Automating Reasoning of MITRE ATT&CK for Predicting Cyber Attack Techniques using Statistical Machine Learning

Rawan Al-Shaer and Dr. Jonathan Spring

Software Engineering Institute
Carnegie Mellon University
Pittsburgh, PA 15213

This material is based upon work funded and supported by the Department of Defense under Contract No. FA8702-15-D-0002 with Carnegie Mellon University for the operation of the Software Engineering Institute, a federally funded research and development center.

DM19-0822

Introduction

•MITRE ATT&CK TTP (Tactics, Techniques, Procedures) are low-level descriptions of adversarial actions.

•Everyone is interested in using ATT&CK for detection, prediction, forensics, and threat hunting because it provides observables for detecting attacks.

•Goal: Characterize the behavior of APT, malware, and software attacks

•Challenges:

1. MITRE ATT&CK is not ordered in a technique level – important for prediction and threat hunting
2. MITRE ATT&CK is not ordered in Kill Chain level – important to understand attacker strategies and constructing TTP Chains

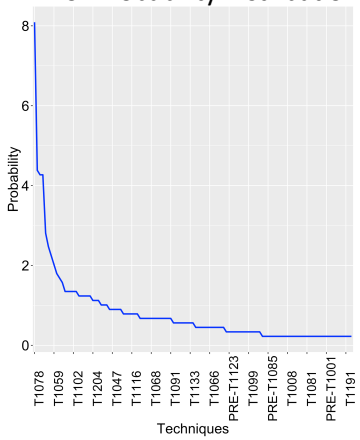
•Hypothesis: Do MITRE ATT&CK techniques exhibit associations, pre-conditions, or post-conditions?

Preliminary Statistical Learning

- Analysis on 55 APT attacks
- Technical Approach

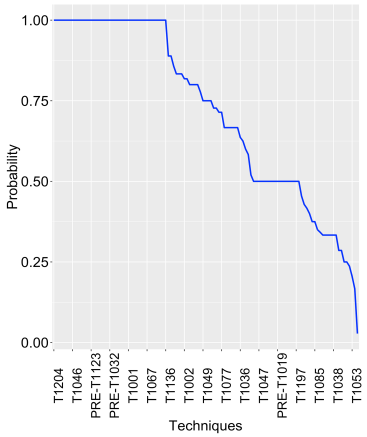
1. Prior Probability
2. Maximum Precondition Likelihood
3. Technique Predictability

Prior Probability Distribution



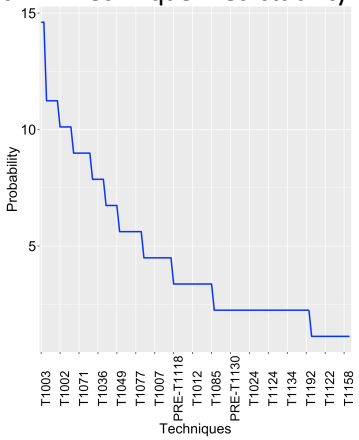
- The prior probability distribution exhibits the **Power-Law distribution.**
- E.g., **16%** of the techniques are used in **55%** of the APT attacks.

Maximum Precondition Likelihood



- **37%** of the techniques exhibit the *maximum precondition likelihood* of 100%.
- **61%** of the techniques exhibit more than a **65%** likelihood of being preconditions.

Technique Predictability



- **10%** of the techniques have a predictability of **80%** relative to the rest of the techniques.
- The absolute **predictability is low** due to the lack of a large set of attack data.

Preliminary Findings

Yes, techniques exhibit associations, pre-conditions, post-conditions ACSAC 2018 Poster Paper

- Problems:

1. Statistical Learning performed only a pairwise analysis which does not create the chain
2. Not scalable: 59,536 pairwise comparisons
3. Chain of 7: cost of computation is 9,366,645,241,008 combinations
4. Partially addressed Challenge 1 (technique ordering), but not Challenge 2 (Kill Chain ordering)

Research Overview

Improving our Preliminary analysis → Statistical Machine Learning

1. To address Challenge 1, we implemented a series of Unsupervised ML along with Statistical Validation
2. To address Challenge 2, we used heuristics along with data mining approaches

Methods Outline

- Data set Collection
- Clustering
  - Distance Metrics
  - Measuring the clusterability
    - Hopkins Statistic + Visual Assessment of cluster Tendency
  - Partitioned Clustering
    - Finding the optimal K clusters
    - K means clustering
    - PAM clustering
    - Fuzzy Analysis clustering
    - Cluster Validation
  - Hierarchical Clustering
    - Finding the optimal K clusters
    - Agglomerative clustering
    - Divisive clustering
    - Cluster Validation
- Statistical Validation
- Heuristics
  - Mapping tactics and techniques to Kill Chain phases
- Sequential Pattern Mining
  - Extracting temporal technique rules with corresponding confidence

Dataset Collection

- Our dataset includes:
  - 66 APT attacks from MITRE
  - 204 Software attacks from MITRE*
  - 498 Software attacks from CTI reports extracted using TTPDrill
    - Subset of 170 were used for analysis*
- Description of our data:
  - Each attack is an observation, with techniques as features
  - Asymmetrical Binary dataset: 1 for technique being observed in an attack, 0 for technique not being observed in an attack

*Results from this data set were not included for compactness

Performing Clustering

- Clustering provides technique associations based on how attacks combined these techniques together
- Challenges Faced:
  - Determining the appropriate distance metric
  - Determining the most effective clustering method
  - Extracting significant technique associations
  - Creating an approach for validation

Distance Metrics

- Clustering is performed using distance measures
- Objectives:
  1. Finding the distance metric appropriate for the type of data
  2. Maintaining interpretability with the distance metric
- Our study on distance metrics determined:
  1. Jaccard Distances (typically used for asymmetrical binary data) are appropriate
  2. Spearman Correlation Distances are appropriate
  3. Euclidean Distances (typically used) are *not* appropriate

Measuring the Clusterability

•Hopkins Statistic: used to assess the clustering tendency of a dataset by measuring the probability that a given dataset is generated by a uniform distribution – tests the spatial randomness of the data

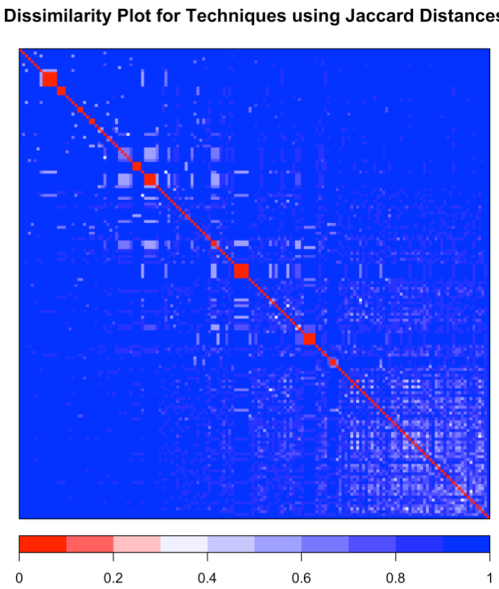
•Interpretability:

- **H = 0.5**: The data set is uniformly distributed and contains no meaningful clusters
- **H ≅ 1**: The data set contains meaningful clusters
- **H ≅ 0**: The data set is regularly spaced (neither clustered nor random)

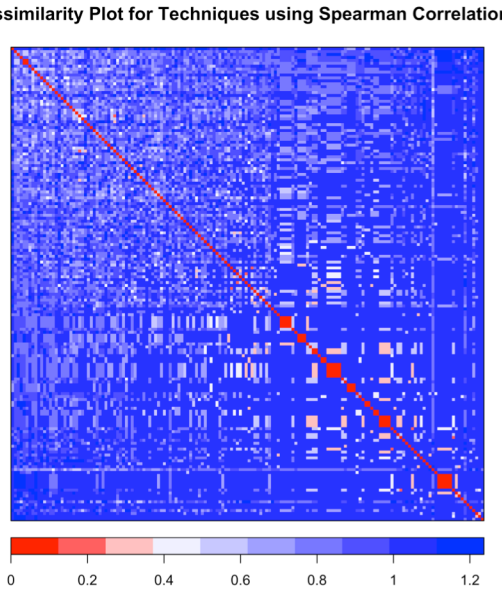
•Visual Assessment of Cluster Tendency: using the Ordered Dissimilarity Matrix to visualize the (dis)similarity between your data

Measuring the Clusterability Results

- Using Jaccard Distances: **H = 0.48**



- Using Spearman Distances: **H = 0.6**



Methods Outline

- Data set Collection
- Clustering
  - Distance Metrics
  - Measuring the clusterability
    - Hopkins Statistic + Visual Assessment of cluster Tendency
  - Partitioned Clustering
    - Finding the optimal K clusters
    - K means clustering
    - PAM clustering
    - Fuzzy Analysis clustering
    - Cluster Validation
  - Hierarchical Clustering
    - Finding the optimal K clusters
    - Agglomerative clustering
    - Divisive clustering
    - Cluster Validation
- Statistical Validation
- Heuristics
  - Mapping tactics and techniques to Kill Chain phases
- Sequential Pattern Mining
  - Extracting temporal technique rules with corresponding confidence

Finding the Optimal K Clusters

Finding the Optimal number of K clusters is important for partitioned clustering, and is beneficial for hierarchical clustering

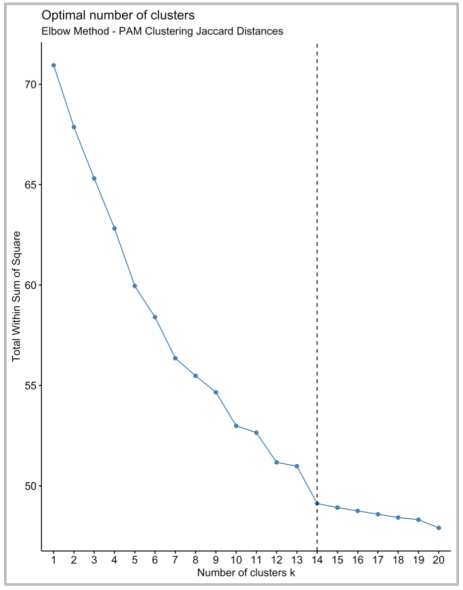
Methods appropriate for our data set:

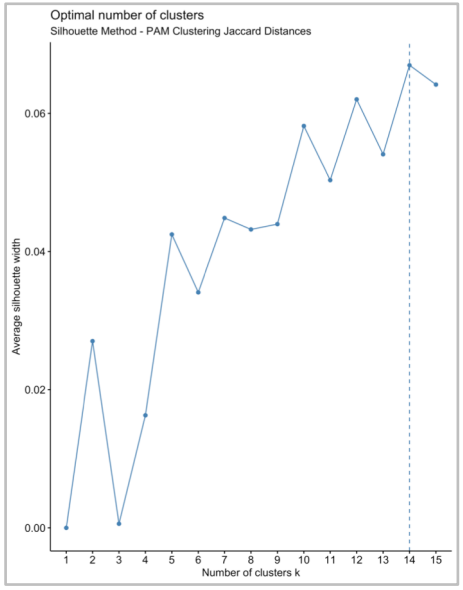
1. **Elbow Method:** Looks at total Within Sum of Squares (WSS) as a function of the clusters.
   a) Choose a number of clusters K so that adding another cluster does not improve WSS much
2. **Silhouette Method:** Measures the quality of clustering by determining how well each object lies within a cluster (measures how close each point in one cluster is to points in neighboring clusters)
   a) Choose maximum average silhouette over range of possible K

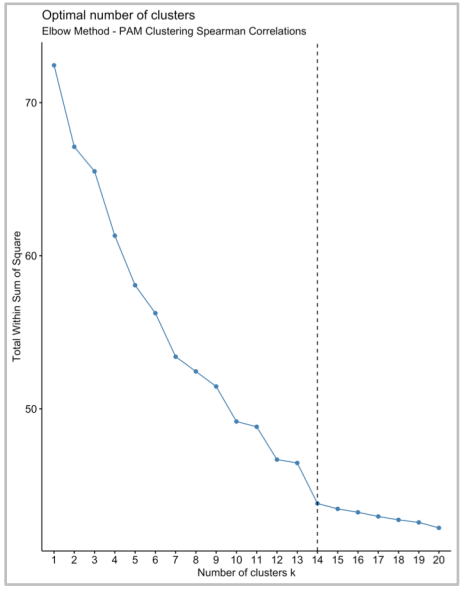
Finding the Optimal K Clusters – Partitioned Clustering Results

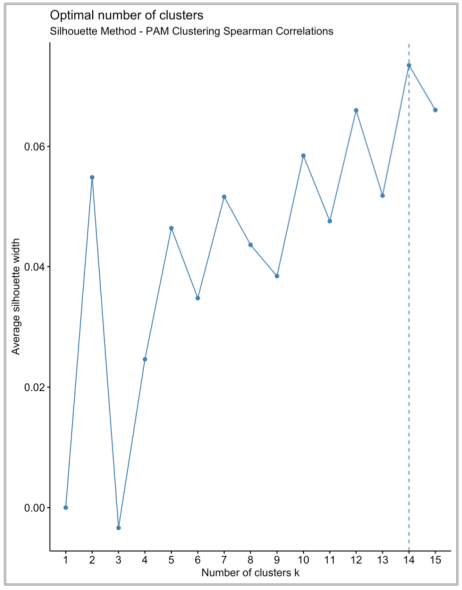
- Jaccard Distances

Spearman Correlation Distances









Partitioned Clustering

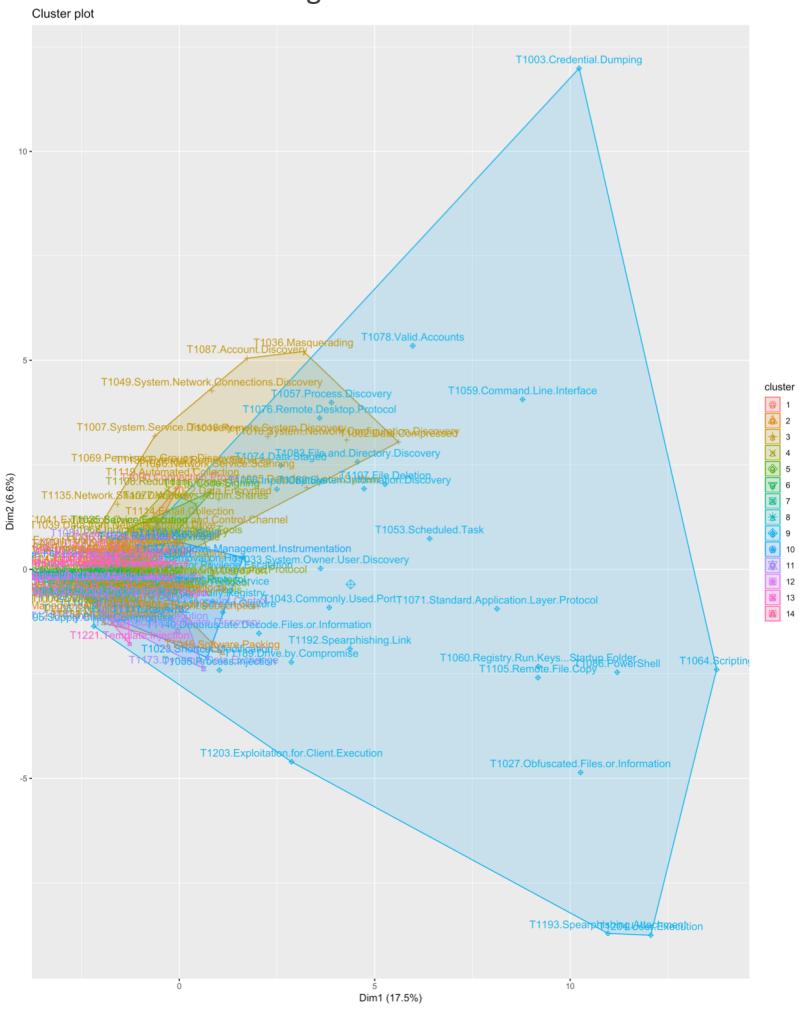
- Defining clusters by constantly iterating and updating a centroid that minimizes total intra-cluster (within-cluster) variation
- **K-means**: The centroid is defined to be the mean of all data points in a cluster

$$W(C_k) = \sum_{x_i \in C_k} (x_i - \mu_k)^2$$

  - Euclidean distances
  - Not appropriate for binary data, creates too many ties and leads to arbitrary decisions in clustering
- **K-medoids (PAM)**: The centroid is defined to be a data point in a cluster
  - Any distance metric can be used

PAM Clustering

PAM Clustering Jaccard Distances



PAM Clustering Spearman Correlation Distances

Cluster Validation

•Methods for Cluster Validation:

1.Silhouette Coefficient: Estimates the average distance between clusters, measures how close each point in one cluster is to points in neighboring clusters

a. Large S_i (close to 1) = Well clustered
b. Small S_i (close to 0) = Observation lies between 2 clusters
c. Negative S_i = Observation in wrong cluster

2.Dunn Index: Measures if the dataset contains compact and well-separated clusters, the diameter of the clusters is expected to be small and the distance between the clusters is expected to be large

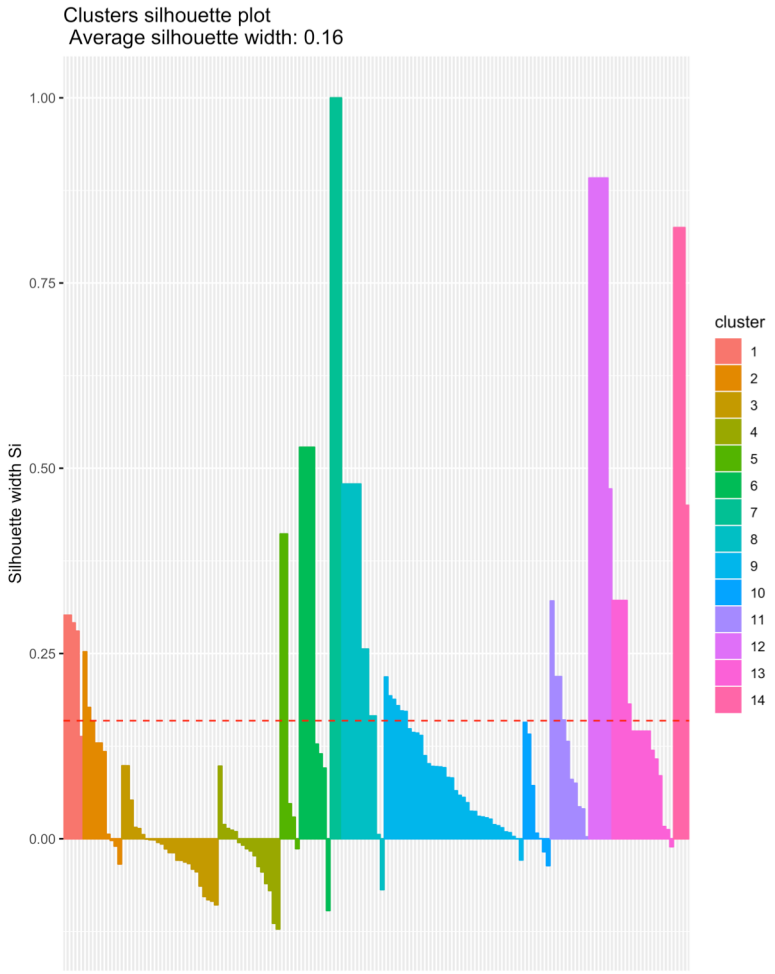
3.Average Distance Within Cluster

4.Average Distance Between Cluster

5.Entropy of the distribution of cluster memberships

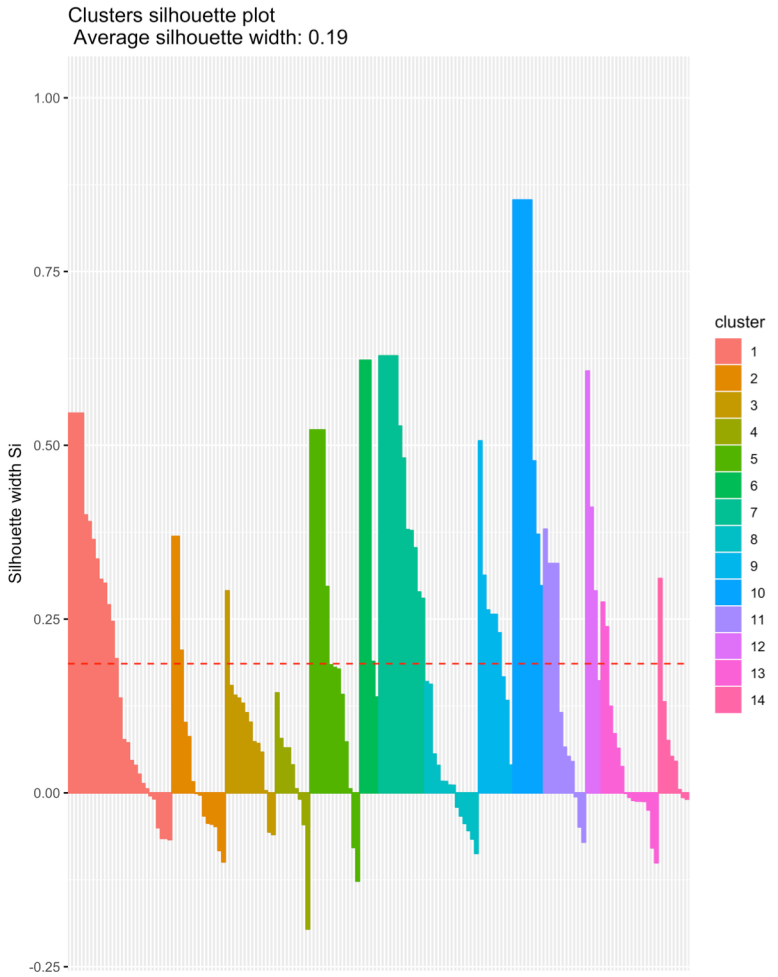
Cluster Validation – PAM Jaccard Distances

- Validation Statistics:
  - Average Silhouette Width: 0.16
  - Dunn Index: 0.5
  - Average Distance Within Cluster: 0.78
  - Average Distance Between Cluster: 0.95
  - Entropy: 2.39



Cluster Validation – PAM Spearman Correlation Distances

- Validation Statistics:
  - Average Silhouette Width: 0.19
  - Dunn Index: 0.25
  - Average Distance Within Cluster: 0.66
  - Average Distance Between Cluster: 0.95
  - Entropy: 2.5



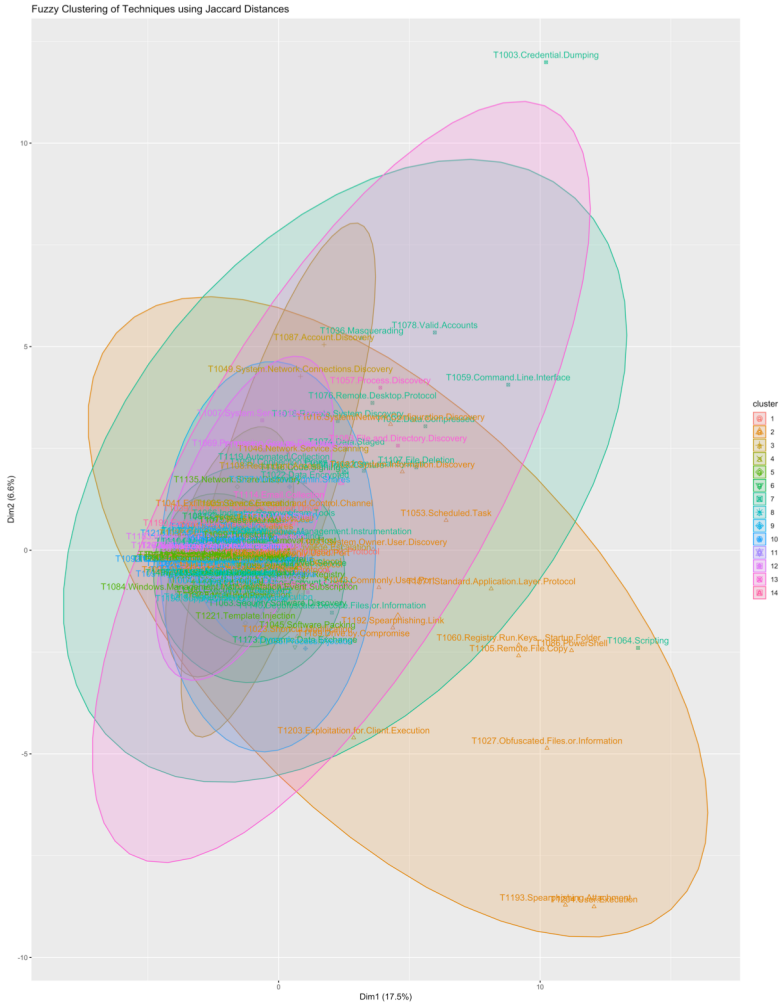
Fuzzy Analysis Clustering

- Soft clustering, in which each element has a probability of belonging to each cluster.
- Each element has a set of membership coefficients corresponding to the degree of being in a given cluster.
- Specifying a degree of fuzziness, you can specify the extent of the memberships and distances of points in a cluster

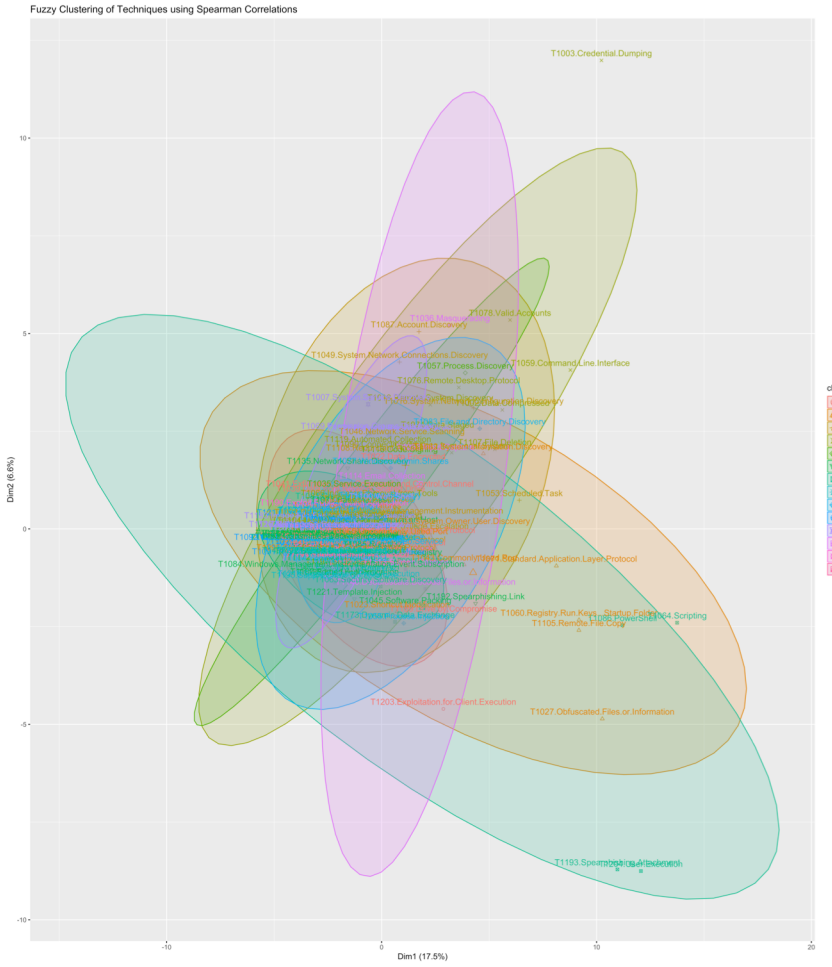
$$\sum_{v=1}^{k} \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} u_{iv}^{r} u_{jv}^{r} d(i,j)}{2 \sum_{j=1}^{n} u_{jv}^{r}}$$

Fuzzy Analysis Clustering

Fuzzy Clustering Jaccard Distances



Fuzzy Clustering Spearman Correlation Distances



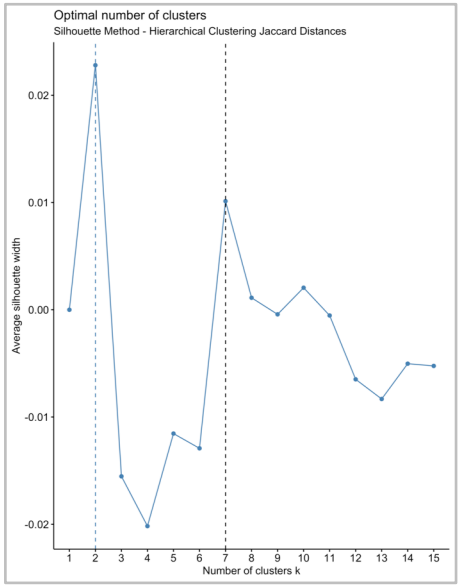
Methods Outline

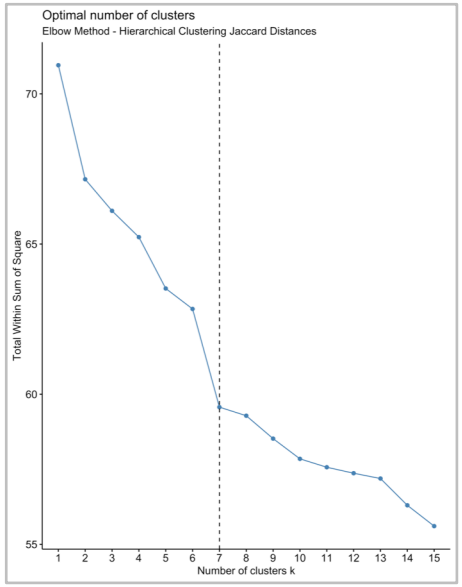
- Data set Collection
- Clustering
  - Distance Metrics
  - Measuring the clusterability
    - Hopkins Statistic + Visual Assessment of cluster Tendency
  - Partitioned Clustering
    - Finding the optimal K clusters
    - K means clustering
    - PAM clustering
    - Fuzzy Analysis clustering
    - Cluster Validation
  - Hierarchical Clustering
    - Finding the optimal K clusters
    - Agglomerative clustering
    - Divisive clustering
    - Cluster Validation
- Statistical Validation
- Heuristics
  - Mapping tactics and techniques to Kill Chain phases
- Sequential Pattern Mining
  - Extracting temporal technique rules with corresponding confidence

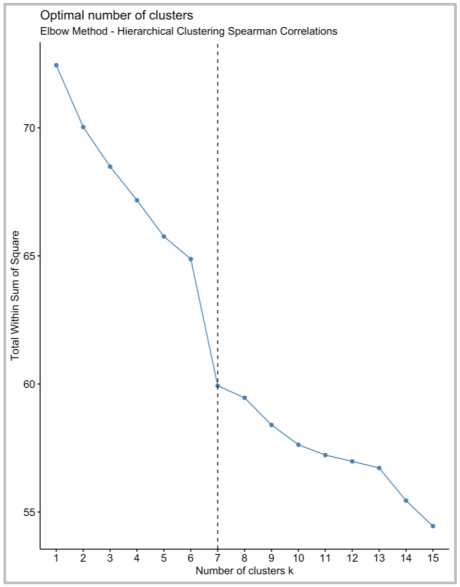
Finding the Optimal K Clusters – Hierarchical Clustering

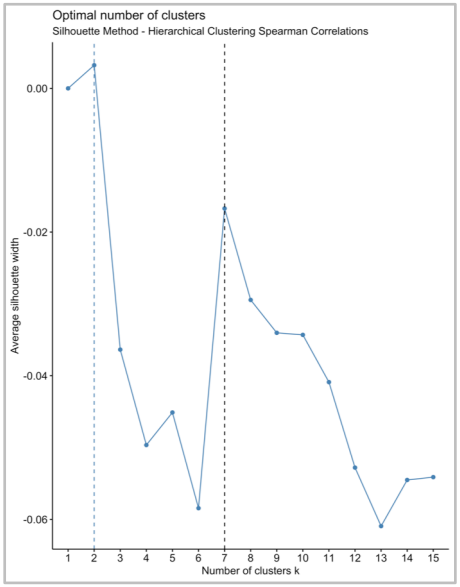
- Jaccard Distances

Spearman Correlation Distances





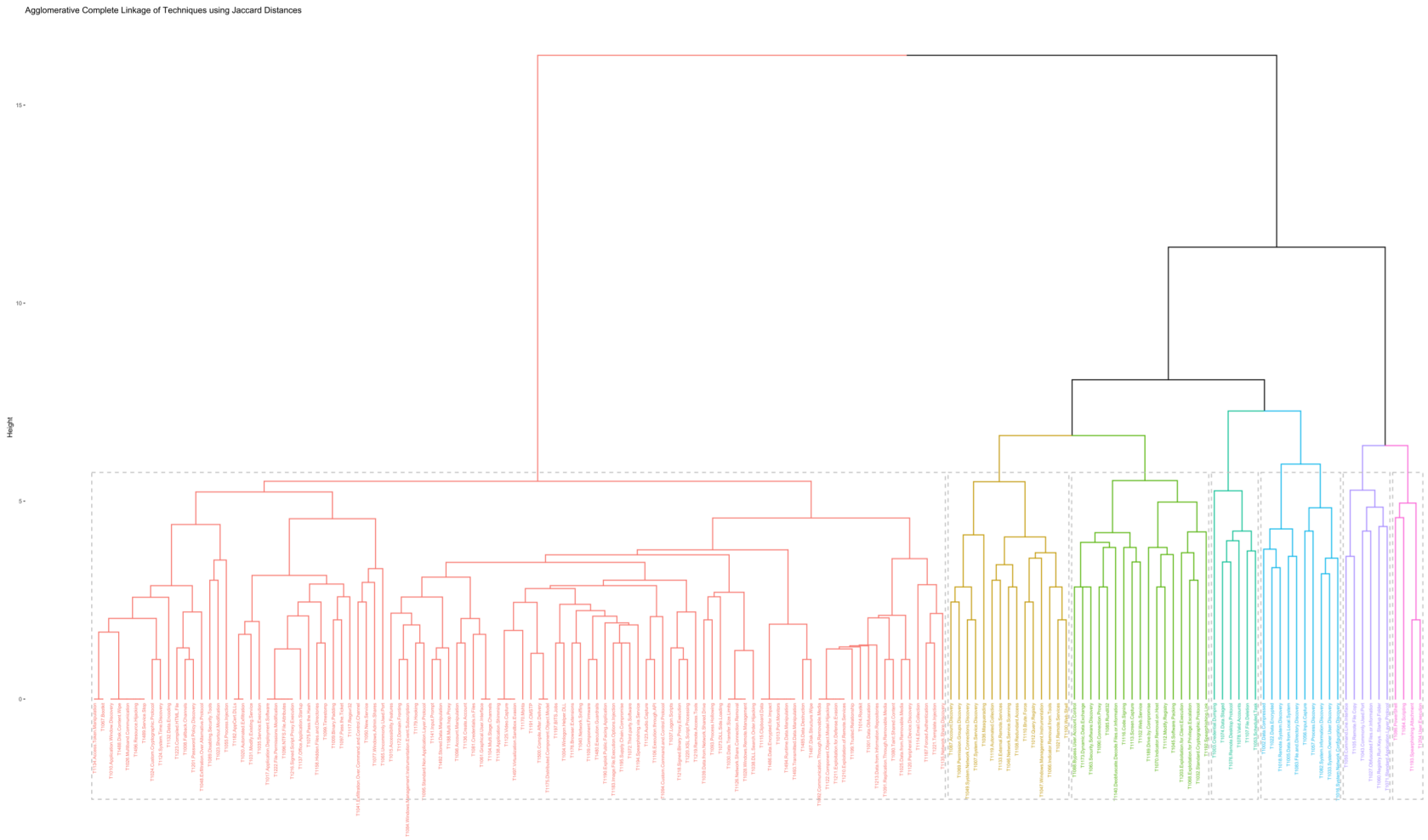




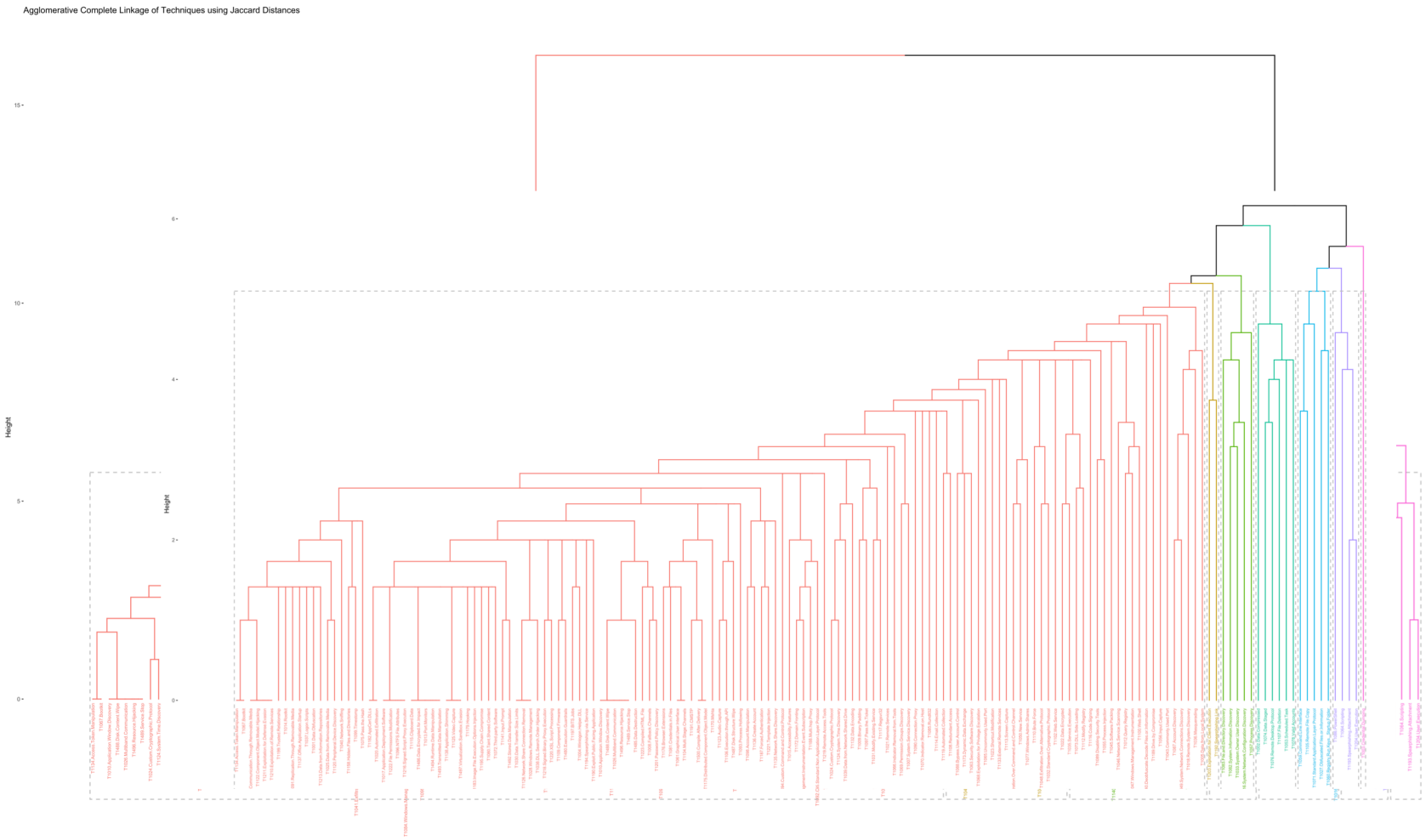
Hierarchical Clustering

- Uses a pairwise distance matrix to find clusters in your data set
- Resulting output is a tree-like representation called a Dendrogram
- Methods
  1. **Agglomerative Hierarchical Clustering**: Bottom up, each object is a single leaf and then clusters with most similarity are combined until all are combined to the root
  2. **Divisive Hierarchical Clustering**: Top down, begins with the root and clusters with most dissimilarity are separated until all objects are leaves

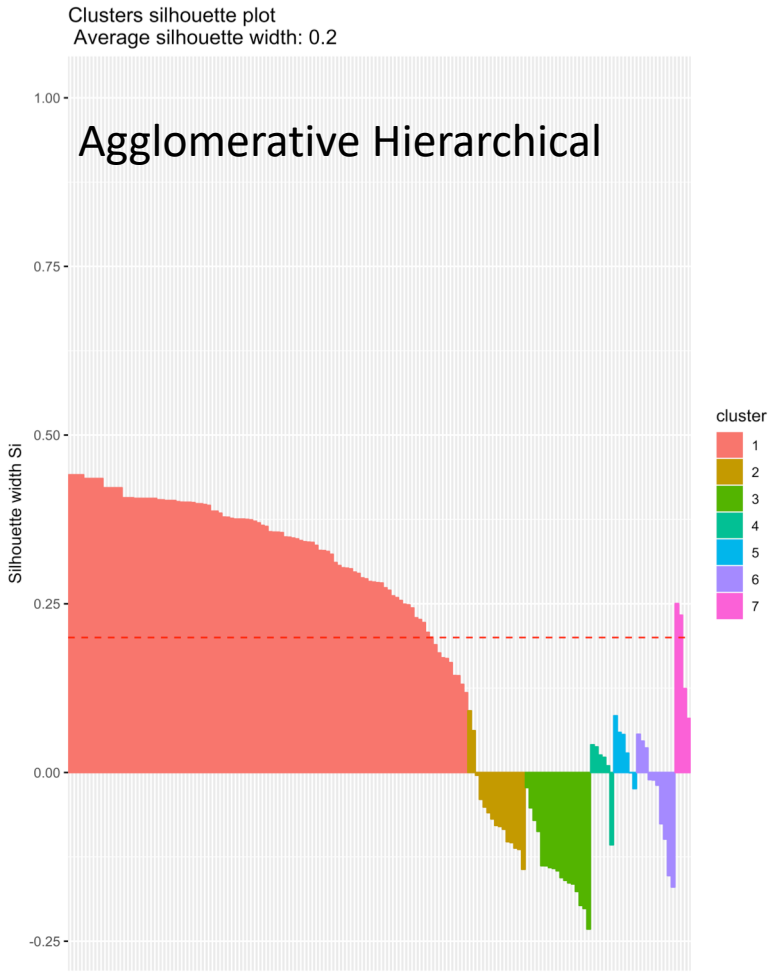
Agglomerative Hierarchical Clustering – Jaccard Distances

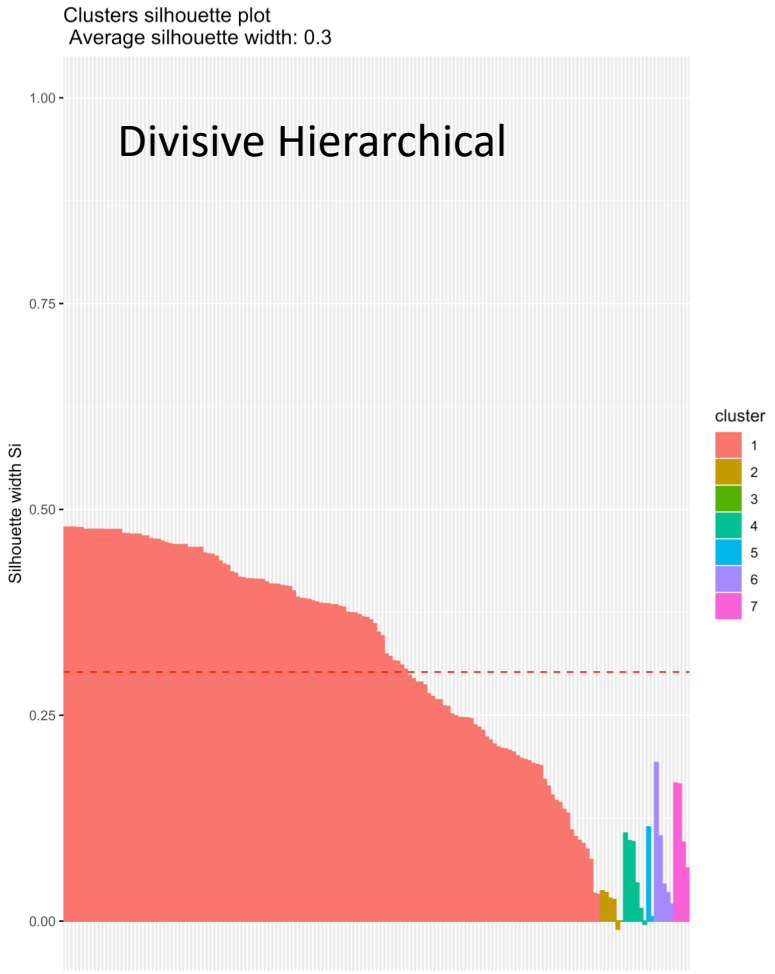


Divisive Hierarchical Clustering – Jaccard Distances

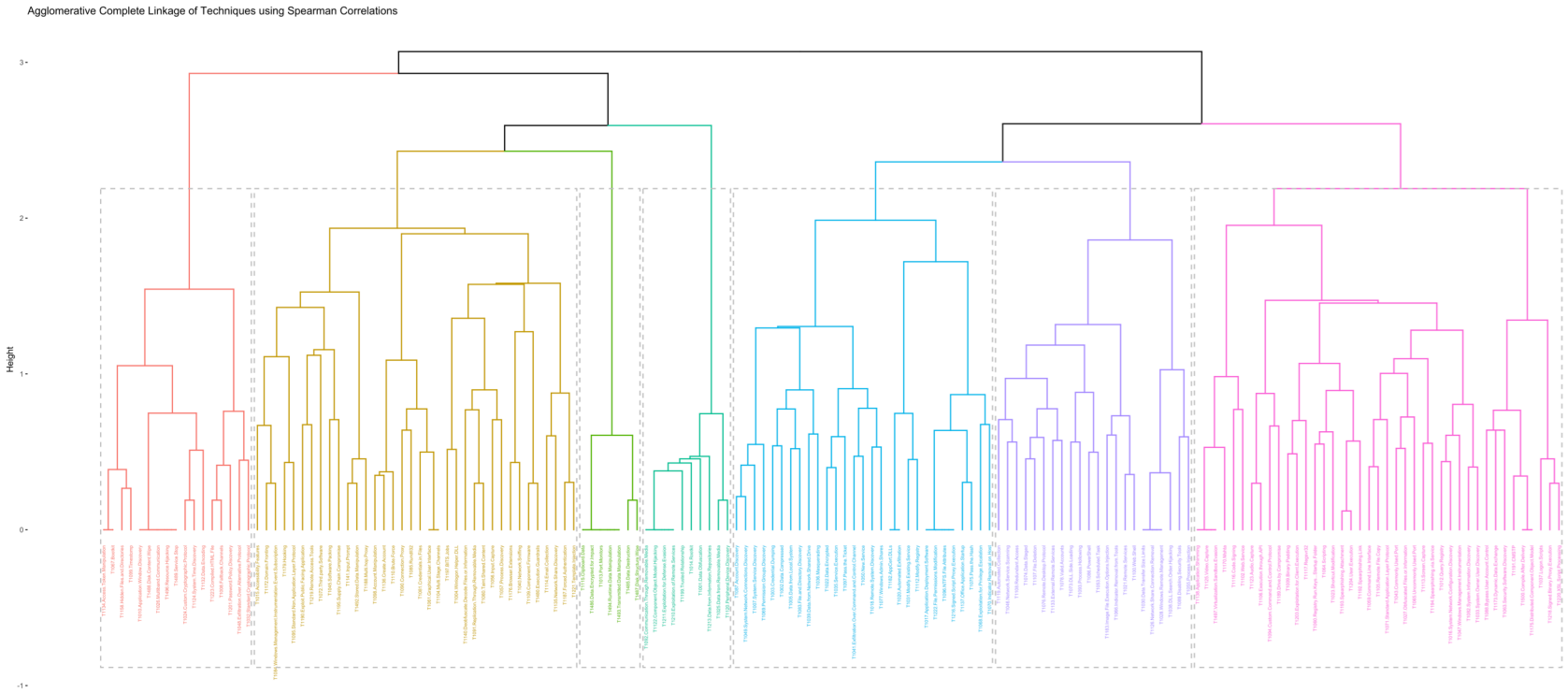


Cluster Validation – Hierarchical Clustering Jaccard Distances

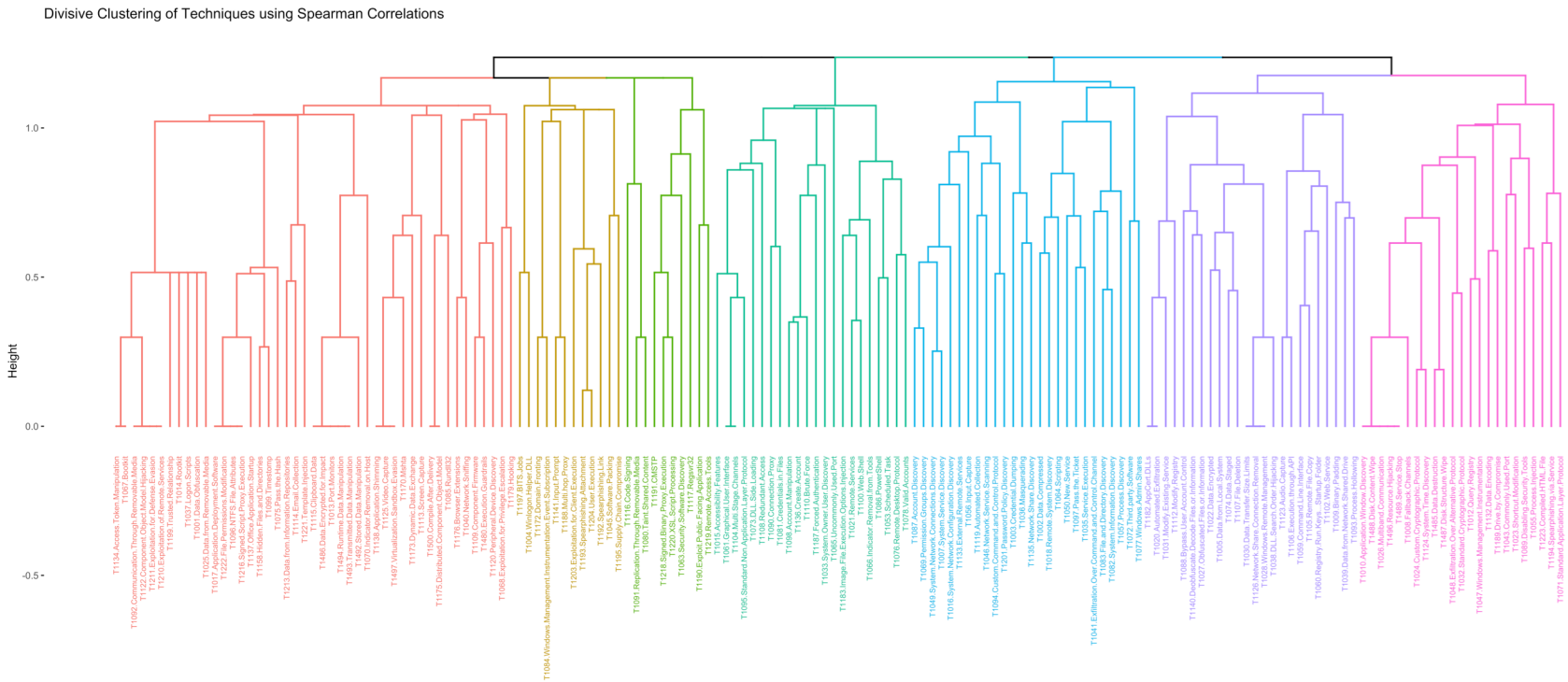




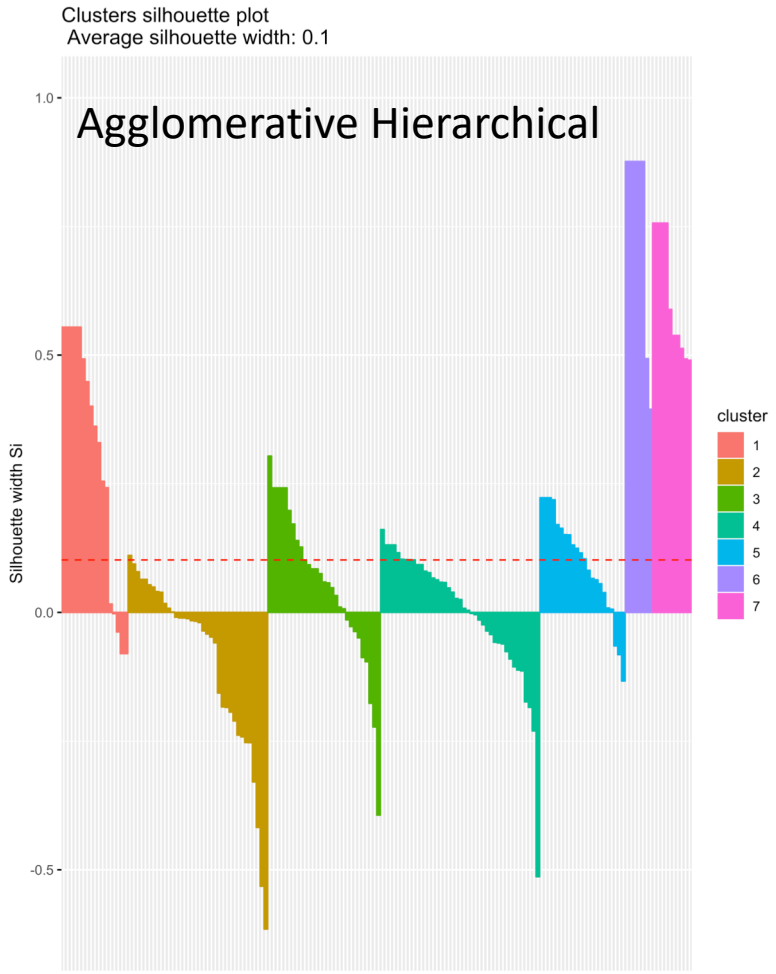
Agglomerative Hierarchical Clustering – Spearman Correlation Distances

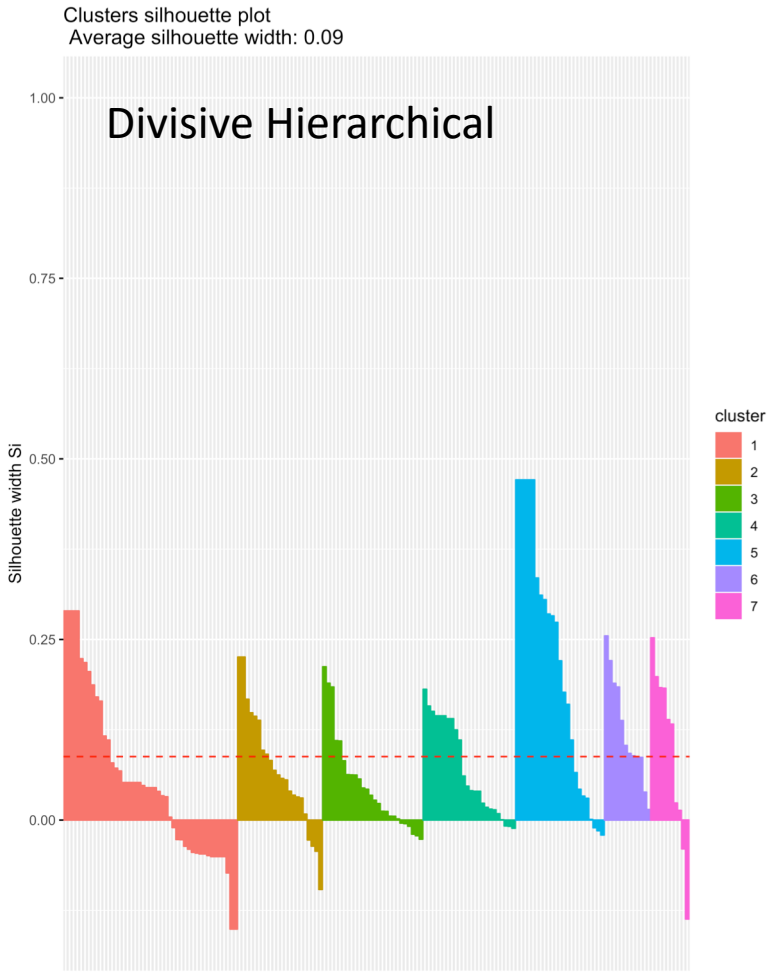


Agglomerative Hierarchical Clustering – Spearman Correlation Distances



Cluster Validation – Hierarchical Clustering Spearman Correlation Distances





Clustering Discussion

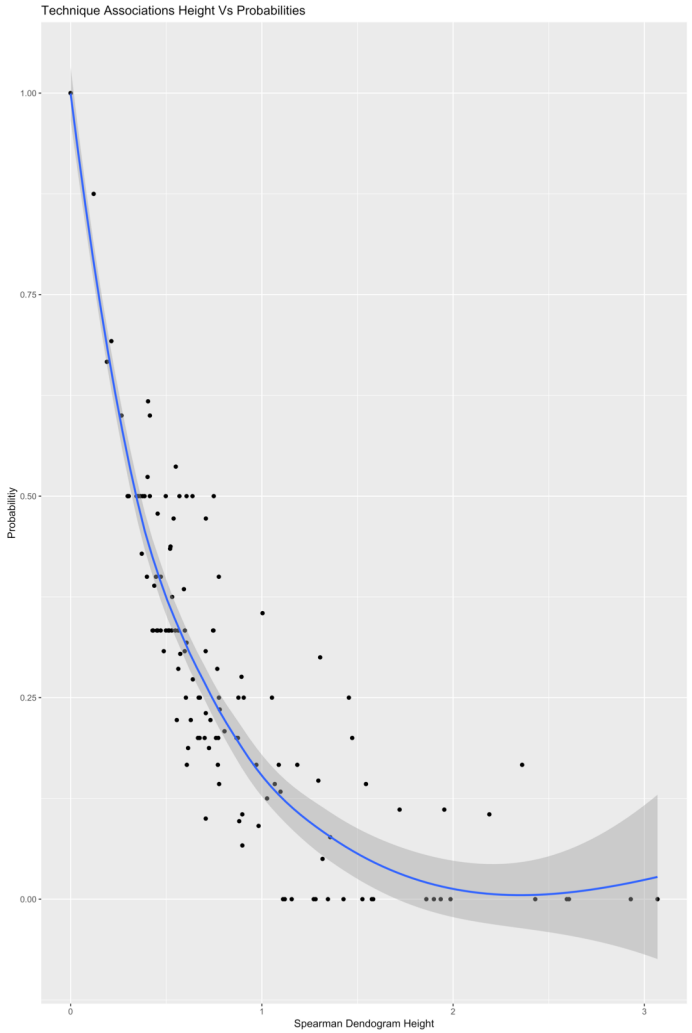
- We determined that **Spearman Correlation Agglomerative Hierarchical Clustering** is the most appropriate type of clustering for our data
- Using Spearman Correlation Distances maintained the interpretability of the results (techniques associated have co-occurred) and yielded better cluster tendency and validation
- From the tree, we are able to subjectively determine technique associations
- How to validate these technique associations and extract the most statistically significant?

Methods Outline

- Data set Collection
- Clustering
  - Distance Metrics
  - Measuring the clusterability
    - Hopkins Statistic + Visual Assessment of cluster Tendency
  - Partitioned Clustering
    - Finding the optimal K clusters
    - K means clustering
    - PAM clustering
    - Fuzzy Analysis clustering
    - Cluster Validation
  - Hierarchical Clustering
    - Finding the optimal K clusters
    - Agglomerative clustering
    - Divisive clustering
    - Cluster Validation
- Statistical Validation
- Heuristics
  - Mapping tactics and techniques to Kill Chain phases
- Sequential Pattern Mining
  - Extracting temporal technique rules with corresponding confidence

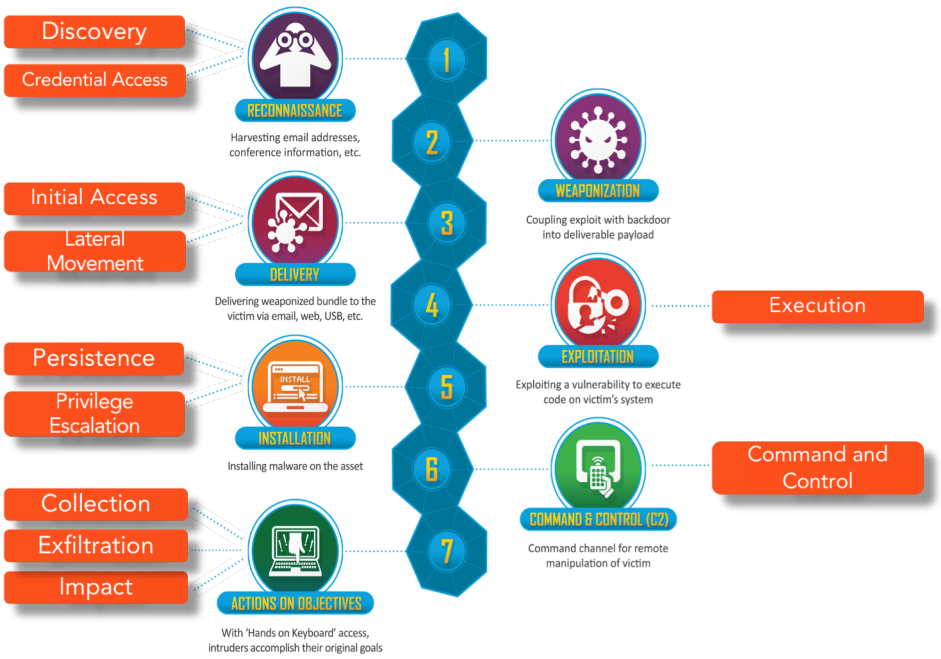
Statistical Validation

- We validated **Spearman Correlation Agglomerative Hierarchical Clustering**
- Used probabilities to extract technique associations
- We determined that as Spearman heights increase, probabilities of technique associations decrease
- Using the cutoff of a probability of 0.6 and Spearman height of at most 0.4, we extracted **21** technique associations of length 2+



Heuristics

- We mapped MITRE tactics and techniques using pre-condition analysis
- This can be used to temporally order technique associations
- We performed sequential pattern mining to also extract technique rules



Sequential Pattern Mining

- After temporally ordering observed attacks, sequential pattern mining created technique rules based on which techniques often showed a temporal order
- Confidence: likelihood that the sequential rule A → B actually occurs among transactions containing item set A, under the constraint that item set A is before B. High confidence implies a high likelihood that B occurs in a future sequence
- Extracted **19** technique rules with confidence of 0.5 or higher

Sequential Pattern Mining – Technique Rules

Technique Sequence Rule	Confidence
T1043.Commonly.Used.Port, T1105.Remote.File.Copy, T1071.Standard.Application.Layer.Protocol => T1078.Valid.Accounts	1
T1192.Spearphishing.Link => T1078.Valid.Accounts	1
T1071.Standard.Application.Layer.Protocol => T1064.Scripting	1
T1189.Drive.by.Compromise, T1193.Spearphishing.Attachment => T1064.Scripting	0.75
T1107.File.Deletion, T1027.Obfuscated.Files.or.Information => T1064.Scripting	0.667
T1003.Credential.Dumping => T1074.Data.Staged	0.6
T1059.Command.Line.Interface => T1036.Masquerading	0.6
T1193.Spearphishing.Attachment, T1192.Spearphishing.Link => T1204.User.Execution	0.5