Nine Years of Observing Traffic Anomalies: Trending Analysis in Backbone Networks

Youngjoon Won
Hanyang Univ.
Seoul, Korea
youngjoon@hanyang.ac.kr

Romain Fontugne
Univ. of Tokyo/JFLI
Tokyo, Japan
romain@hongo.wide.ad.jp

Kenjiro Cho
IIJ
Tokyo, Japan
kjc@iijlab.net

Hiroshi Esaki
Univ. of Tokyo
Tokyo, Japan
hiroshi@wide.ad.jp

Kensuke Fukuda
NII/Presto, JST
Tokyo, Japan
kensuke@nii.ac.jp

***Abstract*—We present the longitudinal trending analysis of traffic anomalies on a trans-Pacific backbone network over nine years. Throughout our analysis, we try to answer several questions: how frequent do such anomalies appear and how long do they last? Does a set of anomalous hosts occur correspondingly? We answer these by applying the state-of-the-art anomaly detectors to (un)anonymized packet traces and look into interesting insights from the long-term analysis. The key observations are as follow. The sources of anomalies are decreasing over the recent years, but take a significant portion of traffic volume during the measurement period (i.e., 0.03% of all IP addresses take upto 30% of traffic volume). The frequency analysis reveals that there is a clear periodicity of anomalies and anomalous host occurrences in various durations. Finally, we find the influences of anomaly detectors to the overall trending and how they differ from each other.**

I. Introduction

Anomalies in Internet traffic are often highlighted when there is a severe damage to network operations. Otherwise, ignorance is a bless since anomalous is not necessarily malicious. For decades, detection/counteraction of such anomalies has been a primary focus in the research community. While we are overwhelmed with numerous short-term traffic snapshot analysis, we question ourselves if there is any pattern for anomaly occurrences and life-time of traffic anomalies in a long term. By definition, a longitudinal study refers to a correlation research that involved repeated observations over long periods, e.g., decades. In this paper, we analyze the traffic from the MAWI repository [1] and provide a longitudinal study of anomalous Internet traffic from 2002 to 2010 over a trans-Pacific backbone link.

We investigate how each detector reveals a temporal trending of anomalies and how they differ, through running both original and anonymized (IP address only) traces against state-of-the-art detectors. MAWILab [2] provides the methodology to automatically label the traffic with anomalies by combining multiple detectors with respect to statistical similarity. It updates daily on top of the MAWI repository. The traces and anomaly reports referring in this paper are available at http://www.fukuda-lab.org/mawilab/.

We answer our questions by analyzing long-term evolution of per-host (IP) statistics and distributions of anomalous hosts. However, this is impossible with the anonymized datasets since there is no guarantee for IP integrity in the daily traces. We use the original datasets for this purpose. The anonymized datasets are sufficient for the alarm reports because we consider that each trace is an independent event. Finally, to investigate any periodicity, we rely on two frequency analyses; the power-spectrum and wavelet analyses.

There are a few head-ups before we proceed. First, the performance of each detector is validated by MAWILab; the detectors were fine-tunned based on the previous results. Our interest lies in identifying any concrete trend from the multiple detector results that may be helpful for understanding their singularities and improving their efficiency. Secondly, there is a weak correlation mapping between host and IP in the large time granularity, e.g., years. A single IP may represent multiple hosts over a long time. Thus, we consider an IP that is closely intacted within short intervals as a single host. The term 'host' and 'IP' are used interchangeably throughout the text.

Long-term observations reveal interesting insights. We make several key observations from our study. Over the 9 years of measurement periods,

- Less than 0.03% of all IP spaces are responsible for repetitive anomalies and upto 30% of traffic volume.
- There is no typical size of anomalies in duration.
- Periodicity (oscillation) of anomalous hosts and alarms exists in both higher and lower frequencies.

To best of our knowledge, this work is a first of its kind to reveal long-term trending analysis of anomalies. The rest of this paper is organized as follows. Section II describes our datasets and anomaly detectors in choice. Long-term observations are explained in Section III. We present related work in Section IV and conclude in Section V.

II. Methodology and Datasets

The MAWI datasets used in our analysis are daily 15-minutes packet traces (Jan. 2002–Dec. 2010, 3285 days) of a link between Japan and US. The traffic is measured everyday from 14:00 to 14:15 JST. We analyze 3074 trace files which contain usually 100–700K unique IP addresses. Note that, 211 days of traces are missing due to the scheduled network maintenance. Except for the three month gap (link update) in June–Aug. 2006, the concern for data continuity is minimal. Each trace is tagged with a summary of anomalies from MAWILab. In a nutshell, MAWILab identifies anomalies using

978-3-901882-50-0

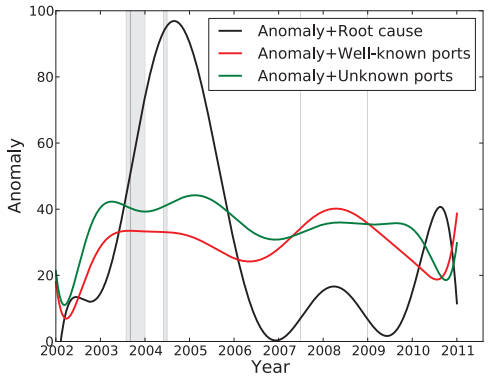


Fig. 1. Anomalies in 2002–2010: The grey regions indicate known anomaly incidents.

a combination of four anomaly detectors and takes advantage of a community mining algorithm to aggregate the detectors results. Thereby, it provides brief anomaly descriptions that contain anomalous IPs representing either the major sources or destinations of the anomaly. Each anomaly is also annotated with one of the three following labels: (1) known root causes, (2) well-known ports, and (3) unknown ports. (1) identifies the exact cause of problem, such as *Sasser* worm. (2) appears to have some involvement with well-known ports traffic; however, it is not clear what it is due to possible port masquerading. Finally (3) contains a large volume of dynamic ports traffic.

At 2003–6, case (1) anomaly rises while the others occur consistently over the last 9 years (Fig. 1). We find that the increase of labeled anomalies follows after the events in early 2000s. Unlike the early years, the 2007 and 2009 events show less of incremental. Each grey region in Fig. 1 represents the following events in chronological order.

- From Aug. 2003 the outbreak of the *Blaster* worm is observed in the MAWI traffic. This worm was spreading through a Windows security hole and has been observed all over the world.
- From Sept. to Dec. 2003, we observed a substantial number of ICMP flows constituting a long-lasting ping flood. The root cause of this anomalous event is undetermined, however, it has seriously impacted the network resources as it represents 34.5% of packets transmitted during this period.
- From June 2004 to June 2005, another worm called *Sasser* is observed in the form of three peaks representing three outbreaks of different variants of this worm.
- After the update of the MAWI link in July 2006, an important traffic against DNS servers is observed. This traffic is particularly intense in the middle of Nov. 2006, for example, the DNS traffic measured on the 2006/11/11 stands for 83% of all packets recorded this day.
- From 2009 the traffic classified as unknown shows a dramatic increase. This traffic is difficult to investigate as it is observed on high/unassigned port numbers (Table I); consequently, the root cause of this event is unclear. However, we notice that a virulent worm called *Conficker* was emerged at the similar period of time. Fig. 2 illustrates more detailed application and anomaly breakdown ratios of the traffic.

TABLE I
HEURISTICS DEDUCED FROM MAIN ANOMALIES [8] AND MANUAL INSPECTION OF THE MAWI ARCHIVE.

Category	Label	Details
Root cause	Sasser	Traffic on ports 1023/tcp, 5554/tcp or 9898/tcp
Root cause	RPC	Traffic on port 135/tcp
Root cause	Ping	High ICMP traffic
Root cause	Other attacks (Flood)	Traffic with more than 50% of SYN, RST or FIN flag. And http, ftp, ssh, or dns traffic with more than 30% of flag SYN
Root cause	Blaster	Traffic on ports 137/udp or 139/tcp
Well-known ports	Http	Traffic on ports 80/tcp and 8080/tcp with less than 30% of SYN flag
Well-known ports	dns, ftp, ssh	Traffic on ports 20/tcp, 21/tcp, 22/tcp or 53/tcp&udp with less than 30% of SYN flag
Unknown ports	Unknown	Traffic that does not match other heuristics

We identify anomalies by using similarity estimators and multivariate combination strategy (SCANN) [2] which aggregates the alarms from the following four detectors. (1) The traffic is split into sketches and modeled using Gamma Modeling (Gamma) technique [3]. The traffic being apart from the computed range is referred to as anomaly. (2) The Hough transform (Hough) is a pattern recognition technique that allows for the identification of a specific shape in a picture. It has been applied to several domains, including backbone traffic [4] and identified a unusual shape within the 2-D scatter plot of traffic. (3) Principal Components Analysis (PCA) highlights the main features of the data. When applying to traffic [5], the PCA determines a normal traffic behavior based on the traffic features. (4) The Kullback-Leibler divergence (KL) technique [6] detects the prominent changes in traffic. Its association rule mining allows for the extraction of set of traffic features by histogram analysis. Our analysis relies on the three unsupervised anomaly detectors (1)–(3), based on distinct statistical techniques. However, the KL is omitted from this paper since its reported alarms do not necessarily contain IPs at all times.

III. LONG-TERM OBSERVATIONS

We mainly focus on the two metrics: *the number of anomalies* and *the number of anomalous IPs*. The number of anomalies is counted by the SCANN-aggregated alarms that may be related to multiple IPs (i.e., DDoS). Also, the number of anomalous IPs is counted by the unique IPs in the alarms from each detector or SCANN. We cross-analyze the two metrics to describe the behavior of anomalies and their contributing hosts.

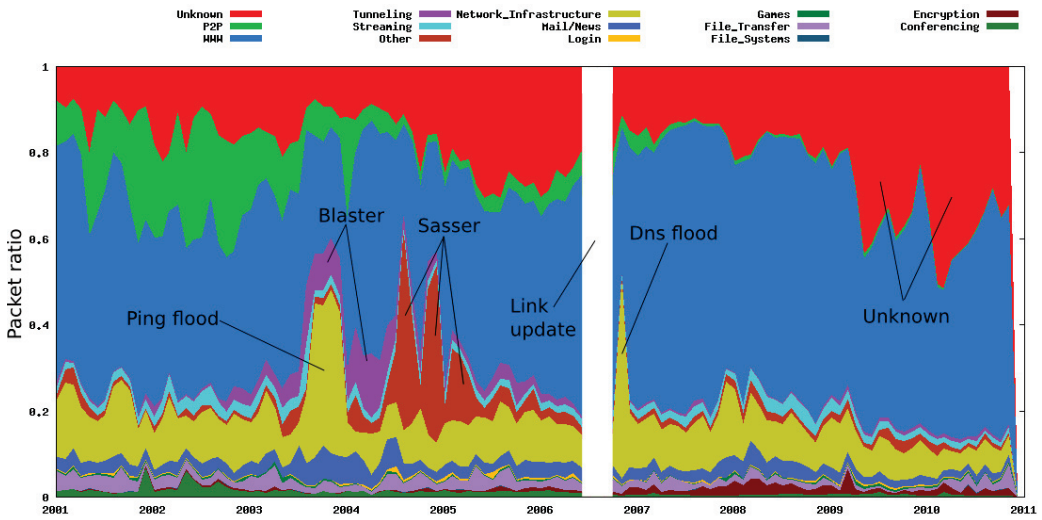


Fig. 2. Application breakdown in 2001-2010: Overview of traffic composition in packet ratio

A. Trending of anomalous IPs

Fig. 3 exhibits the number of anomalous IPs for each detector and how it responds to the overall fluctuation of all IPs. The number of unique IPs reaches upto 700K per day. The troublesome IPs vary depending on the detectors; the Hough and PCA stay within the range of 0–250 and contribute to the earlier peak of 2004–5 in the all IPs'. Meanwhile, the Gamma shows its corresponding peak in 2010 to the all IPs'. The count is higher (i.e. 300) than the other detectors which are less influenced by the increase in the number of all IPs.

The SCANN aggregates the anomalous IPs, where its IP count is significantly less than the number of IPs reported by each detector (i.e. maximum of 250 IPs). It illustrates two interesting and exact opposite phenomena. In 2004, the number of SCANN-aggregated IPs grows as the number of all IPs increases (Fig. 3(c)). The opposite happens in 2010 in which the SCANN shows a decrease and rather steady numbers. Such phenomena can have two possible explanations. First, the anomalies have been evolved to include less number of anomaly-triggering hosts over time. Second, the current popular detectors may not be able to cope with the traffic nature of more recent and smart anomalies. In other words, there is a possibility that we are simply incapable of acknowledging its existence.

B. Appearance of anomalous IPs

Fig. 4 illustrates how anomalous IP occurrences are distributed over our measurement period. The y-axis stands for the top 100 most frequent IPs and the x-axis is over time-series. Each point indicates the corresponding IP appeared at least once on a given day (x). We applied a hierarchical clustering technique to the host data in order to highlight the temporal similarity (i.e., overlap) between hosts. We see that there is a thicker region (more densely concentrated) shared by all three detectors: the upper middle region of 2004–6. Meanwhile, the lower right corner of 2008–10 are shared by the Gamma and PCA only. As we move up on the y-axis, we do not observe significant horizontal-widespread except for the Gamma figure. The first and last occurrences of 70% of each Gamma IP are spaced over a 1500+ days interval which is higher than the other detectors. It is temporally more widespread. Thus, the behavior of reported hosts varies depending on the nature of detectors.

Vertical bands indicate anomalies that were spatially widespread. Likewise, the thicker vertical region indicates higher correlation between the anomalies. We observe very clear spatial-compact regions around 2004–5 in the PCA. They indicate the reported anomaly event in MAWI, the *Sasser* worm traffic from June 2004. Moreover, the vertical white line in the middle indicated the missing dataset of 3 months in 2006.

We find that the most frequent IPs in all three figures share a few common characteristics. They appear densely intacted over the entire monitoring periods which implies their repetitive and aggressive contribution to the anomalies in the short intervals. Fig. 5 illustrates CCDFs of a collection of horizontal disparities, temporal distributions, in consecutive days.

We confirm clear long-tailed distributions close to power-laws, suggesting that there is no typical size of anomalies in consecutive days. Less than 1% of the Gamma-identified lengths terminate within 5-10 consecutive days. 99% of them are just one day long. Similarly, the PCA shows 0.1% within 25 days. The Hough has a long tail distribution of 10-90 consecutive days within its 1%.

We have seen a small set of anomalous IPs appears continuously as a group of relatively small disparities (i.e. tens of days). We now question ourselves whether the anomalous hosts appear randomly over time. Fig. 6(a) illustrates a sample of the autocorrelation function (ACF) of sample Gamma hosts. The other detectors are omitted since they show a similar pattern. As the ACF converges to 0 it indicates randomness,

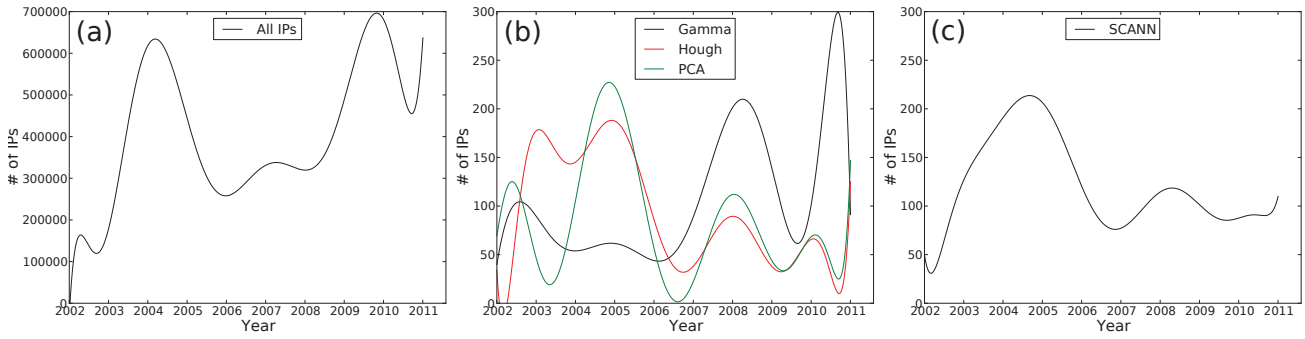


Fig. 3. All IPs (a); Anomalous IPs by Gamma, Hough, PCA (b); by SCANN (c)

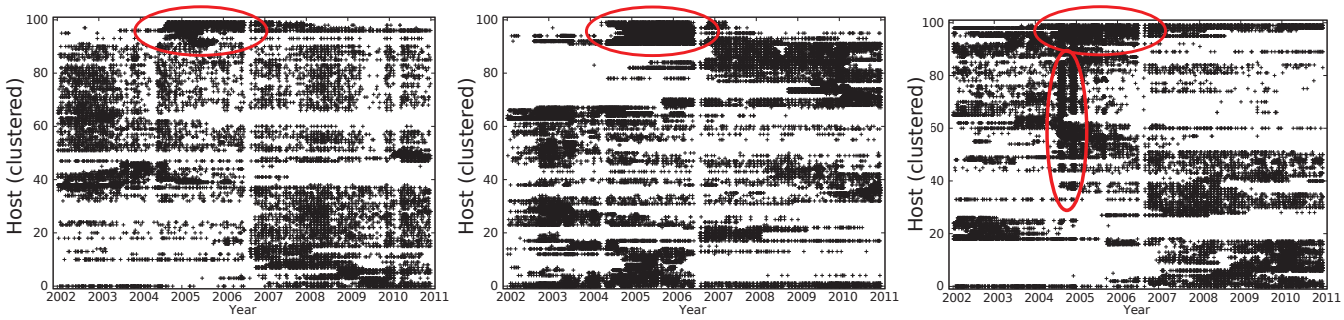


Fig. 4. Footprints of top 100 IPs; Gamma, Hough, and PCA (in order)

however, we find that there is a clear fluctuating pattern before it converges to 0. To find out the exact periodicity in days, we break down the x-axis into the bins of 500 days. The first three bins did not show any cyclic pattern; however, the last two bins reveal about 30 days cycle of anomalous host occurrences (as it corresponds the highest peak of ACFs in the figure). In other words, the anomalous hosts appear in no concrete pattern during the beginning 2/3 of the measurement periods and later they follow a monthly appearance cycle. Thus, the results from the last 1000 days (≅3 years) dominate the overall pattern of anomalous hosts and confirmed the monthly cycle of anomalous hosts in our dataset.

Finally, we find a common set of anomalous IPs shared by all the detectors and analyze their impact on traffic volume. 78 IPs are identified which is 0.03% of all IPs. We filter out the traffic relating to these IPs. Fig. 6(b) illustrates the traffic volume occupied by 78 IPs over time and they are responsible for surprisingly a large portion of traffic (i.e. up to 30%). And they appear consistently almost over the decade. Note that, their traffic in the figure does not necessarily include anomalies only, but legitimate traffic as well. Indeed, the hostnames of some anomalous IPs implied planet-lab nodes, VPN servers, NAT servers and important DNS servers.

C. Frequency of anomalies

We consider the frequency analysis of anomalies and anomalous IPs. We use the the power-spectrum analysis to transform time-dependent data into the frequency domain, to focus on predominant frequencies, such as periodicity in

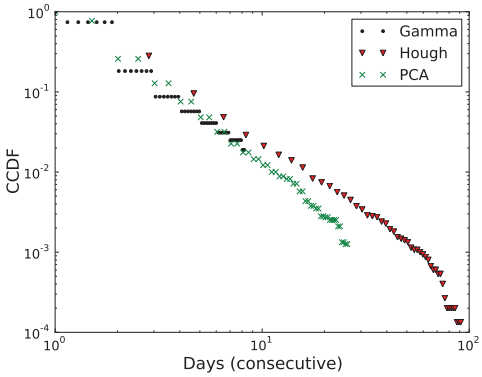


Fig. 5. CCDFs of consecutive days of anomalous IPs appeared per each detector

anomalies. The Discrete Wavelet Transform (DWT) is also a common tool for analyzing localized variation of power within a time series [7]. We use both since the DWT (morlet as basis in our study) provides a method for power-spectrum analysis and is a complementary to any obscurity.

Fig. 7 illustrates periodograms of frequency/periodicity vs. power in the numbers of anomalous IPs over time in log-log scale. The slope in the figures is a power-law with slope 1.0, corresponding to the boundary between stationary and non-stationarity of the original time series. In panel (a), the plots are close to a power-law over frequency though the plots in panel (b) and (c) are characterized by a power-law in the low-frequency part ($>$ 10days) and white-noise in the high-frequency part. The white noise behavior corresponds to non-

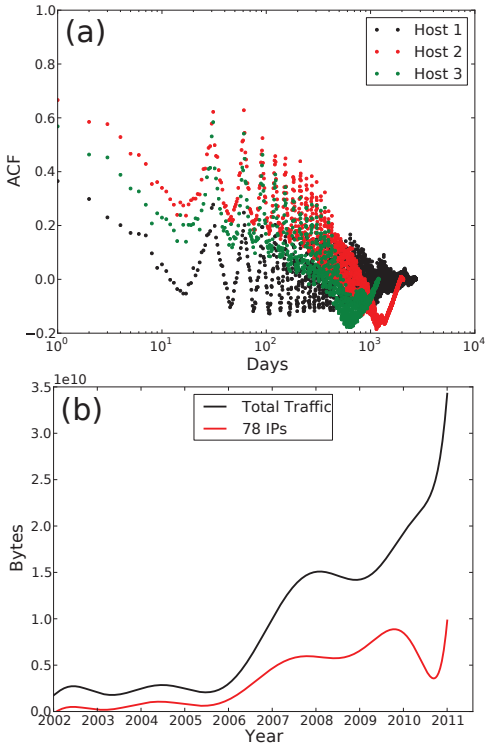


Fig. 6. Randomness of anomalous IP occurrences (a); Traffic volume by 78 IPs (b)

temporally correlated fluctuations in the original time series. Thus, this difference suggests the difference of the temporal structure of the reported anomalies by the three detectors. Furthermore, we confirmed a series of small peaks across the 20–50 days on the x-axis (labelled on top) where the arrows in the figures indicate the maximum peaks of around 50 days in panel (a) and (c). The peaks toward the right are obscure due to the white noise. We do not consider them for any cyclic behavior. Thus, this implies that the number of anomalous hosts reaches a maximum in the cycle of 20–50 days. To see any window-size impact, we try for other day bins, such as 256, 512, and 2048 days. However, in a smaller window-size, we find a clear and shorter period of 20–30 days which builds up the larger cycle we have observed in Fig. 7.

Fig. 8(a) indicates the intrinsic behavior of impulses (arrows) for the alarms, corresponding to approximately 7 days cycle of alarm occurrence. To verify its consistency, we compute ACF of the original time-series data (Fig. 8(b)). The ACF shows the reoccurring peaks in the continuous interval of 7 days over the sample period of 400 days. Furthermore, a higher (> 1.0) value of the slope in the figure suggests the non-stationary nature of the number of alarms, as shown in [8].

Similar to the power-spectrum, Fig. 9(a) illustrates the wavelet power spectrum of anomalous IPs over time. The x-axis is the wavelet location in time and the y-axis is the wavelet period in days. The difference of the colors indicates that of the activities; reddish area corresponds to a stronger density. Each column is binned by 1024 days and continuing from Jan. 2002. Within 10% significance regions, the small red areas indicate that high activities of anomalies during 2002–4 and 2008–9, while 2005–7 (middle) was relative calm. Interestingly, at the end of the year, there are high activities in either higher or lower frequencies except for the end of 2005 and 2010. It is clear that there were high activities in 4–7 days of period during the early years in 2000s. Overall, we can observe near 32 days oscillation at lower frequencies consistently throughout the measurement period (i.e. 2003–5 and 2006–9). It becomes more apparent if we break it down in to other time bins, such as 2056 days. Thus, a monthly cycle of anomalous IPs is the dominant characteristic in the wavelet analysis. This observation also coincides with the power-spectrum result of 30 days cycle. However, the DWT verifies more clear and longer continuity of lower frequency cycles (horizontally wider).

In Fig. 9(b), the alarm signal illustrates strong and frequent activities at the high frequency compared to Fig. 9(a). The strong activities in mid-2004 are respect to the *Sasser* outbreak that have seen in the MAWI. At the end of 2006, we also observe another high frequency activities which relate to the unusual DNS traffic burst. It is likewise event for 2009. Overall, the alarms are highly concentrated over a shorter period, 4–7 days, and do not have a long oscillation. The anomalous IPs and SCANN alarms are occurred in two difference frequencies, low and high frequencies, respectively.

IV. Related Work

Our interest in this topic was motivated by the need for deeper understanding of the MAWILab results. The absence of ground truth limits the scope of such anomaly detection works; thus, we wanted to find the common trends in traffic anomalies and see how numerous detection algorithms [9], [6], [4], [3], [5] coincide within themselves.

A number of papers have discussed the properties of normal traffic [8], [10]. The intend of these works are similar to our own since they also focused on quantitative and temporal characteristics. However, the traffic domain shifts from normal to abnormal in this paper. In regards to abnormal traffic, previous works have been mainly focusing on measurement at the edge of the network, for example, at a single site [11] or a range of IPs [12], [13] and their analysis is restricted to a specific type of anomaly. The goal of our study is beyond previous works as it exhibits long-term properties and frequency analysis of any kinds of anomaly occurring on backbone networks.

V. Conclusion

In this paper, we present the longitudinal trending study of traffic anomalies in a large backbone. Our analysis begins to investigate the existence of any concrete pattern in anomalies and anomalous host occurrences and other long-term properties that we were not aware of from the previous short-term snapshot analysis.

Overall, we observe that only 0.03% of all IPs take a significant traffic volume and appear repetitively for the measurement

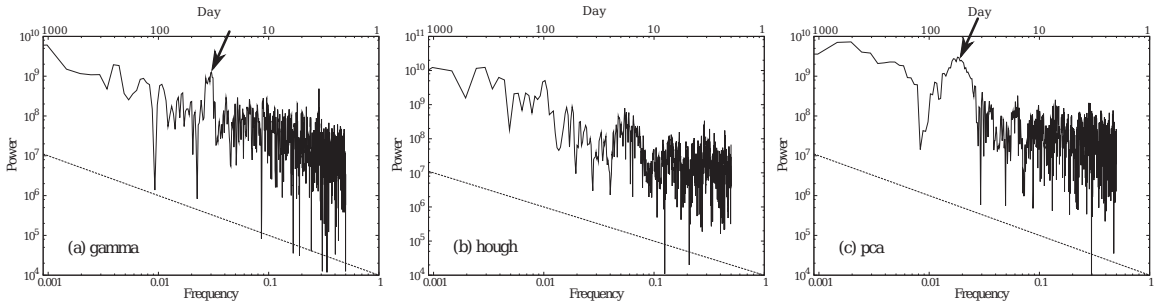


Fig. 7. Frequency and periodicity of anomalous IPs in window size of 2048 days for Gamma, Hough, and PCA (in order)

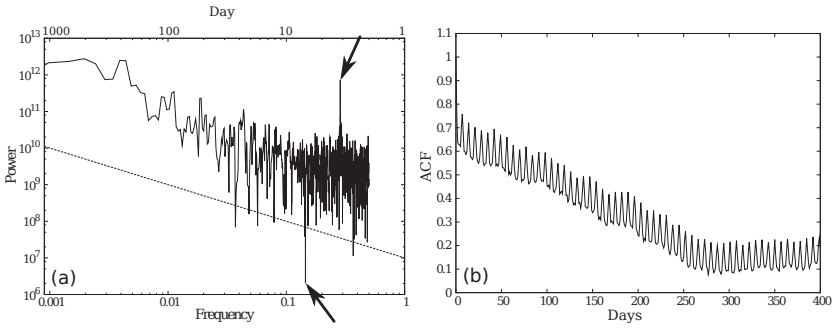
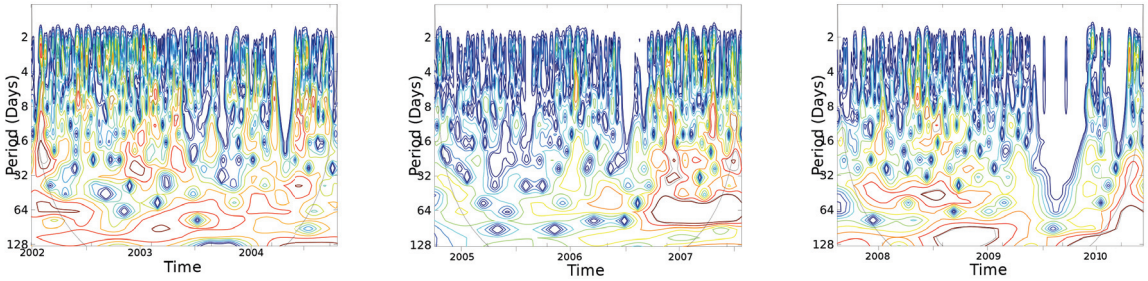
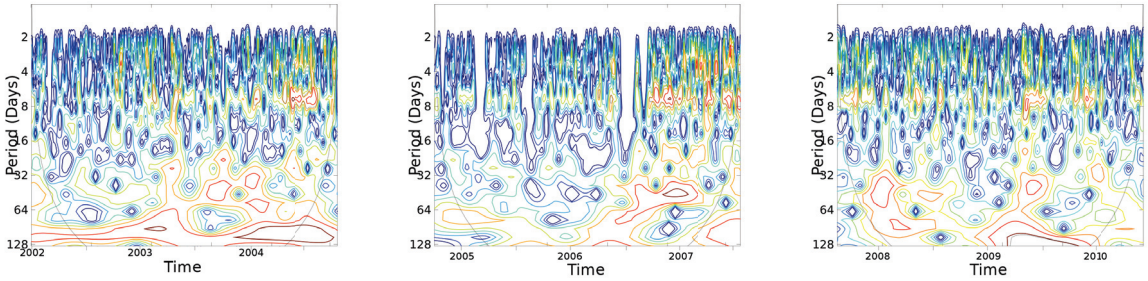


Fig. 8. Frequency and periodicity of SCANN alarms in window size of 2048 days (a); ACF (b)



(a) Anomalous IPs reported by gamma



(b) SCANN alarms

Fig. 9. Wavelet power spectrum in window size of 1024 days

period. The frequency analysis shows that there is a clear periodicity of anomalies and troubled hosts at both low and high frequencies. The anomaly alarms are occurring less than a week (7 days or less) cycle; meanwhile, the anomalous hosts has a cycle of month (30 days). This implies that not every anomalous host contributes to alarm-triggering events and there is no direct correlation between the anomalies and anomalous hosts. Finally, we understand that the impact of anomaly detectors to overall trending and how they differ. In future work, we plan to investigate self-similarity and heavy-tail distributions over a decade long of anomaly traffic only. And we further analyze its contributions to self-similar traffic characteristics in the traffic mix.

REFERENCES

[1] K. Cho, A. Mitsuya, and A. Kato, "Traffic Data Repository at the WIDE Project," in *USENIX 2000 Annual Technical Conf.*, 2000, pp. 263–270.

[2] R. Fontugne, P. Borgnat, P. Abry, and K. Fukuda, "MAWILab: Combining Diverse Anomaly Detectors for Automated Anomaly Labeling and Performance Benchmarking," in *ACM CoNEXT*, 2010.

[3] G. Dewaele, K. Fukuda, P. Borgnat, P. Abry, and K. Cho, "Extracting Hidden Anomalies Using Sketch and Non Gaussian Multiresolution Statistical Detection Procedures," in *ACM SIGCOMM LSAD*, 2007, pp. 145–152.

[4] R. Fontugne and K. Fukuda, "A hough-transform-based anomaly detector with an adaptive time interval," *ACM Applied Computing Review*, vol. 11, no. 3, pp. 41–51, 2011.

[5] Y. Kanda, K. Fukuda, and T. Sugawara, "An Evaluation of Anomaly Detection Based on Sketch and PCA," in *IEEE GLOBECOM*, 2010.

[6] D. Brauckhoff, X. Dimitropoulos, A. Wagner, and K. Salamatian, "Anomaly Extraction in Backbone Networks Using Association Rules," in *ACM IMC*, 2009, pp. 28–34.

[7] C. Torrence and G. Compo, "A practical guide to wavelet analysis," *Bulletin of the American Meteorological Society*, vol. 79, no. 1, pp. 61–78, 1998.

[8] P. Borgnat, G. Dewaele, K. Fukuda, P. Abry, and K. Cho, "Seven Years and One Day: Sketching the Evolution of Internet Traffic," in *IEEE INFOCOM*, 2009.

[9] K. Carter, R. Lippmann, and S. Boyer, "Temporally Oblivious Anomaly Detection on Large Networks Using Functional Peers," in *ACM IMC*, 2010.

[10] P. Loiseau, P. Goncalves, G. Dewaele, P. Borgnat, P. Abrby, and P. Primet, "Investigating self-similarity and heavy-tailed distributions on a large scale experimental facility," *IEEE/ACM ToN*, vol. 18, no. 4, pp. 1261–1274, 2010.

[11] M. Allman, V. Paxson, and N. Weaver, "A Brief History of Scanning," in *ACM IMC*, 2007.

[12] R. Pang, V. Yegneswaran, P. Barford, V. Paxson, and L. Peterson, "Characteristics of Internet Background Radiation," in *ACM IMC*, 2004.

[13] E. Wustrow, M. Karir, M. Bailey, F. Jahanian, and G. Huston, "Internet Background Radiation Revisited," in *ACM IMC*, 2010.