Poster: On Predicting BGP Anomalous Incidents: A Bayesian Approach

Clint McElroy
Indiana University
clmcelro@indiana.edu

Pablo Moriano
Indiana University
pmoriano@indiana.edu

L. Jean Camp
Indiana University
ljcamp@indiana.edu

***Abstract*—Despite multiple efforts to secure the Internet control plane, Border Gateway Protocol (BGP) anomalous incidents have been increasing in both frequency and impact over the past several years. These anomalies are events in which Internet traffic is accidentally or maliciously routed incorrectly. Here, we examine a popular public dataset of BGP anomalies to develop Bayesian Generalized Linear Models (BGLMs) that capture the frequency and impact of BGP anomalies. We find that the daily frequency can be modeled by a lognormal distribution while their impact is better captured by a discrete Laplace distribution. Knowing these distributions can provide insights into the generative mechanisms of these anomalies and inform future predictions.**

I. Introduction

The Border Gateway Protocol (BGP) is a core protocol connecting the Internet infrastructure, allowing the exchange of routing information between autonomous systems (ASes). Attacks that leverage this protocol are also on the rise, affecting Internet traffic. The reason BGP can be manipulated is because ASes rely on initial trust rather than repeated verification, meaning they are susceptible to being hijacked, sometimes for undesirable purposes (including e-crime and nation-state espionage [2]). The goal of our research is to analyze the data obtained from Argus [3] to learn trends on the generative mechanisms of BGP anomalies in terms of frequency and impact. We use that knowledge to build data-driven Bayesian models that can be used to make future predictions.

II. Data

The source of our data is Argus, a BGP anomaly detection and classification system developed at Tsinghua University. Argus constantly correlates control- and data-plane information to detect anomalies while reducing the number of false negatives and false positives. The dataset we explore contains nearly 4.4 million BGP anomalies that were collected between June 3, 2011 and March 25, 2017. With each anomaly detected, a number of parameters are logged, such as date, time, impacted IP prefixes, impacted IP addresses, and anomaly type.

The anomalies captured by Argus are identified as being one of three types: an origin anomaly (OA), adjacency anomaly (AA), or policy anomaly (PA). In this paper, our focus is on OAs, which are often referenced as Multiple Origin AS conflicts [4]. These anomalies are sometimes the result of rogue announcements advertising incorrect traffic routes diverting IP address blocks away from legitimate ASes to the rogue ones.

Of the nearly 4.4 million anomalies, we focus on the 640,870 OAs that impact IPv4 addresses and disregard the 1,671 that impact IPv6 addresses. It should be noted, however, that there are limitations to the completeness and accuracy of this data and we are also reliant on anomalies being accurately identified.

III. Modeling anomaly event trends

Here, we present our approach in modeling these OAs using Bayesian Generalized Linear Models (BGLMs).

A. Daily frequency

Daily frequency refers to the number of OAs detected in one calendar day. Let N denote the daily number of anomalies. Figure 1 shows its complementary cumulative distribution function (CCDF) for the empirical data, which resembles a heavy-tailed distribution. We determine the best distribution that fits this data by examining different candidates, including, lognormal, log-skewed normal, exponential, and power-law. For each candidate, we use maximum likelihood to estimate the best parameters for the candidate distribution. We then perform a Kolmogorov-Smirnov (KS) test to determine the goodness of fit. Figure 1 also shows the fit to the lognormal distribution. The KS test results in $p = 0.49$, meaning we cannot rule out this possibility. For all other distribution candidates, the KS test produces p-values that suggest rejection of the hypothesis as possible generators of the distributions.

With this information, we built a BGLM to model the frequency of the anomalies, given by Equation 1.

$$\begin{array}{rcl} N & \sim & Lognormal(\mu, \tau) \\ \mu & = & \beta_0 + \beta_1 t + \beta_2 t^2 + \cdots + \beta_d t^d \\ \beta_0 & \sim & \mathcal{N}(Median(\log(N)), 1) \\ \beta_i & \sim & \mathcal{N}(0, \frac{1}{Var[t^i]}) \\ \tau & \sim & Gamma(1, 1) \end{array} \qquad (1)$$

Here, μ is the location parameter and τ is the shape parameter of a lognormal distribution. We assume that the location parameter is a polynomial function of time, t, to account for temporal variations in the generation of anomalies. The rest of the model priors have been defined as suggested in [1]. We use the Bayesian Information Criterion (BIC) to determine the best order of the polynomial for μ, which is $d = 0$, indicating

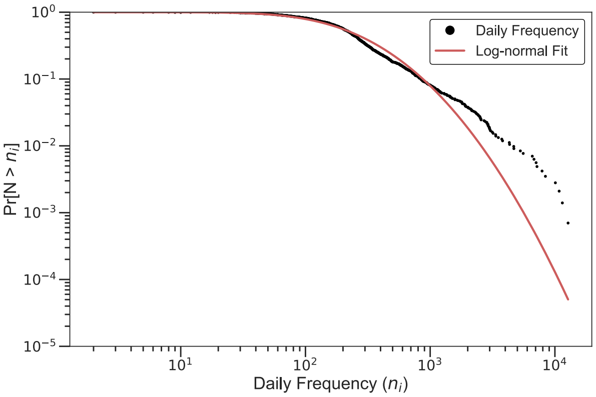

Fig. 1. CCDF plot of daily OA frequency and the fit to a lognormal distribution.

a constant daily frequency of anomalies. We then use Markov chain Monte Carlo (MCMC) to sample from the proposed model and get confidence intervals for its parameters.

B. Prefix length

The prefix length is directly related to the size of the affected IP space. We estimate the impact of BGP incidents by measuring the prefix length, or mask number, and thus the potential to compromise hosts.

We determine the discrete Laplace distribution to be the best fit based upon a goodness of fit Chi-squared test. Let M denote the prefix length. Among other discrete distribution candidates we examined, negative binomial, Poisson, geometric, and hypergeometric. We restrict the range from 0 to 32 for every possible prefix length. Figure 2 shows the discrete Laplace fit to the empirical data.

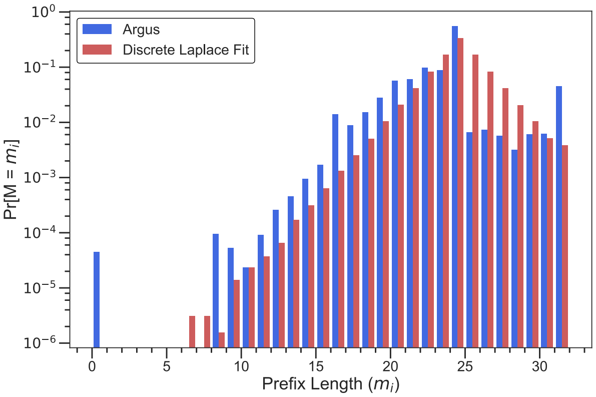

Fig. 2. Distribution of prefix length and its fit to a discrete Laplace distribution.

With this information, we built a BGLM to model the impact of the anomalies, given by Equation 2.

$$\begin{aligned} M &\sim discreteLaplace(\alpha, \sigma) \\ \alpha &= \beta_0 + \beta_1 t + \beta_2 t^2 + \cdots + \beta_d t^d \\ \beta_0 &\sim \mathcal{N}(Median(M), 1) \\ \beta_i &\sim \mathcal{N}(0, \frac{1}{Var[t^i]}) \\ \sigma &\sim Gamma(1, 1) \end{aligned} \tag{2}$$

Here, α is the location parameter and σ is the shape parameter of the discrete Laplace distribution. We use the BIC criteria to determine the order of the polynomial for α, which is $d = 0$, and then MCMC sampling to determine the best parameter values.

In the poster itself, we include additional visualizations of our analysis. In particular, we plot the empirical data against synthetic observations generated through the data-driven models using the best estimates of the parameters for frequency and prefix length using MCMC.

The value of this predictive modeling is in improving the understanding of the generative mechanisms of OAs. The model allows for estimation of the impact of past and, if trends hold, future impact of anomalies.

IV. Predictions

Because of the heavy-tailed nature in the frequency of the OAs, it is not feasible to make precise predictions about the exact number of routing anomalies and their impact. However, our model can assess the likelihood of the occurrence of anomalies and their respective impact, i.e., it can predict the probability of a certain number of anomalies of a specific prefix during a given time-frame. Table I shows the proportion of synthetic generated anomalies for some prefix lengths by the end of 2017 and 2018. A complete estimation for remaining prefix lengths is included in the poster itself. Note that we can say that despite high variability in terms of daily frequencies, almost 17% of OAs in 2017 are affecting /24s prefixes. This tendency does not change much for 2018.

TABLE I
Predictions for March 26, 2017 through the end of 2017 and 2018 by prefix length.

Prefix length	% Chance	
	End 2017	End 2018
8	0	0.002
16	0.263	0.243
24	16.732	16.800
32	0.054	0.072

Acknowledgment

This work was supported in part by NSF award CNS-1566375, Cisco Research support #591000, and the Comcast Innovation Award.

References

[1] A. Gelman and J. Hill, *Data analysis using regression and multilevel-hierarchical models*. Cambridge University Press, New York, NY, USA, 2007, vol. 1.

[2] P. Moriano, S. Achar, and L. J. Camp, “Incompetents, criminals, or spies: Macroeconomic analysis of routing anomalies,” *Computers & Security*, vol. 70, pp. 319–334, 2017.

[3] X. Shi, Y. Xiang, Z. Wang, X. Yin, and J. Wu, “Detecting Prefix Hijackings in the Internet with Argus,” in *Proceedings of the 2012 ACM Conference on Internet Measurement Conference*, Boston, MA, USA, 2012, pp. 15–28.

[4] X. Zhao, D. Pei, L. Wang, D. Massey, A. Mankin, S. F. Wu, and L. Zhang, “An Analysis of BGP Multiple Origin AS (MOAS) Conflicts,” in *Proceedings of the 1st ACM SIGCOMM Workshop on Internet Measurement*, San Francisco, CA, USA, 2001, pp. 31–35.

On Predicting BGP Anomalous Incidents: A Bayesian Approach

Clint M^cElroy, Pablo Moriano, L. Jean Camp
{clmcelro, pmoriano, ljcamp}@indiana.edu
School of Informatics, Computing, and Engineering, Indiana University Bloomington

ABSTRACT

Despite multiple efforts to secure the Internet control plane, Border Gateway Protocol (BGP) anomaly incidents have been increasing in both frequency and impact over the past several years. These anomalies are events in which Internet traffic is accidentally or maliciously routed incorrectly. Estimating both the likelihood of their future recurrence and their impact is of great utility for designing incentive mechanisms. Here, we examine a popular public dataset of BGP anomalies to develop Bayesian Generalized Linear Models (BGLMs) to model the frequency and impact of BGP anomalies. We find that daily frequency can be modeled by a lognormal distribution and impact can be modeled by a discrete Laplace distribution. Knowing these distributions can provide insights into the generative mechanisms of these anomalies and the ability to inform future predictions.

QUESTION

Using this data-driven model, can predictions be made about future BGP anomalies? This is the fundamental question for our research. The implications of being able to predict future incidents is of great utility for designing incentive mechanisms.

METHODS

Data

The source of our data is Argus, an anomaly detection and classification system developed by researchers at Tsinghua University. Argus constantly correlates control- and data-plane information. The data we explore contains nearly 4.4 millions BGP anomalies that were collected between June 3, 2011 and March 25, 2017. Of the nearly 4.4 million anomalies, we focus on the 640,870 origin anomalies (OAs) that impact IPv4 addresses and disregard the 1,671 that impact IPv6 addresses.

Modeling Anomalous Events

Our approach to model these OAs is BGLMs. We perform Kolmogorov-Smirnov (KS) tests to determine the best distribution fit and use the Bayesian Information Criterion (BIC) to determine the best order of the polynomial for the location parameter by randomly sampling the model. We assume the location parameter is a polynomial function of time, *t*. After determining the order of the polynomial, we use Markov chain Monte Carlo (MCMC) to sample from the proposed model to get best parameter estimates.

ANALYSIS & RESULTS

Daily Frequency

Daily frequency is the number of anomalies in one calendar day. Figure 1 shows its complementary cumulative distribution function (CCDF) for the empirical data and the fit to the lognormal distribution. The KS test gives $p = 0.49$.

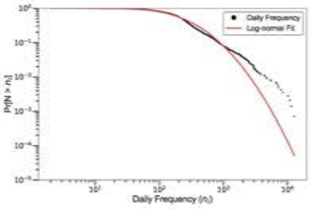

Figure 1: CCDF plot of daily OA frequency and the fit to a lognormal distribution.

Prefix Length

We estimate the impact of BGP incidents by measuring the prefix length. We determine the discrete Laplace distribution to be the best. Figure 2 shows the discrete Laplace fit to the empirical data.

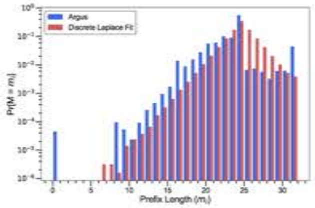

Figure 2: The distribution of prefixes and the fit to a discrete Laplace distribution.

Bayesian Analysis

The BGLM for frequency of the anomalies is given by Equation 1.

$$\begin{aligned} N &\sim Lognormal(\mu, \tau) \\ \mu &= \beta_0 + \beta_1 t + \beta_2 t^2 + \cdots + \beta_d t^d \\ \beta_0 &\sim \mathcal{N}(Median(log(N)), 1) \\ \beta_i &\sim \mathcal{N}(0, \tfrac{1}{Var[t^i]}) \\ \tau &\sim Gamma(1, 1) \end{aligned}$$

Equation 1: Daily frequency Bayesian model.

Here, the lognormal distribution parameters are μ for location and τ for shape. We use the BIC to determine the polynomial for μ, which is $d = 0$, indicating a constant daily frequency of anomalies. Figure 3 shows data generated from the model using best estimates plotted against the empirical data.

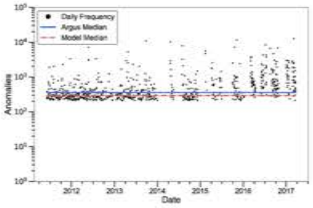

Figure 3: The daily frequency versus the maximum likelihood estimate of the median size.

We construct a BGLM to model the impact, given by Equation 2.

$$\begin{aligned} M &\sim discreteLaplace(\alpha, \sigma) \\ \alpha &= \beta_0 + \beta_1 t + \beta_2 t^2 + \cdots + \beta_d t^d \\ \beta_0 &\sim \mathcal{N}(Median(M), 1) \\ \beta_i &\sim \mathcal{N}(0, \tfrac{1}{Var[t^i]}) \\ \sigma &\sim Gamma(1, 1) \end{aligned}$$

Equation 2: Prefix length model.

Here, the discrete Laplace distribution parameters are α for location and σ for shape. We use the BIC to determine the polynomial for α, which is $d = 0$. Figure 4 shows data generated from the model using best estimates plotted against the empirical data.

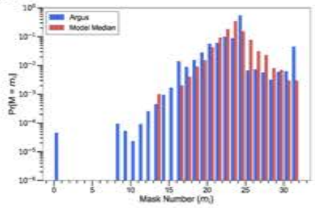

Figure 4: The distribution of impacted mask numbers and the results of randomly sampling our model.

Predictions

Due to the heavy-tailed nature in the frequency of the OA dataset, it is not feasible to make precise predictions about the exact number of routing anomalies and their impact. However, our model can assess the likelihood of the occurrence of anomalies and their respective impact. Note that we can say that despite high variability in terms of daily frequencies, almost 17 percent of OA anomalies in 2017 are affecting /24s prefixes. This tendency does not change much for 2018. Table 1 shows the proportion of synthetic generated anomalies for prefix lengths by the end of 2017 and 2018.

CONCLUSION

This research presented an analysis of BGP anomaly frequency and impact. We built a data-driven model that mimics the underlying dynamics of BGP anomalies and use that to inform future predictions.

Prefix Length	% Chance	
	End 2017	End 2018
8	0	0.002
9	0	0.001
10	0.003	0.004
11	0.009	0.008
12	0.015	0.015
13	0.040	0.030
14	0.064	0.063
15	0.110	0.117
16	0.263	0.243
17	0.485	0.492
18	1.029	1.025
19	2.075	2.043
21	8.238	8.335
22	16.880	16.646
23	33.955	33.950
24	16.732	16.800
25	8.281	8.300
26	4.066	4.036
27	2.066	2.070
28	0.980	1.000
29	0.487	0.502
30	0.236	0.237
31	0.129	0.110
32	0.054	0.072

ACKNOWLEDGMENT

This work was supported in part by NSF award CNS-1566375, Cisco Research support #591000, and the Comcast Innovation Award.

REFERENCES

[1] A. Gelman and J. Hill, Data analysis using regression and multilevel hierarchical models. Cambridge University Press, New York, NY, USA, 2007, vol. 1.

[2] P. Moriano, S. Achar, and L. J. Camp, "Incompetents, criminals, or spies: Macroeconomic analysis of routing anomalies," Computers & Security, vol. 70, pp. 319–334, 2017.

[3] X. Shi, Y. Xiang, Z. Wang, X. Yin, and J. Wu, "Detecting Prefix Hijackings in the Internet with Argus," in Proceedings of the 2012 ACM Conference on Internet Measurement Conference, Boston, MA, USA, 2012, pp. 15–28.

[4] X. Zhao, D. Pei, L. Wang, D. Massey, A. Mankin, S. F. Wu, and L. Zhang, "An Analysis of BGP Multiple Origin AS (MOAS) Conflicts," in Proceedings of the 1st ACM SIGCOMM Workshop