Research of BGP Routing Instability Detection and Analysis

Tang Qi-jie[1,a], Zhao Yu[1], Zhou Yuan[1], Zhou Yun-ting[1], Luo Qi-fu[1], Ji Feng-zhu[1]

[1]Xichang Satellite Launch Center, Xichang Sichuan, China

[a]27359943@qq.com

Keywords: Border Gateway Protocol, BGP routing instability, simulation experiment, Dynamic Data Modeling, hypothesis test.

Abstract. In this paper, we propose an BGP routing instability detection algorithm that is based on statistics and pattern recognition through deeply learning BGP-4 and seriously analyzing the cause, properties, Manifestations of BGP routing instability, and verify its correctness and effectiveness by using SSFNet simulation experiment. BGP routing instability detection algorithm can accurately detect change points in time series and what type of instability sub-time series with reasonable values of parameters.

1 Introduction

BGP routing Instability, Generally called route flaps, significantly contributes to poor end-to-end network performance and degrades the overall efficiency of the Internet infrastructure. Routing Instability, informally defined as the rapid change of network reachability and topology information, has a number of origins, including router configuration errors, transient physical and data link problems, and software bugs [1-3]. All of these sources of network instability result in a large number of routing updates that are passed to the core Internet exchange point. By long-term data collection and dynamic fault injection mechanism in major internet routing switching nodes, the experiments shows [4] that a failed inter-domain connection takes on average 3 minutes [5] to find a new stability route and extremely takes 15 minutes. Routing policy conflict existed in autonomous systems would lead BGP routing tables to flap Continuously [6].

In view of BGP routing instability problem [7] and the serious problems to the internet, this paper will study BGP routing instability problem to BGP router failures and recovery, BGP routing policy conflicts and worm virus attacks based on master understanding about BGP-4 protocol and mathematical statistics and pattern recognition.

2 Design of BGP Routing Instability Detection and Analysis

This paper based on SSFNet protocol simulation software simulates routing instability in a series of network behavior in order to collect original training samples data of BGP instability and to calculate feature extraction transformation matrix for projecting high dimensional space data into low dimensional space data. For BGP routing behavior in a certain period of time, firstly we project sample data to be detected into low dimensional space; Secondly we detect change point using BGP routing instability detection and analysis algorithm for partitioning BGP routing behavior to be detected; Thirdly we design classifier to distinguish status of BGP routing instability and type of BGP routing instability about unknown each sub-time series samples using Fisher linear discriminant.

We select optionally a BGP router to collect BGP update messages with various length of AS_PATH path attributes. Then we select the eigenvectors whose eigenvalue arranged in order of size to constitute a feature extraction transformation matrix using Euclidean distance measure of pattern recognition [8].

3 Simulation and Feature Extraction

We simulate network structure and behavior of 29 autonomous systems using large network simulation software called SSFNet. We obtain raw feature vector data and get transformation matrix for feature extraction using Euclidean distance measure by three simulations which include BGP router failure and recovery test, BGP routing policy conflicts test and worm virus attacks test. Then we analyze BGP routing instability situation making use of the data collected.

We use perl script process to deal with raw data that was collected from BGP router failure and recovery test, BGP routing policy conflicts test and worm virus attacks test to calculate the number of withdrawal messages and various length of announcement messages in every 300-second intervals. And the range of announcement messages length is 1 to 12. According with the Euclidean distance metric feature extraction method and using Matlab tool[9], we get the rank of matrix $\hat{S}_w^{-1}\hat{S}_b$ is 1 and the vector consisting of eigenvalues is

$$(0.5497,-0.0,0.0,-0.0,0.0,-0.0,-0.0+0.0i,-0.0-0.0i,-0.0+0.0i,-0.0-0.0i,0.0,0.0,0.0)^T$$

and there is only one Non-zero eigenvalue in it. In accord with feature extraction methods we get the Feature extraction transformation matrix A is a 13×1 dimensional matrix, and eigenvalue 0.5497 corresponding eigenvector is

$$(-0.0070,-0.9547,0.2925,-0.0341,0.0066,-0.0022,\\ -0.0003,-0.0224,0.0164,-0.0177,-0.0210,-0.0135,0.0098)^T$$

Then, the transformation matrix A is equivalent to the corresponding eigenvalue 0.5497 corresponding eigenvector. Then 13×1 dimensional measurement space through transformation matrix A which acts on 13×1 dimensional raw data vector mapped to 1×1 low-dimensional feature space.

4 BGP Routing Instability Detection and Analysis

This section introduces the algorithm process and design of BGP routing instability detection and analysis in detail. We use SSFNet protocol simulation software to design a simulation test which simultaneously contains BGP router failure and recovery and BGP routing policy conflicts. We analyze how algorithm parameters affect correctness and effectiveness of algorithm through the test.

4.1 Summary about Algorithm of BGP Routing Instability Detection and Analysis

We extract features form announcement messages with various length of AS_PATH path attributes in each time interval and deal the time series with mean during the entire simulation. Because there is a BGP router failure and recovery, BGP routing policy conflicts or worm attack in the simulation test, so BGP routing messages received from a variety of time series can be used to form time series analysis. Firstly, we divide the origin time series into multiple sub-time series using difference of variance of residuals corresponding two sub-time series obeyed the normal overall distribution. Then, the second step is to optimize the segment boundary position. At last, we divide the whole time series into multiple sub-time series and determine the steady state and what type of BGP routing instability state of each sub-time series using Fisher criterion.

The original initialization and optimal treatment of time series need to use difference of variance of residuals corresponding two sub-time series. In case the residual series is white noise sequence, we specifically introduce the initial partition and optimization of time series.

4.2 Initial Division of Time Series

BGP routing instability detection actually detects whether time series has change points. Therefore, BGP routing instability detection problem is changed into a statistical change point detection problem. Solving the online time series segmentation problem need to introduce the following questions:

Question 1. Assuming that sampling time m is a recent change point from sampling time $t > m$ in time series TS, the sample value of time series corresponding sampling time t is v_t, we determine whether v_t is a change point.

If v_t is a change point, then the sub-time series $TS_t = \{v_m, v_{m+1}, \cdots, v_{t-1}, v_t\}$ would be divided into tow sub-sequences: $TS_{t1} = \{v_m, v_{m+1}, \cdots, v_{t-1}\}$ and $TS_{t2} = \{v_t\}$.

We obtain a fitting time series $TS_t^{'}$ and its residuals $E_t = \{\varepsilon_{\rho+1}, \varepsilon_{\rho+2}, \cdots, \varepsilon_t\}$ by imposing an AR model on the time series TS_t data where ρ is order of AR model and residuals obey a zero-mean normal population distribution. Then question 1 is converted to the following question:

Question 2. The residuals E_t corresponding to the Sub-time series TS_t is divided into two sub-residuals E_{t1} and E_{t2} whether it is necessary.

To solve question2, we need to study whether the new nodes v_t is on the significant impact of variance of the residuals E_{t1}. The residuals E_{t1} obey a normal distribution with zero mean: $E_{t1} \sim N(0, \sigma_{t1}^2)$; The residuals E_t corresponding to new sub-time series TS_t added a new node v_t obey a zero-mean normal distribution: $E_t \sim N(0, \sigma_t^2)$. So question2 is converted to study the differences between σ_{t1}^2 and σ_t^2.

Definition 1. The minimum length of sub-time series is ML when time series TS is divided into multiple sub-time series.

Definition 2. Same variance test statistic $F_{i,j}$ is used to determine the difference of sample variance of residuals corresponding to two sub-time series. We construct the same variance test statistic $F_{i,j} = \dfrac{\frac{1}{n_i}\sum_{k=1}^{n_i} \varepsilon_{i,k}^2}{\frac{1}{n_j}\sum_{k=1}^{n_j} \varepsilon_{j,k}^2}$ with $F_{i,j} \sim F(n_i, n_j)$ using the residuals E_i and E_j corresponding to the sub-time series TS_i and TS_j. And n_i, n_j respectively is the length of residuals E_i and E_j.

For the question2, we establish two appropriate AR models for sub-time series TS_{t1} and TS_t, and calculate their residuals. After added a new node into TS_{t1} to form a new sub-time series TS_t, there is a new residual item added into the new residuals corresponding to TS_t. We study the statistic $F = \dfrac{\frac{1}{n}\sum_{k=1}^{n} \varepsilon_k{}^2}{\frac{1}{n-1}\sum_{k=1}^{n-1} \varepsilon_k^{'2}} \sim F(n, n-1)$ whether it fall into the rejection region $(F < F_{1-\alpha/2}(n_1, n_2)) \cup (F > F_{\alpha/2}(n_1, n_2))$, where n is the length of the residuals E_t corresponding to the sub-time series TS_t. If the value of F is not in the rejection region, then we add the node v_t into the sub-time series TS_{t1}. The solution of question 2 is that the residuals E_t corresponding to the sub-time series TS_t is not need to be divided into two sub-residuals. Otherwise, the residuals E_t need to be divided into two sub-residuals E_{t1} and E_{t2}.

4.3 Optimization Of Initial Division of Time Series

During the division of the time series, the first ML data of each sub-time series are added directly into it. Its purpose is for rapid division. To reduce impact on the initial division of the time series on account of abnormal value and improve the accuracy and precision of the division, we need to re-confirm the first ML data of each sub-time series to which sub-time series it belong.

Suppose that the i th sub-time series of the time series TS is TS_i, then we will re-confirm the

first ML data of TS_i to which sub-time series it belong. Firstly, we merge the first $k, 1 \le k \le ML$ data of the first ML data into the $i-1$th sub-time series of the time series TS to obtain a new sub-time series $TS_{i-1}^{(k)}$. Secondly, we establish an AR model for $TS_{i-1}^{(k)}$ and obtain its residuals. The remaining $L_i^{(k)} = L_i - k$ data of the sub-time series TS_i forms a new sub-time series $TS_i^{(k)}$ and L_i is the number of elements in the sub-time series TS_i. Thirdly, we establish an AR model for $TS_i^{(k)}$ and obtain its residuals. Then, We construct the same variance test statistic

$$F_k = \frac{\frac{1}{N_{i-1}^{(k)}} \sum_{j=1}^{N_{i-1}^{(k)}} \varepsilon_j^{\,2}}{\frac{1}{N_i^{(k)}} \sum_{j=1}^{N_i^{(k)}} \varepsilon_j^{'2}} \sim F(N_{i-1}^{(k)}, N_i^{(k)})$$ and $N_{i-1}^{(k)}, N_i^{(k)}$ respectively is the number

of elements in the sub-time series of $TS_{i-1}^{(k)}$ and $TS_i^{(k)}$. With different values of k in the range of $1 \le k \le ML$, we obtain a sequence of statistics $\{F_1, F_2, \mathrm{L}\ , F_{ML}\}$. We calculate the distribution function value of every statistics in the sequence of statistics $\{F_1, F_2, \mathrm{L}\ , F_{ML}\}$ and these values form a sequence $\{P_1, P_2, \mathrm{L}\ , P_{ML}\}$. Then, we calculate the distance between each point in the sequence $\{P_1, P_2, \mathrm{L}\ , P_{ML}\}$ and center point which value is equal to 0.5. If the first ML data of the sub-time series TS_i includes some points belonging to the sub-time series TS_{i-1}, then the variance of residuals corresponding to the new sub-time series $TS_{i-1}^{(k)}$ and $TS_i^{(k)}$ should have a smaller value.

Now, if we merge the points belonging to the sub-time series $TS_i^{(k)}$ into the sub-time series $TS_{i-1}^{(k)}$ that will form two new sub-time series, the ratio of the variance of residuals corresponding to the two new sub-time series is either large or small. If the points should belonging to the sub-time series TS_{i-1} is still in the sub-time series TS_i, the ratio of the variance of residuals is bound to much less than 1 and the corresponding distribution function value is less than 0.5. If the points should belonging to the sub-time series TS_i is still in the sub-time series TS_{i-1}, the ratio of the variance of residuals is bound to much greater than 1 and the corresponding distribution function value is greater than 0.5. We only study the boundary points that make the ratio of the variance of residuals greater than 0.5. We find the maximum deviation $AP_{\max}$ from the center position in the sequence $\{P_1, P_2, \mathrm{L}\ , P_{ML}\}$ and view the subscript of the sub-time series data corresponding to the maximum deviation $AP_{\max}$. Then, the subscript of the optimal boundary point in the sub-time series TS_i is $index$. Finally, we merge the first $index-1$ data of the sub-time series TS_i into the sub-time series TS_{i-1} and the rest of $L_i - index + 1$ data are retained in the sub-time series TS_i.

4.4 State Identification and Instability Classification of Time Series

After the end of the time series segmentation, we do linear discriminant analysis for each sub-time series to determine the stability of sub-time series. We use training sample set to establish Fisher linear discriminant function for better classification of samples data. The basic idea of Fisher criterion is to find a method of classification of the best projection line that makes the sample data of multi-dimensional space projected onto the one-dimensional space with the greatest degree distinction through projection line.

We use simulation test of BGP routing instability to obtain the training sample data and obtain the optimal projection weight vector w^* according to Fisher criterion function. We use the training sample data to determine the boundary threshold y_0. Here, we select y_0 in the Eq.1 as boundary point of the stable and unstable condition:

$$y_0 = \frac{n_1 \tilde{f}_1 + n_2 \tilde{f}_2}{n_1 + n_2}. \qquad (1)$$

Where $\tilde{f}_1$ is the sample mean of sample data of the stable sub-time series that are projected

onto

the one-dimensional space, n_1 is the number of sample data of the stable sub-time series; $\tilde{f}_2$ is the sample mean of sample data of the unstable sub-time series that are projected onto the one-dimensional space, n_2 is the number of sample data of the unstable sub-time series.

When we determine whether sub-time series $TS_i, i = 1,2,\mathrm{L}\ ,k$ is stable with the total number of segmentation k, firstly we project the sample data of each sub-time series onto the one-dimensional space to obtain a new sub-time series through the optimal projection vector w^* with expression $y = \omega^{*T} x$, then we calculate the mean denoted as $\bar{y}$ of the new sub-time series and compare $\bar{y}$ with the boundary threshold y_0. According decision rule Eq.2:

$$\bar{y} \geq or \leq y_0 \Rightarrow TS_i \in \begin{cases} C_1 \\ C_2 \end{cases}. \tag{2}$$

we can determine whether sub-time series TS_i is stable where C_1 is the class of the stable sample and C_1 is the class of the unstable sample.

When we determine the unstable type to which the unstable sub-time series belong, we use the training sample data to obtain the boundary threshold y_0 with the same expression Eq.1.

Where $\tilde{f}_1$ is defined as the sample mean of sample data of the BGP router failures and recovery that are projected onto the one-dimensional space, n_1 is the number of sample data of the BGP

router failures and recovery; $\tilde{f}_2$ is defined as the sample mean of sample data of the BGP routing policy conflict that are projected onto the one-dimensional space, n_1 is the number of sample data of the BGP routing policy conflict.

When we determine the unstable type to which sub-time series $TS_i, i = 1,2,\mathrm{L}\ ,k$ belong with the total number of segmentation k, firstly we project the sample data of each sub-time series onto the one-dimensional space to obtain a new sub-time series through the optimal projection vector w^* with expression $y = \omega^{*T} x$, then we calculate the mean denoted as $\bar{y}$ of the new sub-time series and compare $\bar{y}$ with the boundary threshold y_0. According decision rule Eq.3:

$$\bar{y} \geq or \leq y_0 \Rightarrow TS_i \in \begin{cases} \tilde{C}_1 \\ \tilde{C}_2 \end{cases}. \tag{3}$$

We can determine the unstable type to which sub-time series TS_i belong where C_1 is the class of the unstable sample caused by BGP router failures and recovery and C_1 is the class of the unstable sample caused by BGP routing policy conflict.

5 Conclusion

This paper presents an algorithm of BGP routing instability detection and analysis that is based on mathematical statistics and pattern recognition. The algorithm is used to section time series into many sub-time series and judge the states and unstable type of sub-time series through router update messages collected for statistical analysis and Fisher linear discriminant method. Finally, we verified correctness and efficiency of the algorithm through a simulation.

As the algorithm refers to time series modeling, so model type and model order selection have an important relationship with accuracy of BGP routing instability detection and analysis. In addition, the design quality of SSFNet simulation may affect the accuracy of analytical results. How to design effective simulation will be a part of future research.

References

[1]. Craig Labovitz, G. Robert Malan, Farnam Jahanian. Origins of Internet Routing Instability[C]. In Proceedings of IEEE INFOCOM. 218 - 226 (1999).

[2]. J.L. Sobrinho. On the convergence of path vector routing protocols[C]. IEEE Workshop on High Performance Switching and Routing. 292 - 296 (2001).

[3]. Randy Zhang, Micah Bartell. BGP Design and Implementation[M]. Cisco Press. (2004).

[4]. C. Labovitz, A. Ahuja, A. Bose, and F. Jahanian. Delayed Internet routing convergence[J]. IEEE/ACM Transaction on Networking. 293 - 306 (2001).

[5]. T.G. Griffin, B.J. Premore. An Experimental Analysis of BGP Convergence time[C]. In Proceedings of the Ninth International Conference on Network Protocols. 53 - 61 (2001).

[6]. K. Varadhan, R. Govindan, and D. Estrin. Persistent route oscillations in inter-domain routing[C]. In Proceedings of Computer Networks. 1 - 16 (2000).

[7]. C. Labovitz, G. R. Malan, and F. Jahanian. Internet routing instability[J]. IEEE/ACM Transaction on Networking. 515 - 528 (1998).

[8]. Zhaoqi Bian, Xuegong Zhang. Pattern Recognition [M]. Beijing: Tsinghua University Press. (2004).

[9]. Yanke Bao, na Li. Mathematical Statistics and Data Processing with MATLAB [M]. Shenyang: Northeastern University Press. (2008).