Computer Networks 57 (2013) 475–486

Contents lists available at SciVerse ScienceDirect

Computer Networks

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

CoCoSpot: Clustering and recognizing botnet command and control channels using traffic analysis

Christian J. Dietrich [a,c,*], Christian Rossow [a,b], Norbert Pohlmann [a]

[a] Institute for Internet Security, University of Applied Sciences Gelsenkirchen, Neidenburger Str. 43, 45877 Gelsenkirchen, Germany
[b] VU University Amsterdam, The Network Institute, The Netherlands
[c] Department of Computer Science, Friedrich-Alexander University, Erlangen, Germany

ARTICLE INFO

Article history:
Received 20 December 2011
Received in revised form 15 June 2012
Accepted 20 June 2012
Available online 15 July 2012

Keywords:
Botnet C&C
Botnet detection
Traffic analysis
Network security

ABSTRACT

We present CoCoSpot, a novel approach to recognize botnet command and control channels solely based on traffic analysis features, namely carrier protocol distinction, message length sequences and encoding differences. Thus, CoCoSpot can deal with obfuscated and encrypted C&C protocols and complements current methods to fingerprint and recognize botnet C&C channels. Using average-linkage hierarchical clustering of labeled C&C flows, we show that for more than 20 recent botnets and over 87,000 C&C flows, CoCoSpot can recognize more than 88% of the C&C flows at a false positive rate below 0.1%.

1. Introduction and problem statement

A defining characteristic of a bot is its ability to be remote-controlled by way of command and control (C&C). Typically, a bot receives commands from its master, performs tasks and reports back on the execution results. All communication between a C&C server and a bot is performed using a specific C&C protocol over a certain C&C channel. Consequently, in order to instruct and control their bots, bot masters – knowingly or not – have to define and use a certain command and control protocol. The C&C protocol is thus considered a bot-inherent property.

Historically, bots used cleartext C&C protocols, such as plaintext messages transmitted using IRC or HTTP. However, a C&C channel relying on a plaintext protocol can be detected reliably. Methods such as payload byte signatures as shown by Rieck et al. [20] or heuristics on common C&C message elements such as IRC nicknames as proposed by Goebel and Holz [11] are examples for such detection techniques. To evade payload-based detection, botnets have evolved and often employ C&C protocols with obfuscated or encrypted messages as is the case with Waledac [7], Zeus [6], Hlux [27], TDSS [12], Virut [21] and Feederbot [8], to name but a few. The change towards encrypted or obfuscated C&C messages effectively prevents detection approaches that rely on plaintext C&C message contents.

In this article, we take a different approach to recognize C&C channels of botnets and fingerprint botnet C&C channels based on traffic analysis properties. The rationale behind our methodology is that for a variety of botnets, characteristics of their C&C protocol manifest in the C&C communication behavior. For this reason, our recognition approach is solely based on traffic analysis.

As an example, consider a C&C protocol that defines a specific handshake – e.g., for mutual authentication – to be performed in the beginning of each C&C connection. Each request and response exchanged during this handshake procedure conforms to a predefined structure and

* Corresponding author at: Institute for Internet Security, University of Applied Sciences Gelsenkirchen, Neidenburger Str. 43, 45877 Gelsenkirchen, Germany. Tel.: +49 2099596696.
E-mail addresses: dietrich@internet-sicherheit.de (C.J. Dietrich), rossow@internet-sicherheit.de (C. Rossow), pohlmann@internet-sicherheit.de (N. Pohlmann).

1389-1286/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.comnet.2012.06.019

Table 1
Examples of message length sequences for Virut and Palevo C&C flows.

ID	Family	Message length sequence							
		1	2	3	4	5	6	7	8
1	Virut	60	328	12	132	9	10	9	10
2	Virut	69	248	69	10	9	10	9	10
3	Virut	68	588	9	10	9	10	9	10
4	Virut	67	260	9	10	9	10	9	10
5	Palevo	21	21	30	197	32	10	23	10
6	Palevo	21	21	30	283	21	10	23	10

length, which in turn leads to a characteristic sequence of message lengths. In fact, we found that in the context of botnet C&C, the sequence of message lengths is a well-working example for traffic analysis features. Table 1 shows the sequence of the first 8 messages in four Virut C&C flows and two Palevo[1] C&C flows. Whereas Virut exhibits similar message lengths for the first message (in the range 60–69) and a typical sequence of message lengths at positions 5–8, for Palevo, the first three message lengths provide a characteristic fingerprint.

Leveraging statistical protocol analysis and hierarchical clustering analysis, we develop CoCoSpot[2], a method to group similar botnet C&C channels and derive fingerprints of C&C channels based on the message length sequence, the underlying carrier protocol and encoding properties. Furthermore, we design a classifier that is able to recognize known C&C channels in network traffic of contained malware execution environments, such as Sandnet [22].

The ability to recognize botnet C&C channels serves several purposes. A bot (net)'s C&C channel is a botnet's weakest link [10]. Disrupting the C&C channel renders a bot (net) ineffective. Thus, it is of high interest to develop methods that can reliably recognize botnet C&C channels. Furthermore, driven by insights of our analysis of botnet network traffic, we found that a bot's command and control protocol serves as a fingerprint for a whole bot family. Whereas for example properties of the PE binary change due to polymorphism, we witness that the C&C protocol and the corresponding communication behavior seldom undergo substantial modifications throughout the lifetime of a botnet. From an analyst's perspective, our classifier helps to detect and aggregate similar C&C channels, reducing the amount of manually inspected traffic.

To summarize, our main contributions are twofold:

- We provide a clustering method to analyze relationships between botnet C&C flows.
- We present CoCoSpot, a novel approach to recognize botnet command and control channels solely based on traffic analysis features, namely carrier protocol distinction, message length sequences and encoding differences.

[1] A synonym for the malware family Palevo is Rimecud (Microsoft terminology).

[2] CoCoSpot is derived from spotting command and control.

The remainder of this article is structured as follows. Section 2 sheds light on related work, defines the scope of this article and highlights innovative aspects of our approach. Subsequently, in Section 3, the general methodology as well as the feature space is described. While Section 4 deals with the clustering phase of C&C flows, Section 5 outlines the classifier which is then used to classify unknown flows. In order to evaluate our approach, as described in Section 6, we classified arbitrary network flows emitted from the dynamic malware analysis environment Sandnet as either C&C or Non-C&C and verified the results using two datasets. Finally, we discuss limitations of our approach in Section 7 and conclude in Section 8.

2. Related work

Traditionally, botnet C&C channels have mainly been identified in two ways. First, publicly available blacklists [1,13,19,23] provide lists of known botnet servers by IP addresses or domain names. The drawback of blacklists is that properties like IP addresses or domains are volatile. Bot masters can and do change these often, rendering detection methods based blacklists infeasible. In addition, botnets that rely on a peer-to-peer C&C architecture exhibit quickly changing sets of rendez-vous points. Some bot masters design their bot's bootstrap process to be even more resilient by avoiding any static rendez-vous coordinates, e.g., by using domain generation algorithms where the current rendez-vous point is valid for a very limited time span such as a few hours. Second, botnet C&C channels can be detected by checking for characteristic payload substrings. For example, Botzilla [20], Rishi [11] and rules for the Snort IDS [25] identify C&C channels in network traffic using payload byte signatures for a small set of known botnets. However, most encrypted or obfuscated C&C protocols do not exhibit characteristic payload patterns and undermine existing payload byte signatures.

Consequently, the traditional techniques are unsatisfying and have motivated research for automated and more reliable processes. In that trail of research, BotMiner [14] and BotGrep ([18]) provide approaches to use traffic analysis in order to find botnet C&C channels. However, while BotMiner requires detectable so-called A-plane activity such as spam, DDoS or scanning, *CoCoSpot* does not require any *a priori* or accompanying malicious actions and aims at the *recognition* of C&C channels. For *CoCoSpot*, in order to detect a bot, it is enough to just exhibit C&C communication. The graph-based approach of BotGrep requires botnets with distributed C&C architectures in order to detect them. However, *CoCoSpot* not only works with P2P botnets, but also with botnets exhibiting a centralized C&C architecture.

Jacob et al. present JACKSTRAWS [16], which exploits that certain C&C channels show recognizable system-level behavior in terms of system call traces of Anubis [5]. Particularly, JACKSTRAWS dynamically analyzes malware binaries (e.g., with Anubis) and models system call graphs for known C&C connections. New C&C channels are detected by matching unknown network connections against these graphs. As opposed to JACKSTRAWS,