# Apply Anomaly Grey Forecasting Algorithm to Cyberspace Situation Prediction

Weisong He,Guangmin Hu
School of Communication and Information Engineering
University of Electronic Science and Technology of China,
Chengdu, P.R.China
{weisonghe,hgm}@uestc.edu.cn

Hongmei Xiang
Chongqing College of Electronic Engineering
Chongqing, P.R.China

***Abstract*—In recent years, much research has been devoted to the cyberspace situation awareness; nevertheless, few have investigated the case that the network traffic data collected may include missing values and sufficient network traffic data may not be acquired for privacy protection or the limitation of network storage equipment capacity. Our focus in this position paper is on introducing an anomaly grey forecasting (AGF) method for cyberspace situation prediction under less data little sample, and the experiment with Abilene network Netflow data verify this method.**

***Index Terms*—Cyberspace Situation Prediction, Anomaly Grey Forecasting, Time Distribution Series, Time Distribution Mapping.**

## I. Introduction

Endsley has formally defined Situation Awareness (SA) as the "perception of the elements in the environment within a volume of time and space, the comprehension of their meaning and the projection of their status in the near future". This definition breaks the concept of Cyberspace Situation Awareness(CSA) into three distinct levels: perception of the cyberspace environment, comprehension of the meaning of cyberspace situation information, and projection or prediction of events or actions in the future based on this perception and comprehension [2][3].

Large-scale network traffic modeling and prediction are important for cyberspace situation awareness. But, network traffic is highly dynamic, which exhibits multi-timescale properties (temporal domain). As anomalies (failures, attacks) interfere with cyberspace situation analysis, anomalies are needed to be isolated from normal traffic variation to achieve better modeling and prediction. In this paper, network traffic data is decomposed into two parts by subspace method using Principal Components Analysis (PCA): normal subspace and abnormal subspace. In normal subspace, traffic variations follow certain law, which are predictable and can be modeled, while anomalies in abnormal subspace consist of abrupt changes and are not predictable. Furthermore, it is the abrupt change points in abnormal subspace that is focused in this paper. The abrupt change points are acquired by specifying the threshold. In order to predict cyberspace situation efficiently, the massive network traffic data is needed to be compressed. Shannon entropy is good approach for the information compression of network traffic data. In the following cases: the network traffic data collected may include missing values and sufficient network traffic data can not be accessed for privacy protection or limitation of network storage equipment capacity, how can we predict the cyberspace situation and which time range the abnormal value (i.e. caused by worms)will appear? Anomaly Grey Forecasting (AGF) is proposed for solving these problems.

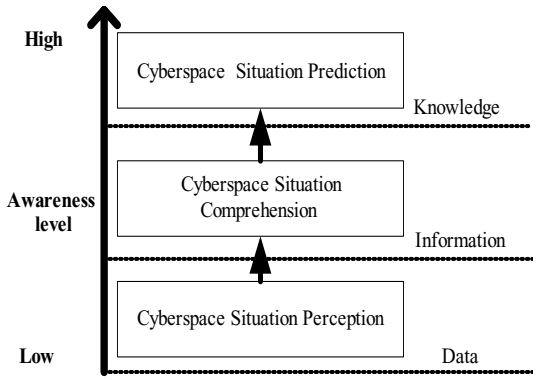


Fig. 1. The Conceptual Model of Cyberspace Situation Awareness.

The contributions of this work are as follows. To begin with, not packet level traffic information but flow level traffic information such as byte counts are considered in our experiment. The abnormal entropy value of byte counts may reflect abnormal behavior. Secondly, the next abnormal time point (i.e. the time point when worms will burst) can be predicted with less network traffic data by using anomaly grey forecasting.

The rest of the paper is organized as follows. In Section II, Shannon entropy and its application are introduced. In Section III, the proposed anomaly grey forecasting method is discussed in detail. In Section IV, the experiment with real network traffic are given and a conclusion in Section V.

## II. Shannom Entropy

The data packets within each bin of five minutes are summarized by a set of aggregate features. A flow is formally defined as a 5-tuple: source address, destination address, source port, destination port, and type of protocol. We focus on seven fields: source address (sometimes called source IP and denoted srcIP), destination address (or destination IP, denoted dstIP), source port (srcPort), destination port (dstPort), counts

of octets, counts of packets, counts of flows, type of protocol. Entropy is a metric that captures the degree of dispersal or concentration of a distribution. Entropy has kept the geometry topology of data distribution so it can have the homomorphism transformation. A wide variety of anomalies will impact the distribution of one of the discussed IP features [4] see as in

$$\begin{aligned} H(X) &= -\sum_{i=1}^{n} P[X = x_i] log P[X = x_i] \\ &= -\sum_{i=1}^{n} E(log P[X = x_i]) \end{aligned} \quad (1)$$

$P[X = x_i]$is the probability of event $x_i \in X$ occurring. For example, the probability of seeing IP 129.173.192.0 is defined to be number of packets using IP 129.173.192.0 divided by the total number of packets in the given time interval.

## III. Anomaly Grey Forecasting Algorithm

Grey theory [5] was pioneered by Deng Julong in 1982. Grey theory deal with uncertainty in less data little sample, incomplete information and devoid of experience.

### A. Basic Concept

*1) Anomaly Value and Series:* Assume series
$x = (x(1), x(2), ..., x(n))$,let $\xi$ as specified value,$c_x$ is coverage of data of $x$.
General: $x(i) \in x \Rightarrow x(i) \in c_x, i \in K = \{1, 2, ..., n\}$;
Special: $x(k_\xi) \in x \Rightarrow x(k_\xi) \notin c_x$.
We refer to $x(k_\xi)$ as anomaly value if and only if
$x(k_\xi) \in c_x, x(k_\xi) > \xi \ or \ x(k_\xi) < \xi$.
$\xi$ is threshold value. $x$ is series which include anomaly value. Series with blanks is that the series eliminate anomaly value. Assume $x$ is series which include anomaly value:
$x = (x(1), x(2), ..., x(n))$,
General: $x(i) \in x \Rightarrow x(i) \in c_x, i \in K = \{1, 2, ..., n\}$;
Special: $x(k_\xi) \in x \Rightarrow x(k_\xi) \notin c_x$,if $x(k_\xi) > \xi$, then $x(k_\xi)$ is upper anomaly value. if $x(k_\xi) < \xi$, then $x(k_\xi)$ is lower anomaly value.

*2) Anomaly Series:* Assume $x$ is series which include anomaly value:
$x = (x(1), x(2), ..., x(n))$,
General: $x(i) \in x \Rightarrow x(i) \in c_x, i \in K = \{1, 2, ..., n\}$;
Special:
$x(t_1), x(t_2), ..., x(t_m) \in x \Rightarrow x(t_1), x(t_2), ..., x(t_m) \notin c_x$,
$x(t_1), x(t_2), ..., x(t_m) > \xi, or \ x(t_1), x(t_2), ..., x(t_m) < \xi$,then $x_\xi$ is anomaly value series.
$x_\xi = (x(t_1), x(t_2), ..., x(t_m)), x_\xi \subset x$.
If $\xi$ is upper threshold value and $x_\xi$ is upper anomaly value series, then $x(t_1), x(t_2), ..., x(t_m) > \xi$;
If $\xi$ is lower threshold value and $x_\xi$ is lower anomaly value series, then $x(t_1), x(t_2), ..., x(t_m) < \xi$.

*3) Time Distribution Series:* Assume $x_\xi$ as anomaly value series,
$x_\xi = (x(t_1), x(t_2), ..., x(t_m))$,then refer to $\tau = (t_1, t_2, ..., t_m)$ as the time distribution series of $x_\xi$.

*4) Time Distribution Mapping:* Assume $M_\tau$ as mapping, if satisfy
$M_\tau : x_\xi \to \tau, M_\tau(x(t_k)) \to t_k$,
we refer to $M_\tau$ as time distribution mapping, refer to $\tau$ as image of $x_\xi$ based on $M_\tau$ ,refer to $x_\xi$ as origin image of $\tau$.

### B. AGF algorithm

The basic form of the AGF algorithm is as follows:
*Step1*. According to origin time series $x$, specify the threshold value $\xi$.
*Step 2*. Construct anomaly value series with $x(t_k)$:
$x_\xi = (x(t_1), x(t_2), ..., x(t_m))$.
*Step 3*. Through time distribution mapping $M_\tau$,we obtain time distribution series $\tau$.
$M_\tau : x_\xi \to \tau, M_\tau(x(t_k)) = t_k, \ \tau = (t_1, t_2, ..., t_m)$.
*Step 4*. Construct *GM(1,1)* of time distribution series $\tau$.
$GM_p \circ AGO : \tau \to (a, b)$,
$GM_{def} \circ AGO : \tau \to \tau^{(0)}(k) + az_\tau^{(1)}(k) = b$,
$IAGO \circ \widehat{GM}_\xi \circ AGO : \tau \to \widehat{t}_{m+\xi}$.
*Step 5*. Predict the anomaly value.

## IV. Experiment

### A. Data Sets

The sampled flow data collected from the backbone networks: Abilene [1]. Abilene is the Internet2 backbone network, connecting over 200 US universities and peering with research networks in Europe and Asia. It consists of 11 Points of Presence (PoPs), spanning the continental US. Sampling is periodic, at a rate of 1 out of 100 packets.

### B. Simulation Test

As analyzing Netflow data is a good compromise for cyberspace situation prediction at network-wide level and packet level in its performance and accuracy, many researcheres from industrial and scientific field attach great importance to this topic. We collect Netflow data during 08:05 a.m. December 18, 2006 and 08:00 a.m. on December 25, 2006 from IPLSng router. The flow level traffic information, byte counts, is collected over each time bin of 5 minutes. After computing the entropy of byte counts for each time bin, all of the entropy make up a time series which is denoted as H(octets). In order to achieve better modeling and prediction, anomaly entropy time series are needed to be isolated from normal entropy time series using PCA. Having finished steps mentioned above, abnormal traffic trends prediction can be done as follows. Firstly, we specify the threshold entropy value (=0.32) to entropy anomaly time series to construct anomaly time distribution series.
$\tau' = (151, 152, 292, 987, 2110)$,
$\tau' \subset \tau = (151, 152, 292, 987, 988, 2110)$.

Secondly, we construct *GM(1,1)* of time distribution series $\tau'$. The result is shown in Fig.2.

The white response of *GM(1,1)* is:

$$\widehat{x}^{(1)}(i) = 130.9814 * e^{0.81941*(i-1)} + 20.0186 \quad (2)$$

TABLE I
CHECKING RESIDUAL ERROR

| *Number* | $x^{(0)}(i)$ | $\hat{x}^{(1)}(i)$ | $\varepsilon^{(0)}(i) = x^{(0)}(i) - \hat{x}^{(1)}(i)$ | *RelativeError(%)* |
|---|---|---|---|---|
| $i = 1$ | 151 | 151 | 0 | 0 |
| $i = 2$ | 152 | 166.2 | -14.2 | -9.4 |
| $i = 3$ | 292 | 377.2 | -85.2 | -29.2 |
| $i = 4$ | 987 | 856 | 131 | 13.3 |
| $i = 5$ | 2110 | 1942.3 | 167.7 | 7.9 |

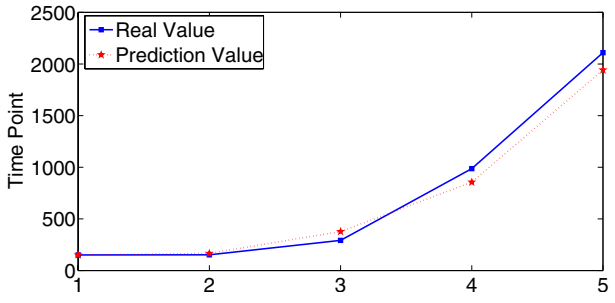


Fig. 2. Prediction Value and Real Value.

Average of positive residual error : $\varepsilon_+(average) = 149.35$, Average of negative residual error : $\varepsilon_-(average) = -49.7$. Variance ratio:

$$c = \sqrt{\frac{S_1^2}{n-1}} / \sqrt{\frac{S_0^2}{n-1}} = 0.1258 \tag{3}$$

The small error probability:

$$\varepsilon^{(0)}(i) = x^{(0)}(i) - \hat{x}^{(0)}(i), (i = 1, 2, ..., n), \tag{4}$$

$$\bar{\varepsilon}^{(0)} = \frac{1}{n}\sum_{i=1}^{n} \varepsilon^{(0)}(i) \tag{5}$$

$$p = \{|\varepsilon^{(0)} - \bar{\varepsilon}^{(0)}| < 0.6745 \bullet S_0\} = 1 \tag{6}$$

So the prediction value should be modified as follows: $t_{prediction} = [\hat{t} + \varepsilon_-(average), \hat{t} + \varepsilon_+(average)]$.

For $c < 0.35, p > 0.95$, the *GM(1,1)* model is good for predicting the time when the next anomaly will appear. According to the forecasting, the next anomaly time point of feature H(octets) is lied in the interval:$[4357, 4556]$.

## V. CONCLUSION

In this paper we apply anomaly grey forecasting method to predict cyberspace situation. In order to process data more efficiently, the immense Netflow data information is compressed by using Shannon entropy. The abnormal entropy time series is separated from normal entropy time series by PCA method. The anomaly grey forecasting method construct anomaly value series which contains the significant traffic spikes that exceeds specific threshold. Through time distribution mapping, anomaly value series is transformed into time distribution series. With *GM(1,1)* of time distribution series, the next time point of anomaly value can be predicted.

We evaluate the method on byte counts anomalies, which are specific instance of flow level traffic anomalies resulting from unusual changes in the flow traffic. We showed how to use anomaly grey forecasting method to predict abnormal byte counts from simple and readily available flow measurement. We quantified the efficacy our method on Netflow data collected from backbone networks, and showed that the anomaly grey forecasting method can successfully predict byte counts anomalies with small error probability and small variance ratio.

Although the evaluation in this paper is in terms of byte counts anomalies, the anomaly grey forecasting method is not specific to them. The strength of the method lies in that it can be applied to less data little sample, incomplete information. This allows it to treat prediction problem under such condition that the network traffic data include missing values (i.e. physical fault, storage limitation, privacy protection) and that insufficient traffic data may contains noisy information. Such an approach can therefore be extended to other types of elements of flow data (i.e. protocol type, port number), and thus other types of flow anomalies. Our going work is centered on the improvement of the prediction precision of anomaly grey forecasting method.

## ACKNOWLEDGMENT

The authors would like to thank Abilene for the providing Netflow data and thank reviewers for their helpful comments. This research supported by Chinese National Science Foundation under grant No.60572092, Program for New Century Excellent Talents in University.

## REFERENCES

[1] http://abilene.internet2.edu.
[2] M.R. Endsley. "Theoretical underpinnings of situation awareness: A critical review". In *Situation Awareness Analysis and Measurement*. Lawrence Erlbaum Associates, Mahawah, New Jersey,USA,2000.
[3] Weisong He, Guangmin Hu, Guangyuan Kan, "Comprehend Network Situation Using Time Series Data Mining and Time Frequency Analysis".In *the International Conference 2007 on Information Computing and Automation*, Chengdu, China.
[4] A.Lakhina, M.Crovella and C.Diot, "Mining anomalies using traffic feature distributions". In *ACM SIGCOMM*(Philadelphia, Pennsylvania, USA, 2005), pp.217-228.
[5] DENG Julong, *Grey Forecast and Grey Decision*, Huazhong University of Science and Technology Press, 2002.