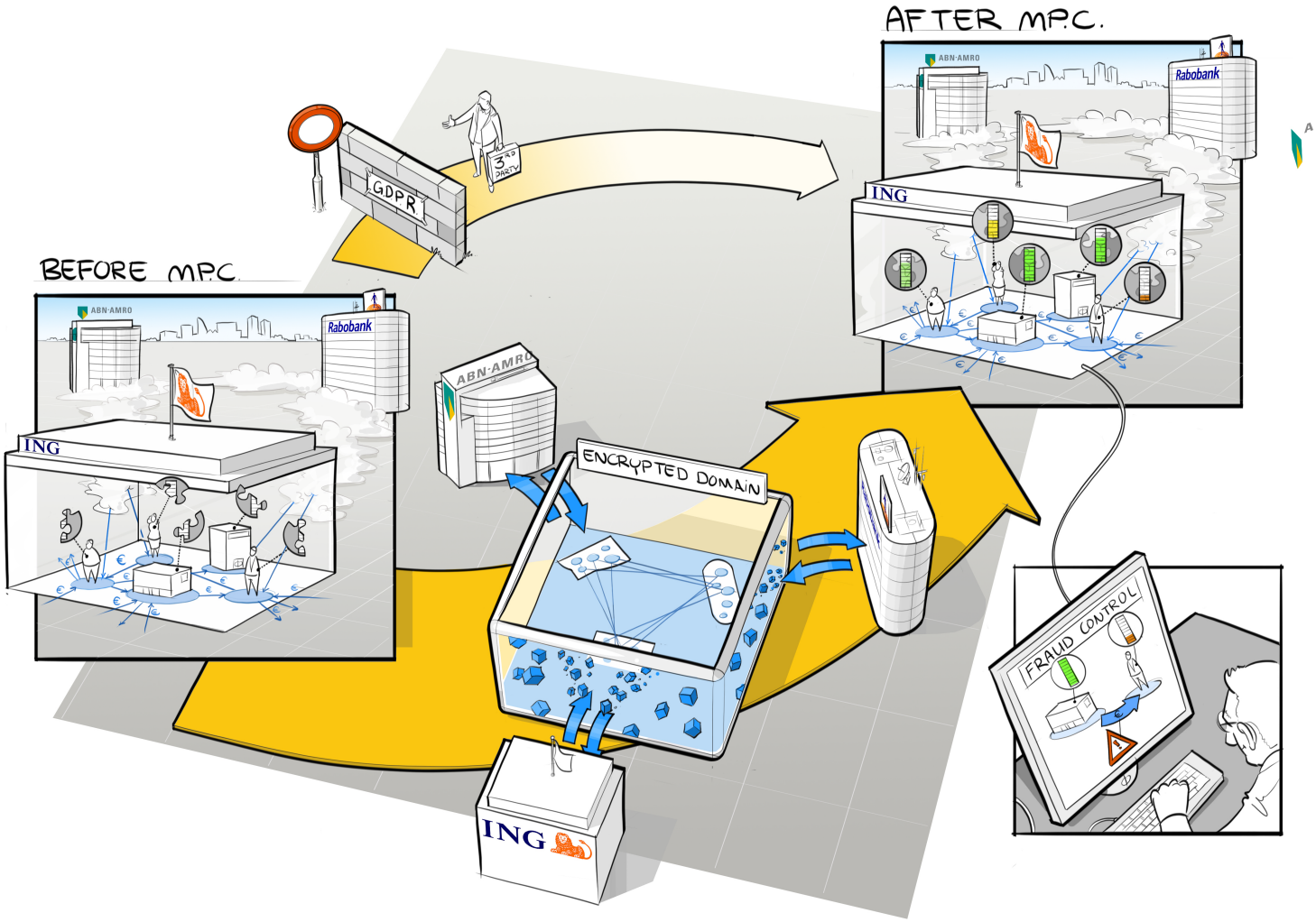
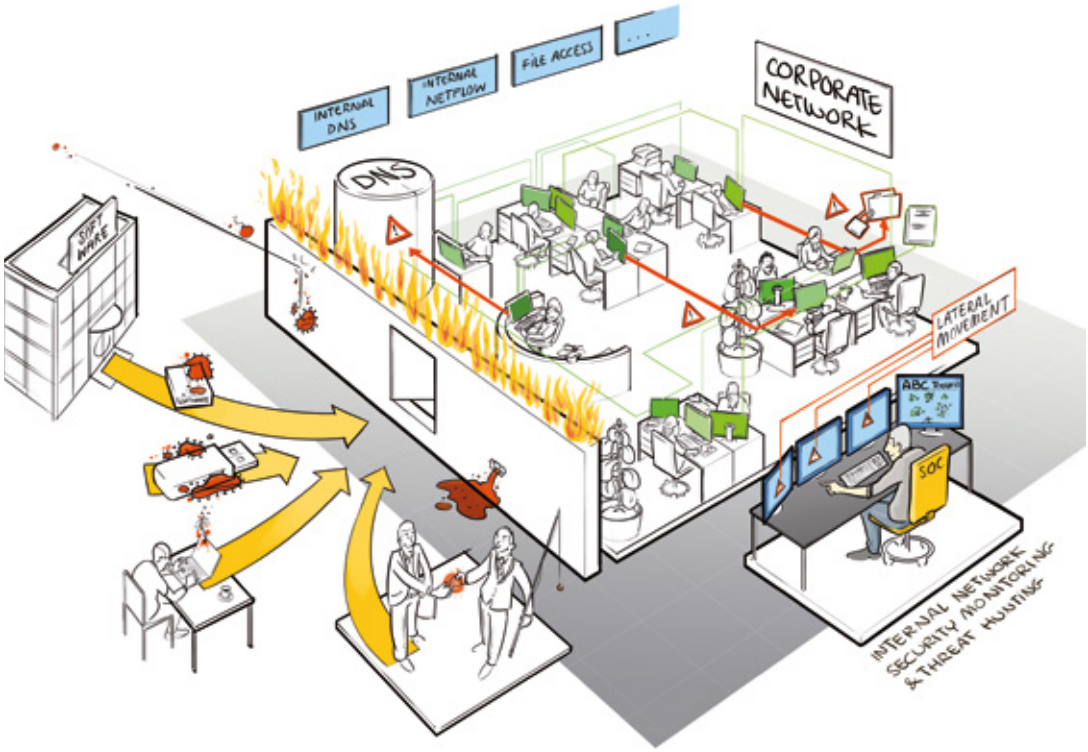
DETECTING LATERAL MOVEMENT WITH ANOMALYBASED CLUSTERING



Detecting Lateral Movement with Anomaly-Based Clustering

Alex Sangers (TNO), Erik Meeuwissen (TNO), Kadir Kalayci (Achmea)



We focus on finding anomalies (i.e. deviations from normal behavior) in internal network communication patterns which may be an indication of lateral movement in the cyber attack kill chain.

Introduction

Cyber attacks are getting more and more advanced and targeted. The traditional eggshell model of a hardened perimeter and a more or less open internal network is no longer sufficient (ClearSky 2018). In this research, we assume that a security breach already took place, and we investigated how it can be detected using internal network traffic (Beukema 2017). In fact, we focus on finding anomalies (i.e. deviations from normal behavior) in internal network communication patterns which may be an indication of lateral movement in the Cyber attack kill chain (Cutler 2010). To this end, the generic Anomaly-Based Clustering (ABC) toolkit has been developed.

It has the following key features:

- Applicable to data commonly available in IT environments (e.g. internal DNS, internal Netflow, file access logs). In fact, we use available data sources to model internal enterprise networks; we specifically focus on which end points (hosts) communicate with each other, and how much. The data does not need to be labeled with cyber attacks.
- Automatic clustering of hosts in the network with similar communication patterns.
- Statistical modeling of normal communication behavior between clusters.

- Detecting when individual end points start to deviate from the baseline of normal behavior. Typically, such deviations may indicate network misconfigurations as well as lateral movement as part of a Cyber attack.
- Internal network visualization and anomaly visualization to support SOC analysts and threat hunters.

The ABC toolkit supports anomaly detection and threat hunting [SQRRL 2018]. For example, by starting from a specific hypothesis of how a successful Cyber attack on an organization could take place, it can be used to search for such an attack. Thus, it can be considered as an extra line of defense to complement real-time detection based on signatures or blacklists.

The remainder of this paper is organized as follows. First, we describe the rationale behind Anomaly Based Clustering. Then, we describe the process steps supported by the ABC toolkit. Next, we describe the experimental validation with a real-life dataset consisting of file access logs. Finally, we end with conclusions.

Rationale behind clustering of hosts

One of the research challenges for anomaly detection in internal network communications traffic is to create a useful model of normal behaviour. As individual hosts typically behave relatively unpredictable, the ABC toolkit models groups of hosts (i.e. clusters) with similar communication behavior. This helps to reduce false positives caused by irregular behaviour of individual hosts. Moreover, it improves scalability such that large networks can be modelled (e.g. more than 100.000 end points).

Process steps supported by the ABC toolkit

When using the ABC toolkit, a threat hunting hypothesis must be developed. To effectively use the ABC toolkit, the hypothesis should consist of attack steps such as lateral movement or internal data exfiltration. The threat hunting scenario might also depend on the availability of data sources on the internal network. The attack (steps) should hypothetically impact available data. The ABC toolkit approach consists of the following six phases:

As individual hosts typically behave relatively unpredictable, the ABC toolkit models groups of hosts (i.e. clusters) with similar communication behavior.

1 Data selection.

Define a threat hunting scenario and find data where deviating communication patterns between source and destination addresses might indicate an attack. The data should consist of information on what systems communicating with what systems with what volume. Some examples of suitable data sources for threat hunting scenarios:

a. Based on Netflow data, e.g. port 1433 for detection of anomalous connections to SQL servers.
b. Based on file access data: detection of anomalous file path attempts.
c. Based on DNS data: detection of anomalous A-queries of hosts to the DNS resolver

2 Data parsing.

Filter, clean and process the input data to a format that can be interpreted by the subsequent modules of the ABC toolkit. Examples of filtering are IP addresses that are expected to behave anomalously by nature, and data that is irrelevant for the threat hunting scenario.

3 Clustering and network visualization.

The selected clustering technique is called Louvain clustering [Blondel 2008]. Louvain clustering is a community detection algorithm that groups hosts that are strongly interconnected between themselves in a cluster, and less strongly connected to other clusters. The clustering is automatically determined based on the communication patterns in the historical (network) data.

4 Clustering and cluster modelling.

The Louvain clustering ensures that hosts within a cluster are strongly interconnected. Traffic within a cluster is assumed to be normal. Traffic between clusters is considered to be interesting to monitor. Each cluster has so-called inter-cluster communication models that captures the normal traffic between that cluster and other clusters. The inter-cluster communication models are statistical models of communication between clusters based on training data.

5 Anomaly detection.

Compare the inter-cluster communication models with new (test) data to find deviating communication patterns between hosts in different clusters. A sudden increase of inter-cluster communication might indicate a host is deviating from its normal behaviour. The ABC toolkit detects three types of anomalies:

a. A host with a statistically significant increased amount of traffic to hosts in another cluster.
b. A host that communicates to hosts in another cluster, although no communication between the corresponding two clusters existed in the inter-cluster communication models.
c. A host communicates to hosts that were not present in the inter-cluster communication models at all.

The output of this module is a prioritized list of (anomalous) hosts.

6 Anomaly inspection.

Zoom into the behaviour of individual anomalous hosts. Both the inter-cluster communication model of the host and the communication during testing are visualized and can be compared. In addition to the visualization, a text file providing the corresponding data rows has been generated to support actionable follow-up.

The process steps are visualized in Figure 1.

Experimental validation for the use case file ac cess logs

To apply the ABC toolkit to several use cases, it is important to understand the modelling and detection approach to determine what use case to develop. The use case should consist a threat hunting scenario including the available data source with (internal) source and destination addresses.

File access data consists of data on at what time what user was trying to access which file location. It is also includes whether the user was reading, writing or executing the file and whether this was a successful attempt or not. We executed an experiment with real-life file access data to identify hosts that had anomalous file access attempts.

1 Data selection.

We used file access data with user ID as source address and file path as destination address. The threat hunting hypothesis is that an adversary is already inside and scanning for commercially or privacy sensitive data that is stored on one or more file servers. Deviations in file access data indicate that a user is attempting to access file paths more often than usual or attempting to access other file paths than usual. This behaviour might be caused by a fraudulent employee or an adversary that is gathering data.

File access data consists of information on whether the user was reading, writing or executing the file and whether this was a successful attempt or not.

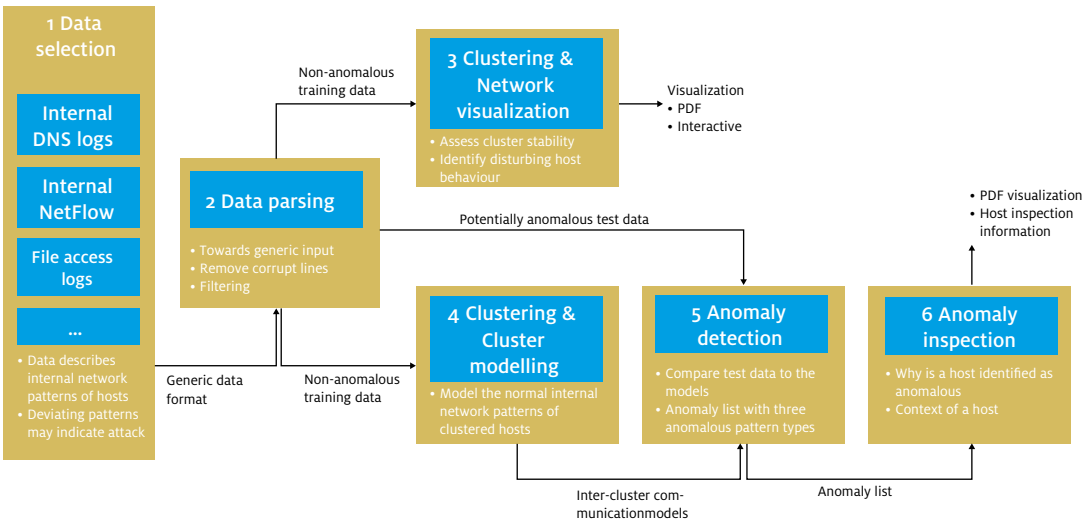


Figure 1: The process steps of the ABC tooling.

Many users only accessed a limited number of file paths and many file paths were accessed by only a limited number of users.

2 Data parsing.

The data was filtered on some anomalous systems that are anomalous due to their functionality, and corrupt data rows were removed. Additionally, the data was formatted to the generic format. The training data is based on file access data of the first half of a working day (12 hours) and the test data is based on the second half of that working day (12 hours).

3 Clustering and network visualization.

The community detection of the file access training data is shown in Figure 3. Each node in the graph represents a user or file path and each colour represents a community. Note that many separate small communities exist, meaning that many users only accessed limited number of file paths and that many file paths were accessed by only a limited number of users. In the middle of the graph a larger number of nodes are connected. There exist a small number of communities within this middle part. This indicates that there are users and file paths connecting all together, but some parts are more connected than others. In Figure 3, we recognize some similar patterns but there are some differences on first sight. Firstly, more nodes are present, indicating more users and/or file paths present in the second data set. Secondly, the middle part that is connected has more nodes but similar number of clusters that are identified.

For the sake of experimentation the clusters were deemed to be sufficiently similar based on visual inspection of these figures.

4 Clustering and cluster modelling.

The first half of the working day, shown in Figure 3, was used as training data to develop a baseline. The baseline consists of inter-cluster communication models that provide a statistical analysis on the communication of hosts between clusters.

5 Anomaly detection.

The inter-cluster communication models as developed in the previous step were compared to the test data as visualized in Figure 1. The result is a list of anomalous user IDs. The most anomalous host was assigned to cluster 683 and has a significantly increased number of file path visits, attempts to access file path that were not visited by its cluster during the training period and it also visits file paths that were not seen in the training period yet.

6 Anomaly inspection.

Zooming into the behaviour of the anomalous host shows what communication was anomalous. The anomalous host was assigned to cluster 683 and that cluster has had some communication to cluster 44 during the training period (Figure 4). During the testing period, the anomalous host in cluster 683 started visiting significantly more file paths belonging to cluster 44, started accessing file paths in cluster 700, which did not occur in training period, and started visiting new file paths that were not even present in training data, represented by cluster 701. This is shown in Figure 5.

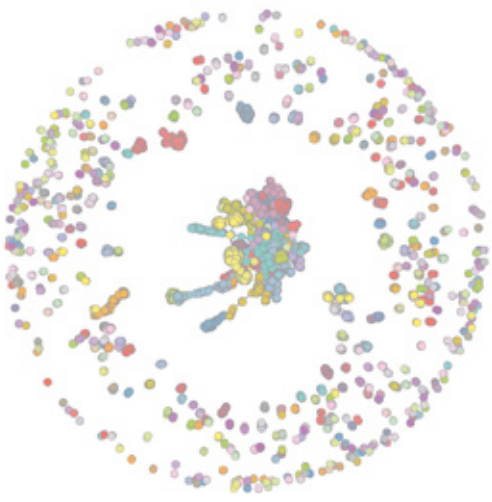


Figure 2: Cluster visualization of file access data of the first half of a working day.

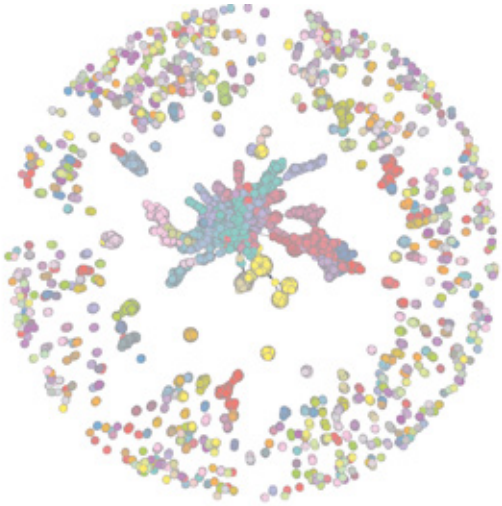


Figure 3: Cluster visualization of file access data of the second half of a working day.

In addition to the visualization, a text file providing the exact file access attempts has been generated to support actionable follow-up.

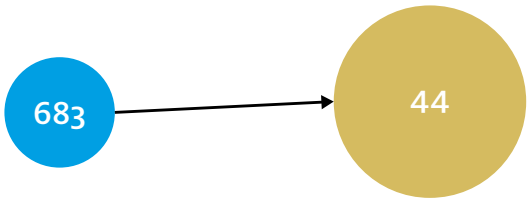


Figure 4: The communication behavior of the host in cluster 683 during the training period.

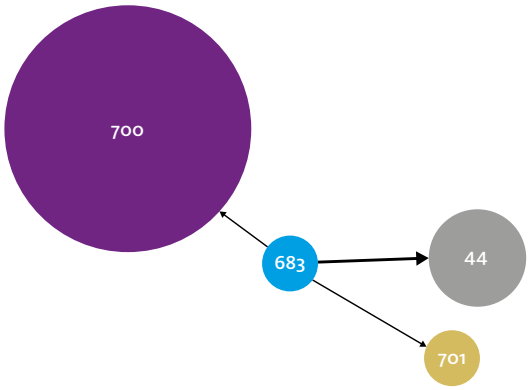


Figure 5: The communication behavior of the anomalous host in cluster 683 during the test period.

The output of the Anomaly Based Clustering approach was further investigated. It appeared that the user was rightly classified as anomalous, but that is was not related to an ongoing Cyber attack.

Conclusion

The ABC toolkit can be used as a threat hunting tool to inspect various internal network data sources (e.g. internal DNS, internal Netflow, and file access data) and is generically applicable lateral movement detection. The ABC toolkit has the following unique properties:

- **The ABC toolkit** has shown to be practically applicable on a real-life file access data set and rightly identifies anomalies.
- **No prior knowledge** about the IT network, infrastructure or systems is required. The toolkit automatically develops the baseline based on observed historical data. If this historical data is updated, internal network changes are automatically taken into account.
- The use of unsupervised machine learning does **not require labelled data** with known threats.,

The (Louvain) clustering enables to automatically model normal communication patterns.

The (Louvain) clustering enables to automatically model normal communication patterns. In contrast with other approaches which use signatures, the ABC toolkit can be used to detect newly developed attack steps due to anomaly detection.

- By modelling behaviour of clusters of hosts instead of modelling the unpredictable behaviour of individual hosts, the number of **false-positives** is **reduced**. In addition, the clustering offers a **scalable** solution for anomaly detection with tens of thousands of hosts. Moreover, the clustering is based on the communication within the whole internal network, so that the outcome is not predictable by attackers.
- The output of the ABC toolkit is **actionable**; it results in a prioritized list of anomalous hosts including what deviating communication patterns triggered the detector. Therefore, it is easy to follow-up on the anomalies by looking into the context and the content for further security investigation.

A next step is to practically evaluate the toolkit during red team activities to further improve its functionality and validate its effectiveness to detect lateral movement. The ABC toolkit has initially been developed for threat hunting; another next step is to extend the ABC toolkit to detect lateral movement (near) real-time.

The ABC toolkit is developed in collaboration with the SRP partners ABN AMRO, Achmea, ING and Rabobank. The experimental software is available “as is”. Parties who are interested to apply the ABC toolkit on their internal network data are invited to contact TNO.

Next step is to practically evaluate the toolkit during red team activities to further improve its functionality.

Bibliography

[Beukema 2017] Beukema, W.J.B., Attema, T., Schotanus, H.A. 2017. "Internal network monitor-." Proceedings of the 3rd International Conference on Information Systems Security and Privacy. SciTePress. 694-703.

[Blondel 2008] Blondel, V.D., Guillaume, J., Lambiotte, R., Lefebvre, E. 2008. "Fast unfolding of communities in large networks." Journal of Statistical Mechanics.

[ClearSky 2018] ClearSky. 2018. Cyber Intelligence Report 2017. Technical Report, ClearSkye Cyber Security Ltd.

[Cutler 2010] Cutler, T. 2010. Anatomy of an advanced persistent threat. Accessed May 2019. http://terrycutler.com/news/securityweek%20-%20anatomy-advanced-persistent-threat.pdf.

[SQRRL 2018] SQRRL. 2018. "A Framework for Cyber threat hunting."