Chapter 3
Analysis, Development and Deployment of Statistical Anomaly Detection Techniques for Real E-Mail Traffic

Gianluca Papaleo
Istituto di Elettronica e di Ingegneria dell'Informazione e delle Telecomunicazioni, Italy & Consiglio Nazionale delle Ricerche, Italy

Davide Chiarella
Istituto di Elettronica e di Ingegneria dell'Informazione e delle Telecomunicazioni, Italy & Consiglio Nazionale delle Ricerche, Italy

Maurizio Aiello
Istituto di Elettronica e di Ingegneria dell'Informazione e delle Telecomunicazioni, Italy & Consiglio Nazionale delle Ricerche, Italy

Luca Caviglione
Istituto di Studi sui Sistemi Intelligenti per l'Automazione, Italy & Consiglio Nazionale delle Ricerche, Italy

ABSTRACT

Even if new interaction paradigms, such as the Voice over IP (VoIP), are becoming popular and widely adopted, the e-mail is still one of the most utilized ways to communicate across the Internet. However, many malicious threats are conveyed via e-mails. Usually, the authors can exploit two different approaches: i) analyzing the logs produced by e-mail servers or ii) reconstruct the e-mail flows by capturing data directly from the network by placing ad-hoc probes. In this vein, this Chapter discusses the analysis, development and deployment of statistical detection techniques aimed at the detection of Internet worms. For what concerns i), they introduce a tool called Log Mail Analyzer (LMA), which allows to overcome the complexity of inspecting multiple logs created from a heterogeneous population of mail servers. In the perspective of ii) they briefly discuss an alternative solution, based on ad-hoc network probes, to be properly placed to collect traffic and then reconstruct the e-mail flow to be monitored. Lastly, the authors introduce a threshold mechanism, based on a simple statistical framework, to automatically detect and identify different worm activities.

DOI: 10.4018/978-1-61350-507-6.ch003

INTRODUCTION

Despite new interaction paradigms, such as the Voice over IP (VoIP), are becoming popular and largely adopted, the electronic mail (e-mail) is still one of the most preferred method to communicate across the Internet. Also it is available for a variety of mobile devices, ranging from Personal Data Assistants (PDAs) to regular gaming consoles, as well as many network appliances. In addition, e-mail is constantly used also while on the road, for instance from cellular phones. As a matter of fact, ad-hoc mechanisms to handle e-mail traffic over the cellular network, such as the push e-mail, dramatically increase its usage and popularity among power users and businessmen.

Anyway, due to its ubiquitous adoption, the e-mail is one of the most used vector to spread malicious programs like viruses and *worms*. This is mainly due to the superimposition of the following reasons: *i*) e-mails can carry attachments, thus allowing malicious software to easily propagate through the Internet without the need of developing specific communication mechanisms to spread the infection. Besides, e-mail traffic, even if filtered, is commonly allowed in the majority of network deployments, or loosely regulated, as conversely happens for peer-to-peer (p2p) file-sharing applications; *ii*) e-mails can be retrieved, organized and sent through dedicated client-interfaces, which are often affected by several security breaches. Thus, they account for infecting the hosting machine and to spread the malicious entity through an automated and hidden e-mail flow; *iii*) an e-mail account is generally required to join modern social-oriented services, e.g., social networks. Therefore, it is possible to collect in a simple manner a huge amount of addresses and personal information, e.g., usernames and passwords. Such stolen accounts are often used to send malicious contents. Furthermore, the increasing diffusion of Web 2.0 applications rises the usage of malicious code within web pages composing the service itself. We cite, among the others, menaces such as the Cross-Site Scripting (XSS) and the Cross-Site Request Forgery (CSRF or XSRF) that can be injected within the Asynchronous Javascript and XML (AJAX) powered contents. Nevertheless, the e-mail infrastructure can be infected from completely different services too. As an example, malicious software can be retrieved from file-sharing networks (Caviglione, 2009). Then, if executed it can attach to the client-interface devoted to manage e-mails as to "hijack" the network infrastructure running the e-mail delivery service. Yet, file-sharing client-interfaces can have built-in malware code in order to spawn *botnets*. Even if they are usually adopted for cycle-stealing software, botnets can also use the e-mail traffic of hosted machines to spread the infection or to produce unsolicited e-mails, e.g., spam messages. Lastly, e-mails are also plagued by other hazards, such as phishing. Even if they are relevant problems, both in terms of traffic and waste of resources, investigations devoted to recognize and prevent such issues are out of the scope of this Chapter. Conversely, it focuses on a statistical framework for the detection of *worms*.

The Literature shows two main approaches to perform worm detection: *misuse* based and *anomaly* based. The first one employs the signature concept, while the second one tries to create a model to characterize users' normal behaviors. Misuse detection lacks the ability of identifying the presence of worms not fitting a pre-defined signature. Conversely, in anomaly detection the system defines an expected network performance and, if there are significant deviations from the given profile, spawns an alarm. Despite the particular techniques, the needed conceptual steps to effectively recognize and take subsequent actions to detect anomalies within e-mail traffic are: *i*) performing monitoring actions by gathering information about the observed e-mail flow and *ii*) investigating the collected data by using some criteria to reveal anomalies (possibly in an automatic way).