RANK-1 SIMILARITY MATRIX DECOMPOSITION FOR MODELING CHANGES IN ANTIVIRUS CONSENSUS THROUGH TIME

Robert J. Joyce
Booz Allen Hamilton
University of Maryland, Baltimore County
USA
joyce_robert2@bah.com

Edward Raff
Booz Allen Hamilton
University of Maryland, Baltimore County
USA
raff_edward@bah.com

Charles Nicholas
University of Maryland, Baltimore County
USA
nicholas@umbc.edu

ANTIVIRUS CORRELATION

- Some groups of AVs known to make correlated labeling decisions

- Conventional wisdom in industry has a few explanations:
  - Copying results of leading vendors
  - Different AV products using the same engine
  - Signature sharing

- All explanations involve “first order” interactions
  - Creates direct link between labeling decisions of two AVs

WHY DOES THIS MATTER?

- Aggregated results from a collection of AVs is much better than using a single AV
  - Detection – is it malicious or benign?
  - Classification – which family does it belong to?

- Correlations between AVs can influence voting and other aggregation approaches

RESEARCH QUESTIONS

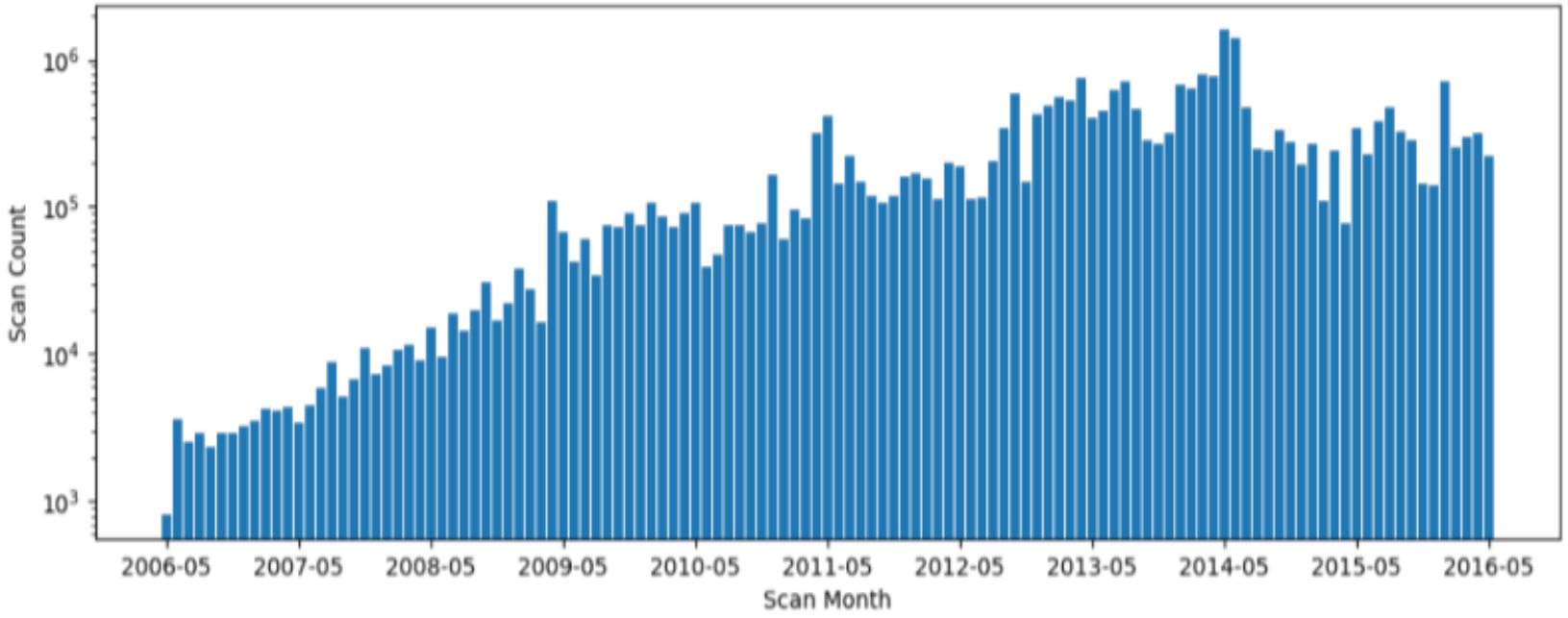
- Are existing assumptions about AV correlation correct?

- Is AV agreement predominantly due to first-order interactions?
  - When detecting files as malware?
  - When classifying malware by family?

- How do first-order interactions between AVs change over time?

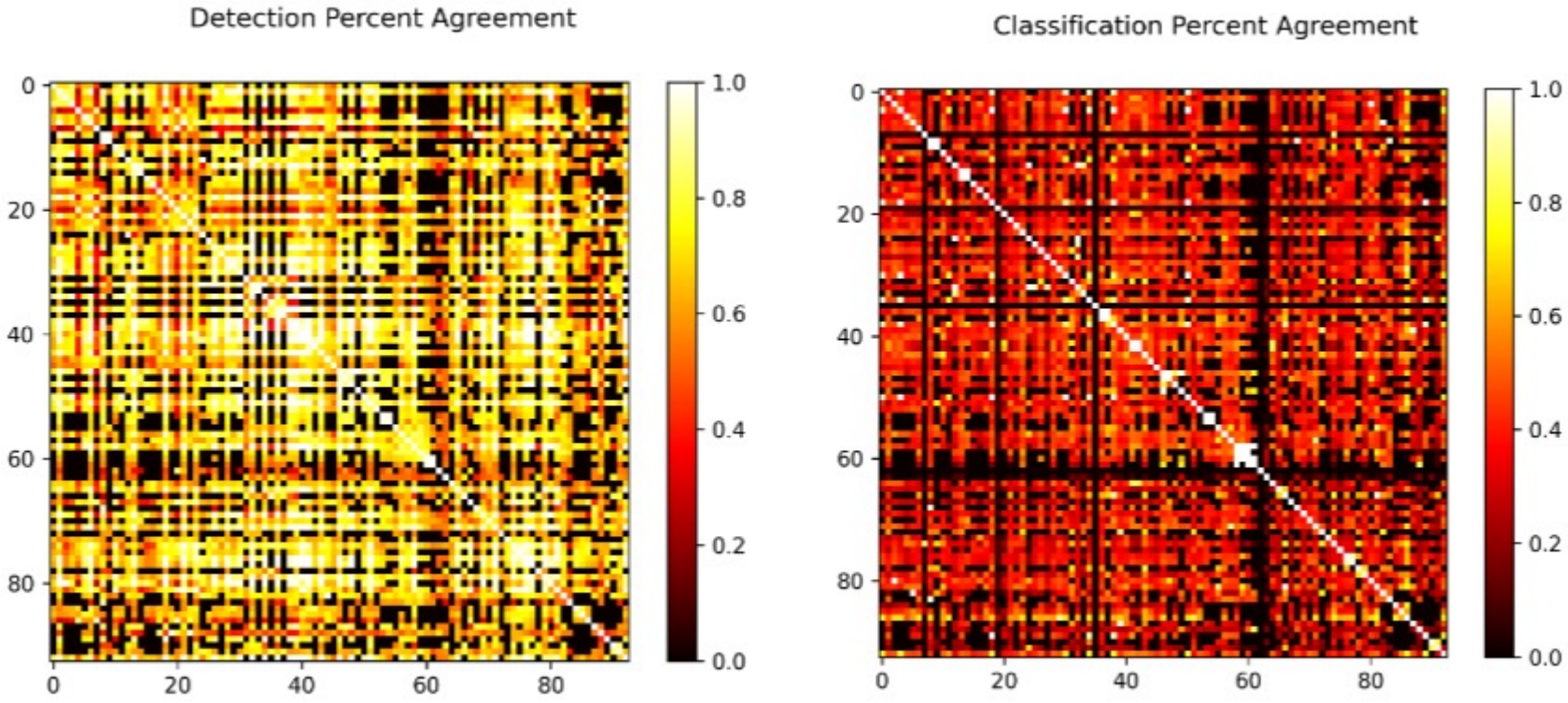
THE DATA

- Using 25,100,286 VirusTotal scan reports over 10-year period
- Malware samples are from the VirusShare dataset
- 93 distinct AVs



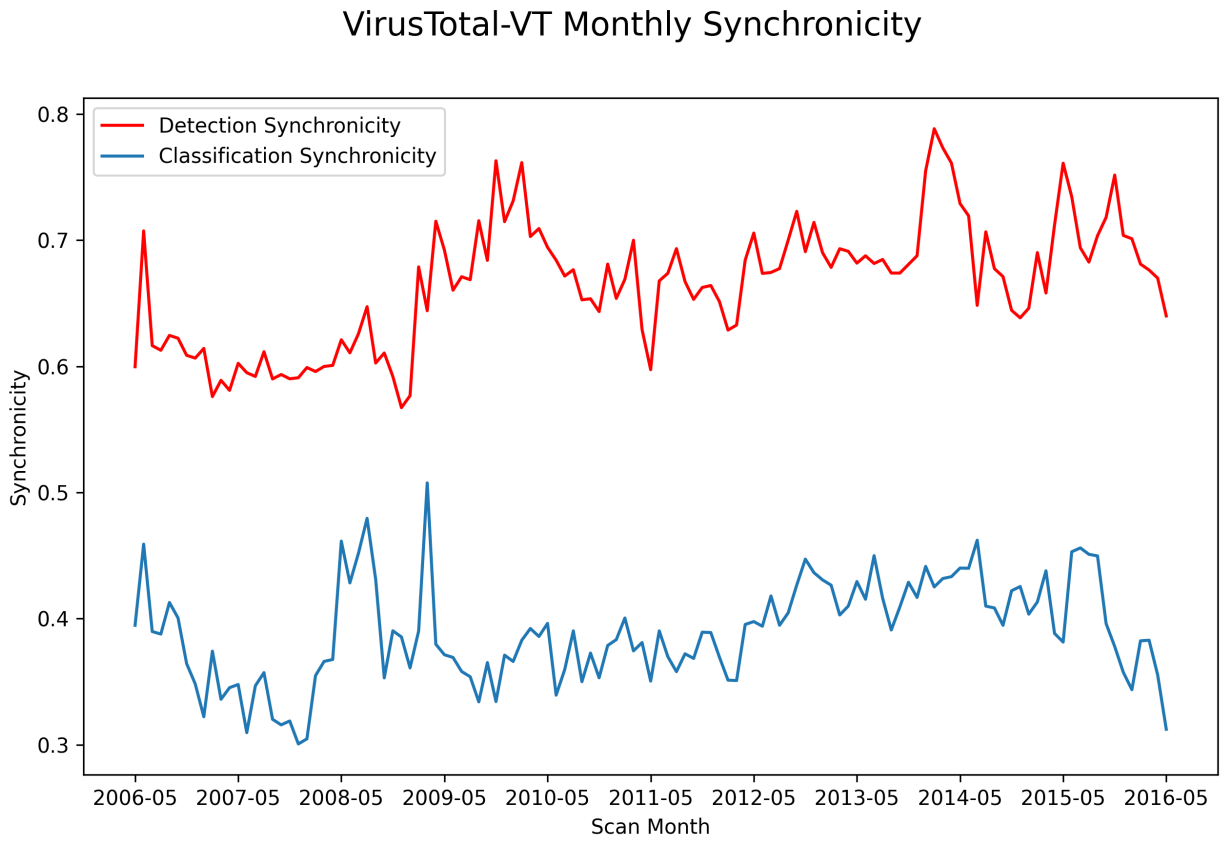
ANTIVIRUS AGREEMENT

- Before we look into first-order interactions, we need to measure how frequently AVs agree with each other



ANTIVIRUS SYNCHRONICITY

- Synchronicity is the average pairwise agreement between all AVs
- Extremely variable on a monthly timescale



THE R1SM DECOMPOSITION

- Rank-1 Similarity Matrix Decomposition

- $D = \sum_1^k triu(\boldsymbol{r}_i \boldsymbol{r}_i^T, 1)$

- Decomposes a similarity matrix D into a sum of k rank-1 outer products with shared, non-negative weights
  - Because the components $\boldsymbol{r}_1, \boldsymbol{r}_2 \ldots \boldsymbol{r}_k$ have rank 1, they manifest first-order interactions between objects

SOLVING THE R1SM DECOMPOSITION

- The R1SM decomposition typically has multiple solutions

- We developed a greedy algorithm for solving R1SM
  - At each iteration, identifies a component that explains as much of the remaining similarity matrix as possible
  - Stops when a component fails to explain a small percentage δ of the similarity matrix ($\delta = 0.1\%$ by default)

- More details in paper!

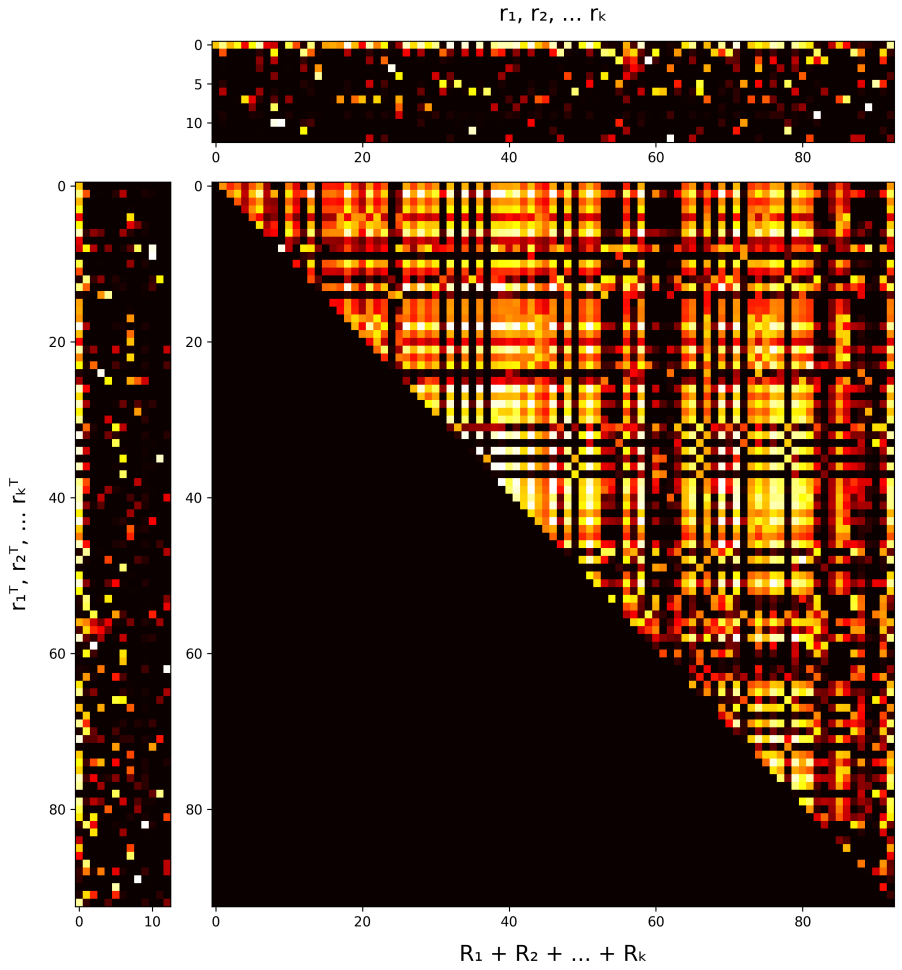
APPLYING R1SM TO ANTIVIRUS SCAN DATA

- Applied R1SM to the detection and classification percent agreement similarity matrices

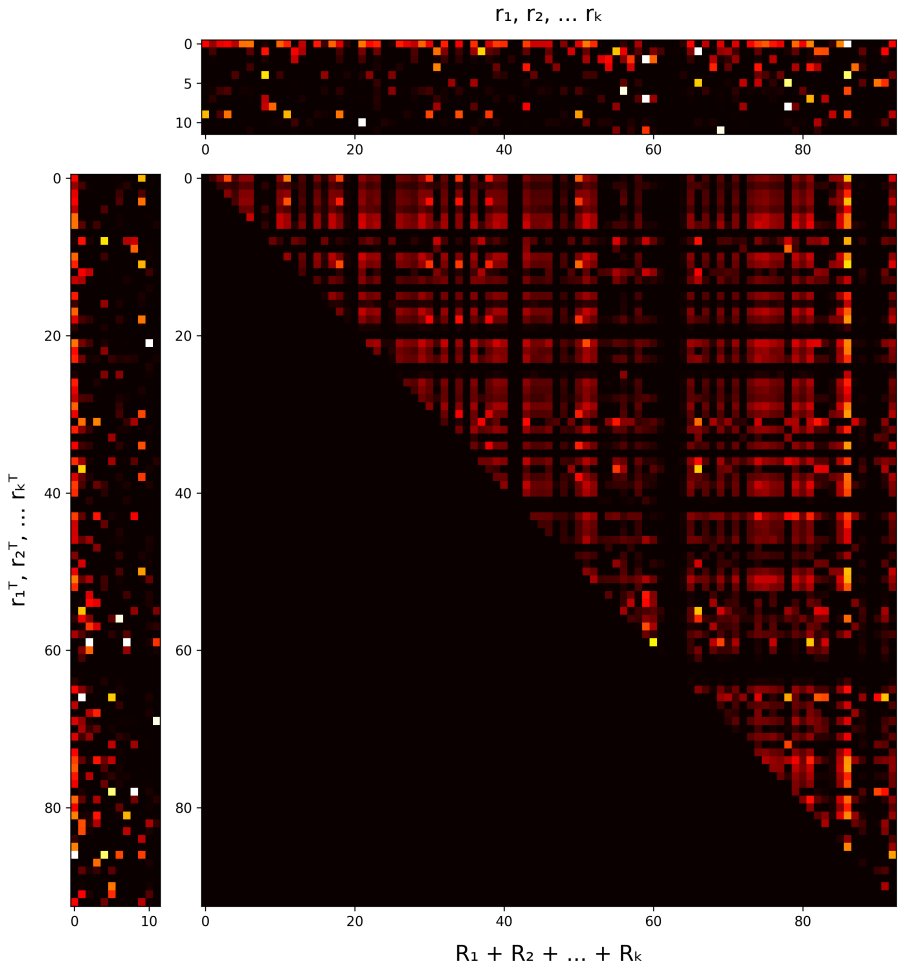
- Detection percent agreement:
  - $k = 16$ components that explain 60.596% of similarity matrix

- Classification percent agreement:
  - $k = 21$ components that explain 58.394% of similarity matrix

APPLYING R1SM TO ANTIVIRUS SCAN DATA



Detection Percent Agreement



Classification Percent Agreement

R1SM-T: EXTENDING R1SM TO TIME-SERIES DATA

- R1SM decomposition of a time-series of similarity matrices
- Shares information across all matrices as a function of their spatial relationships in time
- Implemented as a deep neural network over positional embeddings

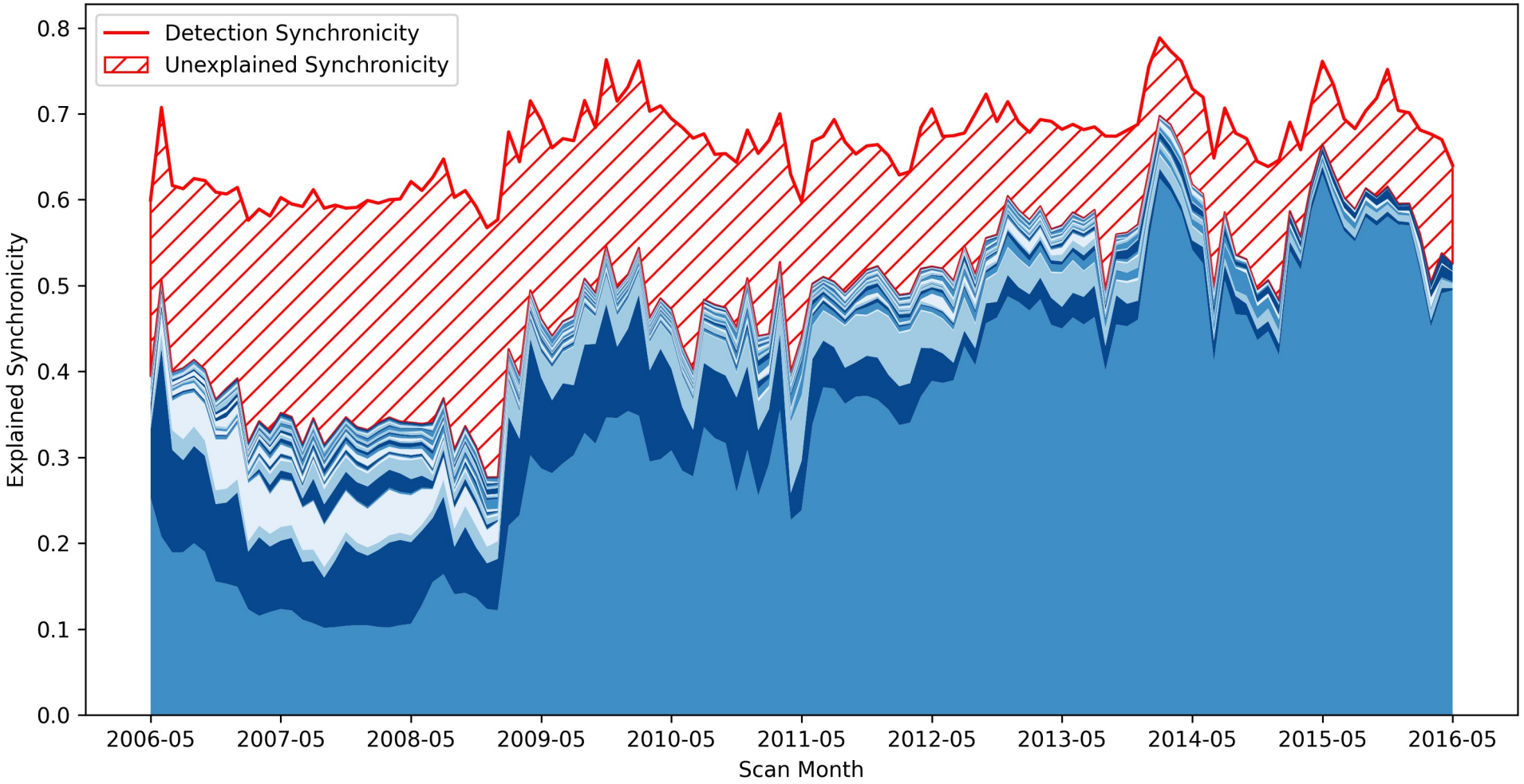
APPLYING R1SM-T TO ANTIVIRUS SCAN DATA

- Decomposed a time-series of similarity matrices representing monthly detection and classification agreement

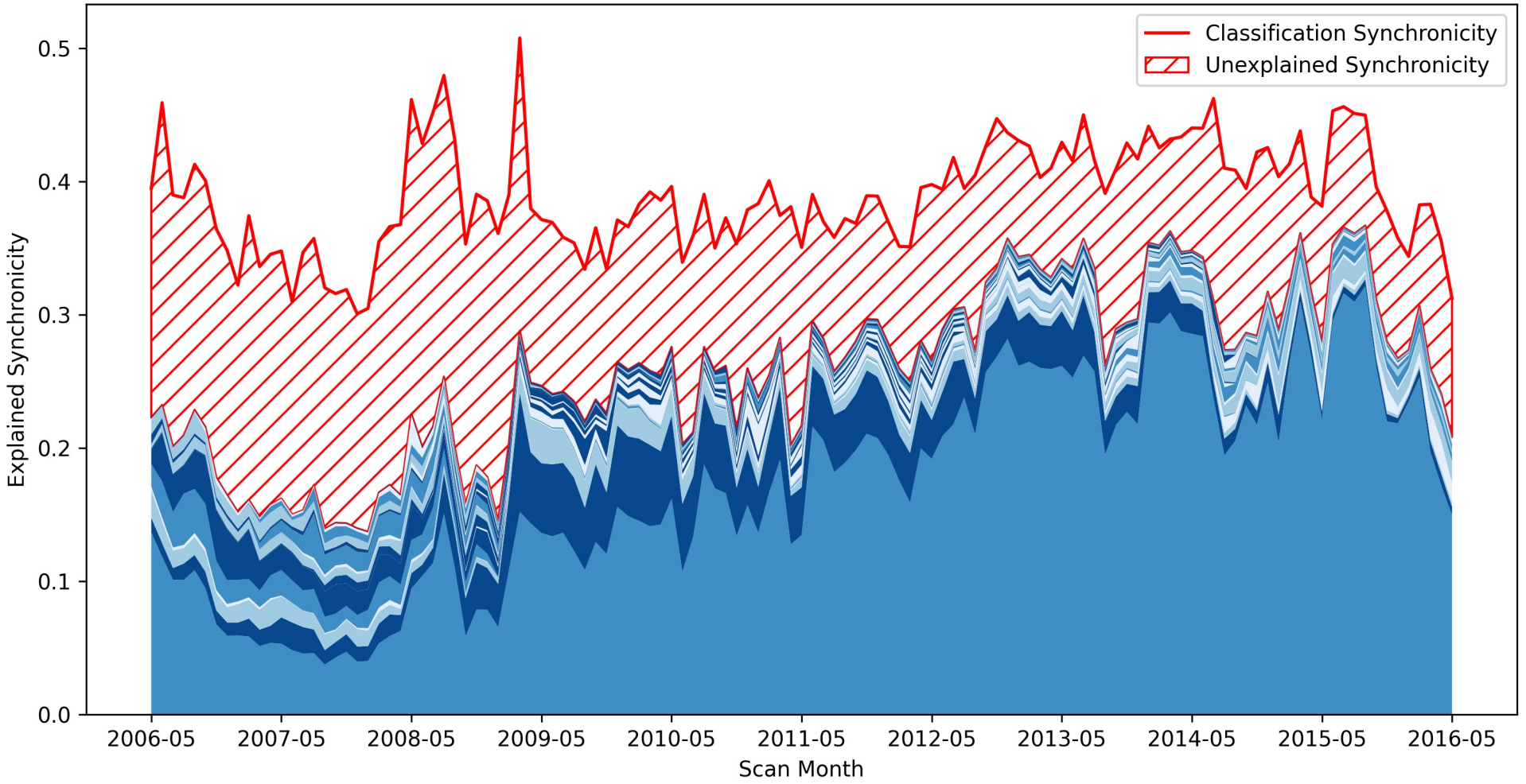
- Detection percent agreement:
  - $k = 26$ components that explain 73.709% of time-series

- Classification percent agreement:
  - $k = 26$ components that explain 67.196% of time-series

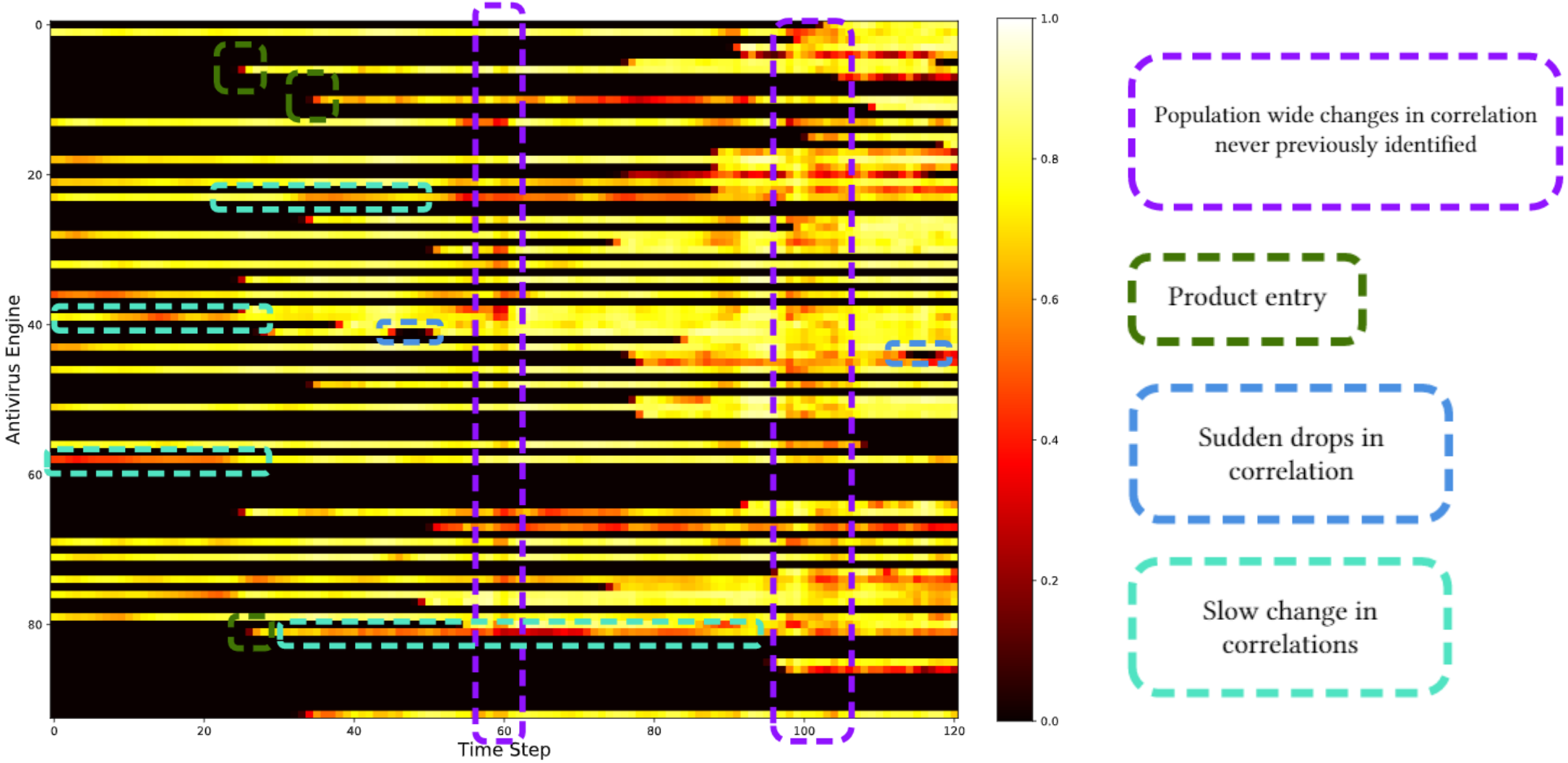
R1SM-T MONTHLY EXPLAINED DETECTION SYNCHRONICITY



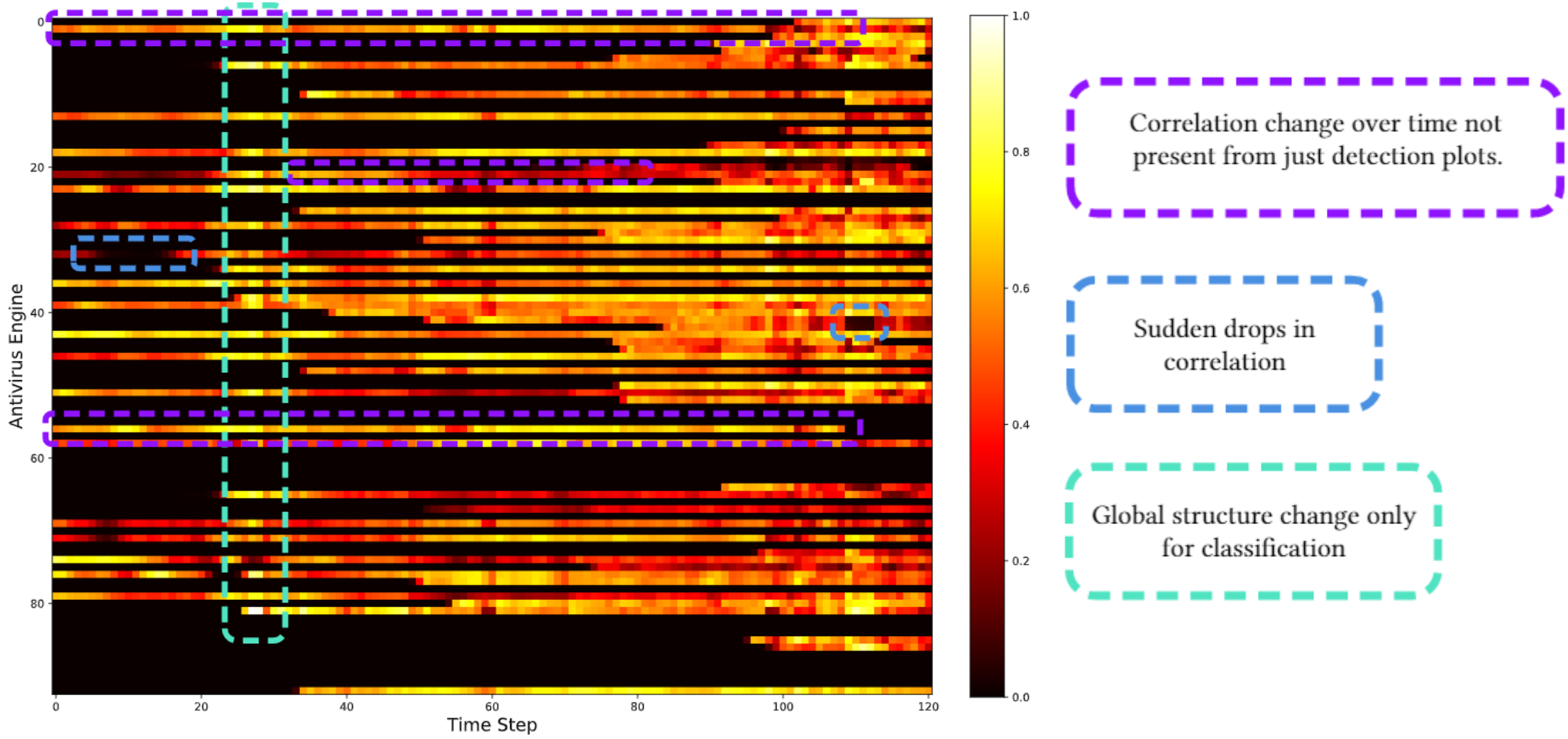
R1SM-T MONTHLY EXPLAINED CLASSIFICATION SYNCHRONICITY



R1SM-T DETECTION TIME-SERIES COMPONENT 1



R1SM-T CLASSIFICATION TIME-SERIES COMPONENT 1



CONCLUSIONS

- First-order interactions alone are not sufficient for modeling the complex interactions between AVs
  - We do not fully understand causes of AV correlation
  - Relationships between AVs more volatile than previously thought

- Future AV aggregation approaches should weight voters as both a function of correlation and time