A Artifact Appendix

A.1 Abstract

In our paper, “Uninvited Guests: Analyzing the Identity and Behavior of Certificate Transparency Bots”, we curated an extensive dataset of web requests originating from bots monitoring Certificate Transparency (CT) logs. In total, we recorded over 1.5 million requests from CT bots, originating from 31,898 unique IP addresses. To assist in the understanding and further exploration of this previously-unexplored population of bots, we are releasing our dataset and domain generation script to researchers.

We observed that CT bots can be subdivided into distinct groups based on the types of hosts they target, each with varying behaviors. Using our provided dataset, one can analyze these subsets of CT bots, including the populations of each group and characteristics of the web requests they transmit.

A.2 Artifact check-list (meta-information)

- **Publicly available:**
  - Dataset: https://zenodo.org/record/6677235#.YrH-o3jMJes
  - Domain generation script: https://zenodo.org/record/6818616#.YsxXLy-B0iY
- **Data licenses:** Licenced under *Creative Commons Attribution 4.0 International*
- **DOI:**
  - Dataset: 10.5281/zenodo.6677235
  - Domain generation script: 10.5281/zenodo.6818616

A.3 Description

A.3.1 How to access

Download our dataset and domain generation script using the Zenodo links in Section A.2.

A.3.2 Hardware dependencies

N/A

A.3.3 Software dependencies

- Python3
- Python Pandas library

A.3.4 Data sets

N/A

A.3.5 Models

N/A

A.3.6 Security, privacy, and ethical concerns

N/A

A.4 Installation

N/A

A.5 Evaluation and expected results

In our paper, we studied the behavior and identity of Certificate Transparency (CT) bots, curating a dataset consisting of over 1.5 million web requests from bots consuming CT logs. We found that this previously-unexplored population of web bots can be sub-divided into groups that target specific types of hosts based on the content of their domains names–with each subset exhibiting unique behaviors.

In addition to providing our full dataset of CT bot web requests, we have also included an example analysis script that can be used to reproduce a number of general statistics and a table listed in our paper. To do this, download our dataset and analysis script from the Zenodo repository listed in Section A.2. Next, using a Python3 interpreter, install the Python Pandas library (listed in the included *requirements.txt* file). Finally, run the example analysis script and review the outputted results on the terminal, which include the population sizes of each CT bot subset as well as a table listing the most common file paths requested by bots in each group.

To assist the research community in reproducing our CtPot system, we are also releasing a Python script that can be used to generate pseudo-random subdomains to be advertised on CT through the generation of TLS certificates. After downloading the script from the Zenodo link listed in Section A.2, simply execute it using a Python3 interpreter. A unique pseudo-random subdomain will then be printed onto the terminal.

A.6 Version

Based on the LaTeX template for Artifact Evaluation V20220119.