A Functional Approach to Scanner Detection

Robert McAndrew
Colorado State University
rmcand@colostate.edu

Manaf Gharaibeh
Colorado State University
manafgh@colostate.edu

Haonan Wang
Colorado State University
wanghn@stat.colostate.edu

Stephen Hayne
Colorado State University
stephen.hayne@colostate.edu

Christos Papadopoulos
Colorado State University
christos@colostate.edu

ABSTRACT

Detecting scanning in Internet traffic is a well-studied topic with no single, definitive approach. Among the proposed methods are two which are widely accepted, but with known limitations: one based on a static fanout ratio, and another on principal component analysis (PCA). We introduce a two-step procedure based on Functional PCA and k-means clustering which we argue provides significantly better robustness and data-driven applicability. We validate and compare using synthetic datasets with "ground truth" about anomalies on FTP and HTTP port traffic flows; our method identifies all scanners. We also compare approaches using NTP flow data prior to a reflective DDoS attack in 2014, providing a real-world example to illustrate the deficiencies of existing approaches and how they are addressed by our functional framework procedure. Lastly, we discuss insights into the traffic that cannot be obtained by the previous methods.

CCS CONCEPTS

• **Security and privacy → Intrusion detection systems**; • **Networks → Network security**;

KEYWORDS

Network Flow Analytics, Network Scans, DDoS, NTP

ACM Reference Format:
Robert McAndrew, Manaf Gharaibeh, Haonan Wang, Stephen Hayne, and Christos Papadopoulos. 2017. A Functional Approach to Scanner Detection. In *AINTEC '17: AINTEC '17: Asian Internet Engineering Conference , November 20–22, 2017, Bangkok, Thailand*. ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3154970.3154976

1 INTRODUCTION

Bad guys are continually scanning networks as they try to find vulnerable systems. They wish to exploit these with a variety of attacks ranging from distributed denial of service (DDoS) to botnet farming to individual system compromise. High profile networks in companies, banks, governments, and service providers are targeted. The adversaries want to not only steal data for use or sale, but also disrupt the operations of their victims and impact their reputation.

Just as there are those working on the next attacks, our community must strive to prevent them. Our goal is to be able to pre-empt an attack, yet we do not attempt to detect attacks specifically; we analyze the activity prior to them. It is possible that our technique could be used to detect attacks and support mitigation, but that is not the focus here.

Several approaches to find scanners use heuristic-based methods [5], Machine Learning [6], or Statistical-based filtering [20], with varying degrees of success. For a survey of scanning, see the works of [1] and [3].

We add to this collection by presenting a method to detect 'outliers,' or unusual behaviors in a network, in two steps: (1) functional principal component analysis (FPCA) and (2) k-means clustering. With a synthetic data set that includes ground truth regarding scanning activity, we show that our approach is well-suited for identifying these behaviors. To further discuss application of the procedure and interpretation of results, we study actual NTP traffic flows during the three months leading to a ditributed reflective DDoS attack. Throughout, we compare our proposed method against previously published approaches for scanner detection: a static fanout ratio and a technique based on the standard principal component analysis (PCA). We demonstrate improvements our *functional* framework provides over both.

2 RELATED WORK

Anomaly detection methods can be classified into (1) signature based and (2) profile-based [6]. Signature-based uses prior knowledge about characteristics of the anomaly of interest to identify suspects. These methods have several concerns, e.g., the need for labeled data, an external supervisor, and prior results from anomalies. One well-known signature-based approach was proposed by [1], which focused on detecting TCP scanners. They consider the count of a remote host's connection attempts to access local hosts' services that result in established connections (i.e., good service fanout) and those that did not result in established connections (bad service fanout). A remote host is classified as a scanner if it has a service fanout of at least four and a ratio of bad to good service fanout of at least two. A similar technique is used by [5]. We refer to this method as the "static fanout ratio" later on in the paper.

Profile-based methods create representative "normal" traffic behavior, and anomalies are detected by deviations from this profile. While there may be higher false alarm rates, profile-based methods are more promising due to their data-driven flexibility and they

AINTEC '17, November 20–22, 2017, Bangkok, Thailand

ACM ISBN 978-1-4503-5551-3/17/11…$15.00
https://doi.org/10.1145/3154970.3154976

may also detect unknown anomalies [2, 17]. PCA is a widely used profile-based method. [12] has investigated detecting traffic anomalies in DDoS data, where scanners are embedded as a subset. They use PCA to decompose network traffic into two components. The anomalous subspace, which is noisier and contains the significant traffic spikes, is separated from the normal, which is dominated by predictable traffic. An individual observation is deemed an anomaly if its projection to the anomalous subspace is large. We refer to this technique hereafter as the "subspace method". [6] proposes a two-stage approach, using (1) PCA to identify potential anomalies, and (2) a meta-heuristic to group them.

[19] criticizes the use of PCA and presents four issues pertaining to (i) false positive rates, (ii) traffic measurement aggregation, (iii) normal subspace pollution, and (iv) correct anomaly identification. We add to these concerns, by noting that the subspace approach needs to choose which principal components represent "normal" behavior, and which ones represent "abnormal" behavior. We will demonstrate later in the paper that some traffic captures do not lend themselves to this partition/selection, i.e., all principal components contain abnormal behaviors, and thus this approach is not usable.

Clustering is another example of a profile-based method. [14] clustered *all* traffic, comparing the centers of known "normal" traffic clusters to the centers of actual traffic, to try and determine if the actual traffic is not normal. Unfortunately, this approach has only been applied to Simple Network Management Protocol (SNMP) objects, not network flows, and requires known normal traffic data.

[7] applies clustering techniques to characterize DDoS attack traffic (k-means, CLARA, and Self Organizing Maps). k-means was found to be the most accurate for attack detection because attack traffic displays strong similarity as opposed to the heterogeneity of normal traffic. However, their "attack" cluster still mixed legitimate traffic in with malicious (between .4% and 2.04%). We believe this phenomenon can be eliminated by clustering only demonstrated "outliers," not all traffic.

3 DATASET DESCRIPTIONS

3.1 DARPA Synthetic Dataset

The Defense Advanced Research Projects Agency (DARPA) 2009 intrusion detection dataset, available via Impact Cyber Trust [9], is synthetically created to emulate traffic between a /16 subnet (172.28.0.0/16) and the Internet. The dataset is a full packet capture in pcap format and spans a period of 10 days between the 3rd and the 12th of November 2009, aggregated hourly. Our analysis uses the FTP port and HTTP port traffic, which include scanning activity during normal traffic and DDoS attacks, respectively. We focus on the scan security events in this paper, but believe in the method's ability to perform in both situations.

The synthetic data set of FTP packet flows between source and destination IPs over 227 hours includes a multitude of security events, all scans or failed scans, carried out by 4 individual addresses. There are 216 remaining sources which are classified as exhibiting "normal" behavior. For the HTTP port, scans as well as attacks (DDoS and failed attempts) are instigated by 112 individual addresses. The remaining 1981 IPs constitute normal behavior.

3.2 CSU NTP Dataset

The real-world dataset we use in this paper was also previously used to study and characterize the NTP amplified DDoS attack and their impact on a local network [5]. The main wave occurred in late 2013 to early 2014 and peaked at 1% of all global Internet traffic on February 11, 2014. A significant increase in scanning activities was seen from a darknet (unused IP address space) operated by Merit Network [15], preceding the DDoS attacks by a week. This suggests there are likely scanners in the traffic flows prior to the actual attack. Our data was extracted from Argus files, which contain the flow data for all traffic collected from a vantage point at the Colorado State University (CSU) border router. Raw data is the temporal count of packets sent between a source and destination IP pair, specifically for those involving the NTP port. Both inbound and outbound (with respect to the source address) flows are observed hourly for six different two-week (approximately) periods from October 2013 to January 2014, just prior to the beginning of the attack. Due to privacy concerns, the last octet of each IP address was anonymized with a different key every two weeks (period).

4 METHODOLOGY

Our approach identifies scanner behavior in two steps. First, the data is modeled using Functional Principal Component Analysis (FPCA), and a three standard deviation threshold is applied to identify outlier IPs. Second, we employ k-means clustering to group the resulting sources. With this stratification, we can not only understand the activity on a network better, but also investigate common behaviors in each cluster and identify potential scanners, as well as other possible anomalies.

Our method requires time-series of measurements capturing aspects of Internet communications. As the method works in two steps, two features of interest are necessary: one should be able to be viewed as functional data for FPCA, and the other should be scalar or multi-dimensional to allow k-means clustering. In our case, we use information from source and destination IP pairs in our search for scanners. Particularly, a time series measuring the number of destination IPs contacted by individual sources is used in the first step, and the proportion of these which are met with a response is used for clustering, defined as at least one packet returned by the destination. With this, the outliers identified in the first step will be sources that contact an 'unusual' amount of destinations with respect to the rest of the data set. Then, these are clustered by their proportions of response.

Certainly, other features or aggregations can be used, such as traffic measurements by input links or ingress routers, but this specific scheme is chosen in order to detect scanners. Just as in the static fanout ratio, we look for sources that contact a large number of destinations with a low proportion of responses. After the second step of our procedure is completed, the clusters centered around lower values can be inspected to find exactly this behavior.

4.1 FPCA & Outlier Detection

Our research methodology relies on a technique that generalizes the idea of PCA to manipulate functional data [18]. In PCA, data is transformed to a space spanned by orthogonal vectors, known as principal components, in such a way as to maximize the variance

of the linear transformation. FPCA is similar, but from a functional perspective. We observe a $n \times T$ matrix whose entry with coordinates (i, t) is the number of destinations contacted by the ith source IP during the tth time interval. The method models the functional data as a mean curve plus a linear combination of orthogonal curves. The model is written as

$$Y_i(t_{ij}) = X_i(t_{ij}) + \epsilon_{ij} = \mu(t_{ij}) + \sum_{k=1}^{r} a_{ik}\psi_k(t_{ij}) + \epsilon_{ij}, \quad (1)$$

where t_{ij} is the time of the jth observation from the ith source, $Y_i(t_{ij})$ is the observation, $X_i(t)$ is the true trajectory for each source, and ϵ_{ij} are the independent and identically distributed measurement errors with $\mathbb{E}[\epsilon_{it}] = 0$ and $Var[\epsilon_{it}] = \sigma^2$. $\mu(t)$ is the mean function, and $\psi_k(t)$ represents the kth orthogonal curve which is known as an eigenfunction. We think of these as new dimensions on which our observations live. A benefit of FPCA is that the eigenfunctions are constructed so that each explains as much of the variance in the original data set as possible. That is, eigenfunction 1 is the direction of highest variability in the data, eigenfunction 2 the second, and so on. Thus, observations that are extreme with respect to these new dimensions can be thought of as extreme in the sense of the original data.

Further, we assume the covariance matrix has eigenvalues λ_k for $k \in [r]$, and orthogonal decomposition given by:

$$Cov[Y_i(t_{ij}), Y_i(t_{ij'})] = \sum_{k=1}^{r} \lambda_k \psi_k(t_{ij})\psi_k(t_{ij'}), \quad (2)$$

The number of eigenfunctions r in the above equations is a parameter which must be determined, and we apply both the Akaike information criterion (AIC) and Bayesian information criterion (BIC) toward this end [13]. While the formulation of these are similar, their differences and trade-offs are significant [23]. Primarily, the choice of method depends on a notion of the "true model." BIC is better suited for a simpler, finite dimensional truth, while AIC is better for more complex and non-parametric models [22]. As information about this data's true model is only hypothesis, both information criterion are applied and results are compared.

To determine outliers, we calculate FPCA scores for our data. Given by

$$a_{ik} = \int (X_i(t) - \mu(t))\, \psi_k(t) dt, \quad (3)$$

these are the projections of the data onto the eigenfunctions, which represent their locations along each new dimension. Every data curve has one score for each eigenfunction. When a rule is set to define an outlier with respect to a single eigenfunction, we classify a source IP as such if it meets the criteria on at least one dimension. In application, we use the threshold-based rule of being more than three standard deviations away from the mean score. In implementing FPCA, we follow the method of [24], known as the Principal Analysis by Conditional Expectation (PACE) algorithm, gathered in the package with the same name [16].

4.2 Clustering with k-Means

k-means aims to separate the m outliers into k clusters so that the sum of squares within clusters is minimized. We use the algorithm of [8]. Initial centers are chosen at random from the data and are refined by the algorithm. These random starts are carried out multiple times in order to investigate the sensitivity of the method to choice of initial centers; our data is not sensitive to the choice of centers.

To select the number of clusters, the "elbow method" is used, in which the fraction of variance explained (FVE, y-axis) by the clusters is graphed against the number of clusters (x-axis). Inspecting the marginal gain associated with the inclusion of each additional cluster, the "elbow" is identified by the point at which this quantity decreases abruptly. This elbow may not be well-defined in some cases [11], resulting in a small set of potential cluster amounts, to be investigated further; in our application, this is well-defined.

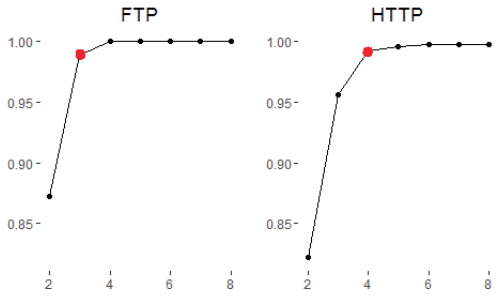

Figure 1: Elbow Plots for DARPA Clustering

5 APPLICATION & DISCUSSION

5.1 Analysis of DARPA Data

Functional Approach - For the FTP-port traffic, both AIC and BIC selection methods indicate ten eigenfunctions, which leads to nine source IPs identified as outliers (based on the common three standard deviation threshold). These are separated into three clusters of two, two, and five IPs, centered at 0, .656, and 1, respectively. Elbow plots are shown in Figure 1, with the point colored red indicating the optimal number of clusters. Among the set of outliers are all four "ground truth" scanner IPs and upon clustering, all of these addresses appear in those centered around low and mid proportions of successful contacts. While the other five IPs identified are not considered scanners in the "ground truth", we find them to be outliers all captured in the cluster centered around the highest proportion of successful contacts.

For the HTTP port, both AIC and BIC select 18 eigenfunctions for the model. 282 source IPs are identified as outliers, and this set includes all 112 addresses that correspond to the security events in the data. Results of the clustering are summarized in Table 1, with the quantity in parentheses representing the number of ground truth (GTruth) anomalies captured in that cluster. In both cases, our method finds the ground-truth anomalies, thus demonstrating correct anomaly identification, but also finds more outliers. We will discuss this finding and its implications in more detail in §5.3.

Static Fanout Ratio - A summary of the results from applying the static fanout ratio to the DARPA datasets is given in Table 2. For the FTP-port, two IPs are detected, both true scanners, and the other two went undetected. For the HTTP-port, five IPs are caught,

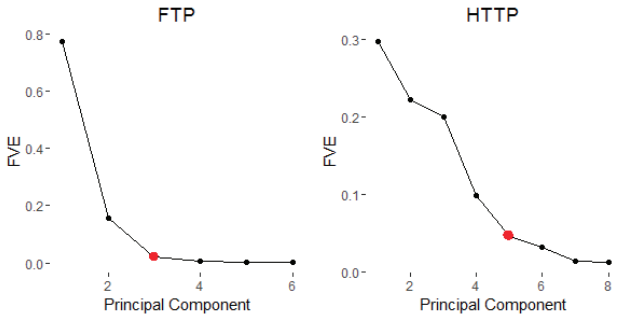

Figure 2: Scree Plots for Subspace Method

two of which are scanners. The remaining 110 true scanners and DDoS events are not identified. From comparison to the ground truth, we see that our method significantly outperforms the static fanout ratio.

Subspace Approach - In applying the subspace method of [12], we use their recommendation for determining the normal and anomalous components, which is also based on the three standard deviation cutoff. An obstacle presents itself immediately: three scores exceed the threshold on the first principal component. It follows that this, and all subsequent dimensions, comprise the anomalous subspace. Thus, the normal subspace is empty, all source addresses are fully anomalous, and further analysis becomes nonsensical.

In an attempt to avoid this issue, we try to determine the dimension of the normal subspace by investigating scree plots, shown in Figure 2. These illustrate the eigenvalues for the first principal components [10]. Just as in our selection of the number of clusters, we look for the elbow in the scree plot, and retain that number of dimensions for the normal subspace (three for the FTP data and five for the HTTP data). No changes are made to any other steps of the procedure. A summary of the results is shown in Table 3, and we can see that the method performs the same as the static fanout ratio in the case of the FTP-port, i.e., it detects two of the four ground-truth scanners. The method identifies many more IPs when applied to the HTTP data; all anomalies are detected with the exception of one scanner. Less source addresses are detected than in our method, but FPCA does catch this undetected anomaly.

The lack of applicability to this synthetic data set when using the basic and recommended standard deviation cutoff is concerning, along with other aspects previously mentioned in §2. §5.2 expands on these issues and how the functional approach improves upon them.

5.2 FPCA+Clustering vs. Subspace & PCA

We showed that the subspace method is not useable when a relatively large anomaly is found in the first principal component. This is not an isolated case - we also find this to be a problem in the real-world NTP data. It is an extreme example of what [19] refers to as 'normal subspace pollution:' a relatively large anomaly captured by one of the first principal components. Due to this and the reasons detailed in the remainder of this section, the subspace method is not discussed in our NTP analysis.

Table 1: FPCA+Clustering - DARPA Data

	FTP		HTTP	
Cluster	Center	Detected (GTruth)	Center	Detected (GTruth)
1	0	2 (2)	.005	55 (3)
2	.656	2 (2)	.348	20 (1)
3	1	5 (0)	.654	87 (0)
4			.949	120 (108)

It is important to note that the purveyors of this method report that many procedures can be used to separate the principal components into the normal and anomalous subspaces [12]. Going beyond a standard deviation cutoff, [19] investigates other statistical techniques with less than favorable results, motivated by sensitivity of the false-positive rate with respect to dimension of the normal subspace.

Both the functional and subspace methods seek to separate normal from anomalous, but do so from different perspectives. The subspace approach decomposes the space in which traffic exists into the two categories. The assumption that the dimensions capturing the largest portion of the data's variance (the first few principal components) constitute "normal" behavior is not one our model makes, and we do not assert this is separable from the anomalous. Instead, the outlier IPs, selected based on source information, are clustered using destination information. This is exactly what makes FPCA well-suited for our analysis; it has the ability to detect functional patterns even though they may not be extreme with respect to magnitude or some other specific feature.

Another benefit of FPCA is its ability to reduce the dimension of the data we are working with. Rather than deal with infinite-dimension functional space, we can investigate a finite number of projections onto components that capture meaningful features of the data, i.e., the decreasing variance in the eigenfunctions. Viewing our observations in this way simplifies how to define an outlier in the functional sense. In the subspace method, *all* principal components are retained, carrying much more information for computations.

Further, the functional framework has some advantages over traditional PCA in that it is more flexible. FPCA does not depend on the structure or size of the input data; that is, it can be collected from regular or irregular longitudinal data. Consider each flow measured at different time points as a vector. PCA would require all vectors to have the same length and each corresponding value to be measured at the same time point, but these two constraints are not required when using FPCA. This is better for Internet traffic data, as not all IPs will necessarily have flows captured at every time for various reasons (e.g., outages). In FPCA, the eigenfunctions are smooth, which we believe fits the general nature of changes in flow data.

There are other practical benefits that FPCA has over PCA. For example, viewing the smoothing of FPCA as an imputation technique, [4] shows that it outperforms both probabilistic and Bayesian PCA through simulation study. The success of the functional version suggests that our method is preferred to smoothing the data prior to application of PCA. Further, [21] demonstrates that PCA

Table 2: Static Fanout Ratio - DARPA Data

	FTP	HTTP
IPs Detected vs. (Ground Truth)	2 (2)	5 (2)
Scanners Undetected	2	110

Table 3: Subspace Method - DARPA Data

	FTP	HTTP
IPs Detected vs. (Ground Truth)	2 (2)	152(111)
Scanners Undetected	2	1

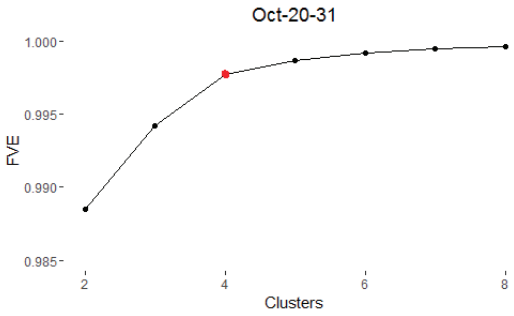

Figure 3: Elbow Plot for NTP Outlier Clustering

has difficulty with high dimensional sparse data - the estimate of the covariance matrix resulting from PCA is poor even if numerical results are stable; FPCA overcomes this issue. Recall that the covariance matrix, given by Equation 2, is directly related to the eigenvalues in our model, and their estimates are used to determine the variance captured by each eigenfunction (or principal component). In methods that use eigenvalues for decisions regarding which components to investigate (such as scree plots), poor estimates can skew results. As studied in [19], the subspace method is very sensitive to how many principal components comprise the normal subspace. Our method uses the FPCA implementation of [24], which was developed specifically for sparse data.

5.3 Analysis of NTP Data

Due to the anonymization of IP addresses every two weeks, we analyze the NTP traffic flows as six individual datasets. Due to space limitations, our figures are constructed using only the data from the October period. Table 4 provides a summary of the numerical features resulting from FPCA analysis. In each, the AIC and BIC selection criteria agree on the number of eigenfunctions, providing the optimal model. That is, the number of eigenfunctions on which we search for outliers is appropriate - we do not need to consider any more to find outliers. Based on elbow plots, shown in Figure 3, each period admits four clusters with the exception of December 29 to January 11. Table 5 contains a summary of the cluster centers and outliers within each.

We use R 3.3.1 on a basic AMD processor system. Running FPCA takes no more than five seconds on the largest set of NTP data (period 1, with about 900 source IPs). By default, R performs these calculations on a single core; parallelizing the algorithm will only increase the speed of the procedure. The process of performing k-means clustering comes at a low cost due to an efficient algorithm for implementation, but also because it is only carried out using the outliers detected in the first step, a small subset of the data.

Recall that all source IPs identified from the results of FPCA are outliers with respect to the number of destinations contacted. That is, the pattern by which they send requests deviates significantly from the rest of the traffic in that period. In both the DARPA and NTP analysis a large quantity of outliers are detected. This does not necessarily imply that these are scanners, or that they all deviate in a similar fashion, so the clustering helps us better understand our set of outlier IPs by using destination information. Specifically, we base the clustering on the source address' proportion of successful contacts, where a success is defined to be at least one packet sent in response by the contacted destination. Since scanners are probing networks for vulnerable hosts unknown to them, we anticipate them to meet failed contacts, and this characteristic is expected to differ for "normal" behavior or even other anomalies.

Similarities arise in the results across all periods. A cluster centered around a large proportion is always present, capturing many outliers. The remaining IPs are spread throughout low and mid-range clusters of smaller sizes. We examined the behavior within each and found that four common patterns of activity had emerged. We label these as follows:

- Blatant Scanners (Cluster 1) - Source addresses that contact a large number of destinations with a low proportion of response. These are all captured in the cluster centered around the smallest value.
- List Keepers (Clusters 2 and 3) - Source addresses that contact many destinations consistently throughout the period, with a varying proportion of successful requests. Thus, this behavior is spread throughout clusters centered around the 'mid-range' values. We think of these as potential scanners checking the status of a list of possibly vulnerable destinations.
- Stealthy Scanners (Cluster 3) - Source addresses contacting few destinations sparsely throughout the period and a generally low proportion of response. We think of these as true scanners looking for vulnerable destinations while trying not to attract attention. With the sometimes small number of requests, this behavior appears in a cluster with a larger center.
- Good Guys (Cluster 4) - Source addresses with a consistently large number of requests sent to a few destinations, and proportion of successful contacts close to 1, which is typical behavior of appropriate NTP interactions. These are all captured by the cluster centered around the largest value.

Certainly, these are not the only types of outlier behavior within Internet traffic, but these are the ones we observe in this data. The largest cluster of outliers in our NTP traffic consists of "good guys," which can be thought of as known entities conducting business as usual. Being able to discover these systems and cluster them analytically makes it easier to verify them, and create a consistent

Table 4: Summary of NTP Data and Results by Period

Period	Hours	Source IPs	FPCA+clustering Count	Fanout Ratio Count
Oct 20-31	288	869	74	15
Nov 1-15	190	713	82	32
Nov 17-25	88	516	76	37
Dec 1-14	195	601	80	27
Dec 15-28	194	658	99	5
Dec 29 - Jan 11	233	686	83	6

Table 5: Summary of NTP Clustering Results by Period

	Oct 20-31		Nov 1-15		Nov 17-25		Dec 1-14		Dec 15-28		Dec-29 - Jan-11	
Cluster	Center	Count	Center	Count	Center	Count	Center	Count	Center	Count	Center	Count
1	.074 (×)	6	.012	10	.011	17	.126	14	.045	4	.006	3
2	.199 (△)	4	.261	12	.298	14	.261	8	.895	1		
3	.949 (○)	11	.609	11	.525	10	.557	12	.966	7	.954	4
4	.999 (+)	53	.998	49	.995	35	.997	46	1.00	87	1.00	76

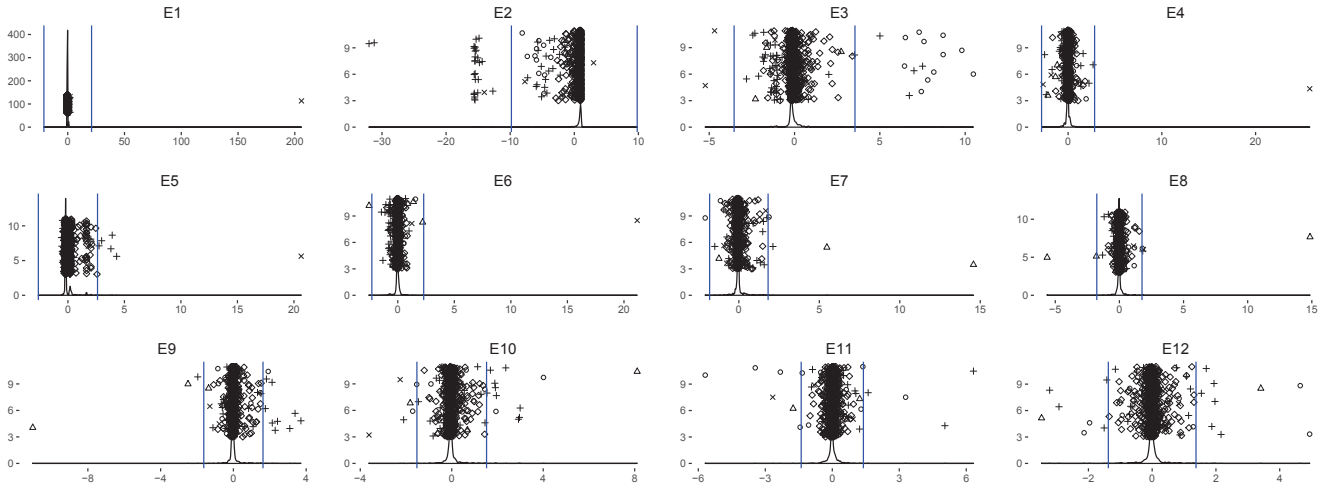

Figure 4: Oct 20-31 NTP Scores with Density Estimates and Cutoff

list for monitoring over time. With this list in hand prior to an attack, one might be able to find a way to let those sources through during mitigation, thus improving service levels. The same goes for IPs thought to be "bad guys" - these can be blocked if necessary. This list would need to be inspected regularly, as malicious users could spoof in response to this strategy.

Figure 4 shows the scores and density estimates for the eigenfunctions considered in the Oct 20-31 period, along with vertical lines representing the three standard deviation cutoff. The shape of the points in the plots reflect which cluster the IP with the corresponding score exists in; an acting legend is located in the second column of Table 5. We note that there is at least one outlier on each eigenfunction, particularly the first, indicated by the single large score. The associated IP is a member of the cluster centered around the lowest proportion, making it a "blatant scanner" (illustrating the concept of 'normal subspace pollution.'). This illustrates the concept of 'normal subspace pollution.' Members of this cluster appear as outliers up to the sixth eigenfunction, reinforcing our belief that the dimensions capturing large variance of the data are not necessarily 'normal.'

Another interesting result seen in Figure 4 is the location of IPs in the cluster centered around .199 (denoted by △ in the graph). As seen in the DARPA FTP-port analysis, the mid-range clusters capture malicious behavior, making them important to investigate. These

IPs are not detected until the sixth and later eigenfunctions, suggesting that even dimensions that capture relatively small amounts of variance are important to consider when searching for outlier IPs.

We note that other data choices could be made, e.g., use FPCA on packet counts received by a destination and cluster on ratio of packets returned to those sent; this would identify which addresses provide amplification. Regardless of the characteristics chosen, our method will also find outliers that do not appear to be "bad guys." Rather than label these as false-positives that indicate the need for refinement of our method, we propose that this mechanism has significant practical importance. Since nothing is known about ground truth, we can only say the outliers found are unusual behaviors when compared to the rest. If some of the activity is not considered malicious, these results provide administrators with aspects of the network to further investigate and determine why they are identified as outliers. The clustering step streamlines this secondary analysis, and knowing the centers of the clusters gives some indication of where to begin searching for true anomalies. For example, in scanner detection, we look at the outliers with low proportions of success, as this is expected of scanner behavior.

5.4 Comparison to Static Fanout Ratio

The last column of Table 4 summarizes the results of applying the static fanout ratio to the NTP datasets. In each period, we have no more than two addresses that the fanout ratio identifies that FPCA+Clustering does not. While investigating these IPs to understand why they were not identified by our method, we see that they contact a small number of destinations only a few times throughout the period; behavior that we classify as "stealthy scanners." It is unknown if these are true scanners, but they warrant closer analysis.

This sort of stealthy behavior is what one would look for in the pursuit of scanners, and we want our method to identify as much of it as possible. In comparing the sources with these features that were detected in FPCA, a difference in the timing of their contacts is noticeable. The IPs that FPCA misses send packets to a handful of destinations only one or two times throughout a period, making them rather difficult to detect in the first step of our method, which is based on the time series of contacts. The NTP datasets are aggregated into hourly buckets, so using a finer resolution may improve our ability to detect these behaviors, but our method provides another possible solution. In the set of "stealthy scanners" that were identified by FPCA, we observe IPs with similar activity in the same subnet as the missing source address; we could focus on these subnets. Of course, such commonalities may not always be true, but our method provides a mechanism through which they can easily be investigated.

Turning to the additional outliers found by our method, we see IPs that closely resemble and are in the same subnet as those of the fanout ratio. The static method fails if the data does not meet its exact threshold, and there are IPs whose fanout just misses the fixed cutoff. Our method finds these IPs and is dynamic & data-driven, because it considers all of the traffic when determining outliers, and needs no prior knowledge to build a rule. FPCA searches for abnormal patterns in the data observed. Scanners are expected to exhibit this with respect to usual traffic and normal processes, but not always in the same manner. Thus, the adaptability of the functional approach is better suited for detection than a strict cutoff.

6 CONCLUSIONS

With motivation of detecting scanners, we propose a new method to identify outlier behavior and patterns in Internet traffic flows based on FPCA and k-means clustering. We demonstrate its ability to catch all security events within two synthetic datasets equipped with ground truth. Further, our method achieves this when existing PCA-based and static fanout ratio approaches do not. We also demonstrate our method using real-world NTP traffic flows prior to a reflection DDoS attack in the beginning of 2014. This illustrates the network information that can be gained from analyzing outliers from the method.

ACKNOWLEDGMENTS

This work was supported in part by the Department of Homeland Security Science and Technology Directorate (D15PC00205).

REFERENCES

[1] Mark Allman, Vern Paxson, and Jeff Terrell. 2007. A Brief History of Scanning. In *Proceedings of the 7th ACM SIGCOMM Conf. on Internet Measurement (IMC '07)*. ACM, New York, NY, USA, 77–82. https://doi.org/10.1145/1298306.1298316

[2] Monowar H Bhuyan, Dhruba Kumar Bhattacharyya, and Jugal K Kalita. 2014. Network anomaly detection: methods, systems and tools. *Ieee communications surveys & tutorials* 16, 1 (2014), 303–336.

[3] Elias Bou-Harb, Mourad Debbabi, and Chadi Assi. 2014. Cyber scanning: a comprehensive survey. *Ieee communications surveys & tutorials* 16, 3 (2014), 1496–1519.

[4] Jeng-Min Chiou, Yi-Chen Zhang, Wan-Hui Chen, and Chiung-Wen Chang. 2014. A functional data approach to missing value imputation and outlier detection for traffic flow data. *Transportmetrica B: Transport Dynamics* 2, 2 (2014), 106–129.

[5] Jakub Czyz, Michael Kallitsis, Manaf Gharaibeh, Christos Papadopoulos, Michael Bailey, and Manish Karir. 2014. Taming the 800 Pound Gorilla: The Rise and Decline of NTP DDoS Attacks. In *2014 Internet Measurement Conf. (IMC '14)*. ACM, New York, NY, USA, 435–448. https://doi.org/10.1145/2663716.2663717

[6] Gilberto Fernandes, Luiz F Carvalho, Joel JPC Rodrigues, and Mario Lemes Proenca. 2016. Network anomaly detection using IP flows with principal component analysis and ant colony optimization. *Journal of Network and Computer Applications* 64 (2016), 1–11.

[7] Badis Hammi, Mohamed Cherif Rahal, and Rida Khatoun. 2016. Clustering methods comparison: Application to source based detection of botclouds. In *Security of Smart Cities, Industrial Control System and Communications, 2016 Intntl. Conf. on*. IEEE, 1–7.

[8] John A Hartigan and Manchek A Wong. 1979. Algorithm AS 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C* 28, 1 (1979), 100–108.

[9] Impact Cyber Trust. 2009. DARPA Scalable Network Monitoring Program Traffic. https://www.impactcybertrust.org/dataset_view?idDataset=303. (2009).

[10] Gibbs Y Kanyongo. 2005. Determining the correct number of components to extract from a principal components analysis: A Monte Carlo study of the accuracy of the scree plot. *Journal of Modern Applied Statistical Methods* 4, 1 (2005), 13.

[11] David J Ketchen and Christopher L Shook. 1996. The application of cluster analysis in strategic management research: an analysis and critique. *Strategic management journal* 17, 6 (1996), 441–458.

[12] Anukool Lakhina, Mark Crovella, and Christophe Diot. 2004. Diagnosing network-wide traffic anomalies. In *ACM SIGCOMM Computer Communication Review*, Vol. 34. ACM, 219–230.

[13] Yehua Li, Naisyin Wang, and Raymond J Carroll. 2013. Selecting the number of principal components in functional data. *J. Amer. Statist. Assoc.* 108, 504 (2013), 1284–1294.

[14] Moisés F Lima, Lucas DH Sampaio, Bruno B Zarpelao, Joel JPC Rodrigues, Taufik Abrao, and Mario Lemes Proenca Jr. 2010. Networking anomaly detection using dsns and particle swarm optimization with re-clustering. In *GLOBECOM 2010*. IEEE, 1–6.

[15] Merit Network. 2016. Merit. https://www.merit.edu/. (2016).

[16] Hans-Georg Muller and Jane-Lin Wang. 2015. PACE Package for Functional Data Analysis and Empirical Dynamics. http://www.stat.ucdavis.edu/PACE/. (2015).
[17] Animesh Patcha and Jung-Min Park. 2007. An overview of anomaly detection techniques: Existing solutions and latest technological trends. *Computer networks* 51, 12 (2007), 3448–3470.
[18] James O Ramsay and Bernard W Silverman. 2002. *Applied functional data analysis: methods and case studies*. Vol. 77. Citeseer.
[19] Haakon Ringberg, Augustin Soule, Jennifer Rexford, and Christophe Diot. 2007. Sensitivity of PCA for traffic anomaly detection. *ACM SIGMETRICS Performance Evaluation Review* 35, 1 (2007), 109–120.
[20] Pourya Shamsolmoali and Masoumeh Zareapoor. 2014. Statistical-based filtering system against ddos attacks in cloud computing. In *Advances in Computing, Communications & Informatics, Conf.* IEEE, 1234–1239.
[21] Han Lin Shang. 2014. A survey of functional principal component analysis. *Advances in Statistical Analysis* 98, 2 (2014), 121–142.
[22] Jun Shao. 1997. An asymptotic theory for linear model selection. *Statistica Sinica* (1997), 221–242.
[23] Scott I Vrieze. 2012. Model selection and psychological theory: a discussion of the differences between the AIC and the BIC. *Psychological methods* 17, 2 (2012), 228.
[24] Fang Yao, Hans-Georg Müller, and Jane-Ling Wang. 2005. Functional data analysis for sparse longitudinal data. *J. Amer. Statist. Assoc.* 100, 470 (2005), 577–590.