompressed, an efficient mechanism to parse this data will have to be used.

References

[1] K. Butler, T.R. Farley, P. McDaniel, and J. Rexford. A Survey of BGP Security Issues and Solutions. *Proceedings of the IEEE*, 98(1):100 –122, January 2010.

[2] A. Elmokashfi, A. Kvalbein, and C. Dovrolis. BGP Churn Evolution: a Perspective from the Core. In *INFOCOM, 2010 Proceedings IEEE*, pages 1 –9, March 2010.

[3] Y. Rekhter, T. Li, and S. Hares. A Border Gateway Protocol 4 (BGP-4), January 2006.

[4] IANA. Number Resources, February 2011.

[5] DARPA. Internet Protocol, September 1981.

[6] V. Fuller and T. Li. Classless Inter-domain Routing (CIDR): The Internet Address Assignment and Aggregation Plan, August 2006.

[7] Khin Thida Latt, Yasuhiro Ohara, Satoshi Uda, and Yoichi Shinoda. Analysis of IP Prefix Hijacking and Traffic Interception. *International Journal of Computer Science and Network Security*, 10(7):22–31, July 2010.

[8] Stephen Kent, Charles Lynn, Joanne Mikkelson, and Karen Seo. Secure border gateway protocol (s-bgp. *IEEE Journal on Selected Areas in Communications*, 18:103–116, 2000.

[9] Xiaoliang Zhao, Dan Pei, Lan Wang, Dan Massey, Allison Mankin, S. Felix Wu, and Lixia Zhang. Detection of Invalid Routing Announcement in the Internet. *Dependable Systems and Networks, International Conference on*, 0:59, 2002.

[10] J. Karlin, S. Forrest, and J. Rexford. Pretty Good BGP: Improving BGP by Cautiously Adopting Routes. In *Network Protocols, 2006. ICNP '06. Proceedings of the 2006 14th IEEE International Conference on*, pages 290 –299, November 2006.

[11] Geoffrey Goodell, William Aiello, Timothy Griffin, John Ioannidis, Patrick McDaniel, and Aviel Rubin. Working Around BGP: An Incremental Approach to Improving Security and Accuracy of Interdomain Routing. In *In Proc. NDSS*, 2003.

[12] Xin Hu and Z.M. Mao. Accurate Real-time Identification of IP Prefix Hijacking. In *Security and Privacy, 2007. SP '07. IEEE Symposium on*, pages 3 –17, May 2007.

[13] Mohit Lad, Dan Massey, Dan Pei, Yiguo Wu, Beichuan Zhang, and Lixia Zhang. PHAS: a prefix hijack alert system. In *Proceedings of the 15th conference on USENIX Security Symposium - Volume 15*, Berkeley, CA, USA, 2006. USENIX Association.

[14] Christopher Kruegel, Darren Mutz, William Robertson, and Fredrik Valeur. Topology-Based Detection of Anomalous BGP Messages. In Giovanni Vigna, Christopher Kruegel, and Erland Jonsson, editors, *Recent Advances in Intrusion Detection*, volume 2820 of *Lecture Notes in Computer Science*, pages 17–35. Springer Berlin / Heidelberg, 2003.

[15] Soon-Tee Teoh, Kwan-Liu Ma, S. Felix Wu, Dan Massey, Xiao-Liang Zhao, Dan Pei, Lan Wang, Lixia Zhang, and Randy Bush. Visual-Based Anomaly Detection for BGP Origin AS Change (OASC) Events. In Marcus Brunner and Alexander Keller, editors, *Self-Managing Distributed Systems*, volume 2867 of *Lecture Notes in Computer Science*, pages 269–298. Springer Berlin / Heidelberg, 2003. 10.1007/978-3-540-39671-$0_1$4.

[16] Ke Zhang, Amy Yen, Xiaoliang Zhao, Dan Massey, S. Felix Wu, and Lixia Zhang. On Detection of Anomalous Routing Dynamics in BGP. In Nikolas Mitrou, Kimon Kontovasilis, George N. Rouskas, Ilias Iliadis, and Lazaros Merakos, editors, *NETWORKING 2004, Networking Technologies, Services, and Protocols; Performance of Computer and Communication Networks; Mobile and Wireless Communications*, volume 3042 of *Lecture Notes in Computer Science*, pages 259–270. Springer Berlin / Heidelberg, 2004. 10.1007/978-3-540-24693-$0_2$2.

[17] Soon Tee Teoh, Ke Zhang, Shih-Ming Tseng, Kwan-Liu Ma, and S. Felix Wu. Combining visual and automated data mining for near-real-time anomaly detection and analysis in BGP. In *Proceedings of the 2004 ACM workshop on Visualization and data mining for computer security*, VizSEC/DMSEC '04, pages 35–44, New York, NY, USA, 2004. ACM.

[18] B. Aditya Prakash, Nicholas Valler, David Andersen, Michalis Faloutsos, and Christos Faloutsos. BGP-lens: patterns and anomalies in internet routing updates. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '09, pages 1315–1324, New York, NY, USA, 2009. ACM.

[19] Jianning Mai, Lihua Yuan, and Chen-Nee Chuah. Detecting BGP anomalies with wavelet. In *Network Operations and Management Symposium, 2008. NOMS 2008. IEEE*, pages 465 –472, 2008.

[20] Jian Zhang, Jennifer Rexford, and Joan Feigenbaum. Learning-based anomaly detection in BGP updates. In *Proceedings of the 2005 ACM SIGCOMM workshop on Mining network data*, MineNet '05, pages 219–220, New York, NY, USA, 2005. ACM.