Progess of the Project

白宗民

2023/7/26

Outline

- **Graphormer**
- **Graph - Data Analysis**
- **TRAM**
- **Future Work**

Graphormer

Data Format

- A **jsonl** file:

edge_index (sequence)	edge_attr (sequence)	y (sequence)	num_nodes (int64)	node_feat (sequence)
[[0, 1, 1, 2, 2, 3, 3, 4, 4, 5, 5, 6, 6, 7, 7,…	[[0, 0, 1], [0, 0, 1], [3, 0, 1], [3, 0, …	[0]	24	[[6, 0, 3, 5, 2, 0, 1, 0, 0], [5,…
[[0, 1, 1, 2, 1, 3, 1, 4, 4, 5, 5, 6, 6, 7, 6,…	[[1, 0, 0], [1, 0, 0], [1, 0, 0], [1, 0, …	[0]	10	[[7, 0, 1, 5, 0, 0, 1, 0, 0], [15…

- **Edge_index:** contains the indices of nodes in edges, stored as a list containing two parallel lists of edge indices `edge_index = [[1,2,1], [2,3,3]]`
- **Labels:** list or an integer contain the corresponding techniques
- **Nodes_nums:** total number of the nodes
- **Node_feat:** contains the available features of each node (if present)
- **Edge_feat:** contains the available features of each edge (if present)

Data Format

- Try to input the data with different format → tried 8 versions

VincentPai/for-graphormer-new
Viewer • Updated about 21 hours ago • 1

VincentPai/for-graphormer-v6
Viewer • Updated 1 day ago

VincentPai/for-graphormer-v5
Viewer • Updated 2 days ago • 3

VincentPai/for-graphormer-v4
Viewer • Updated 4 days ago • 4

VincentPai/for-graphormer-v3
Viewer • Updated 4 days ago • 1

VincentPai/for-graphormer-v2
Viewer • Updated 5 days ago • 4

VincentPai/for-graphormer
Viewer • Updated 5 days ago • 1

VincentPai/repo_name
Preview • Updated 5 days ago

- My data version2

edge_index (sequence)	node_feat (sequence)	edge_attr (sequence)	y (sequence)	num_nodes (int64)
[["422696", "650081"], ["650081", "9"]]	[[0], [0], [0]]	[[0], [0]]	["0"]	3

Data Format

- Official format:

edge_index (sequence)	edge_attr (sequence)
[[0, 1, 1, 2, 1, 3, 3, 4, 4, 5, 5, 6, 6, 7, 7, 8, 7, 9, 9, 10, 10, 11, 11, 12, 12, 13, 13, 14, 14, 15, 15,…	[[0, 0, 0], [0, 0, 0], [1, 0, 1], [1, 0, 1], [0, 0, 1], [0, 0, 1], [0, 0, 0], [0, 0, 0], [0,…

- My last format:

y (sequence)	num_nodes (int64)	node_feat (sequence)	edge_attr (sequence)	edge_index (sequence)
["0"]	3	[[483679, 21, 799842]]	[[0], [0]]	[[0, 1], [1, 2]]

- y is label

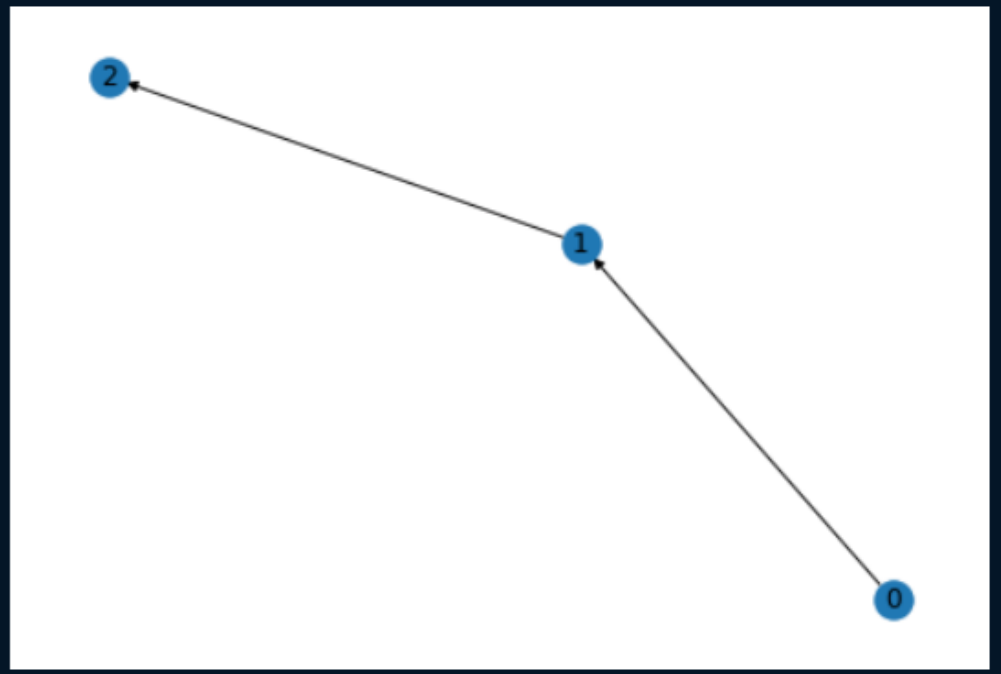
```
DatasetDict({
    train: Dataset({
        features: ['y', 'num_nodes', 'node_feat', 'edge_attr', 'edge_index'],
        num_rows: 2959563
    })
    validation: Dataset({
        features: ['y', 'num_nodes', 'node_feat', 'edge_attr', 'edge_index'],
        num_rows: 986521
    })
    test: Dataset({
        features: ['y', 'num_nodes', 'node_feat', 'edge_attr', 'edge_index'],
        num_rows: 986521
    })
})
```

Train:Validation:Test = 3:1:1

Data Format

- My data in a Directed Graph:

```
{'y': ['0'], 'num_nodes': 3, 'node_feat': [[445528, 24, 740662]], 'edge_attr': [[0], [0]], 'edge_index': [[0, 1], [1, 2]]}
```



- Preprocessing

```
from transformers.models.graphormer.collating_graphormer import preprocess_item, GraphormerDataCollator
dataset_processed = dataset.map(preprocess_item, batched=False)
```

Data Format

- My data after preprocessing:

```
DatasetDict({
    train: Dataset({
        features: ['y', 'num_nodes', 'node_feat', 'edge_attr', 'edge_index', 'input_nodes',
        num_rows: 2959563
    })
    validation: Dataset({
        features: ['y', 'num_nodes', 'node_feat', 'edge_attr', 'edge_index', 'input_nodes',
        num_rows: 986521
    })
    test: Dataset({
        features: ['y', 'num_nodes', 'node_feat', 'edge_attr', 'edge_index', 'input_nodes',
        num_rows: 986521
    })
})
```

```
'attn_bias', 'attn_edge_type', 'spatial_pos', 'in_degree', 'out_degree', 'input_edges', 'labels'],
```

```
'attn_bias', 'attn_edge_type', 'spatial_pos', 'in_degree', 'out_degree', 'input_edges', 'labels'],
```

```
'attn_bias', 'attn_edge_type', 'spatial_pos', 'in_degree', 'out_degree', 'input_edges', 'labels'],
```

Still has some BUGGGGGGGGs

```
File "/workdir/home/euni/anaconda3/lib/python3.9/site-packages/transformers/models/graphormer/collating_graphormer.py",
line 112, in __call__ batch["attn_bias"][ix, : f["attn_bias"].shape[0], : f["attn_bias"].shape[1]] = f["attn_bias"]
RuntimeError: The expanded size of the tensor (2) must match the existing size (4) at non-singleton dimension 1.
Target sizes: [2, 2]. Tensor sizes: [4, 4]
```

Graph - Data Analysis

Target

- Constructing the **directed graph** of every Attack Patterns (167 APs)
  - Connecting the source and the destination
  - Recording the **# of relations** with the same source and destination
  - Exclude T1046_5a4 (1022 triplets) and T1005_720 (13801 triplets)
    - Final result would contain 165 Aps

- Connecting all the **related neighbor** nodes in a **single hop**
  - Labelling them with different color

Data Preprocessing

```
483679 39363 14
483679 39363 2
483679 39363 6
483679 362399 21
483679 362399 10
483679 362399 24
```

+

```
registry benign
registry benign
registry benign
file benign
file benign
file benign
```

→

Src, Dest, Rel, Label

```
565544,78056,11,1548.002_665432a
565544,78056,17,1548.002_665432a
335532,662677,23,T1135_deeac480-
262572,255488,23,T1016_921055f4-
262572,255488,23,0
```

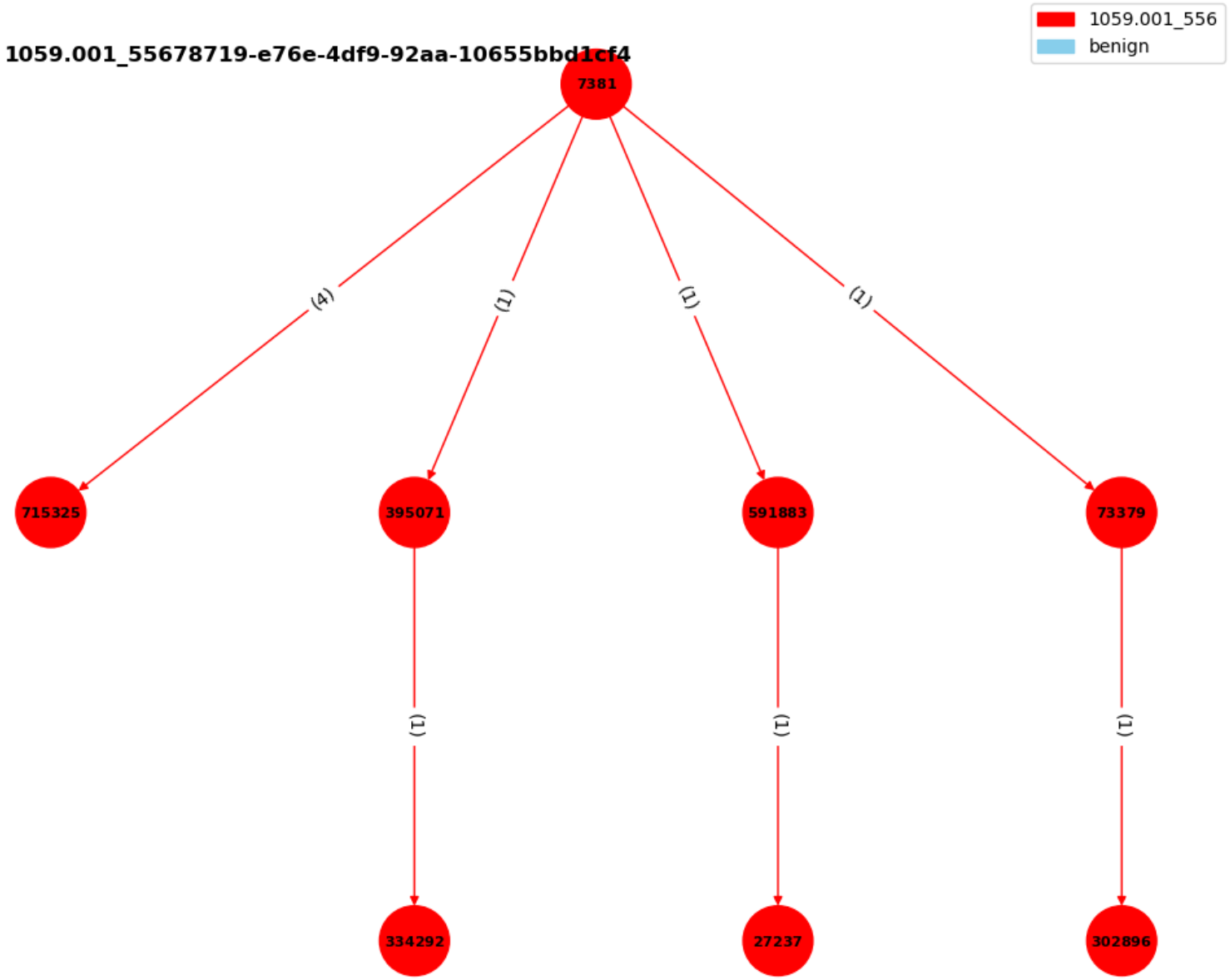
- Benign → set to 0
- Filtering the T1046_5a4 and T1005_720

- Packages be used in my graphing.ipynb

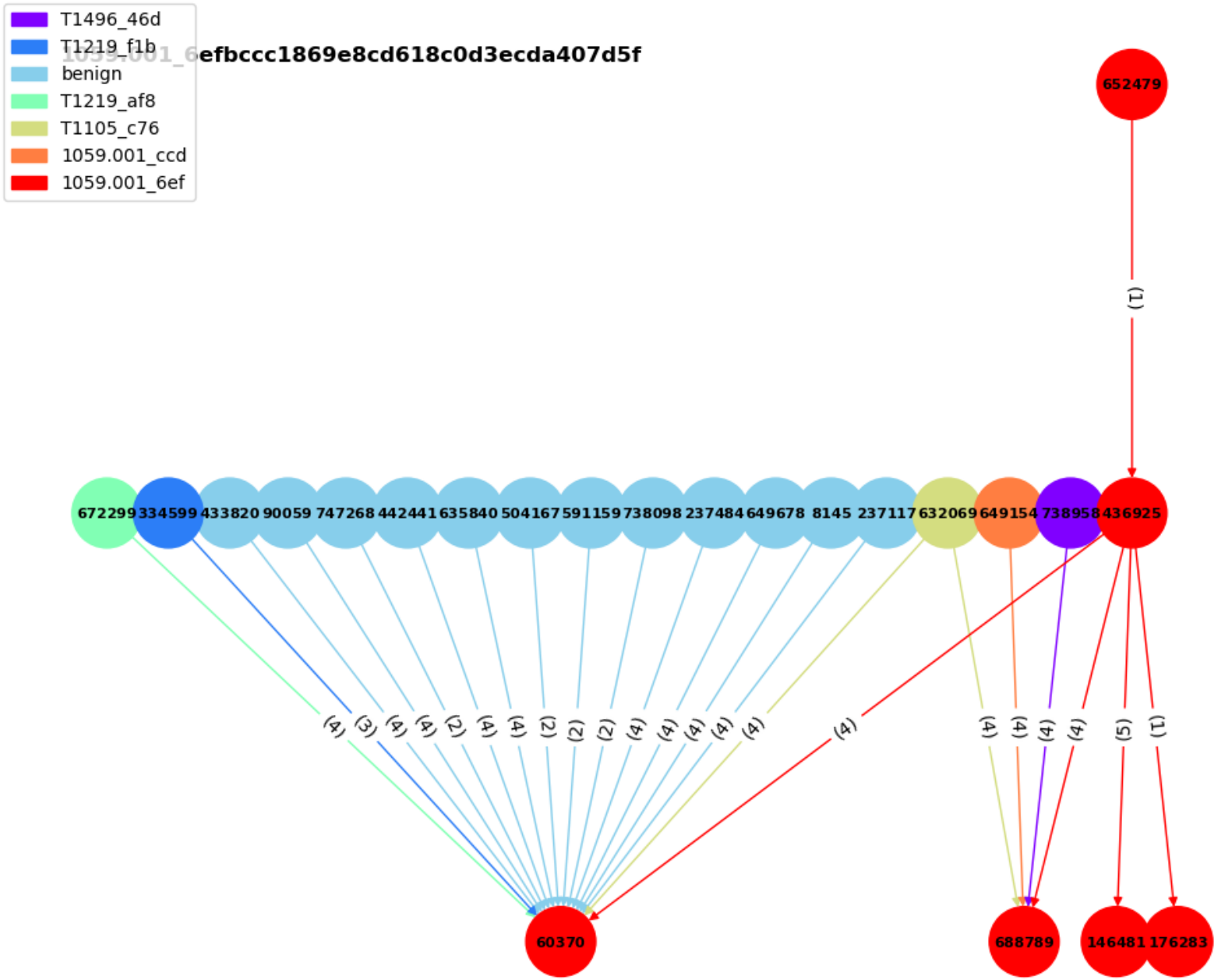
```python
import os
import numpy as np
import networkx as nx
import matplotlib.cm as cm
import matplotlib.pyplot as plt
from networkx.drawing.nx_agraph import graphviz_layout
import matplotlib.patches as mpatches
```

Example 1

- Main graph is red
- Number on the edges is the # of the relations in the pair
- No other related AP nodes
- No other related benign nodes



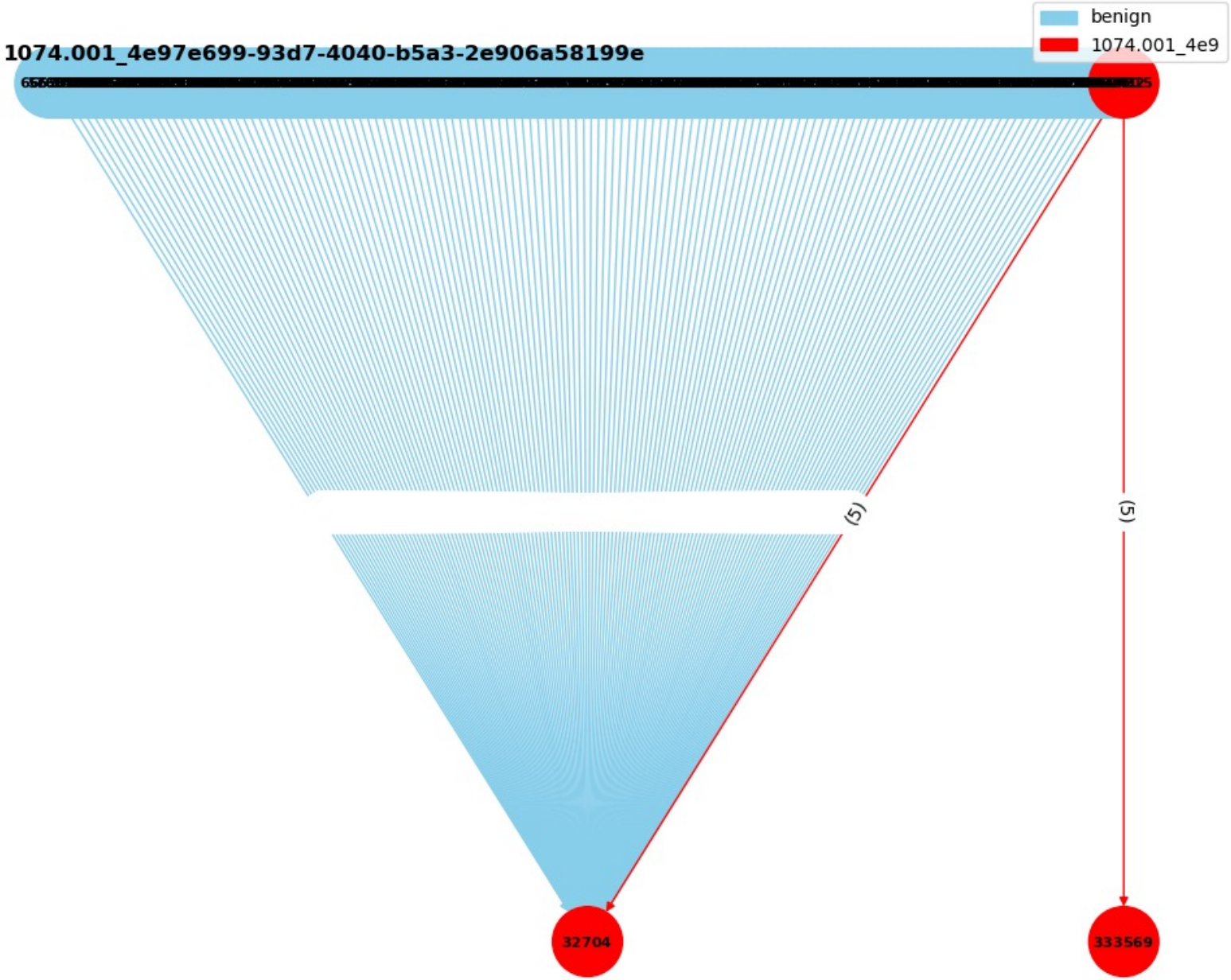
Example 2



- Main graph is red
- Number on the edges is the # of the relations in the pair
- 12 related benign nodes
- Other 5 related Aps

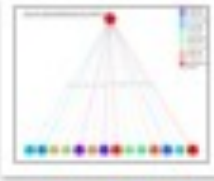
Example 3

- Main graph is red
- Number on the edges is the # the relations in the pair
- A lot of related benign nodes
- No related Aps

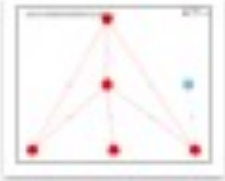


Overall

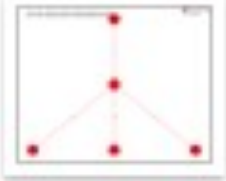
- Many of them have no neighbor nodes.



1518.001_b8453a5fe06b2...d3a.png



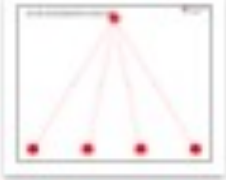
1546.013_f9a968af61d36...e17.png



1547.001_0dbdf1a2a87e71...fac.png

1547.004_0856714c9810...8a0.png

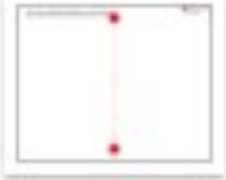
1547.004_aa147165f6c11...8ce.png



1547.009_501af516bd8b2...861.png

1562.002_6a8d25d65a7...1e6a.png

1562.002_94f51bf01a703...669.png

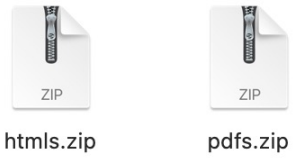


1562.004_5b93df032e23...77d1.png

TRAM

CTI report

Data Format



- PDFs (2.16 GB):
  - Successfully uploaded: 111 files
  - Unsuccessfully uploaded: 19 files
  - Haven’t exported
- HTMLs (15.94 GB):
  - I haven’t tried
  - Maybe directly try on the USB

Job: Analyze Malware-Madness-EXCEPTION-edition.pdf By: djangoSuperuser on 2023-23-25 15:23:16 UTC		Error		
Job: Analyze Hive-Analysis-Study.pdf By: djangoSuperuser on 2023-23-25 15:23:16 UTC		Error		
Bootstrap Training Data By: pipeline (manual) on 2022-06-04 01:05:13 UTC	Analyze Export	Accepted	Accepted: 12588 Reviewing: 0 Total: 12588	
Report for MOLERATS-IN-THE-CLOUD-New-Malware-Arsenal-Abuses-Cloud-Platf.pdf By: djangoSuperuser on 2023-07-25 15:22:16 UTC	Analyze Export Download	Reviewing	Accepted: 0 Reviewing: 112 Total: 112	
Report for Suspected-Iran-Nexus-TAG-56-Uses-UAE-Forum-Lure-for-Credenti.pdf By: djangoSuperuser on 2023-07-25 15:22:24 UTC	Analyze Export Download	Reviewing	Accepted: 0 Reviewing: 120 Total: 120	

Future Work

Future Work

- **Graph - Data Analysis**
  - Tackle the extreme cases
  - Trace back to more than one hop (till the end)
  - Different entity with different shape
- **Graphormer**
  - Input the data
  - Train the model
- **TRAM**
  - Figure out the reasons for the errors
  - Try to export
  - Try HTMLs

Thanks!!