Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/cose

Computers & Security

An empirical comparison of botnet detection methods

S. García [a,b,*], M. Grill [b], J. Stiborek [b], A. Zunino [a]

[a] ISISTAN Research Institute – CONICET, Faculty of Sciences, UNICEN University, Argentina
[b] Agents Technology Group, Department of Computer Science and Engineering, Czech Technical University in Prague, Czech Republic

ARTICLE INFO

Article history:
Received 21 October 2013
Received in revised form
29 April 2014
Accepted 27 May 2014
Available online 5 June 2014

Keywords:
Botnet detection
Malware detection
Methods comparison
Botnet dataset
Anomaly detection
Network traffic

ABSTRACT

The results of botnet detection methods are usually presented without any comparison. Although it is generally accepted that more comparisons with third-party methods may help to improve the area, few papers could do it. Among the factors that prevent a comparison are the difficulties to share a dataset, the lack of a good dataset, the absence of a proper description of the methods and the lack of a comparison methodology. This paper compares the output of three different botnet detection methods by executing them over a new, real, labeled and large botnet dataset. This dataset includes botnet, normal and background traffic. The results of our two methods (BClus and CAMNEP) and BotHunter were compared using a methodology and a novel error metric designed for botnet detections methods. We conclude that comparing methods indeed helps to better estimate how good the methods are, to improve the algorithms, to build better datasets and to build a comparison methodology.

1. Introduction

It is difficult to estimate how much a new botnet detection method improves the current results in the area. It may be done by comparing the new results with other methods, but this has already been proven hard to accomplish (Aviv and Haeberlen, 2011). Among the factors that prevent these comparisons are: the absence of proper documentation of the methods (Tavallaee et al., 2010), the lack of a common, labeled and good botnet dataset (Rossow et al., 2012), the lack of a comparison methodology (Aviv and Haeberlen, 2011) and the lack of a suitable error metric (Salgarelli et al., 2007).

Although the comparison of methods can greatly help to improve the botnet detection area, few proposals made such a comparison (García et al., 2013). As far as we know, only three papers (Wurzinger et al., 2010; Zhao et al., 2013; Li et al., 2010) made the effort so far.

Obtaining a good dataset for comparisons is difficult. Currently, most detection proposals tend to create their own botnet datasets to evaluate their methods. However, these datasets are difficult to create (Lu et al., 2009) and usually end up being suboptimal (Shiravi et al., 2012), i.e. they lack some important features, such as ground-truth labels, heterogeneity or real-world traffic. These custom datasets are often difficult to use for comparison with other methods. This is because each

* *Corresponding author*. ISISTAN Research Institute–CONICET, Faculty of Sciences, UNICEN University, Argentina.
E-mail addresses: sebastian.garcia@isistan.unicen.edu.ar, eldraco@gmail.com (S. García), grill@agents.fel.cvut.cz (M. Grill), jan.stiborek@agents.felk.cvut.cz (J. Stiborek), alejandro.zunino@isistan.unicen.edu.ar (A. Zunino).
http://dx.doi.org/10.1016/j.cose.2014.05.011
0167-4048/

method is usually focused on different properties of the dataset. The problem is to find a good, common and public dataset that can be read by all methods and satisfy all the constrains.

The difficultly to compare detection methods goes beyond the dataset. The lack of good descriptions of the methods and error metrics contribute to the problem. As stated by Rossow et al. (2012), the error metrics used on most papers are usually non-homogeneous. They tend to use different error metrics and different definitions of error. Moreover, the most common error metrics, e.g. FPR, seems to be not enough to compare botnet detection methods. The classic error metrics were defined from a statistical point of view and they fail to address the detection needs of a network administrator.

The goal of this paper is to compare three botnet detection methods using a simple and reproducible methodology, a good dataset and a new error metric. The contributions of our paper are:

- A deep comparison of three detection methods. Our own algorithms, CAMNEP and BClus, and the third-party algorithm BotHunter (Gu et al., 2007).
- A simple methodology for comparing botnet detection methods along with the corresponding public tool for reproducing the methodology.
- A new error metric designed for comparing botnet detection methods.
- A new, large, labeled and real botnet dataset that includes botnet, normal and background data.

We conclude that the comparison of different botnet detection methods with other proposals is highly beneficial for the botnet research community because it helps to objectively assess the methods and improve the techniques. Also, that the use of a good botnet dataset is paramount for the comparison.

The rest of the paper is organized as follows. Section 2 shows previous work in the area. Section 3 describes the CAMNEP detection method. Section 4 shows the BClus botnet detection method. Section 5 describes the BotHunter method. Section 6 describes the dataset and its features. Section 7 describes the comparison methodology, the public tool and the new error metric. Section 8 shows the results and compares the methods and Section 9 presents our conclusions.

2. Previous work

The comparison of detection methods is usually considered a difficult task. In the case of botnets it is also related to the creation of a new dataset. The next Subsections describe the previous work in the area of comparison of methods and the area of creation of datasets.

2.1. *Comparison of methods*

The comparison of a new detection method with a third-party method is difficult. In the survey presented by García et al. (2013), where there is a deep analysis of fourteen network-based botnet detection methods, the authors found only one paper that made such a comparison. The survey compared the motivations, datasets and results of the fourteen proposals. It concludes that it is difficult to compare the results with another proposal because the datasets tend to be private and the descriptions of the methods tend to be incomplete.

Another analysis of the difficulty of reproducing a method was described by Tavallaee et al. (2010), where they state that there is an absence of proper documentation of the methods and experiments in most detection proposals.

One of the detection proposals that actually made a comparison with a third-party method was presented by Wurzinger et al. (2010). The purpose of the paper is to identify single infected machines using previously generated detection models. It first extracts the characters strings from the network to find the commands sent by the C&C and then it finds the bot responses to those commands. The authors downloaded and executed the BotHunter program of Gu et al. (2007) on their dataset and made a comparison. However, the paper only compares the results of both proposals using the TPR error metric and the FP values.

The other paper that made a comparison with a third-party method was presented by Zhao et al. (2013). This proposal selects a set of attributes from the network flows and then applies a Bayes Network algorithm and a Decision Tree algorithm to classify malicious and non-malicious traffic. The third-party method used for comparison was again BotHunter. There is a description of how BotHunter was executed, but unfortunately the only error metric reported was a zero False Positive. No other numerical values were presented.

The last proposal that also compared its results with a third-party method was made by Li et al. (2010). This paper analyzes the probable bias that the selection of ground-truth labels might have on the accuracy reported for malware clustering techniques. It states that common methods for determining the ground truth of labels may bias the dataset toward easy-to-cluster instances. This work is important because it successfully compared its results with the work of Bayer et al. (2009). The comparison was done with the help of Bayer et al., who run the algorithms described in Li et al. (2010) on their private dataset.

Regarding the creation of datasets for malware-related research, Rossow et al. (2012) presented a good paper about the prudent practices for designing malware experiments. They defined a prudent experiment as one being correct, realistic, transparent and that do not harm others. After analyzing 36 papers they conclude that most of them had shortcomings in one or more of these areas. Most importantly, they conclude that only a minority of papers included real-world traffic in their evaluations.

2.2. *Datasets available*

Regarding botnet datasets that are available for download, a deep study was presented in Shiravi et al. (2012) about the generation of datasets. It describes the properties that a dataset should have in order to be used for comparison purposes. The dataset used in the paper includes an IRC-based Botnet attack,[1] but the bot used for the attack was developed by the authors and therefore it may not represent a real botnet behavior. This dataset may be downloaded with authorization.

[1] http://www.iscx.ca/datasets.