Network Anomaly Detection Using a Commute Distance Based Approach

Authors: Nguyen Lu Dang Khoa and Tahereh Babaie
Supervisor: Professor Sanjay Chawla
SCHOOL OF IT, FACULTY OF ENGINEERING & INFORMATION TECHNOLOGIES

Introduction

• Finding unusual and large changes in network traffic.

• Anomalies: intentional attacks (e.g. a DDoS attack, a viruses spread), an unusual network traffic (e.g. flash crowds).

Background

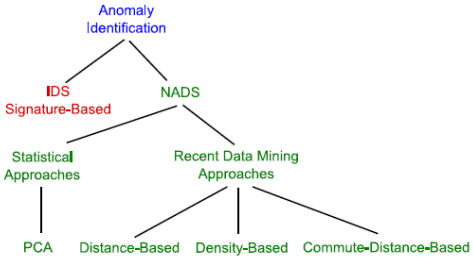
Intrusion detection system (IDS):

• Signature-based method: using a pre-defined set of patterns describing previous anomalous events to identify future anomalies.

• Can only find known attacks and new attack patterns need to be updated over time.

Network anomaly detection system (NADS):

• Try to identify new or previously unknown abnormal traffic behaviours.



Anomaly Detection Using Principal Component Analysis (PCA)

PCA:

• Data in a high dimensional space is transformed to a lower dimensional subspace which captures most of the data variance.

• The new subspace has a smaller number of uncorrelated variables called Principle Components (PC).

• Typically first few PCs account for most of the variance in the original data.

• Computation includes the eigenvalues decomposition of the data covariance matrix.

Anomaly detection using PCA:

• First few PCs:
- Capture most of variance in the dataset
- Be strongly related to one or more of the original variables

• Last few PCs:
- Represent the linear combination of original variables with very small variance
- The data tend to result in similar small values in those PCs.
- Any observation that largely deviates from this tendency for the last few PCs is likely to be an anomaly.

• Subspace approach: based on decomposition of a main space of data into: $x_i = x_i^N + x_i^A$

-Normal subspace: projection of the data on the first few PCs $x_i^N = PP^T x_i = Cx_i$

- Anomalous subspace: projection of data on the last few PCs $x_i^A = (I - C)x_i$

-Network traffic instances are considered anomalies if $\|x_i^A\|^2$ is greater than a chosen threshold.

Anomaly Detection Using Commute Distance

Commute distance:

• The commute time: the expected number of steps a random walk starting at i will take to reach j once and go back to i for the first time.

• Commute distance between two nodes can capture both the distance between them and their local neighborhood densities.

• Can be calculated from the pseudo-inverse of the graph Laplacian matrix

$$c(i,j) = V_G(l_{ii}^+ + l_{jj}^+ - 2l_{ij}^+)$$

Anomaly detection using commute distance:

• Construct the mutual k-nearest neighbor graph from the dataset

• Compute the graph Laplacian matrix L and its pseudoinverse L^+

$$c(i,j) = V_G(l_{ii}^+ + l_{jj}^+ - 2l_{ij}^+)$$

• Find top N outliers using commute distance based technique with pruning rule

Experimental Results

Datasets:

• Small-size NICTA wireless mesh network

• Abilene backbone network

Comparison:

• PCA

• EDOF: Euclidean distance-based method

• LOF: density-based method

• CDOF: commute distance based method

Results and Analysis:

• **NICTA dataset:** PCA found anomalies but was very sensitive to parameters changes.

• **Abilene dataset 1:**

Table I: False positives in the top 30 detected anomalies compared with the top 50 benchmark anomalies

	False positive with top 50 benchmark				
Top 30 detected	PCA	EDOF	LOF	CDOF	Average
PCA	28	28	30	27	28.25
EDOF	2	0	0	2	1.00
LOF	7	4	1	1	3.25
CDOF	2	0	1	0	0.75

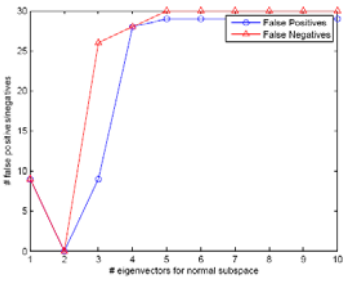
Table II: False negatives in the top 50 detected anomalies compared with the top 30 benchmark anomalies

	False positive with top 30 benchmark				
Top 50 detected	PCA	EDOF	LOF	CDOF	Average
PCA	28	30	30	30	29.50
EDOF	4	0	0	3	1.75
LOF	4	2	0	2	2.00
CDOF	4	0	0	0	1.00

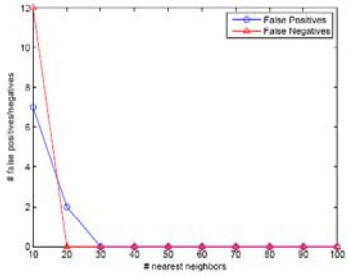
• **Abilene dataset 2:**
-EDOF, LOF, and CDOF all found only one anomaly if the threshold based on anomaly scores was used while PCA could not find it.
- PCA falsely returned 935 time bins as anomalies. An analysis on the data shows that there was a very large and dominating traffic volume appearing in time bin 1350.
- The large anomaly skewed the normal subspace forming by the first few PCs and consequently increased the false positives.

• **Parameters sensitivity:**

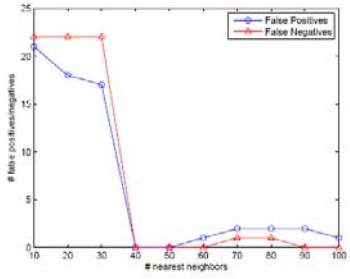
- Varied parameters in all the methods.
- PCA was very sensitive to changes of parameters used.



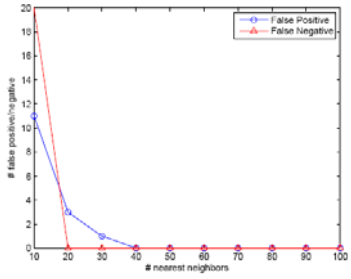
(a) PCA



(b) EDOF



(c) LOF



(d) CDOF

Conclusion

• Propose the use of commute distance to discover anomalies in network traffic data.

• Address the weaknesses of PCA in detecting anomalies in computer network domain.

• The commute distance based approach has a lower false positive and false negative rate in anomaly detection compared to PCA and typical distance-based and density-based approaches.

References

1. Nguyen Lu Dang Khoa and Sanjay Chawla, "*Robust outlier detection using commute time and eigenspace embedding*," in Proceedings of PAKDD 2010, pp. 422–434.
2. Nguyen Lu Dang Khoa, Tahereh Babaie, Sanjay Chawla, and Zainab Zaidi, "*Network Anomaly Detection Using a Commute Distance Based Approach*," in Proceedings of ICDMW 2010 (to appear).

THIS RESEARCH IS SPONSORED BY CMCRC
Capital Markets CRC Limited