# Evading Anomaly Detection through Variance Injection Attacks on PCA
## (Extended Abstract)

Benjamin I.P. Rubinstein[1], Blaine Nelson[1], Ling Huang[2], Anthony D. Joseph[1,2], Shing-hon Lau[1], Nina Taft[2], and J. D. Tygar[1]

[1] UC Berkeley
[2] Intel Research, Berkeley

**Abstract.** Whenever machine learning is applied to security problems, it is important to measure vulnerabilities to adversaries who poison the training data. We demonstrate the impact of variance injection schemes on PCA-based network-wide volume anomaly detectors, when a single compromised PoP injects chaff into the network. These schemes can increase the chance of evading detection by sixfold, for DoS attacks.

## 1 Motivation and Problem Statement

We are broadly interested in understanding vulnerabilities associated with using machine learning in decision-making, specifically how adversaries with even limited information and control over the learner can subvert the decision-making process [1]. An important example is the role played by machine learning in dynamic network anomography, the problem of inferring network-level Origin-Destination (OD) flow anomalies from aggregate network measurements. We ask, can an adversary generate OD flow traffic that misleads network anomography techniques into misclassifying anomalous flows? We show the answer is yes for a popular technique based on Principal Components Analysis (PCA) from [2].

The detector operates on the $T \times N$ link traffic matrix $\mathbf{Y}$, formed by measuring $N$ link volumes between PoPs in a backbone network, over $T$ time intervals. Figure 1 depicts an example OD flow within a PoP-to-PoP topology. Lakhina et al. observed that the rows of the normal traffic in $\mathbf{Y}$ lie close to a low-dimensional subspace captured by PCA using $K = 4$ principal components. Their detection method involves projecting the traffic onto this normal $K$-dimensional subspace; large (small) residuals, as compared with the $Q$-statistic, are called *positive* anomalies (*negative* normal traffic).

## 2 Results and Future Work

Consider an adversary launching a DoS attack on flow $f$ in week $w$. Poisoning aims to rotate PCA's $K$-dimensional subspace so that a false negative (FN) occurs during the attack. Our *Week-Long* schemes achieve this goal by adding high variance chaff at the compromised origin PoP, along $f$, throughout week $w-1$. Figure 2 presents results for two chaff methods. Both methods add chaff $c_t$ to

R. Lippmann, E. Kirda, and A. Trachtenberg (Eds.): RAID 2008, LNCS 5230, pp. 394–395, 2008.

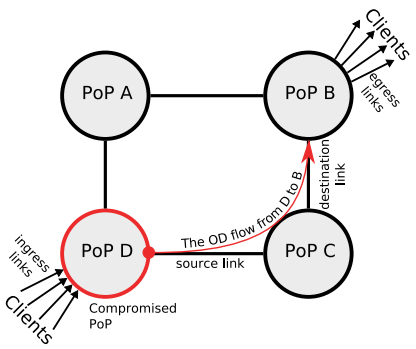


**Fig. 1.** Point-of-presence (PoP)-level granularity in a backbone network, and the links used for data poisoning

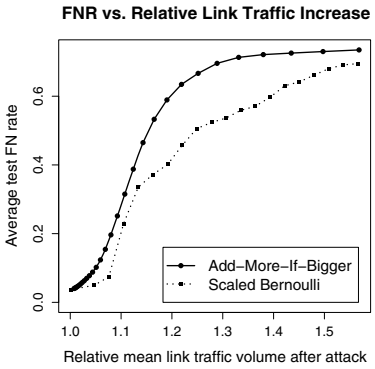


**Fig. 2.** *Week-Long* attacks: test FNRs are plot against the relative increase to the mean link volumes for the attacked flow

each link in $f$ at time $t$, depending on parameter $\theta$: *Scaled Bernoulli* selects $c_t$ uniformly from $\{0, \theta\}$; *Add-More-If-Bigger* adds $c_t = (y_o(t) - \overline{y_o})^\theta$ where $y_o(t)$ and $\overline{y_o}$ are the week $w-1$ origin link traffic at time $t$ and average origin link traffic, respectively. We evaluated these methods on data from Abilene's backbone network of 12 PoPs. For each week 2016 measurements were taken, each averaged over 5 minute intervals, for each of 54 virtual links—15 bi-directional inter-PoP links and the PoPs' ingress & egress links.

The attacker's chance of evasion is measured by the FN rate (FNR). We see that the *Add-More-If-Bigger* chaff method, which exploits information about origin link traffic, achieves greater FNR increases compared to *Scaled Bernoulli*. The baseline FNR of 4% for PCA on clean data, can be doubled by adding on average only 4% additional traffic to the links along the poisoned flow. The FNR can be increased sixfold to 24% via an average increase of 10% to the poisoned link traffic. In our *Boiling Frog* strategies, where poisoning is increased slowly over several weeks, a 50% chance of successful evasion can be achieved with a modest 5% volume increase from week-to-week over a 3 week period [3].

We have verified that simply increasing the number of principal components is not useful in protecting against our attacks [3]. In future work we will evaluate counter-measures based on Robust formulations of PCA, and will devise poisoning strategies for increasing PCA's false positive rate.

## References

1. Barreno, M., Nelson, B., Joseph, A.D., Tygar, J.D.: The security of machine learning. Technical Report UCB/EECS-2008-43, UC Berkeley (April 2008)
2. Lakhina, A., Crovella, M., Diot, C.: Diagnosing network-wide traffic anomalies. In: Proc. SIGCOMM 2004, pp. 219–230 (2004)
3. Rubinstein, B.I.P., Nelson, B., Huang, L., Joseph, A.D., Lau, S., Taft, N., Tygar, J.D.: Compromising PCA-based anomaly detectors for network-wide traffic. Technical report UCB/EECS-2008-73, UC Berkeley (May 2008)