USENIX Security '25 Artifact Appendix: A limited technical background is sufficient for attack-defense tree acceptability

Nathan Daniel Schiele and Olga Gadyatskaya

A Artifact Appendix

A.1 Abstract

Study description

Attack-defense trees (ADTs) are a prominent graphical threat modeling method that is highly recommended for analyzing and communicating security-related information. Despite this, existing empirical studies of attack trees have established their acceptability only for users with highly technical (computer science) backgrounds while raising questions about their suitability for threat modeling stakeholders with a limited technical background. Our research addresses this gap by investigating the impact of the users' technical background on ADT acceptability in an empirical study.

Our Method Evaluation Model-based study consisted of $n = 102$ participants (53 with a strong computer science background and 49 with a limited computer science background) who were asked to complete a series of ADT-related tasks. By analyzing their responses and comparing the results, we reveal that a very limited technical background is sufficient for ADT acceptability. This finding underscores attack trees' viability as a threat modeling method.

Artifact description

This artifact includes the models (attack-defense trees) created by the participants in the study tasks, the study responses to perception questions (both Likert and short answer), the code used to statistically evaluate those survey responses, the study question text and images, the qualitative evaluation rubric for self-drawn ADTs, and the lecture plan and slides. With these artifacts, it should be possible to verify our results, organize training on attack-defense trees, and develop future studies of attack-defense tree acceptability.

A.2 Description & Requirements

A.2.1 Security, privacy, and ethical concerns

Our experiments are not destructive in any way. Our research involved human subjects, however all data in the artifact has been anonymized with no means to reconstruct original participants. Participants are identified with a unique code (`LT`/`HT`)-`##` in order to enable tracking of the same participant across different components (from image of ADT to survey responses). We described our handling of ethics concerns in Section 9 of our paper.

In order to perform the Holm-Bonferroni (HB) correction, all statistical tests need to be collected and corrected. For verification of the representativeness of our sample of the overall group of students, we statistically analyzed the overall course grade data of participants compared to non-participants. We ethically cannot release the underlying grade data for either participants and especially non-participants. We provided the original (non-corrected) results of performing these tests in the jupyter notebook `Statistics.ipynb` such that the results can be corrected (and subsequently affect the other corrections). As such, corrections can be independently performed but these tests cannot be independently performed due to ethical concerns.

A.2.2 How to access

The artifact is uploaded in its entirety to Zenodo and can be found using the following DOI: https://doi.org/10.5281/zenodo.14717342

A.2.3 Hardware dependencies

None.

A.2.4 Software dependencies

Our results were calculated with a Python virtual environment using Python version 3.12.1 with pip version 25.0.1. The following versions of jupyter and associated jupyter packages were used, however any version at least this version should sufficient:

- `IPython` – 8.31.0
- `ipykernel` – 6.29.5
- `jupyter_client` – 8.6.3
- `jupyter_core` – 5.7.2

A.2.5 Benchmarks

None.

A.3 Set-up

To run the statistics experiments, both the `Survey Data.csv` and `Statistics.ipynb` files should be downloaded and placed in the same directory. The Jupyter notebook should be run in a Python virtual environment with at least the versions of Python and the jupyter packages listed in Section A.2.4.

A.3.1 Installation

The first cell of the Jupyter notebook installs dependent packages. The cells are designed to be run in order.

A.3.2 Basic Test

A simple test would be to run all cells of `Statistics.ipynb`. If all cells run without error, the artifact is likely to be functioning correctly.

A.4 Evaluation workflow

A.4.1 Major Claims

(C1): There is no statistically significant difference in the actual effectiveness, perceived ease of use, perceived usefulness and intention to use of ADTs between users with a limited technical background and users with a highly technical background.

(C2:) Creative elements of the drawn of ADTs does not appear to be affected by the user's technical background.

A.4.2 Experiments

(E1): *[Statistical Testing Acceptability]: The statistical analysis of survey responses*

How to: For every hypothesis, we conducted the Brunner-Munzel test to assess if there is a statistically significant difference between the two groups. If no statistically significant difference is found, we performed the two one-sided test (TOST) to determine if there was a statistically significant equivalence between the two groups. All results are subsequently corrected using the Holm-Bonferroni correction. The measurement questions per hypothesis are as follows:

H_1
- `SS-Q3`
- `leaf_node_avg`
  * `SS-Q2`
  * `SS-Q7`
  * `ADT1 # atk leaf nodes`
- `def_node_avg`
  * `SS-Q8`
  * `SS-Q13`
- `attack_vector_avg`
  * `SS-Q14`
  * `SS-Q18`
- `LoA_avg`
  * `SS-Q9`
  * `SS-Q12`
  * `ADT1 # def leaf nodes`

H_{2-1}
- `ADT2 # def leaf nodes`
- `ADT2 # def nodes`
- `ADT2 # atk leaf nodes`
- `ADT2 # atk nodes`
- `ADT2 # and (atk)`
- `ADT2 #or (atk)`
- `ADT2 LoA`

H_{2-2}
- `ADT3 Cohesive`
- `ADT3 Clear`
- `ADT3 Concise`
- `ADT3 Complete`

H_3
- `ADT1 Multi parent nodes`
- `ADT3 Multi parent nodes`
- `ADT1 Multi refinement`
- `ADT3 Multi refinement`
- `ADT1 multi countermeasure`
- `ADT2 multi countermeasure`
- `ADT3 multi countermeasure`
- `ADT1 single child (atk)`
- `ADT2 single child (atk)`
- `ADT3 single child (atk)`

H_4
- `LS-ADT1-L1`
- `SS-Q5`
- `SS-Q10`
- `SS-Q15`
- `SS-Q19`

H_5
- `LS-ADT3-L3`

H_6
- `LS-ADT1-L5`
- `LS-ADT2-L2`
- `LS-ADT3-L1`

H_7
- `LS-ADT2-L1`
- `LS-ADT3-L2`
- `SS-Q6`
- `SS-Q11`
- `SS-Q16`
- `SS-Q20`

H_8
- `LS-ADT3-W3 Yes`
- `LS-ADT3-W3 Communication`
- `LS-ADT3-W3 Analysis`
- `LS-ADT3-W5 Yes`

H_9
- `ADT3 # def leaf nodes`
- `ADT3 # def nodes`

- `ADT3 # atk leaf nodes`
- `ADT3 # atk nodes`
- `ADT3 # and (atk)`
- `ADT3 #or (atk)`
- `ADT3 LoA`
- `ADT3 and:or ratio`

Preparation: Download the `Survey Data.csv` file. Additionally, download the `Statistics.ipynb` file and place it in the same folder. The columns of the survey data are labeled with the measurement questions. The data should be collected into the comparison groups (HT vs. LT). For hypothesis H_1, preprocessing the data to create average correctness values is required.

Execution: For each of the hypotheses listed above, perform the Brunner-Munzel test. If the test is not statistically significant (p-value > 0.05), perform the two one-sided test (TOST). Correct all tests using the Holm-Bonferroni correction, and perform new TOST tests if previous results are determined to be no longer statistically significant.

Note: This is what the code in the jupyter notebook `Statistics.ipynb` does. These tests will be run in the manner described here by running all cells o the notebook.

Results: The corrected p-values should align with the results in the original paper presented in Tables 2–7. These results are used to justify claims C1 and C2.

A.5 Version

Based on the LaTeX template for Artifact Evaluation V20231005. Submission, reviewing and badging methodology followed for the evaluation of this artifact can be found at https://secartifacts.github.io/usenixsec2025/.