International Journal of Information Security and Privacy
Volume 16 • Issue 1

A Proposal to Distinguish DDoS Traffic in Flash Crowd Environments

Anderson Aparecido Alves da Silva, SENAC, Brazil & IPT, Brazil & UNIP, Brazil & USP, Brazil

Leonardo Santos Silva, IPT, Brazil

Erica Leandro Bezerra, USP, Brazil

Adilson Eduardo Guelfi, UNOESTE, Brazil

Claudia de Armas, USP, Brazil

Marcelo Teixeira de Azevedo, USP, Brazil

https://orcid.org/0000-0003-1676-7380

Sergio Takeo Kofuji, USP, Brazil

ABSTRACT

A flash crowd (FC) event occurs when network traffic increases suddenly due to a specific reason (e.g., e-commerce sale). Despite its legitimacy, this kind of situation usually decreases the network resource performance. Furthermore, attackers may simulate FC situations to introduce undetected attacks, such as distributed denial of service (DDoS), since it is very difficult to distinguish between legitimate and malicious data flows. To differentiate malicious and legitimate traffic, the authors propose applying zero inflated count data models in conjunction with the correlation coefficient flow (CCF) method – a well-known method used in FC situations. The results were satisfactory and improve the accuracy of CCF method. Furthermore, since the environment toggles between normal and FC situations, the method has the advantage of working in both situations.

KEYWORDS

Binominal Negative, Distributed Denial of Service (DDos), Flash Crowd, Poisson, Zero Inflated Model

INTRODUCTION

Distributed Denial of Service (DDoS) attacks are among the worst threats to network systems due to their diversity and level of sophistication (Zargar et al., 2013) and (Hoque et al., 2015). The literature on this topic presents several methodologies to deal with that issue; each one is suited to a specific situation (Wang et al., 2018) and (François et al., 2012). A very particular case is the DDoS attack that occurs in Flash Crowd (FC) environments. FC are situations where the Internet network flow suffers a sharp increase in traffic and resource utilization due to a legitimate reason (an e-commerce sale, for example). Such behavior causes performance degradation on the network according to Oikonomou and Mirkovic (2009), Li et al. (2008), Ari et al. (2003), Paxson and Floyd (1994), and Beitollahi and Deconinck (2014).

DOI: 10.4018/IJISP.2022010104

Technically it is quite challenging to distinguish FC traffic from malicious behavior (Yu et al., 2009), (Jung et al., 2002) and (Oikonomou & Mirkovic, 2009). However, few of these works analyze scenarios where traffic toggles between normal and FC behaviors - we consider normal traffic, the legitimate network flow without (or with few) malicious packets and not in an FC situation.

Despite the profusion of methodologies, Yu et al. (2011) developed a Correlation Coefficient Flow method (CCF), which has the merit of supporting both regular and FC environments. However, the CCF method has some known vulnerabilities: it depends on the number of botnets close to the users' number; furthermore, it was not tested in an FC scenario with a DDoS attack.

To improve the accuracy of the CCF method, we propose using count data models (e.g., Poisson and Negative Binomial (BINEG)), which can be used in situations where it is crucial to know the number of times some independent events occur (Park et al., 2006) and (Guerin et al., 2015). In addition to the distinction between legitimate traffic and DDoS, our contribution includes analysis of (1) regular traffic, (2) DDoS over regular traffic, (3) FC traffic, and (4) rare samples with DDoS attacks over FC traffic. To test our proposal, we experiment with a real dataset and compare the results applying the core of the CCF method developed by Yu et al. (2011).

THEORETICAL FOUNDATION AND RELATED WORKS

In this section, the conceptual basis that guides the research will be identified and the main works that sought to deal with the same problem described, highlighting what differentiates our research from the works already carried out.

Count Data Distributions

The count data distributions such as Poisson and BINEG serve to determine the number of event occurrences within a discrete period since these events are independent. Generally, these distributions are used when the sample n is large, and the probability of occurrence p of an event is low (Heckert et al., 2002).

Poisson

Poisson regression models are frequently used to analyze count data. A random variable Y with integer values $y=\{0,1,2,\ldots\}$ and an average number of occurrences $\mu>0$ has a Poisson distribution with probability (Ridout & Hinde, 1998) and (Dobsonh & Barnett, 2018):

$$P\left\{Y=y\right\}=\frac{e^{-\mu}\mu^{y}}{y!} \tag{1}$$

An important issue in the Poisson distribution is that variance is equal to mean: $E(Y)=var(Y)=\mu$. The parameter μ is also used to model the effect of independent variables in the response variable Y through regression. Let $Y=(Y_1, \ldots, Y_n)$ be independent random variables where Y_i is the i^{th} event of n_i and $\theta=(\theta_1, \ldots, \theta_n)$ is the vector of parameters of the distribution, the expected value of Y_i is $E(Y_i)=\mu_i=n_i\theta_i$ and the model dependence of θ_i on the independent variables is: $\theta_i = e^{x_i^T\beta}$. Therefore, the GLM is (Dobsonh & Barnett, 2018):

$$E\left(Y_i\right)=\mu_i=n_i e^{x_i^T\beta} \tag{2}$$

Negative Binomial (BINEG)

In a series of independent trials with Bernoulli distribution, the BINEG distribution indicates the number of faults before the k^{th} success. A BINEG regression model differs from a Poisson model to