網路流量異常偵測分析-以 TWAREN 為例

陳品瑄　陳俊傑　梁明章
財團法人國家實驗研究院國家高速網路與計算中心
{hsuan, jjchen, liangmc}@narlabs.org.tw

摘要

雖然網路帶來許多便利性，但同時也潛藏著許多攻擊。目前網路攻擊問題日益被企業和大眾所認識並且關注，大規模的掃描攻擊、SYN Flooding 、DDoS 網路攻擊等已經成為了網路安全最大的威脅，同時也是網路管理者們心頭最大的一根刺，網路安全問題儼然已成為關鍵議題。本研究收集在 TWAREN 高品質學術研究骨幹網路上的 NetFlow 資料為實驗研究資料分析，結合資料探勘和異常偵測過濾分析，來建構網路流量異常偵測分析系統，透過網路特徵屬性特性，選取關鍵網路攻擊特徵統計分析，使其能降低計算複雜度以便縮減資料提高效率，找出其行為規律性等規則，藉以分辨是否為駭客入侵或攻擊的行為，找出異常網路流量，進而追蹤分析異常，不需查看封包的 payload，就能找出異常行動或潛在的攻擊行為，找出重大嫌疑者，並且對於異常攻擊 IP，快速定位攻擊者所在地(國家、城市)，利用攻擊來源 IP 經緯度以視覺化方式呈現在攻擊地圖上，以便管理者可以快速追蹤攻擊來源。

關鍵詞：NetFlow、TWAREN、資訊安全、網路攻擊、網路流量。

1. 前言

隨著網路及科技的快速發展，網際網路應用愈來愈普及的環境下，網路安全也成了現今網路應用和系統的重要議題。網路攻擊是個難以防範的威脅，只要主機連上網際網路，就有機會被駭客鎖定為攻擊目標的可能。網路攻擊通常具有多種形式，最基本的目的是要讓頻寬消耗殆盡，讓一般的使用者無法使用某台機器或網路資源。同時在資訊量不斷以驚人速度快速成長的情況下，如何將大量的資訊做適當的整理並快速分析，從中挖掘出有用的知識，便是當今重要的課題。在網路攻擊中，單從個別的 NetFlow 紀錄很難去判斷存取的是正常的使用者或是攻擊的駭客。TWAREN[1] 目前為台灣最主要的研究網路，連接全球研究網路及各縣市主要國立大學區網中心(GigaPOP)。正因如此，我們並沒有透過骨幹網路查看所收送的資料封包內容的權力，所以只能透過網路流量和流量特徵屬性數據資料分析。因此本研究以 TWAREN 骨幹網路 NetFlow 為實驗資料，建立一個實驗環境，不需查看封包的 payload，結合資料探勘 WEKA 工具對蒐集到的資料集做初步分析，再透過自行開發流量異常功能相關統計分析模組過濾分析找出重大嫌疑者，透過視覺化攻擊地圖界面呈現給使用者。

圖 1 TWAREN 骨幹圖

2. 相關議題

2.1 WEKA

WEKA[2]為 University of Waikato 的 Machine Learning Group 所開發的開放原始碼軟體，由 JAVA 開發的資料探勘軟體，具有資料的 Pre-Processing、 Classification, Regression、Clustering、Association rules、Visualization 等各類的功能及演算法實作。此工具目前普遍被運用在學術與商業用途上。WEKA 是使用 JAVA 語言編寫的資料探勘機器學習開放式原始碼軟體，因此幾乎可以使用在任何平台，如 Linux、Windows 等作業系統。同時也提供了多種機器學習演算的方法和具備資料探勘實驗完整的流程，包括輸入數據、評估模型、視覺化的輸出資料。具備多樣化的功能，可提供使用者針對不同的問題選擇最適當的演算法。圖2為 Weka 軟體開啟選單之主要介面，支援資料輸入檔案格式:

- ARFF : ARFF 是一種 WEKA 專用的檔案格式，副檔名的格式為.arff
- CSV 格式: .csv
- C4.5格式: .data / .names
- 序列化的實例格式: .bsi

並提供5種主要的應用程式供使用者選擇:

- Exployer : 主要使用的圖形使用者介面，包括預先處理、分類、分群、關聯、屬性選擇以及提供視覺化的界面。
- Experimenter : 實驗工作環境，使用者可以用來比較不同的實驗環境和演算法及管理演算法方案之間的收集統計檢驗資料。
- Knowledge flow :使用者可以自定資料處理流程的方式和順序，同時它也具有可以拖放的

介面。並且支持以增量方式來處理大量資料。

- SimpleCLI : WEKA 提供了一個簡單的命令列介面，因此，使用者可以在沒有具備命令列的作業也可以在系統中直接執行 WEKA 命令。
- Workbench :包含了其他應用程式的組合。

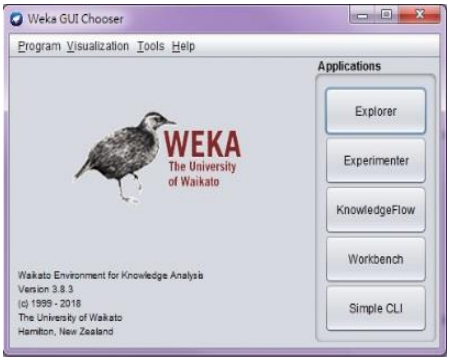

圖 2 WEKA 選單介面

2.2 入侵偵測系統（Intrusion Detection System, IDS）

現今各式各樣攻擊模式與手法不斷地出現，同時目前整體網路流量大幅增加下，相對亦加深網路即時監測的難度。入侵偵測系統(Intrusion Detection System, IDS)，是目前最常用於保護網路安全，避免遭受到外來的網路攻擊，主要收集主機系統內訊息或網路流量，再進行分析比對，以判別是否遭受入侵攻擊。從分析技術來看，IDS 主要可分為誤用偵測（Misuse）及異常偵測（Anomaly）兩種，誤用偵測是使用特徵比對（signature based）的方式，現今常見的入侵偵測方式大多也是以比對特徵或是以資料探勘為基礎來建構系統。目前主要使用的兩種技術說明如下:

- 特徵比對（Misuse detection）
  特徵比對(Misuse Detection)的重點在於如何定義與建立大量的攻擊特徵，判斷方式是依照已知的攻擊特徵資料相同或類似時，則判斷為入侵行為，但是如果攻擊行為不存在資料庫中，就無法偵測，因此必須不斷更新資料庫的特徵資料，才能有效的偵測新的攻擊行為。網路攻擊特徵選取重點在於如何利用多種特徵之間的關連性,選擇正確且合適的特徵，特徵選取好壞影響分類的好壞,分類好壞又影響網路流量分析的正確性跟效能
- 異常偵測(Anomaly Detection)
  異常偵測(Anomaly Detection)則是建置一個正常行為的系統模式或是定義正常的標準值，如果超出先前訓練得到的正常統計範圍之外，則判斷為入侵行為，但是由於正常狀態的不易界定，難以辨別良好流量及惡意流量，容易產生誤判，誤判率將大幅提高。

3. 系統架構與實作

3.1 實驗資料

本研究使用 TWAREN 骨幹網路上路由器收集到的 NetFlow 資訊作為實驗數據資料來源，資料中每條紀錄代表一個 NetFlow 資料。每一筆 flow 是依相同的來源 IP 位址(source IP address)、來源埠號(source port number)、目的 IP 位址(destination IP address)、目的埠號(destination port number)、協定種類(protocol type)、服務種類(type of service)、及路由器輸入介面(router input interface)的封包資訊，透過以上七個欄位的封包，來判斷這個封包是否屬於任何已記錄的 Flow，有的話則將新收集到的封包的相關流量資訊整合到對應的 Flow 記錄中，其 bytes 數和 packets 數都會累計記入該 flow 中，如果找不到封包對應的 Flow 記錄，便產生一個新的 Flow 記錄來儲存相關的流量資訊。在 NetFlow 資料取得方式，TWAREN NOC 維運團隊開發使用以 ElasticSearch Cluster 即時儲存 NetFlow 資料，透過 ELK Stack 來確保高效率且穩健的運作，整合不同資料來源(如 Syslog、NetFlow、SNMP Trap)。架構是基於 ElasticSearch(E)、LogStash(L)[1]、Kibana(K)所組成如圖 3 所示。可以利用基於 HTTP Protocol，以 JSON 為資料交互格式的 RESTful API 來取得資料。可以使用 JAVA 或其他語言使用 RESTful API 利用 9200 port 和存取 Elasticsearch，也可以使用 curl 命令存取操作 Elasticsearch。

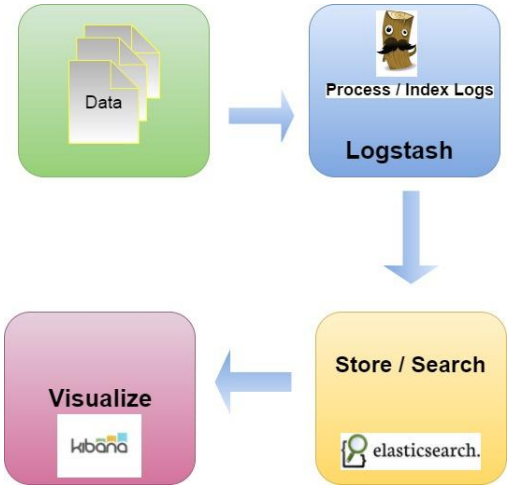

圖 3 ELK

為了確保資料安全， NetFlow 資料也會定期備份。NetFlow 的資料是編成二進位(binary)的形式儲存，而不是以純文字的方式，無法直接了解儲存的內容，因此，要取得資料分析時，必須先透過 nfdump 轉換 NetFlow 之資料型態，而其轉換的依據，則是根據 Cisco 的 NetFlow 定義而定，如圖4 所示。

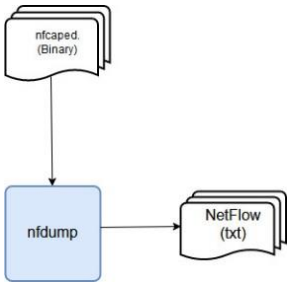

圖 4 Netflow Processing

轉換後每筆 NetFlow 數據資料屬性與說明如表1所示。

表1.NetFlow 格式

屬性	說明
Src IP Addr	來源 IP 位址
Dst IP Addr	目標 IP 位址
Src Pt	來源通訊埠
Dst Pt	目的端通訊埠
Date	時間戳記
first seen	Flow 起始時間戳記
Duration	Flow 結束時間戳記
Flags	連線旗標
Tos	封包服務類型
Packets	該 Flow 累計傳送的封包數量
Bytes	該 Flow 累計傳送的 Byte 數量
Flows	Flow 總數

3.2 系統架構

系統架構如圖5所示。本系統建置實驗環境並使用 Perl 程式及相關的 Library 開發分析相關模組。系統在 TWAREN 骨幹網路上收集到的 NetFlow 資料，以 TWAREN 骨幹網路下的區網單一學校為研究對象，資料會先透過 PreProcess Module 處理 NetFlow 資料，因為產生資料或是轉移資料時皆有可能造成資料格式錯誤或導致資料遺失等問題發生，因此先清洗過濾，並轉換資料格式形態，將 NetFlow 資料轉成 WEKA 支援的檔案格式。因為 NetFlow 資料量龐大，所以每增加一個特徵屬性，就會增加計算時間和效能，為了避免不必要的特徵屬性影響分析結果，因此如何以最少的特徵屬性來代表整筆 NetFlow 資料也是一個重要的課題。因此，為了避免不必要的特徵屬性影响分群結果，我們同時也從原有的 NetFlow 特徵中，選取網路攻擊關鍵特徵，使其能降低計算複雜度，增進系統效能，以達到落在同群的屬性彼此會有較高的相似度，而落在不同群的屬性則會有較高的相異度。研究中使用跨平台的免費軟體 WEKA[2]，將資料代入 WEKA 中運算，統計分析網路流量特徵屬性，藉由 WEKA 由雜亂無章的資料內萃取所需資訊，推導過程是採用訓練及測試綜合方式推導，WEKA 會先將部份的資料雖機抽樣，進行訓練模式推導，再將剩餘資料代入訓練模式，計算出測試資料。利用 K 平均值（K-means）分群演算法對 NetFlow 資料進行統計分析，再調整其操作參數，檢視其是否達到可接受的準確率。K 平均值演算法是一種常用的分群分析演算法。該演算法將 n 個資料都分到 K 個群之中，使得分群結果滿足，同一分群中的資料相似度較高，而不同分群中的物件相似度較小。

K-means 主要概念步驟:

輸入: K 個 clusters 及 n 個資料

輸出 : 滿足標準 k 個 clusters

1. 從 n 個資料中選取 k 個為群心
2. 以 Euclidean distance 作為相似度測度，也就是將每個資料和 k 個群心計算 Euclidean distance
3. 將資料歸類到距離最近的群心
4. 每個資料歸類到距離最近的群心所屬 cluster 後，再去重新更新群心
5. 重複步驟3至步驟5後，直到 clusters 不再變動

分群後的 arff 資料，再利用 Post-process 模組轉換分群資料及資料處理，再透過分析模組，分析統計排序找出 TOP-N 的可疑目標和異常臨界值比對過濾。過濾出的重大可疑 IP，並透過 IP 所取得的經緯度位址視覺化的呈現在網站上，利用視覺化攻擊地圖呈現攻擊來源及相關資訊提供給維運人員。

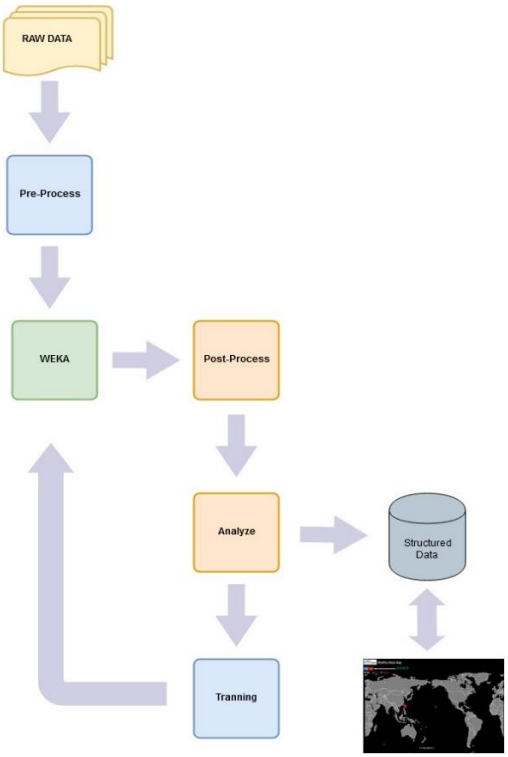

圖 5 系統架構

4. 實驗結果與分析

4.1 評估指標

在 TWAREN 骨幹網路中，我們並沒有辦法也不需要明確捉到每一個網路異常使用者，在骨幹維運上首先關注的是會真正影響到 TWAREN 骨幹或各學校連線單位網路大量流量的嫌疑者，也就是，在網路真正異常使用衡量上是以異常使用量為主要偵測要點。因此，我們會再進行調整設定臨界值(Threshold)與其比較過濾，並透過 TWAREN 骨幹維運經驗值設定臨界值，來決定是否要將此使用者視為嫌疑者。同時，本系統也會將偵測分析到的攻擊者也將參考本中心吸引駭客發動攻擊誘捕系統（Honey Pot）所蒐集攻擊者 IP 等惡意來源名單結果輔佐驗證。

4.2 實驗結果

研究中使用 WEKA 工具載入 NetFlow 資料後，並調整相關參數值，每筆 NetFlow 資料的各個屬性都可以透過 Weka GUI 界面工具個別表示和統計分析如圖6、圖7所示。

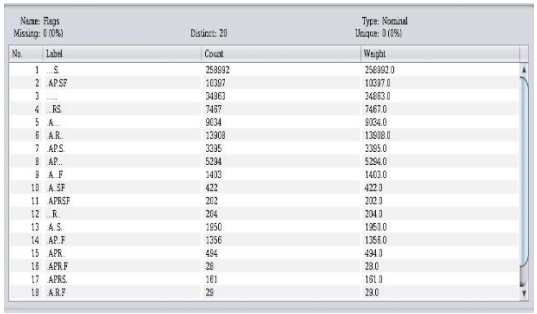

圖 6 Source Port Preprocess

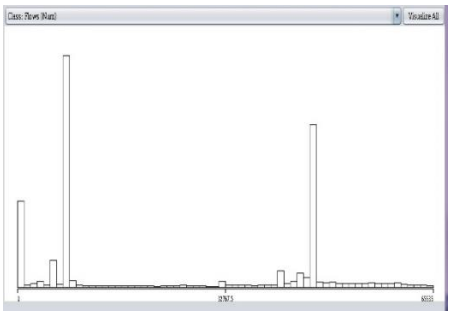

圖 7 Source Port Analysis

然後在 WEKA Cluster 面板中選擇使用 K-means 演算法並設定相關參數，分群的結果可透過分群面板得知。分群結果以表格形式呈現，每一行對應屬性名稱，列則對應分群結果。並且透過分群的面板可知每個分群的數量及百分比。分群結果，透過 WEKA 產生 ARFF 格式資料如圖8。其主要分為兩個部分，Header 及 Data。Header 區塊主要由 @RELATION 及@ATTRIBUTE 組成。@RELATION 與檔名的意義相同，主要用於辨識資料集意義； @ATTRIBUTE 則用來宣告在此資料集中，每個樣本資料的各項特徵屬性名稱及其表示格式。本研究中 NetFlow 資料包含了15種屬性。

```
@relation Neflow_clustered

@attribute Instance_number numeric
@attribute Src_IP_Addr {223.68.210.154}
@attribute Dst_IP_Addr {140.133.32.8,140.133.32.16,140.133.32.2,140.133.32.0,140.133.32.54}
@attribute Src_Pt numeric
@attribute Dst_Pt numeric
@attribute Date {2019-02-28}
@attribute first_seen {17:31:21.569,17:31:21.571,17:31:21.574,17:31:21.575,17:31:21.577}
@attribute Duration numeric
@attribute Proto numeric
@attribute Flags {....S.}
@attribute Tos numeric
@attribute Packets numeric
@attribute Bytes numeric
@attribute Flows numeric
@attribute Cluster {cluster0,cluster1}

@data
```

圖8 ARFF 格式

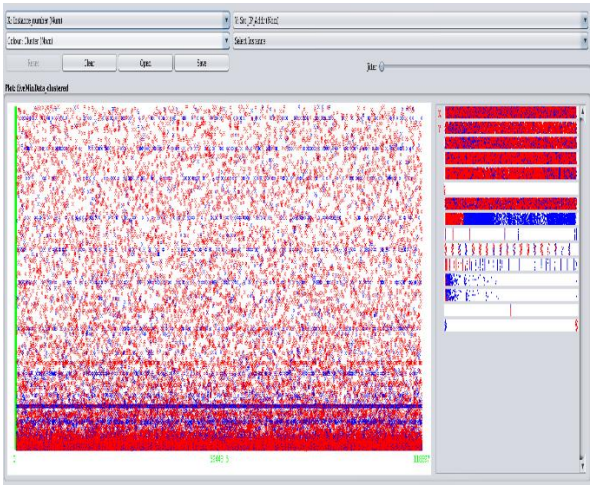

圖 9 分群結果視覺化

以本系統資料分析結果就有發現在攻擊流量中，藉由特徵屬性行為統計分析明確找出重大嫌疑者 IP 有 SYN flood 攻擊，這種攻擊主要是利用 TCP 三向交握的機制來做攻擊。也就是攻擊者惡意的只送出 flag 旗標為 SYN 封包到受駭者主機，但卻沒有等待後續的連結，也就是並沒有相對應數量的結束 flag，導致受駭者儲存太多等待連結資訊而超過負荷，停止提供服務。

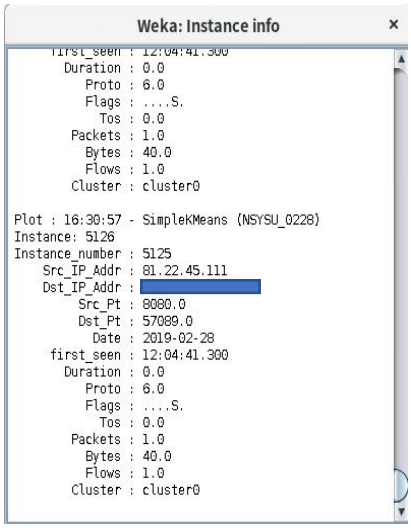

圖 10 Instance info

為了維運者維運需求，以達到提供的高效維運和管控，提供視覺化的監控畫面，取得偵測異常 IP 經緯度，即時定位攻擊者所在地(國家、城市)，以網路地圖方式呈現攻擊來源，將網路攻擊視覺化如圖11，並透過圖形不同的大小呈現影響嚴重性等級，並也可以透過點選可以取得包括攻擊的來源、受駭的目的地、通訊協定的分析等資訊。

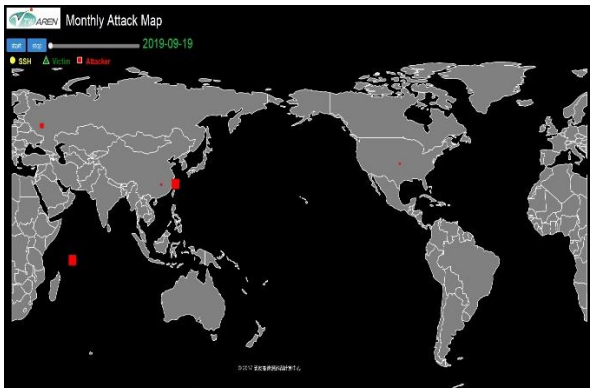

圖11 網路攻擊視覺化

5. 結論與未來展望

利用網路可以拉近人與人之間的關係，增加彼此的互動關係，但同時也有可能因為遭受網路攻擊而使服務中斷或資料被竊取。網路攻擊的常見方式是利用大量異常的連線或破壞封包來佔用網路系統可用的資源，藉以阻擋正常的使用連線，甚至癱瘓網路服務。

在網路攻擊中，單從個別的 NetFlow 紀錄很難去判斷存取的是正常的使用者或是攻擊的駭客，並且，我們也沒有透過骨幹網路收送的資料封包、查看封包內容的權力。因此，本研究藉由收集 TWAREN 骨幹網路上的 NetFlow 資料，透過網路流量和屬性數據資料分析，結合資料探勘技術工具 WEKA 來做初步分析網路流量特徵，辨識正常流量和攻擊流量，進而追蹤分析異常，再以循序相關分析找出頻繁的使用者行為，模式，經解讀這些行為模式後，統計分析排序找出異常使用行為模式，再藉由維運經驗臨界值及用量過濾，找出嚴重影響到 TWAREN 骨幹網路上的可疑 IP。同時為了維運者方便查看，開發視覺化的工具就有其必要性，因此，提供網路攻擊地圖視覺化讓網管人員更容易知道攻擊相關資訊，提高網路可性，以確保大多數合法用戶的使用。當然未來仍有許多可努力的方向，包括如何實際佈署到 TWAREN 骨幹網路上，以及偵測透過其它屬性或是其它網路攻擊行為模式規則建立分析，有效利用網路流量的進出分析來輔助網路管理者能即時防制不法的攻擊，以達到網路安全控管機制。

參考文獻

[1] TaiWan Advanced Research and Education Network. (http://www.twaren.net/)。

[2] WEKA, http://www.cs.waikato.ac.nz/ml/weka.

[3] CVE,http://cve.mitre.org/

[4] Manjula Suresh, R. Anitha, “Evaluating Machine Learning Algorithms for Detecting DDoS Attacks”, in 4th international Conference on Advances in Network Security and Applications(CNSA), pp. 441-452, 2011.

[5] S.Behal, G.K.Ahuja, “Detection of DDoS Attacks using Weka Tool: A Case Study”, Proceedings of National Conference on Computing, Communication & Electrical Systems,Nov. 2017.

[6] Y. Gu, K. Li, Z. Guo, and Y. Wang, “Semi-Supervised K-Means DDoS Detection Method Using Hybrid Feature Selection Algorithm”, IEEE Access, vol. 7, pp. (15 pages), January 2019.