Data Article

Dataset for Botnet group activity with adaptive generator

Dandy Pramana Hostiadi\textsuperscript{a,b}, Tohari Ahmad\textsuperscript{a,*}

\textsuperscript{a} Department of Informatics, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia
\textsuperscript{b} Department of Informatics, Institut Teknologi dan Bisnis STIKOM Bali, Bali, Indonesia

\textbf{Article history:}
Received 5 August 2021
Accepted 27 August 2021
Available online 28 August 2021

\textbf{Keywords:}
Botnet dataset
Bot group activities
Bot correlation activity
Infrastructure
Network security

\textbf{A B S T R A C T}

This dataset represents activity data of bot group and normal activities in a binetflow file. The dataset is generated according to the simulation process of the CTU dataset by extracting activity patterns in each activity type. It is simulated with a modeling approach to producing a new dataset as botnet group activity data. Furthermore, it consists of 13 datasets with different bot group activity scenarios, each containing the number of bots and various activity types in net flows form whose total activity duration is 8 h. The presented dataset has 14 net flow features, and one activity label feature is used to mark either bot or normal activity. This dataset is useful for research in detecting a periodic bot group activity and its intensity and is correlated with each type of bot activity.

\textcopyright{} 2021 The Authors. Published by Elsevier Inc.
This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)

* Corresponding author.
\textit{E-mail address:} tohari@if.its.ac.id (T. Ahmad).

https://doi.org/10.1016/j.dib.2021.107334

2352-3409/© 2021 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)
## Specifications Table

| Subject                      | Cryptography and Cybersecurity                                                                 |
|------------------------------|-------------------------------------------------------------------------------------------------|
| Specific subject area        | Activity Detection, Intrusion Detection System, Anomaly detection                                |
| Type of data                 | Binetflows                                                                                        |
| How data were acquired       | Data are obtained based on the extraction of activity patterns in the CTU dataset, from scenario 1 to scenario 13. An activity pattern is a type of activity that interrelated to each other in a sequential and correlated activity within a scenario, either bot or normal. A system modeling is developed using the Python programming language to get the pattern extraction results. |
| Data format                  | Analyzed (binetflow)                                                                               |
| Parameters for data collection| The dataset is the adoption of activity patterns from the CTU dataset. The patterns include that of bot activity and normal host activity. |
| Description of data collection| The data were collected by simulating botnet activities. This new dataset is generated based on some parameters for identifying the bot group activity pattern. These parameters are total activity times, number of bots in the specified scenarios, type of bot activity scenarios, number of normal activities, and intensity of the botnet activities. |
| Data source location         | Institution: Institut Teknologi Sepuluh Nopember                                                    |
|                             | Country: Indonesia                                                                                 |
| Data accessibility           | Repository name: Mendeley                                                                            |
|                             | Direct URL to data: https://doi.org/10.17632/4vftxh97m8.1                                           |
|                             | Repository name: Zenodo                                                                             |
|                             | Direct URL to source code: https://doi.org/10.5281/zenodo.5151133                                   |
| Related research article     | D. P. Hostiadi, W. Wibisono and T. Ahmad, “B-Corr Model for Bot Group Activity Detection Based on Network Flows Traffic Analysis,” KSII Transactions on Internet and Information Systems, vol. 14, no. 10, pp. 4176–4197, 2020. |
|                             | DOI: https://doi.org/10.3837/tiis.2020.10.014                                                       |

## Value of the Data

- We provide a dataset representing a botnet group activity with a series of interrelated activities and their intensity.
- This dataset is useful for network security research, specifically to evaluate the performance of the method in detecting botnet activities, which can be a bot group activity, bot attack activity correlations, bot attack activity scenarios, and knowledge databases for botnet activity scenarios.
- Moreover, it is generated based on various activity scenarios, which provide the respective knowledge databases. This dataset contains 14 standard features and activity labels to indicate the type of the corresponding bot activity.

### 1. Data Description

This botnet group activity dataset is in the form of a binetflow file, which adopts bot activity patterns found in the CTU dataset [1]. It is done by extracting bot activity patterns according to 13 bot activity scenarios extracted and simulated through modeling [2]. This dataset aims to present bot activity patterns in groups analyzed based on their intensity and linkages between activities. Unlike other bot datasets [3,4], it is precisely implemented for periodic activity patterns correlated with bot activities.

The 13 scenarios are designed by specifying the parameters shown in Table 1. Details of each scenario and reports of bot activity in each of them are shown in Tables 2 and 3, respectively.

The dataset is presented separately into 13 binetflows files. An example of the dataset in scenario one is shown in Fig. 1. The first line of the dataset in Fig. 1 describes 14 features with standardized writing formats [5]. Those are the start time, duration, protocol, source IP address, source port address, direction of transaction, destination IP address, destination port address,
specifically for the tensity.

The state of transaction, source TOS byte value, destination TOS byte value, total transaction packet count, total transaction bytes, source packet transaction and one feature as activity label.

The number of bot activity labels in each scenario varies, wherein in total, there are 23 types. The activity type in each scenario relates to each other, known as a bot group activity. For example, in the first scenario, 23 activity types develop a series of group activities, as depicted in Fig. 2.

The presented bot group activity pattern in the dataset is periodic with its corresponding intensity. This characteristic has made it different from the CTU dataset. Here, periodic means that the bot activity pattern appears in some following segments, while intensity means that there is any bot activity in every segment. These periodic and intensity characteristics are appropriate for evaluating bot group detection with a time-based segmentation approach [6–9]. An example of the scenario bot group activity of dataset 9 is provided in Fig. 3. It shows both the accumulated and hourly-based activity patterns. Furthermore, this dataset’s activity is more stable, specifically in an hour-periodic time, than that of the CTU dataset, as depicted in Figs. 4 and 5.

Table 1
Parameter dataset.

| Name of Parameter | Set value | Description |
|-------------------|-----------|-------------|
| Time Duration     | 8 (hours) | Total duration time for net flows traffic. |
| Number of Bots    | (dynamic for each scenario) | The number of bots exists in each scenario dataset. It is adjusted based on the number of CTU datasets. |
| Type flows        | (dynamic for each scenario) | The number of bot activity types relates to one another as a stage of the bot activity scenario. |
| Flows intensity   | 1000 (times) | The intensity of each type of bot flow activity. |
| Number of normal hosts | (dynamic for each scenario) | The number of hosts with normal activity flows in the dataset is adjusted based on the number of normal hosts from the adopted CTU dataset. |
| Number of normal flows | (dynamic for each scenario) | The number of normal activity flows in the dataset. It is obtained based on the number of normal activities from the adopted CTU dataset. |
| Normal percentage | (dynamic for each scenario) | Percentage of normal activity flows. The comparison between normal and total generated record flows. |
| Bot percentage    | (dynamic for each scenario) | Percentage of bot activity flows. The comparison between the bot and total record flows. |

Table 2
Dataset scenario description.

| Scenario number | Time Duration | Number of bots | Bot flows | Normal hosts | Normal Flows | Total flows |
|-----------------|---------------|----------------|-----------|--------------|--------------|-------------|
| 1               | 8 h           | 1              | 23,000 (1.09%) | 342,740 | 2089,224 (98.91%) | 2112,224 |
| 2               | 8 h           | 1              | 24,000 (1.64%) | 252,263 | 1441,182 (98.36%) | 1465,182 |
| 3               | 8 h           | 1              | 2000 (0.07%) | 240,780 | 2903,611 (99.93%) | 2905,611 |
| 4               | 8 h           | 1              | 11,000 (1.52%) | 66,013 | 713,388 (98.48%) | 724,388 |
| 5               | 8 h           | 1              | 19,000 (20.45%) | 10,346 | 73,917 (79.55%) | 92,917 |
| 6               | 8 h           | 1              | 6000 (1.17%) | 46,627 | 506,021 (98.83%) | 512,021 |
| 7               | 8 h           | 1              | 9000 (10.78%) | 5958 | 74,473 (89.22%) | 83,473 |
| 8               | 8 h           | 1              | 14,000 (0.49%) | 252,162 | 2857,217 (99.51%) | 2871,217 |
| 9               | 8 h           | 1              | 220,000 (13.98%) | 180,554 | 1353,304 (86.02%) | 1573,304 |
| 10              | 8 h           | 1              | 60,000 (6.10%) | 89,915 | 924,369 (93.90%) | 984,369 |
| 11              | 8 h           | 3              | 120,000 (38.75%) | 3729 | 18,964 (61.25%) | 30,964 |
| 12              | 8 h           | 3              | 9000 (3.28%) | 33,613 | 265,186 (96.72%) | 274,186 |
| 13              | 8 h           | 1              | 19,000 (1.01%) | 209,865 | 1857,489 (98.99%) | 1876,489 |
Fig. 1. Dataset files and content.

| Scenario Dataset | Date/Time | Content |
|------------------|-----------|---------|
| scenario_dataset_1 | 1/24/2021 2:13 PM | dataset_result.binetflow, text_document, dataset_result.graph | BINETFLOW File, Text Document, PNG File |
| scenario_dataset_2 | 1/24/2021 2:14 PM | dataset_result.botnet_only_binetflow, dataset_result.graph | BINETFLOW File, PNG File |
| scenario_dataset_3 | 1/24/2021 2:17 PM | dataset_result.graph_accumulated.png, dataset_result.normal_only_binetflow | PNG File, BINETFLOW File |
| scenario_dataset_4 | 1/24/2021 2:19 PM | dataset_result.graph.png, dataset_result_without_label_binetflow | PNG File, BINETFLOW File |
| scenario_dataset_5 | 1/24/2021 2:20 PM | running_set.jpg | JPG File |
| scenario_dataset_6 | 1/24/2021 2:21 PM | |
| scenario_dataset_7 | 1/24/2021 2:22 PM | |
| scenario_dataset_8 | 1/24/2021 2:23 PM | |
| scenario_dataset_9 | 1/24/2021 2:24 PM | |
| scenario_dataset_10 | 1/24/2021 2:25 PM | |
| scenario_dataset_11 | 1/24/2021 2:26 PM | |
| scenario_dataset_12 | 1/24/2021 2:27 PM | |
| scenario_dataset_13 | 1/24/2021 2:28 PM | |

Fig. 2. Examples of correlated types of bot activity.
### Table 3
Bot description.

| Scenario number | Number of bots | Bot flows | Bot IP | Start time | End Time |
|-----------------|----------------|-----------|--------|------------|----------|
| 1               | 1              | 23,000    | 147.32.84.165 | 00:00:00 | 08:00:00 |
| 2               | 1              | 24,000    | 147.32.84.165 | 00:00:04 | 08:00:00 |
| 3               | 1              | 2000      | 147.32.84.165 | 00:01:30 | 07:59:11 |
| 4               | 1              | 11,000    | 147.32.84.165 | 00:00:11 | 07:59:39 |
| 5               | 1              | 19,000    | 147.32.84.165 | 00:00:02 | 07:59:55 |
| 6               | 1              | 6000      | 147.32.84.165 | 00:00:13 | 07:59:59 |
| 7               | 1              | 9000      | 147.32.84.165 | 00:00:01 | 07:59:54 |
| 8               | 1              | 14,000    | 147.32.84.165 | 00:00:02 | 07:59:51 |
| 9               | 10             | 220,000   | 147.32.84.165 | 00:00:03 | 07:59:55 |
|                 |                |           | 147.32.84.191 | 00:00:01 | 07:59:49 |
|                 |                |           | 147.32.84.192 | 00:00:03 | 07:59:56 |
|                 |                |           | 147.32.84.193 | 00:00:02 | 07:59:59 |
|                 |                |           | 147.32.84.204 | 00:00:05 | 07:59:58 |
| 10              | 10             | 60,000    | 147.32.84.195 | 00:00:01 | 08:00:00 |
|                 |                |           | 147.32.84.205 | 00:00:03 | 07:59:58 |
|                 |                |           | 147.32.84.206 | 00:00:00 | 07:59:53 |
|                 |                |           | 147.32.84.207 | 00:00:02 | 07:59:56 |
|                 |                |           | 147.32.84.208 | 00:00:00 | 07:59:54 |
|                 |                |           | 147.32.84.209 | 00:00:00 | 07:59:54 |
| 11              | 3              | 120,000   | 147.32.84.165 | 00:00:14 | 07:59:51 |
|                 |                |           | 147.32.84.191 | 00:00:23 | 07:59:41 |
|                 |                |           | 147.32.84.192 | 00:00:20 | 07:59:48 |
|                 |                |           | 147.32.84.193 | 00:00:01 | 07:59:49 |
|                 |                |           | 147.32.84.204 | 00:00:07 | 07:59:56 |
| 12              | 3              | 9000      | 147.32.84.165 | 00:00:02 | 07:59:45 |
|                 |                |           | 147.32.84.191 | 00:00:02 | 07:59:40 |
|                 |                |           | 147.32.84.192 | 00:00:01 | 07:59:32 |
| 13              | 1              | 19,000    | 147.32.84.165 | 00:00:01 | 07:59:59 |

**Fig. 3.** Example of bot group activity dataset scenario 9.
2. Experimental Design, Materials and Methods

The process of generating the dataset is illustrated in Fig. 6, which comprises inputting data, modeling, generating data, and producing output data in the binetflow file format.

First, 13 scenarios of the CTU dataset [1] are adopted. This dataset is the input for data modeling [2] to detect either normal or bot activity. Next, the modeling result is stored in a knowledge base.

In this knowledge base, each bot activity is marked according to its bot activity type and classified based on the similarity between its IP address and the corresponding attack scenario. Equivalently, each normal activity is also marked and is classified based on its source IP address.

In the process of generating the bot group dataset, some parameters are defined. Those parameters are depicted in Fig. 7, which can be described as follows.

- total time duration: the total activity time (in hours)
- number of bots: the specified number of bots (counted based on the IP bots)
- type of bot flow: the type of activities that relates to the activity data in the knowledge base
- number of normal activities: the number of normal activities taken from the respective knowledge base
Fig. 6. The process of generating the bot group dataset.

Fig. 7. Examples of set parameters in the bot group activity dataset scenario 9.

Fig. 8. Examples of the results of generating the bot group activity dataset in scenario 9.
Fig. 9. Examples of the description of bot group activity dataset scenario 9.

The set of parameters shown in Fig. 7 leads to a bot group dataset in Fig. 8, whose description is given in Fig. 9. The label of bot and normal activities depends on the feature labels in each knowledge database.

**Ethics Statement**

The work does not involve the subject of humans, animals, or data from social media platforms.

**CRediT Author Statement**

Dandy Pramana Hostiadi: Data curation, Data Analysis, Methodology, Software, Writing-Original draft preparation, Validation; Tohari Ahmad: Supervision, Conceptualization, Reviewing and Editing.

**Declaration of Competing Interest**

The authors declare that they have no known competing financial interests or personal relationships which have or could be perceived to have influenced the work reported in this article.

**Acknowledgments**

This work is supported by the NCC (Net-Centric Computing) Laboratory of Department of Informatics, Institut Teknologi Sepuluh Nopember. Sincere gratitude to Dr. Waskitho Wibisono, who supported and directed the completion of this work.

**References**

[1] S. Garc, M. Grill, J. Stiborek, A. Zunino, An empirical comparison of botnet detection methods, Comput. Secur. 5 (2014).
[2] D.P. Hostiadi, T. Ahmad, W. Wibisono, A new approach to detecting bot attack activity scenario, in: Proceedings of the 12th International Conference on Soft Computing and Pattern Recognition (SoCPaR 2020), 2021, pp. 823–835.

[3] N. Koroniotes, N. Moustafa, E. Sitnikova, B. Turnbull, Towards the development of realistic botnet dataset in the Internet of Things for network forensic analytics: bot-IoT dataset, Futur. Gener. Comput. Syst. 100 (2019) 779–796.

[4] D. Zhao, et al., Botnet detection based on traffic behavior analysis and flow intervals, Comput. Secur. 39 (PARTA) (2013) 2–16.

[5] H. Debar, The IDMEF: RFC 4765, Mycol. Res. (2007).

[6] D.P. Hostiadi, W. Wibisono, T. Ahmad, B-Corr model for bot group activity detection based on network flows traffic analysis, KSII Trans. Internet Inf. Syst. 14 (10) (2020) 4176–4197.

[7] H. Choi, H. Lee, H. Lee, H. Kim, Botnet detection by monitoring group activities in DNS traffic, in: CIT 2007 7th IEEE Int. Conf. Comput. Inf. Technol., 2007, pp. 715–720.

[8] H. Choi, H. Lee, H. Kim, BotGAD: detecting botnets by capturing group activities in network traffic, in: Proc. Fourth Int. ICST Conf. Commun. Syst. Softw. Middleware, 2009, pp. 1–8.

[9] D.P. Hostiadi, T. Ahmad, W. Wibisono, A new approach of botnet activity detection model based on time periodic analysis, in: 2020 International Conference on Computer Engineering, Network, and Intelligent Multimedia (CENIM), 2020, pp. 315–320.