Language Modeling for Anomalous Network Activity Detection

Elliott Skomski, Josh Loehr, Robin Cosbey, Brian Hutchinson

Computer Science Department, Western Washington University

Overview

Motivation: Network analysts need to identify potential security incidents. Large computer networks make manual inspection intractable. Traditional automated methods rely on costly feature aggregation and don't provide insight into why events are flagged.

Goal: Achieve highly accurate, interpretable anomaly detection with minimal feature processing using deep learning and natural language processing techniques.

Background

- **Aggregate Features:** user activity counted or averaged over user-days.
- One 108-dimensional aggregate feature vector per user, per day.

- **Baseline Models**
  - Principal Components Analysis (PCA): dimensionality reduction followed by reconstruction.
  - Isolation Forest: Tree-based decision algorithm for detecting outliers.

time	src_user	dst_user	src_pc	dst_pc	auth_type	logon_type	auth_orient	success?
1	C625@DOM1	U147@DOM1	C625	C625	Negotiate	Batch	LogOn	Success

Figure: Example LANL log line.

Language Model

Intuition: log lines are like sentences in a language—we can build a language model to generate probability distributions over sequences of words.

Given a log line of words $x_1, x_2, \ldots, x_T$, we want to predict the word x_t at time t. To do this, we find the probability of word x_t at time t given all preceding words: $P(x_t|x_1 x_2 \ldots x_{t-1})$.

We use recurrent neural networks to generate these probability distributions.

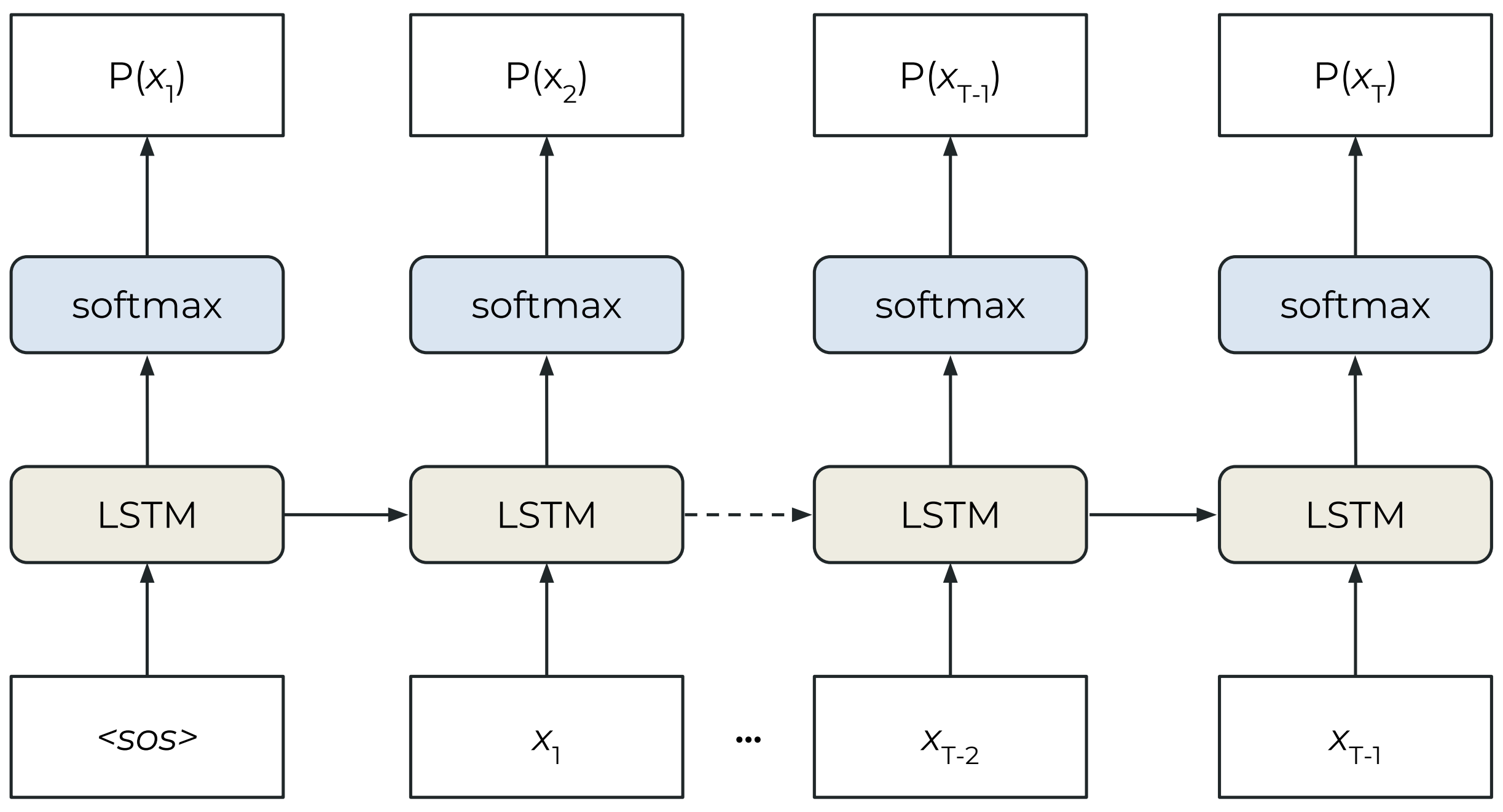

Figure: Recurrent neural network language model.

- Language model learns grammar of "normal" log lines.
- Unusual log lines won't be properly replicated.
- Anomaly score is sum of cross entropy losses over all T words.
- Since model operates on log lines directly, no aggregation is required.

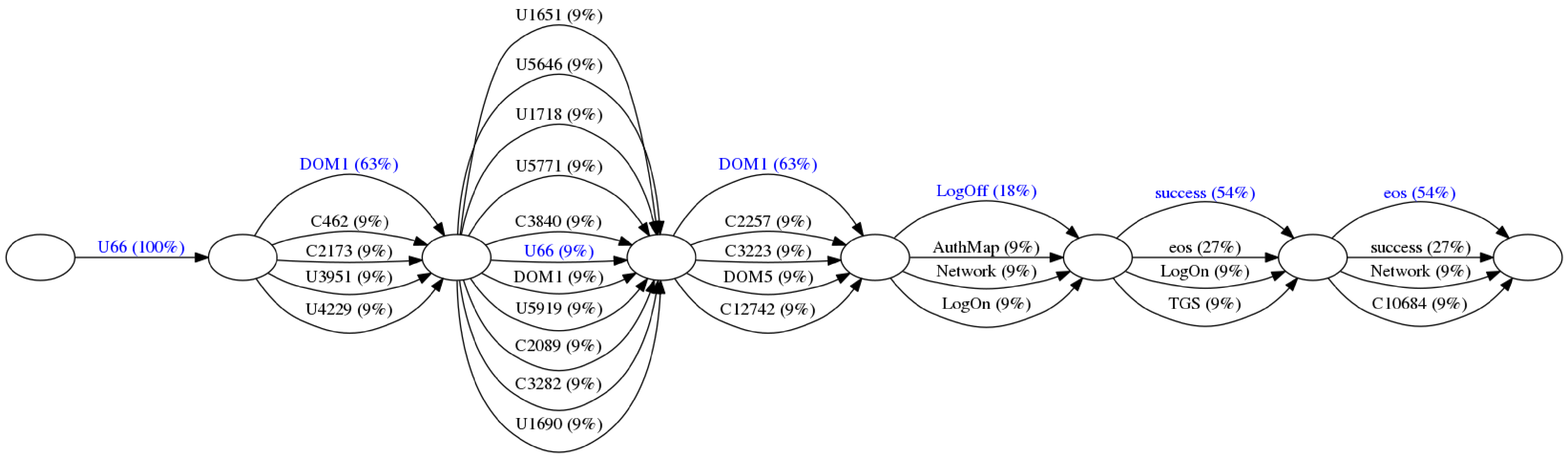

Figure: Interpreting the model's decision with token probabilities.

Acknowledgements

Many thanks to the work and contributions of our past and present collaborators: Aaron Tuor, Ryan Baerwolf, Nicolas Knowles, Nicole Nichols, and Rob Jasper.

Experimental Setup

- **LANL Cyber Security Dataset:** over one billion event log lines collected over 58 consecutive days.

Field	Example	# unique labels
time	1	5011198
source user	C625@DOM1	80553
dest. user	U147@DOM1	98563
source pc	C625	16230
dest. pc	C625	15895
auth. type	Negotiate	29
logon type	Batch	10
auth. orient	LogOn	7
success	Success	2

Figure: Authentication log fields and statistics

Results and Analysis

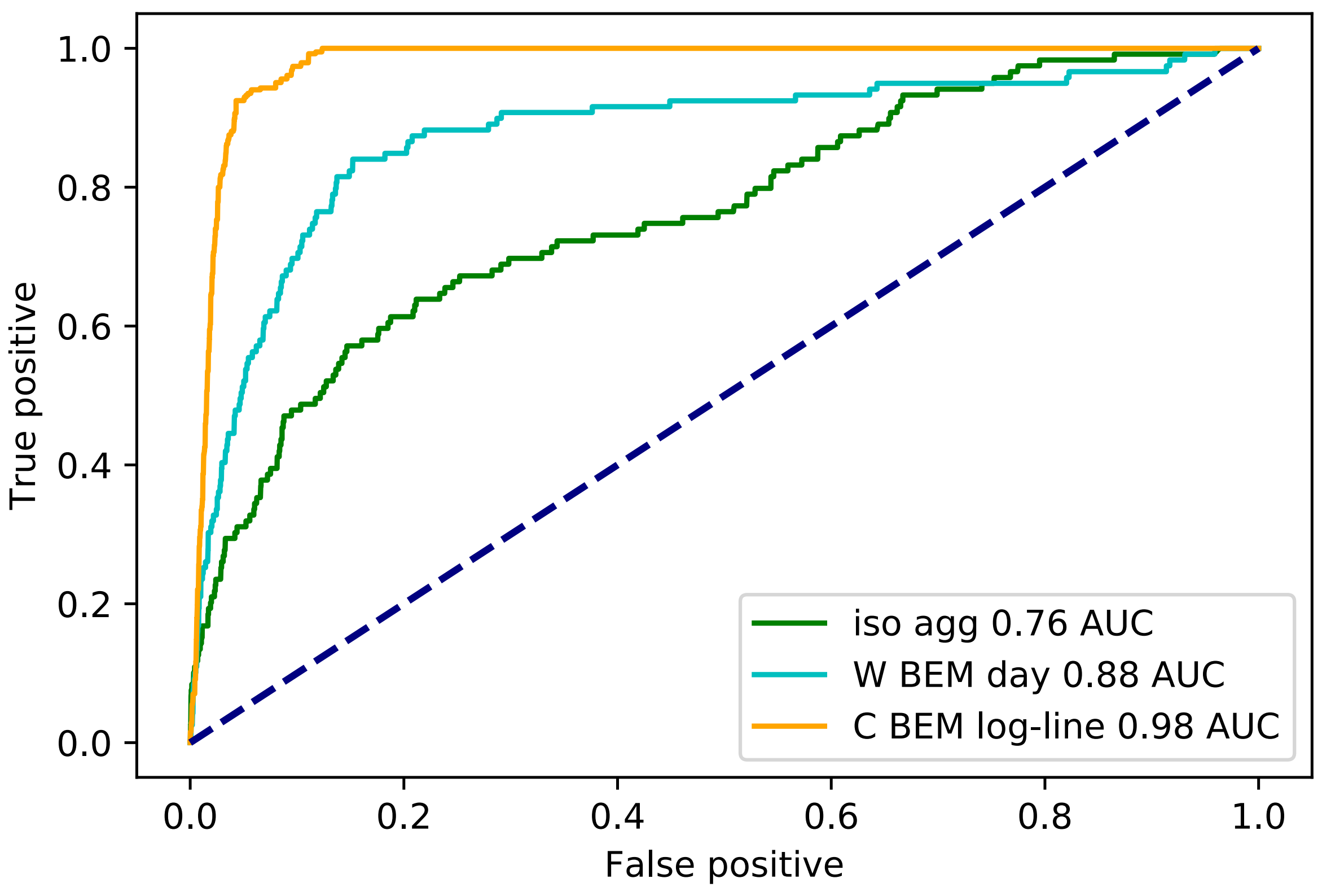

Figure: ROC curves for best baseline, word-level, and character-level language models.

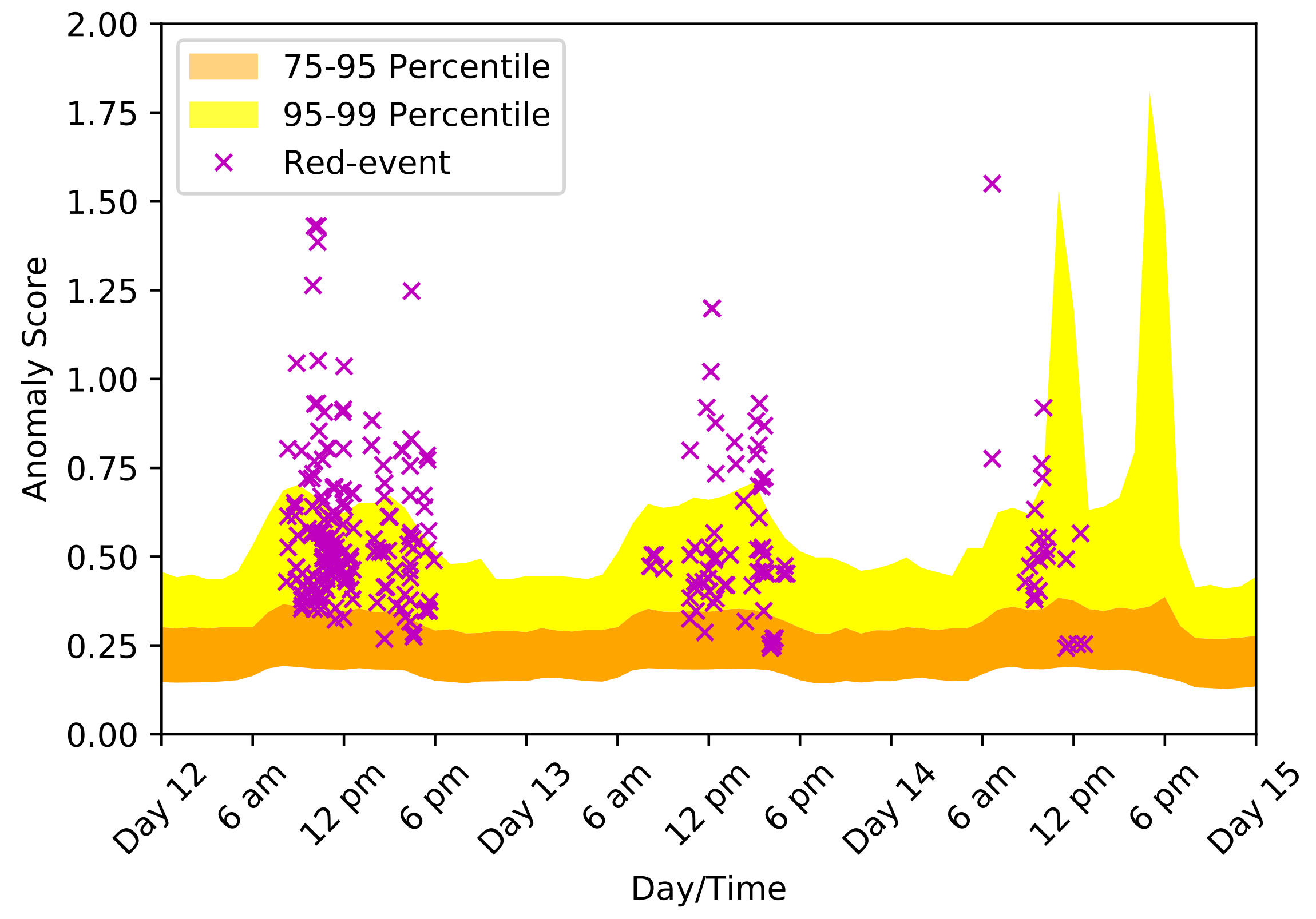

Figure: Character-level red-team log-line anomaly scores in relation to percentiles over time.

Conclusions and Future Work

- Perform granularity analysis for fair baseline comparison.
- Obtain results on other datasets.
- Explore methods of interpretability.