Hierarchical Incident Clustering for Security Operation Centers

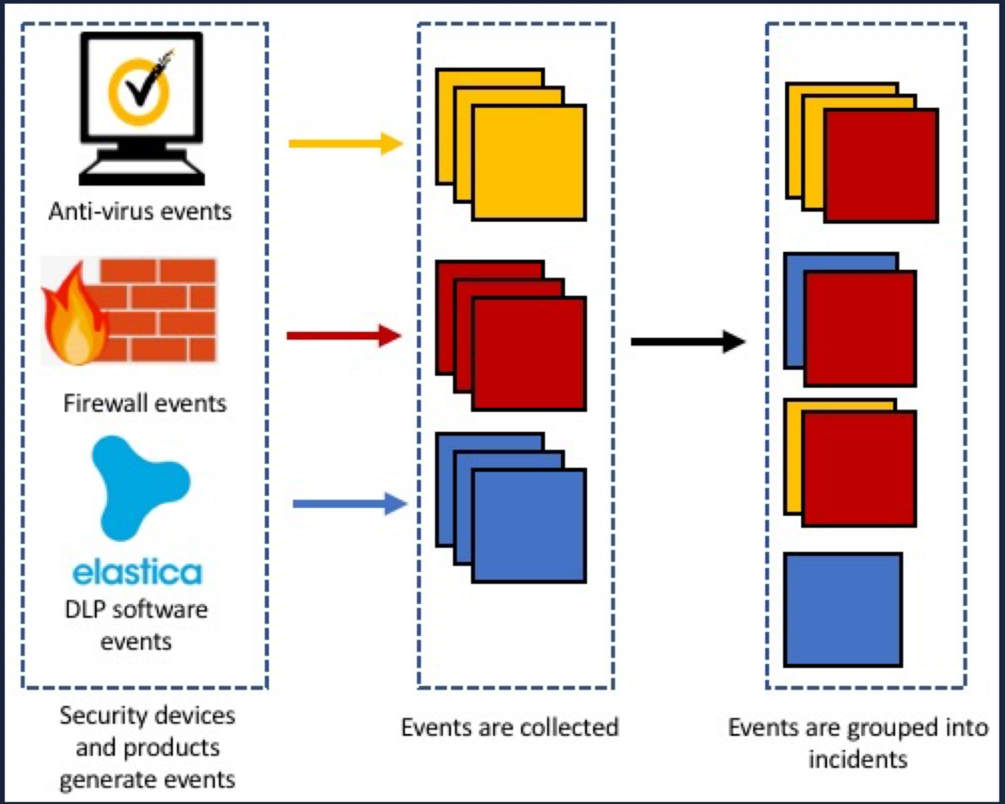
David Silva, Matteo Dell'Amico, Mike Hart, Kevin Roundy & Daniel Kats

Symantec Research Labs

Clustering Security Incidents

Symantec™

Security Operation Centers & Incidents



Security Operation Centers & Incidents

- **Tedious & error prone job** for analysts!
  - Long lists of similar incidents
- Sundaramurthy et al., SOUPS 2016:
  - “I am not learning anything new in my current job […] I feel that the SOC is not doing any real threat detection”
  - “The procedures were turning us into robots […] all the analysts were doing was to click and fill in data”
- Our mission: **cluster that data**!
  - Analysts can pick up a group of similar incidents and **act on all them at once**
  - **Free up analysts’ time** so they can focus on the important and rewarding tasks

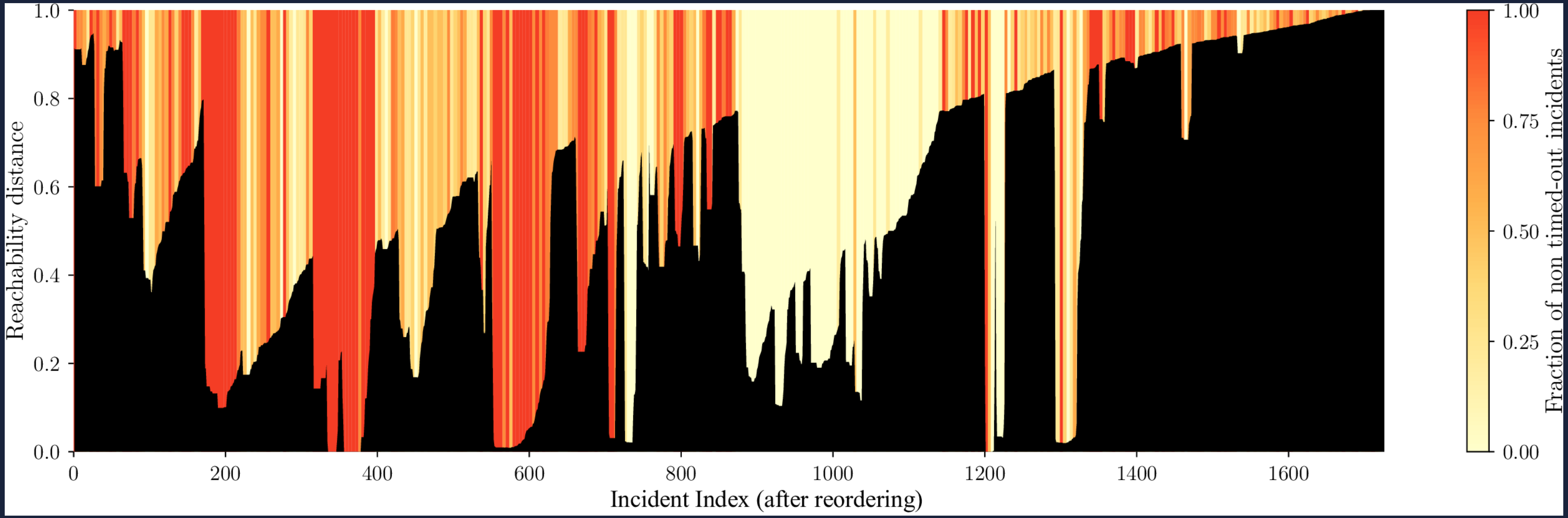
Our Design

1. Based on an **arbitrary distance function**
   - **Easy to update** when we get better ideas
2. Does not force **isolated incidents in an unrelated cluster**
   - Based on **density-based clustering** solutions that have this property
3. It is **hierarchical**
   - Allows the analysts to navigate between **clusters within clusters**
4. It is **scalable**

Our Solution: Distance Function

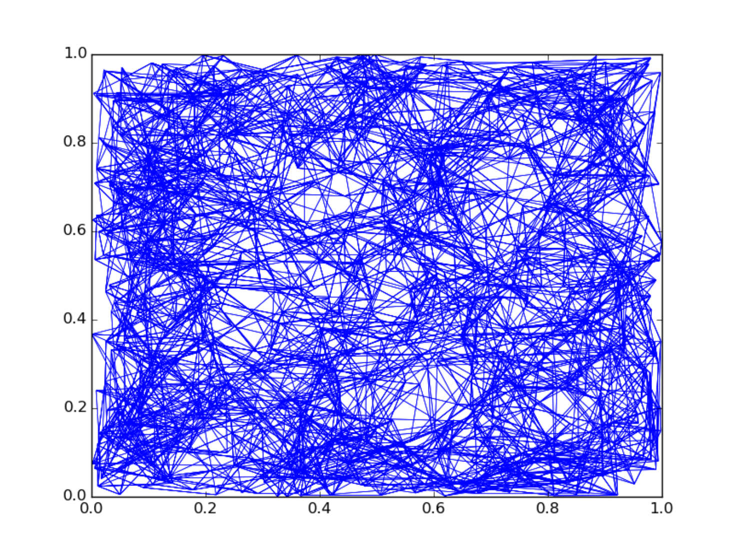
- We consider incidents as **bags** (i.e., multisets) **of events**
- **TF.IDF** (term frequency-inverted document frequency) normalization to discount the importance of very common events (e.g., failed login due to a wrong password)
- **Generalized Jaccard Distance** on the resulting multiset

Our Solution: Clustering Algorithm



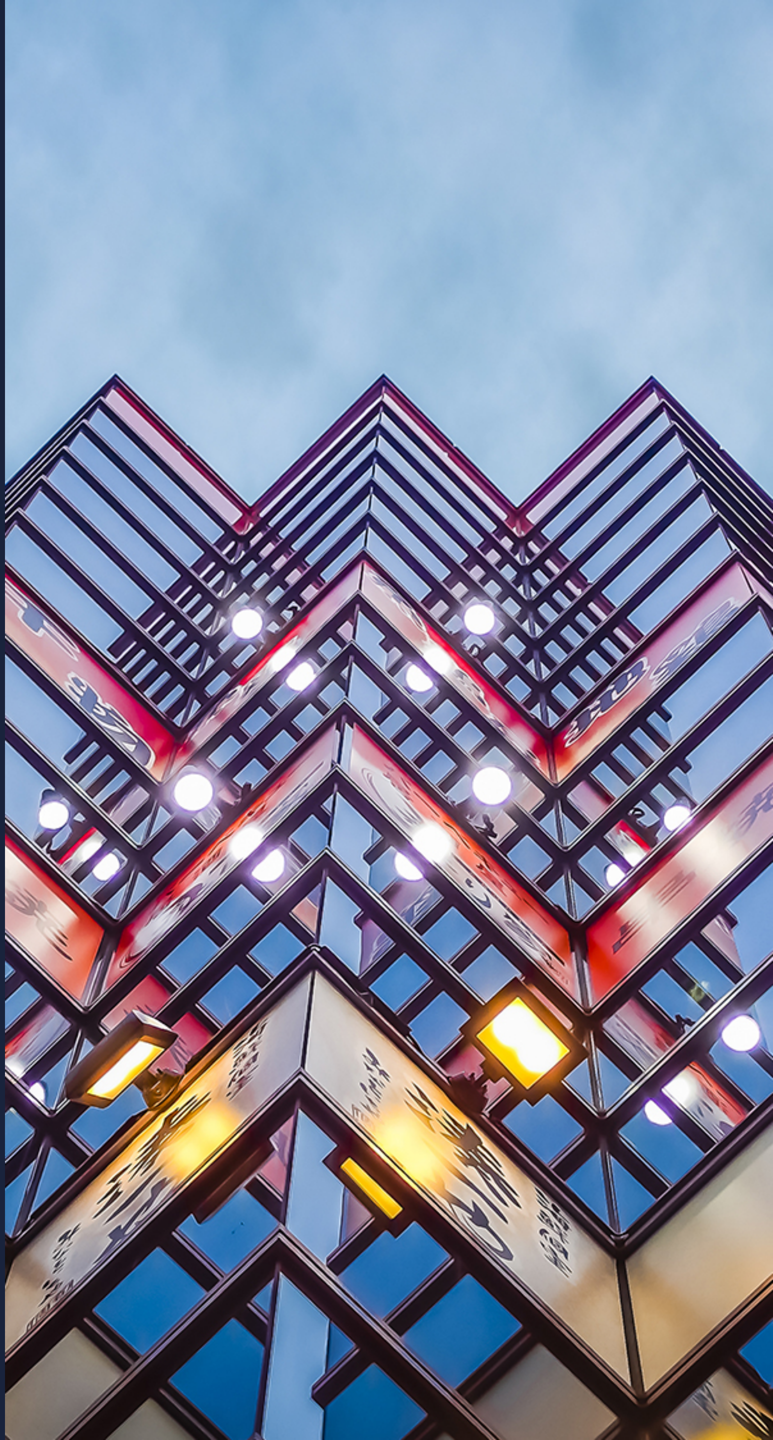
- Based on OPTICS (Ordering Points to Identify the Cluster Structure) (SIGMOD ’99)
  - Points are ordered, putting close ones nearby
  - In this graph, valleys are clusters
    - …and valleys within valleys are hierarchical clusters

Our Solution: Scalability



- For a generic distance function, OPTICS would need **$O(n^2)$ calls to the distance function** (compare all against all) to cluster n objects
  - Obviously wouldn’t scale with large datasets
- Solution based on **NN-DESCENT** (WWW’11)
  - An approximate algorithm to discover similar items for arbitrary distance functions
  - Neighbors found by NN-DESCENT are passed to OPTICS

The GUI



Symantec™

Symantec™

Thank You!