AMoC Hackathon Tasks 2021

1. Longitudinal Visualisation Task

Darknet markets remain a challenging issue in terms of monitoring and mitigating their practices, for example: buying and selling of drugs worldwide. While research often analyses these markets, there remains a gap in understanding the impacts of adverse events in darknet markets following, for example, DDoS attacks or targeted site take downs.

In this task we are looking for an innovative visualisation representing the datasets over time, with particular attention to how significant events might affect the nature of posting on the forums.

1.1. Data

For this task we make use of over 2.5 million posts drawn from over 100,000 users from 40 cybercriminal communities, drawn from a large dataset collected between 2013 and 2015. In particular, we targeted discussion forums within this collection, which acted as support areas for underground marketplaces dealing in a number of different illicit goods. Communities ranged from successfully established markets with thousands of users (though not all were always active posters) to small sites that never moved beyond a handful of initial users.

1.2. Instructions for Accessing the Data

The "Fora" folder has 40 csv-files, each containing data from a specific DNM forum. After loading a target forum file, the community name, user_id, threat_id, date, subject, category, body and quotes can be accessed.

To access a user's metadata and his/her messages posted on the targeted forum file, the code snippet below can be used.

In [1]:

```
import pandas as pd
df = pd.read_csv("fora/abraxas.csv")  # because the first positive user it's
print(len(df))
print(df.columns.values.tolist())
```

```
276300
['community', 'user_id', 'thread_id', 'date', 'subject', 'category', 'body',
'quotes']
```

Additionally, each user's registered information (community, user_id, title, first_seen) can be accessed in the "measure_impact/fora_registedusers.csv" using the code below.¶

```
In [3]:
reg_users = pd.read_csv("measure_impact/fora_registedusers.csv")
train_positive_users = pd.read_csv("reidentification/train_positive_users.csv

user_df = reg_users[reg_users['user_id'].str.match( train_positive_users.iloc
print(user_df)
```

```
            community                                           user_id  \
422   Abraxas Forums  d89ddca020452f38a5f5741628dcdaa24daff3bbfb23b8...

      title           first_seen
422  Vendor  2015-05-27 18:09:50
```

1.3. Evaluation

The panel will evaluate the results for this task against the following qualitative criteria:

- Suitability of visualization chosen (30%)
- Effectiveness of visualization in communicating long term trends (30%)
- Ability to understand multiple facets of the longitudinal data (20%)
- Quality of the submitted code (20%)

1.4. Code Submission

We kindly ask you to submit the following:

- the software you built for this task;
- a README file containing requirements and external resources needed to run your system.

You can choose freely among the available programming languages and among the operating systems. Please upload your software to the "Outputs" folder created for your team in Microsoft Teams.

Deadline for submission is Wednesday 10 February 2021 - 17:00 GMT.

Note: By submitting your software you agree to make your code available under CC-BY-NC license for use by researchers, including the AMoC team.

2. DNM User Re-identification Task

In recent years, Darknet Markets (DNMs) and other environments offering anonymity are becoming increasingly popular among criminals with a high degree of computer literacy and forensic awareness. Although none of such anonymisation techniques is entirely bulletproof, they can easily complicate or even block cybercrime investigations by law enforcement. In such cases, the communications produced on such underground forums can be one of few clues to a cyber offender's identity.

In this task we are looking for novel approaches to automatically re-identify cyber offenders using multiple identities across different underground platforms. More specifically:

- given a large training dataset comprising of DNM users of whom "ground truth" information is available on the different identities they have on one or more Darknet forums, their communications and metadata,
- design a system that is able to group all identities (one or more) of each user in a new dataset (the test dataset).

2.1. Data

For this task we make use of over 10,000 users from different cybercriminal communities who produced a minimum of 5 messages, drawn from the dataset described above. The training dataset provided for this task contains two files:

- train_positive_users.csv – contains the users who have more than one identity. Each line represents a new user with his/her matching identities.
- train_negative_users.csv – contains the users who have only one identity in this dataset.

The test dataset ("reidentification/test_dataset.csv") contains about 3,000 users without any "ground truth" information about multiple user identities. None of the users from the training dataset is included in the test dataset. Additionally, as the training dataset is highly skewed, we ensured that the test dataset shows a similar distribution of positive and negative users as the training dataset.

2.2. Instructions for Accessing the Data

Load positive and negative users for the train dataset from train_positive_users.csv and test_dataset.csv, respectively.

In [4]:

```
train_positive_users = pd.read_csv("reidentification/train_positive_users.csv
train_negative_users = pd.read_csv("reidentification/train_negative_users.csv
```

Within the positive users dataset, each list of multiple pairs indicates the different user-ids used in different communities, but they are actually the same user, and we labeled it as 1.

In [5]:

```
print(len(train_positive_users))
print(train_positive_users.iloc[0])
```

```
148
0                                         Abraxas Forums
1    d89ddca020452f38a5f5741628dcdaa24daff3bbfb23b8...
2                                      Silk Road 2 Forums
3    41e3879649d4f2710e06465dfd2b97b91eff035bc1137a...
```

```
4                                       The Hub Forums
5    62abf80ac979659c1e15ca7971c3f5cf13b507b431bf84...
6                                                    1
Name: 0, dtype: object
```

The negative users dataset contains a list of pairs of community- and user-ids. None of these users have more than one identity in this dataset (as far as we know).

```
In [6]:
print(len(train_negative_users))
print(train_negative_users.head(10))
```

```
11264
                    0                                                  1
0  Silk Road 2 Forums  1418a2a6f57fae847b12adf89e58634119b4737a18fa17...
1  Silk Road 2 Forums  7bbe565e056ac5bbc59506879d539b55e2f0454e02bc03...
2  Silk Road 2 Forums  35dd524ac275b00b8a7433e207f5c39dbce7406bd75ed4...
3      Nucleus Forums  0fbf005930c4c82ee5eaa7b7767a8502391fddf21fa298...
4  Silk Road 2 Forums  655e96ea4b5a9decfae59249d710093e10dc2998cffae5...
5    Silk Road Forums  5f9630d407a6251eab75853629a2094862ff25585585d0...
6    Silk Road Forums  294043793266704155265fcaf30015fcc811e7720978bb...
7  Silk Road 2 Forums  12544dfc90e5a4fbe597269714434d5391d13a56b66d71...
8    Evolution Forums  4f26c9572c7707c97c9aeac36cbf1c3767df863ac1ea11...
9    Silk Road Forums  133ef36ec57f7d2e4a83a476da6bcbd5d000a99a0e2628...
```

To access a user's metadata and his/her messages posted on the targeted forum file, please see the code snippet provided above. To select a target user's relevant information from a community, especially posted messages in the 'body' column, the following code can be used.

```
In [7]:
selected_df = df[df['user_id'].str.match( train_positive_users.iloc[0][1] )]
print(len(selected_df))
print(selected_df['body'][:5])
```

```
40
259361    The notification (1) stays on no matter what w...
259363     What's that does to do with a notification sh...
259364    Sorted (by itself, or by someone interfering y...
274396    A pop-up from within the Pidgin came up, sayin...
274414    Yes, it's normal not to get a tracking number ...
Name: body, dtype: object
```

2.3. Evaluation

Once you have developed your approach for this task, your software can be tested on the test dataset ("reidentification/test_dataset.csv"). After the hackathon, the ground truth labels for the test dataset will be made available.

To evaluate your systems’ performance, we will calculate mean average precision, recall and F1 score based on the labels produced by your system on the test dataset.

3. Output and Code Submission

For this task, we ask you to submit the following:

- one csv-file containing all positive users and one csv-file with all negative users of the test dataset, similar to the train_positive_users.csv and train_negative_users.csv provided for this task;
- the software you built for this task;
- a README file containing requirements and external resources needed to run your system.

Again, you can choose freely among the available programming languages and among the operating systems. Please upload your csv-files and software to the "Outputs" folder created for your team in Microsoft Teams.

Deadline for submission is Wednesday 10 February 2021 - 17:00 GMT.

Note: By submitting your software you agree to make your code available under CC-BY-NC license for use by researchers, including the AMoC team.

4. Report

Finally, we ask you to submit an overview of your approach to solving both AMoC Hackathon tasks and your results in a short report (up to 4 pages) using the IEEE conference paper templates that can be accessed here: https://www.ieee.org/conferences/publishing/templates.html. Please upload your report to the "Outputs" folder created for your team in Microsoft Teams.

Deadline for submitting the report is Thursday 11 February 2021 - 17:00 GMT.

5. Good Luck!

The panel will meet on Tuesday 16 February 2021 to deliberate on which team should win the prize. The results will be communicated to all teams via email following the panel meeting.