Unwanted Traffic Characterization on IP Networks by Low Interactive Honeypot

Alisson Puska and **Michele Nogueira** and **Aldri Santos**
NR2 - Federal University of Paraná - Brazil
Email: {aapuska, michele, aldri}@inf.ufpr.br

***Abstract*—The increasing amount of unwanted traffic on the Internet consumes the available bandwidth on any network connected to it. Despite efforts to address this issue, it is still a challenge to differentiate unwanted traffic. Due to lack of knowledge or investment, organizations fail to implement security policies, such as BCP 38, which helps blocking the flow of unwanted data. This paper presents a method based on low-interaction honeypots and network telescopes for identification and classification of unwanted traffic on IP networks. Our method aims to be simple and support low cost of deployment. An evaluation employed traces of real environments to show the method effectiveness. Results offer useful information about unwanted traffic, reaching a private network in a simple manner and with the reduced cost to block it.**

I. Introduction

The amount of unwanted data traffic on the Internet has grown in the past years [1]. Spams, scanners, worms and brute force attacks are examples of such unwanted traffic that organizations receive daily. Ordinary characteristics of the Internet such as anonymity, freedom of access and disregard of source help to increase the amount of unwanted traffic. Data transmissions due to undesired requests consume network resources, wasting time and money of companies and institutions. Hence, identifying unwanted traffic may improve the usage of services and network resources of an organization.

Recognizing unwanted traffic within the data flow received by an organization is challenging. Feitosa et al [1] highlight the need for previously manual identification of the unwanted and the desired on the network flows. The Internet Background Radiation (IBR) is a type of unwanted traffic among the private network natural flows [2]. The IBR indicates traffic destined to unused and unreferenced public IP addresses. As traffic destined to unused IP addresses should not exist, all these flows are considered anomalous and unwanted. This kind of traffic refers to misconfiguration and exploitation attempts on IP networks. The nature of IBR (traffic to unassigned IP addresses) makes easy to detect it inside the network flow.

Two commonly used techniques for measuring the IBR are: network telescopes and announcement of unused IP addresses on the Internet [3]. Network Telescopes (NT) employs unused public IP addresses together with packet filters to measure and classify the IBR [4]. It uses packet filters, honeypots and machine learning techniques to characterize the traffic behavior. In order to compute the enormous quantity of flows, this technique considers samples of the traffic passing by distributed collection points. The collection points establishes ranges of unused public IP addresses scattered on different locations for an accurate measurement. The management complexity of the collection points and the waste of the finite public IP addresses are downsides of the network telescope techniques. Moreover, the sample-based outcomes are generally imprecise [1]. Other technique announces unused public IP addresses on the Internet to re-route their traffic to a collection point [4]. Besides the unessential Internet traffic generated, the announcement technique needs permission from IANA, making it not suitable in larger scales.

This work presents a method inspired on network telescopes to determine undesired data coming from the Internet towards a private network in order to identify its amount and help to establish rules to block it. The proposed method aims for simplicity of configuration and low cost of deployment. For that, it applies a low interactivity honeypot to collect information on the unwanted traffic (IBR). This honeypot employs a unused public IP addresses to capture the IBR reaching the private network. The collected traffic is classified and grouped by its periodicity, kind and source. In contrast to other solutions for unwanted traffic classification, such as [5], [6], [2], our method uses the nature of the IBR to detect and characterize its amount entering a private network (personal, corporative or academic). The filters established for classification make the solution adaptable for any network size. The method was applied on real traces collected by the honeypot from PoP-PR of RNP-Brazil [7]. The evaluation results revealed the accessed services and their sources and periods, as well as they enabled to infer the profile of unwanted accesses behavior. Such information can be used to create rules for blocking the unwanted traffic on the private network.

This paper is organized as follows. Section II presents the related work. Section III details the proposed method for characterizing accesses and identifying unwanted behavior. Section IV and V show the method assessment and its results, respectively. Section VI concludes the work.

II. Related Work

There are some methods in the literature that employs honeypots for identifying unwanted traffic on private networks. Tiwari and Jain et al. [5] adopt distributed honeypots in order to recognize unwanted traffic in different zones of a network. Their method implements Perl scripts to analyse logs and classify undesirable flows. It was compared with an intrusion detection system (IDS) and traffic filter rules (packet filter) in order to validate the method. Krishnamurthy [6] adopts mobile honeypots for finding undesirable traffic near its source. The method employs a proxy system to redirect the traffic toward

different IP addresses. Even though both methods show good results, they demand high cost and complexity of deployment.

Salles-Loustau et al. [8] apply high interactivity honeypots for identifying attacks. In order to determine the access behavior, the method analyses logs through grouping and classification. However, the honeypot only collects data from accesses that correspond to SSH service. Goebel et al. [9] investigate malware dissemination in an academic network using low interactivity honeypot. The method works with data mining techniques for detecting the rate of worm's propagation and the number of malwares variations in the network. Although this work aims to characterize the evolution and the capacity of dissemination of worms, it does not classify the behavior of accesses over a period of time.

There are still other works in literature that adopt honeypots and similar techniques for traces classification [2], [10]. These techniques also analyse data collected by honeypots. Their focus consists in characterizing the behavior of the background radiation on the Internet. Hence, they are different from this proposed method that aims to define the amount of data received by an organization network in order to block it.

III. Background and Method

The honeypot technique emulates one or many services of a network. It is implemented in a controlled environment to collect malicious traffic information, such as inappropriate accesses and attack attempts. This technique interacts with the accesses offering detailed information about traffic behavior, different from packet filters or machine learning techniques that only observe the traffic passing through the network [1].

Honeypots are categorized in high and low interactivity [11]. A high interactivity honeypot emulates all aspects and services of network. This allows a detailed collecting of access behavior by raising the complexity of implementation and cost of deployment. A low interactivity honeypot simulates parts of network infrastructure and services, collecting basic information on accesses with a simple implementation and a low deployment cost. Thus, our method employs a low interactivity honeypot to collect detailed information on the accesses and achieve the low deployment cost and configuration complexity.

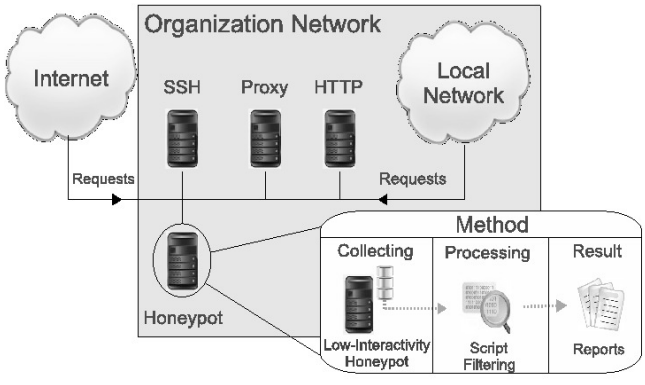

Fig. 1: Method Phases

The prevention honeypot technique employs a centralized approach that can be implemented as a service in an existing server for lower cost or deployed in a different computer for a better security. The application of honeypots simulates the same services on the network and analyse the inside and outside unwanted traffic. To detect the unwanted traffic, the honeypot IP addresses must be configured without any disclosure or DNS references. Once there is no IP address disclosure or references, all accesses to the honeypot services are unwanted. In addition, our method uses at least one public IP address to identify incoming data flows from the Internet. This configuration prevents the occurrence of false positives and provides the IBR traffic destined to the private network.

The method consists of two phases: the collecting data phase and the processing data phase, as illustrated in Fig. 1. The first one detects, measures and monitors the unrequested data. This is accomplished by the honeypot that takes advantage of the IBR nature to identify unwanted requests and data. As IBR refers to misconfiguration and exploit attempts, the real services on the private network could be receiving accesses from the same sources, consuming its resources. With the IBR detected, the second phase identifies the amount of access received and its sources, creating reports. To do this, our method applies AWK scripts for filtering the collected data and generating reports. Therefore, the reports can be applied to block the traffic from the IBR sources.

This method classifies and groups all collected data in a simple manner, creating reports on network and services. For that, a filtering mechanism separates the honeypot records and groups each access by the destination IP address. The source, frequency and the accessed services are parameters for the filtering. The filters create groups separating the access on the records by the IP address of source and the period of time. This processes generate reports about the amount of unrequested accesses received by the private network for different destinations. The next step groups the data for each destination, source and period of time to identify the behavior of unrequested traffic received. An AWK script counts the number of entries on the records ordered by the source IP address, finding the sources with more accesses. This step results in two reports about the amount of unrequested accesses received by services and the amount of unrequested data received by different destinations. These reports generated by the method identify the sources of IBR access, enabling the support for blocking this unwanted traffic.

Further, the method infers a profile of access behavior by source IP addresses. For that, tree metrics are applied: *amount of accesses in a time period (TAA), amount of accesses by source IP address in a time interval (AAI), access mean by source on time periods (AMS)*. The first metric gathers all accesses for a time period. The second metric splits the accesses in regular intervals of time and classifies them by source IP. The last metric estimates the accesses mean by a period of time for each source IP. The time intervals are divided in periods for limiting the scope of the inference of the access behavior. Equation 1 represents the calculation of the Access Mean by Source (AMS) of a source IP:

$$AMS = \frac{\sum_{i=1}^{N} AAI_i}{N} \quad (1)$$

AAI means the amount of accesses for each *Time Interval i* and the *N* is the amount of time intervals (total time). This metric enables an approximation of the accesses mean for intervals in each period by source IP address. It allows the identification of the accesses behavior, like recurrent accesses and burst accesses attempts on the period of time.

IV. Evaluation

The Brazilian National Computer Emergency, Response end Treatment Team (CERT.BR) coordinates the Brazilian Consortium of Honeypot (CBH). This consortium has several network telescopes with low interactive honeypots working together with IDS's. They are distributed around the country collecting information about attacks. Thus, the aim of the CBH is to evaluate the data collected to establish security policies and technical standards for network security. CBH shares security statistics of the Internet in Brazil at CERT.BR and it gives information about the amount of access to the services of all honeypots in the consortium. The Point of Presence on Paraná (PoP-PR) of the National Research Network (RNP) has a low interactive honeypot that belongs to CBH [7].

This evaluation employs the honeypot record files provided by PoP-PR that present all access between 01/01/2012 to 12/31/2012. Each line of the log files contain the tuple <access time/date, source IP address and port, destination IP address and port, amount of data received, attacker operating system>. The method was applied on the traces to classify the undesired incoming data flows. Each data source was grouped and divided by time period of access. To assist the identification of the access behavior, the periods comprehend to months and days to the intervals, they are applied on metrics *AAI* and *AMS*.

```
BEGIN {
      Africa = 0;
      Day = "01-01";
}
$4 ~"^41\.|^102\.|^105\.|^154\.|^196\.|^197\."{
  Africa ++ ;
  DayAux = substr($1,6,5);
  if(day != DayAux){
    printf DayAux" "Africa"\n" ;
    DayAux=Day;
    Africa =0;
  }
  DayAux=Day;
} END{ }
```

Fig. 2: AWK Filtering Script and POSIX Regular Expression

The honeypot configuration on PoP-PR follows the proposed method. The network has more than 250 public address, and those IP addresses are simulated by the low interactive honeypot (honeyD) [12]. PoP-PR server contains services collecting data together with the honeypot. It stores information of each access (e.g. IP address, port) and its interactions with the services. The interactions were not considered on the method application.

The method filters split the access by IP source address for each day. Further, in order to ease the inference process, the accessed source IP addresses were clustered in five groups according to IANA's delegation [13]. Table I shows the groups of the address ranges in the respective agencies. Each one of these five groups contains a set of IP address range for each zone. Fig. 2 illustrates the filtering script taking into account the IP addresses of the Africa group. This work does not identify the set of IP address belonging to each country in the group, only counts the total access of IANA groups.

TABLE I: IP Address Grouping

Group	Region	Address
AFRINIC	Africa	041 102 105 154 196 197
APNIC	Asia e Pacific	001 014 027 036 039 042 043 049 058 059 060 061 101 103 106 [110...126] 133 150 153 163 171 175 180 182 183 202 203 210 211 218 219 220 221 222 223
ARIN	North America	003 004 007 008 009 012 013 [015...020] 023 024 032 034 035 038 040 044 045 047 048 050 052 054 056 [063...076] [096...100] 104 107 108 128 129 130 131 132 [134... 140] [142...144] [146...149] 152 [155...162] [164...170] 172 173 174 184 192 198 199 [204...209] 216
LACNIC	Latin America end Caribbean	177 179 181 186 187 189 190 191 200 201
RIPE	Europe, Middle East	002 005 025 031 037 046 051 062 [077...095] 109 141 145 151 176 178 185 188 193 194 195 212 213 217

The inference phase applies the metrics defined in Section III on the filtered accesses. The period specified in the first metric (TAA) comprehends twelve months to determine the amount of accesses on the year (record's period). Each source IP address into this period was compared with the ones in Table I and accounted for each group, resulting on the *Total Amount of Access in a Period (TAA) report*. For the second metric (AAI), the intervals correspond to the number of days in each month in order to give evidences on the behavior of the access on the period (year). In this way, AAI provides a *Cumulative Distribution of Access by Group report*. On the third metric (AMS), an interval of an hour and a period of a day were defined to estimates the access mean. AMS offers a daily input rate generating a report on the *Distribution of Access by Day*. In this way, the inference tries to identify evidences on the behavior of the accesses in RNP network.

V. Results

The use of a unused public IP address enabled the identification of the amount of undesired data flows from the Internet. After applying the method, the filters showed sequential accesses in short periods of time to the services of the honeypot. The low interactivity honeypot helped to identify patterns in the access behavior. For example, SSH or Telnet accesses attempts are made sequentially in time periods from 2 to 15 seconds throughout a day. These accesses could be reaching the real services on the network and consuming its resources.

The achieved reports show a constant behavior over the months. However, it was detected significant variations on the number of accesses to Asia, Europe and North America in March, April, May and December. Preliminary analysis shows evidences of worms activities on specific days of these months. An extensive investigation is needed to determine the cause of this rise on the number of accesses.

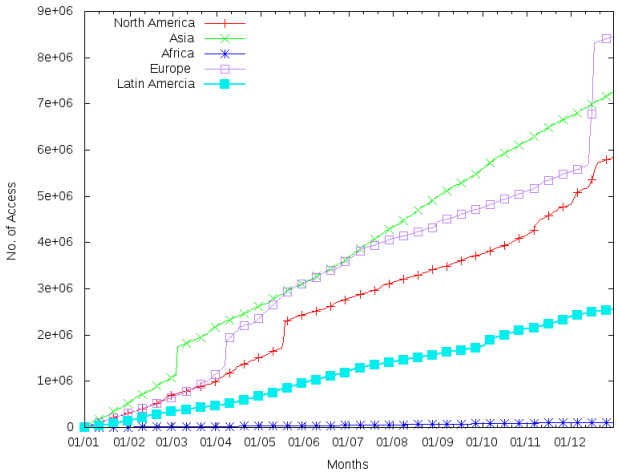

Fig. 3: Cumulative Distribution of Access by Group (TAA)

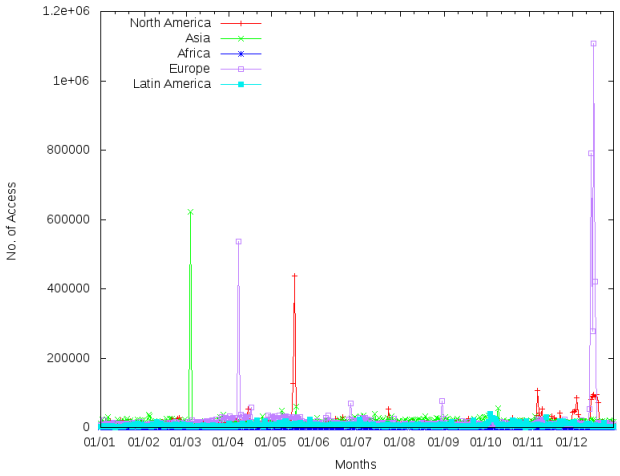

Fig. 4: Distribution of Access by Day (AMS)

Fig. 3 presents the accumulated access number of each one of five groups. The total access number in 2012 was 24.201.219, divided in: Asia 7.250.727, Europe 8.452.571, Africa 108.94, Latin America 2.546.651 and North America 5.841.970. Europe, Asia and Latin America had the greatest access number, which were expected because of the number of IP address allocated to these groups.

TABLE II: SSH Access Grouping by Day

Source	Service	Nm. Access	Date
174.136.35.43	SSH	443	01/01/2012
174.136.35.43	SSH	547	01/02/2012
184.107.214.138	SSH	1459	01/01/2012
184.107.214.138	SSH	0	01/02/2012
124.31.204.99	SSH	251	01/02/2012

Table II offers a part of the amount of access filtered by source IP address and service. It represents the information acquired with the first and second metrics. This table shows the amount of access for each service. Due to the restriction of space, the table has only a part of the filtered traces.

Fig. 4 shows the variations throughout the time period (year). It was observed abrupt changes in the number of requests for the months of March, April, May and December. In March, there was an increase of the number of accesses in the Asia group, which had peaked at 600 thousand accesses in the first half of the month. In April, the number of accesses to the group of Europe increased to more than 500 thousand hits, its access peak was also in the first half of this month. In May, there was an increase of the number of hits over 400 thousand requests from North America group and its accesses peak was at the end of the second half of this month. In December, the European group had an increase over 1,000,000 hits and its accesses peak was between days 15 and 20. Finally, an analysis of the periods showed an increase of the number of scans ports (445, 80 and 8080), indicating the behavior of worm attacks.

VI. Conclusion

This paper proposed a method to identify unwanted data flow in a simple and low cost manner. The method applies a low interactivity honeypot with a public unused IP address to create groups of correlated data. The honeypot stores information on each IBR access in the private network, thereby providing information on the unwanted accesses from the same sources of the IBR traffic. An evaluation showed the method effectiveness to report unwanted network accesses within a period of time, and thus supporting safety policies.

References

[1] E. Feitosa, E. Souto, and D. Sadok, "Internet unwanted traffic: Concepts and solutions," *Textbook of Minicourses of the VIII Brazilian Symposium on Information Security and Computer Systems*, pp. 91–137, 2008.

[2] R. Pang, V. Yegneswaran, P. Barford, V. Paxson, and L. Peterson, "Characteristics of internet background radiation," in *4th ACM SIGCOMM conference on Internet measurement*. ACM, 2004, pp. 27–40.

[3] E. Wustrow, M. Karir, M. Bailey, F. Jahanian, and G. Huston, "Internet background radiation revisited," in *10th ACM SIGCOMM conference on Internet measurement*. ACM, 2010, pp. 62–74.

[4] D. Moore, C. Shannon, G. M. Voelker, and S. Savage, *Network telescopes: Technical report*. Department of Computer Science and Engineering, University of California, San Diego, 2004.

[5] R. Tiwari and A. Jain, "Design and analysis of distributed honeypot system." *International Journal of Computer Applications*, vol. 55, 2012.

[6] B. Krishnamurthy, "Mohonk: Mobile honeypots to trace unwanted traffic early," in *ACM SIGCOMM Net. Troublesh.*, 2004, pp. 277–282.

[7] POP-PR, "Parana's RNP Point of Presence," http://pop-pr.rnp.br/ - Access on: 11/2013.

[8] G. Salles-Loustau, R. Berthier, E. Collange, B. Sobesto, and M. Cukier, "Characterizing attackers and attacks: An empirical study," in *Dependable Computing (PRDC), 2011 IEEE 17th Pacific Rim International Symposium on*. IEEE, 2011, pp. 174–183.

[9] J. Goebel, T. Holz, and C. Willems, "Measurement and analysis of autonomous spreading malware in a university environment," in *Detection of Intrusions, Mal., and Vuln. Assessment*. Springer, 2007, pp. 109–128.

[10] W. T. Strayer, D. Lapsely, R. Walsh, and C. Livadas, "Botnet detection based on network behavior," in *Botnet Detec.* Springer, 2008, pp. 1–24.

[11] E. Peter and T. Schiller, "A practical guide to honeypots," *Washington Univerity*, 2011.

[12] N. Provos, "A virtual honeypot framework." in *USENIX Security Symposium*, vol. 173, 2004.

[13] IANA, "Internet Assigned Numbers Authority," http://www.iana.org - Access on: 11/2013.