**Project-2 Report**

1. What model or algorithm you use?

**Ans:** 我使用scikit-learn這個python套件內的decision tree model。

1. What features/rules you used for your model?

**Ans:** 在每個檔案內，我隨機挑選100個data當作一個單一的entry，去統計這些data的\_score值、source ip/port, destination ip/port的種類有沒有超過指定的threshold，若有超過threshold則標記成1，反之則為0，而他們的label則根據他們所在的檔案決定。以下提供一個例子：

在檔案DDoS.json中我挑選了100個data，這些data的\_score值以1為最多，且source ip的種類超過20種，source port的種類超過10種，destination ip的種類沒有超過20種，destination port的種類沒有超過10種，則這100筆data會形成1個feature\_entry = [1,1,1,0,0], label = “DDoS”。

以上這個步驟皆會在每個檔案中做至少50次來形成足夠多的dataset。

1. Why do you select them?

**Ans:** 檔案裡面有很多feature可以選，但是我發現\_score=1的只有其中兩個檔案，其他都是=0，所以我把它涵蓋進來。另外我觀察在這些網路攻擊中，source ip/port, destination ip/port似乎是最重要的，例如在DDoS中destination port就一定是22，而在RDP\_bruteforce中，destination port就一定是3389。 而套用在Test集上的效果也不錯，所以我就這樣選了。

1. Anything interesting things you find or problems you encounter.

**Ans:** 在這個問題中比較困難的地方可能是我模型中feature\_entry的產生會需要自訂threshold來表示說可能destination port的種類過多或是過少，這個threshold有點難選擇出來，不過我實驗了幾個值後目前對IP類型使用總數20%的threshold和port類型使用總數20%的threshold，應該是比較好的一個值。

1. Result/Accuracy

**0**: IP\_scan, **1**: port\_scan, **2**: DDoS, **3**: RDP\_bruteforece, **4**: C&C

