**Web Bot Detection Dataset**

Automated programs (bots) are responsible for a large percentage of website traffic. These bots, called web bots, vary in sophistication based on their purpose, ranging from simple automated scripts to advanced web bots that have a browser fingerprint, support the main browser functionalities, and exhibit a humanlike behaviour. Advanced web bots are especially appealing to malicious web bot creators, due to their browserlike fingerprint and humanlike behaviour that reduce their detectability. Malicious purposes of web bots include, but are not limited to, content scraping, vulnerability scanning, account takeover, distributed denial of service attacks, marketing fraud, carding, spam, buying all the available stock of specific limited products to later resell at higher price (i.e., scalper bots).

Thus, web servers must be equipped with the tools to detect such malicious web bots. For that, state-of-the-art approaches, both in academia and in commercial solutions, propose the combination of rule-based techniques with machine learning based methods [1]. For the latter, the browsing behaviours of visitors are used to train machine learning models that distinguish web bots from human visitors [1].

Early machine learning based web bot detection methods examined the web logs of the visitors, while more recent approaches use the mouse movements that the visitors perform [1]. The lack of datasets that include mouse movements of humans browsing the web (alone or in combination with the respective web logs) motivated the creation of this dataset which contains the web logs and mouse movements of (i) humans, (ii) *moderate web bots* that have a browser fingerprint and (iii) *advanced web bots* that have a browser fingerprint and also exhibit a humanlike behaviour. This dataset can be used to research web bot detection and evasion techniques that use and/or combine web logs with mouse movements.

The dataset was collected using a web server hosting web pages crawled from Wikipedia (<https://www.wikipedia.org/>) and consists of two parts, each used for the different evaluation phases of [1], respectively:

1. For the first evaluation phase, the web server used hosted 61 web pages from five different categories/topics crawled from Wikipedia, while 50 human sessions were generated by a closed set of participants, i.e., the authors of [1]; in each session, the authors visited the web server for an adequate (not predefined) period of time to generate sufficient data for our experiments.
2. For the second evaluation, an expanded version of the same web server was used; this web server hosted a total of 110 web pages from 11 categories/topics (including the content used in the first version of the web server) crawled again from Wikipedia. In this case, 28 users were asked to visit this web server and to create two sessions each (resulting in a total of 56 human sessions). Each user was instructed to spend about 15–20 minutes per session.

In both evaluation phases, we created the same amount of *moderate web bot* and *advanced web bot* sessions as the sessions generated from the humans. Details about the behaviour of these bots can be found in [1].

The dataset consists of two folders with names “phase1” and “phase2”, one for each of the two evaluation phases.

1. The “phase1” folder used for the first evaluation phase contains two subfolders:
2. The “*annotations*” folder which has two additional subfolders:
   1. “*humans\_and\_moderate\_bots*”: this folder contains the annotations of the human and moderate web bot sessions
   2. “*humans\_and\_advanced\_bots*”: this folder contains the annotations of the human and the advanced web bot sessions.

Each of these folders has two files, one having the annotations of the data used for training (called “train”), and one for testing (called “test”). The “*train*” and “*test*” files have two space-separated columns, the first that contains the PHP session ID of visitors and the second that has the respective annotation (i.e., human, moderate bot, and advanced bot).

1. The “*data*” folder which contains two subfolders:
   1. The “*web\_logs*” folder that contains the web logs of humas and (moderate and advanced) web bots. These are based on the apache2 log format (<https://httpd.apache.org/docs/2.4/logs.html>) but also include the PHP session ID in each request.
   2. The “*mouse\_movements*” folder which contains two subfolders:
      * “*humans\_and\_moderate\_bots*”: this folder contains the data of the human and moderate web bot sessions
      * “*humans\_and\_advanced\_bots*”: this folder contains the data of the human and the advanced web bot sessions.

Each of these subfolders (i.e., the “*humans\_and\_moderate\_bots*” and the “*humans\_and\_advanced\_bots*”) contains one folder for each session with the name of the folder being the session ID. Inside each folder there is the “*mouse\_movements.json*” JSON file that contains the following fields:

* + - * *session\_id*: The PHP session ID
      * *total\_behaviour*: The mouse movement actions of the visitor. These actions can be either move, right click, left click or middle click. The move is written as m(x,y), where the x and y indicate the coordinates of the mouse move. The right, left and middle click are written as c(l), c(r), and c(m) respectively
      * *mousemove\_times*: the timestamp of when each mouse move was performed in the form of {(t0), (t1), …, (tn)}, where n is the total number of points over which the mouse hovered in each web page.
      * *mousemove\_total\_behaviour*: the coordinates of the current mouse point in the form of {(x0, y0), (x1, y1), …, (xn, yn)}, where n is the total number of points over which the mouse hovered in each web page.

1. The “phase2” folder used for the second evaluation phase contains two subfolders:
2. The “*annotations*” folder which has two additional subfolders:
   1. “*humans\_and\_advanced\_bots*”: this folder contains the annotations of the human sessions and advanced web bot sessions used in the first part of the second evaluation phase
   2. “*humans\_and\_moderate\_and\_advanced\_bots*”: this folder contains the annotations of the human sessions and the advanced web bot sessions used in the second part of the second evaluation phase.

Each of these folders contains a file with two space-separated columns, one that contains the PHP session ID of visitors and the second that indicates whether this session is a human, a moderate bot, or an advanced bot.

1. The “*data*” folder, which contains two subfolders:
   1. The “*web\_logs*” folder that contains the web logs of humas and bots. These are based on the apache2 log format (<https://httpd.apache.org/docs/2.4/logs.html>) but also include the PHP session ID in each request.
   2. The “*mouse\_movements*” folder that contains the mouse movements that humans, and moderate and advanced web bots performed during their browsing sessions. These are exported collections from a MongoDB instance (in a JSON format) and can be re-imported directly to a new MongoDB instance. The fields that these files contain are:
      * *session\_id*: The PHP session ID
      * *mousemove\_client\_height\_width*: the client’s browser height and width during each mouse move in the form of {(height0, width0), (height1, width1), …, (heightn, widthn)}, where n is the total number of points over which the mouse hovered in each web page.
      * *mousemove\_times*: the timestamp of when each mouse move was performed in the form of {(t0), (t1), …, (tn)}, where n is the total number of points over which the mouse hovered in each web page.
      * *mousemove\_total\_behaviour*: the coordinates of the current mouse point in the form of {(x0, y0), (x1, y1), …, (xn, yn)}, where n is the total number of points over which the mouse hovered in each web page.
      * *mousemove\_visited\_urls*: the current URL that the mouse move is performed on in the form of {(url0), (url1), …, (urln)}, where n is the total number of points over which the mouse hovered in each web page.