資安監控巨量資料分析-以 G-SOC 建置為例

高天助 劉培文 趙偉傑 沈裕翔 劉上菱 李兆文 李國禎 毛敬豪 朱宇豐
財團法人資訊工業策進會 資安科技研究所

摘要

隨著進階針對性攻擊日益嚴重，現有的資安監控所蒐集資料的縱深與廣度不夠，無法偵測到針對式攻擊的潛在特徵，現有監控的來源（如：防火牆、入侵偵測系統及防毒軟體等）已不足以呈現攻擊特徵。為了能提升資安監控的效益，擴大收納日誌的資訊，並深入收納偵測事件類型，方能提供後續樣態分析更多的行為屬性參考， 在此同時，需考量到監控的成本，包含：資料倉儲資源的需求、分析運算平台的需求以及跨網路資料傳輸需求等。現有資安監控與事件管理平台，缺乏橫向擴充的能力，因此關聯事件無法關聯時間超過天與週以上的事件，而異質事件的樣態與關聯規則亦缺乏有效探尋的方法。本研究試著從資安監控切入，說明資安分析即服務之監控平台(Security Analytics as a Service)架構，並著重於技術分析的三個核心模組：Self-Inspection Modules、Threat Recognition Modules 與 Prioritization, Predictions Decisions Modules。本研究透過政府資安 G-SOC 二線監控平台-收容各個不同來源 SOC 資料，作為分析的案例進行探討。

關鍵詞：資安監控、資料分析、異質性、巨量資料、跨時域

壹、前言

現在資安威脅已從大規模非針對式的攻擊(如：掃描、阻斷服務等)，轉變成具備長時間潛伏、樣態不顯著的針對式攻擊行為，由於針對是攻擊缺乏充足樣本，造成無論在特徵偵測上或是在行為分析上的困難，而資安監控更不易單從網路行為找出針對式攻擊的特徵，往往只能從惡意程式的動靜分析培養出連線中繼站的黑名單、或是從靜態分析中找出惡意程式的特徵戳記，並且配置到入侵偵測系統產生事件，然而中繼站黑名單更新速度快速，透過網域指向的黑名單變動性更高，難以藉由純粹比對涵蓋所有攻擊的可能性。因此，資安監控需持續擴大日誌與事件搜集與分析的範圍，並且加深其分析與建模的深度，因此需要對於異質事件提供傳輸、聚合、倉儲、關聯分析與塑模以及風險威脅整合判斷的系統化機制。

由於攻擊與威脅越趨複雜，資安營運中心(Security Operation Center, SOC)監控的日誌資料種類已不侷限在資安網路防護設備所偵測的事件，更需著重在網路設備日誌、應用服務伺服器的日誌以及企業在雲端上的日誌資訊，其中，根據 Gartner 的報告指出，在 2017 年大約有 25%的企業流量將直接繞過現在企業網路的界限，而直接是連到雲端服務系統上，例如：企業可能使用 Google Gmail，因而客戶端的郵件服務將會繞道溢出企業的控管。此外，社群媒體的興起，企業或組織成員大量使用社群媒體，也有許多外部的

資安最新威脅弱點透過社群媒體散布，因此進行資安監控亦需涵蓋資安社群媒體的資訊萃取與分析。在未來 SOC 運作上，資料來源除了既有的資通安全設備所產生的事件外，更需結合各式應用層級的服務、雲端應用的服務及社群媒體關於內部使用以及外部威脅的情資進行整體的分析。

在 SOC 運作分析需求，包含：大量黑名單比對、跨網路與主機惡意行為關聯、跨來源與組織威脅關聯、跨時間與區域威脅關聯、社群情報預警、跨非(半)結構化與結構化關聯等需求項目。在大量黑名單比對方面，黑名單往往是阻擋的第一層防線，而黑名單隨著攻擊威脅變動性與日劇增，如何動態索引與維護大量的黑名單資訊，成為重要的議題。此外，在跨網路與主機惡意行為關聯、跨來源與組織威脅關聯與跨時間與區域威脅關聯三個需求方面，透過所蒐集到的歷史事件與各類型服務日誌，利用不同時間範圍及對應的各項日誌屬性組合，關聯出核心分析規則。而外部情資更可提升 SOC 防禦時的因應能力並降低突發威脅所造成的衝擊，因此在社群情報預警與跨非(半)結構化與結構化關聯分析上，需結合不同結構的資料類型，並透過不同量化的分析方式進而統整出合適的資安情資資訊。

由於 SOC 除了考量複雜的異質性資料，更需要處理所謂大量日誌產生後續倉儲、傳送蒐集以及關聯分析的議題，而這符合巨量資料[15]所定義資料的四個特性，包括：大量(Volume)、快速性(Velocity)、多樣性(Variety)及複雜性(Complexity)。應用巨量資料可規模化的儲存與運算框架至 SOC 服務，可增加事件與日誌蒐集的類型、提供不受時間區間限制的分析模型、找出不同範圍、概念層級與跨來源的威脅樣態。為了要達到這樣的目標，其核心系統架構需能支援大量、結構化、半結構化與非結構化的資料類型，且資料存入快速索引並且能依照不同面向進行快速提取與屬性萃取，而當事件要進行關聯時，可針對不同時間、不同面向進行運算，關於 SOC 與巨量資料分析的需求如表一所示。

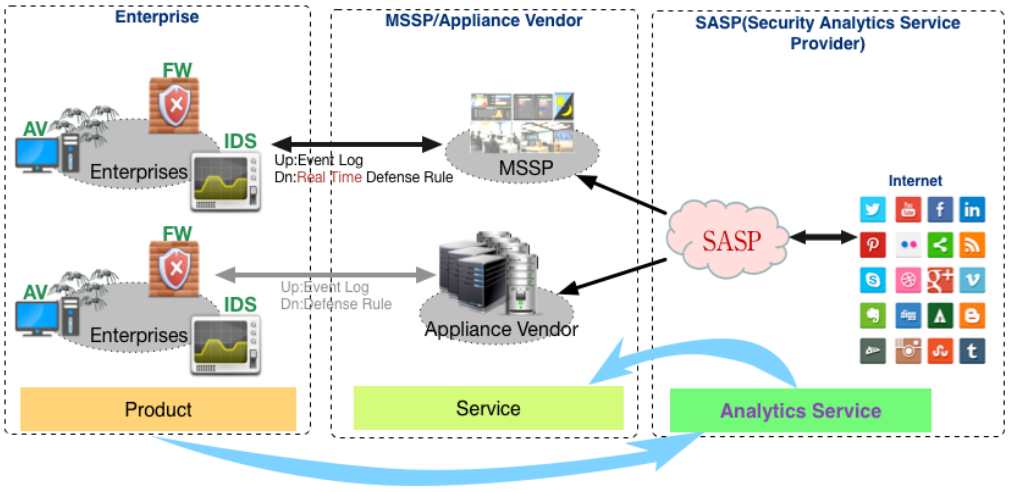
表一：SOC 與巨量資料分析的需求分析表

	大量性	快速性	多樣性	複雜性
大量黑名單比對	高度需求	高度需求	低度需求	低度需求
跨網路與 主機惡意行為關聯	高度需求	中度需求	中度需求	高度需求
跨來源與 組織威脅關聯	高度需求	中度需求	中度需求	中度需求
跨時間與 區域威脅關聯	高度需求	中度需求	中度需求	中度需求
社群情報預警	中低度需求	中低度需求	中高度需求	高度需求
跨非結構化、半結構化與結構化關聯	高度需求	低度需求	高度需求	高度需求

本研究應用巨量資料用於 SOC 的分析場景，依威脅分析的階段，定出不同的場景，依照相關的分析演算法，如：分群、分類、協同過濾、關鍵屬性萃取等演算法[16]，建構出關於 SOC 分析的三個技術分析三個核心模組，分別為 Self-Inspection Modules、Threat Recognition Modules 與 Prioritization, Predictions Decisions Modules，本研究的核心在於建構出 SOC 事件關聯場景、核心模組與運算演算分析的關聯性，並以政府 G-SOC 二線監控為例，呈現應用本研究所提出的資安監控巨量資料分析架構的實際運作方式以及成效，本研究第二章將對於相關文獻進行探討，第三章描述資安監控巨量資料分析架構，第四章則說明於 G-SOC 二線監控實際應用案例，最後第五章為結論。

貳、相關文獻

過去資安業者包含資安網通設備產品商以及資安服務提供商，其中資安產品商以銷售實體防火牆、入侵偵測系統或是防毒軟體等，而服務供應商則提供資安監控、弱點掃描等工作，而 Security Analytics as a Service 的架構同時需收納與傳送資訊給資安業者，並且結合社群媒體以及網際網路的情資，透過整體核心分析模組，產生網路情資給資安業者與服務提供商。



圖一：資安營運中心系統架構圖

為達到各式的分析方式，以下整合相關核心演算模組，藉由不同分析模組的特性，例如：以統計或是機率分布為基礎的異常偵測機制、以序列分析為基礎的異常偵測機制或是以圖形結構為基礎的異常偵測機制，這些偵測機制則可應用於資安監控巨量資料分析不同的分析場景以及分析階段。

Yen 等學者於 2013 提出一個名為 Beehive 的系統[14]，這個系統對雜亂的日誌提供多樣化的分析機制，由於企業環境的資訊系統配置具差異性，因此，Beehive 補強了以戳記比對為基礎的防禦方式，進而採用分析日誌中的可疑行為，透過不同的資料分析方式

塑模與呈現，該機制需區分在企業環境下的日誌，是屬於政策違反類型或是潛在的攻擊惡意行為，該研究除了採用群聚、分類以及相關塑模方法外，而其自動化的資料分析 3 個步驟，首先對於資料進行過濾、正規化透過網路設定相關屬性，接著從 15 種不同的日誌類型中，產生出屬性，最後透過不同的異常偵測塑模方式，找出異常的事件或行為，此研究對於大量資安分析的流程具有完整的描述。

而在異常偵測核心機制方面，巨量資料分析仍需整合多個不同特性的異常偵測模組，方能對於不同類型的攻擊行為進行偵測。Relational Pseudo-Anomaly Detection (RPAD) 學習正常行為的模型透過已觀察資料的樣本，並視這些資料為非異常資料，並且建構相同數量偽異常資料。偽異常資料是依照每個屬性的邊緣機率所構成的聯合機率所產生，當給予一個新的資料，RPAD 結合分類器的預測並考量到偽異常分佈來決定該資料是否為異常，此方法因為運算簡單因此非常具備效率[7][8]。Relational Density Estimation(RDE) 與 RPAD 不同之處在於假設屬性在獨立情況下的聯合機率，每一個邊際機率分佈透過估計核心密度 (Kernel density estimator)，並且其聯合機率被假設為邊際機率的乘積，而異常資料則會被呈現在聯合機率較低的資料點上。

混合式高斯模型(Gaussian Mixture)假設各個屬性的分佈屬於高斯分佈（類似常態分佈），若資料經過混合式高斯分佈乘積後，機率若偏低，則該筆資料偏離了大部份正常的行為，因此可判斷出該資料為異常，通常會採用 EM 演算對於該混合機率進行估算[2][10]，該模型適合對於符合高斯分佈的資料進行評估，然而若在混合高斯模型中，各單一高斯模型之間仍會存在一些差距，因此組合式的混合高斯模型被用來改善這個問題[3][13]，因此我們可透過混合高斯模型去學習使用者的正常行為。

對於序列分析來說，VSM[1]是一個結合 n-gram 與有限集合向量對於序列化資料進行異常偵測的一個演算機制，透過 VSM 這個演算模組，可對於固定序列的事件行為，考量事件順序，找出偏離共同行為的序列化資料[6]。時間為基礎的異常偵測 (Temporal-based AD) 著重對於不同時間解析度的資料進行塑模[9]，並透過序列分析的方式輔以過濾演算法找出具備高敏感度的資料。

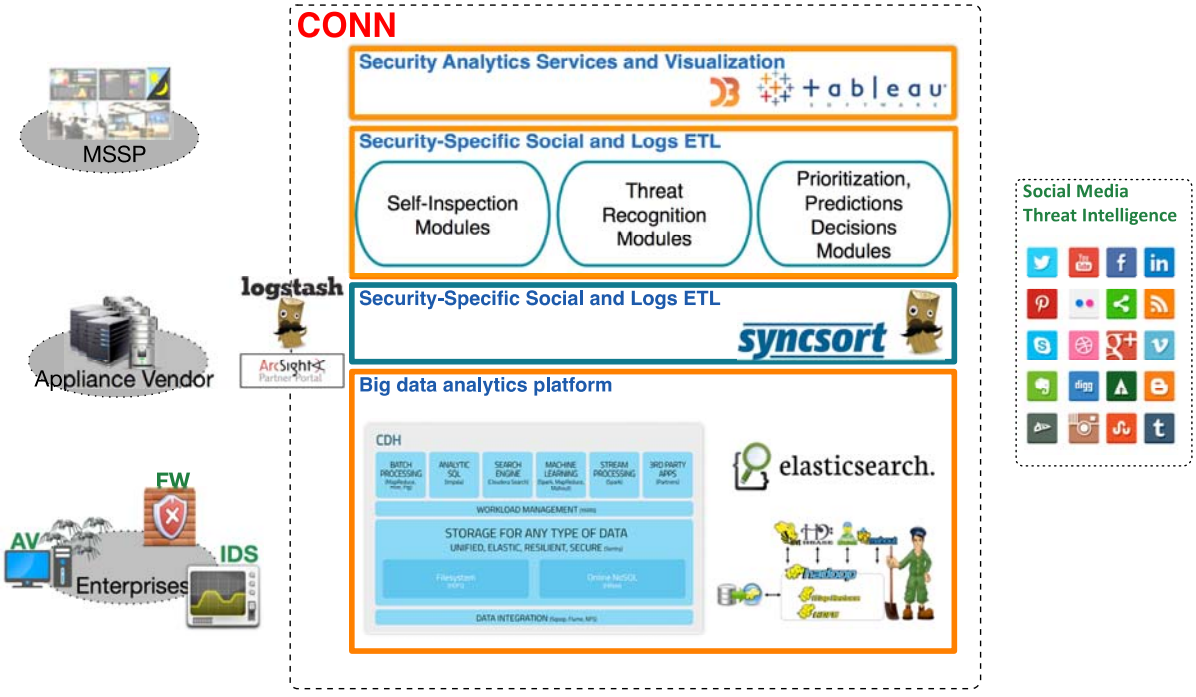
異常行為亦可透過網路關聯資料的結構所發現，分析靜態資料的社群群聚性，亦為一種在圖形結構上網路異常偵測的常見方法，此種特性常出現在具備群聚性的惡意社群行為中[12]，例如：互相哄抬的惡意垃圾郵件帳號，因此找出靜態社群群聚特性，為找出結構化惡意行為的一個重要指標，然而當在處理大量且快速演變的資料時，靜態社群偵測方法會跟不上持續進入資料，原因在於要持續進行較耗成本的運算。因此大串流的資料持續進入並且改變社群資料時，通常進入的資料只會改變暨有資料的群聚性，因此透過拆解掉原本的靜態圖形群組透過移除受到新進入資料而改變的節點，透過這樣的方式大幅降低了節點的數量，進而持續追蹤資料的群聚性並找出重大結構的改變[11]。STINGER 是一個動態圖形結構，擅長用於維護與呈現時間性與語意性的資料圖形結構，該圖形結構可具備百萬至億個連結，而每個連結可給予類別、權重、時間戳記以及實體的識別子。這個資料結構是和處理大量的圖形分析，STINGER 亦為一個動態圖形分析平

台，並可動態地維護圖形的結構，例如新增節點或刪除節點，並且支援快速地重新計算各種圖形的指標[11]。

在圖論的資料結構中，Betweenness centrality (中介中心度)往往是用來評估一個節點重要性的一個重要指標，其他過任兩點的最短路徑所經過該節點的次數進行計算，並對於節點進行所被經過的次數作為重要性的指標[4]。然而中介中心度在效能上因為要計算所有節點的最短路徑，因此當節點數呈指數成長時，其效能會大幅的衰減，故有研究採用有限制性寬度優先的最短路徑計算方式，借此大幅提升運算的效能，借由此種方法，確實大幅降低計算最短路徑的次數[5]。該種評估機制可被用來分析潛在洩露資料或是具備弱點的節點。

參、資安監控巨量資料分析

在資安監控巨量資料分析架構中，如圖二所示。在該架構中，包含具備可規模化資料倉儲與運算的分析的巨量資料分析平台(Big Data Analytics Platform)、提供異質性資料萃取轉換與載入(Extraction Transform and Load, ETL)的資安社群與日誌 ETL 中介層、資安監控資料分析核心模組以及資安服務及視覺呈現模組。其中，我們將著重在異質性資料日誌與資安社群 ETL 與資安監控資料分析核心模組兩個模組。



圖二：資安營運中心巨量資料分析系統架構圖

在資安監控資料分析核心模組中，可依照分析所採用方法的特性分為三個元件，分別為：自我檢測(Self-inspection)元件、威脅感知(Thread Recognition)元件與脅評估與預測

(Prioritization and Prediction)元件。這三個元件，除了其分析演算法性質不同，於 SOC 事件分析的階段也會有所不同，三個元件詳細資訊詳述如下。

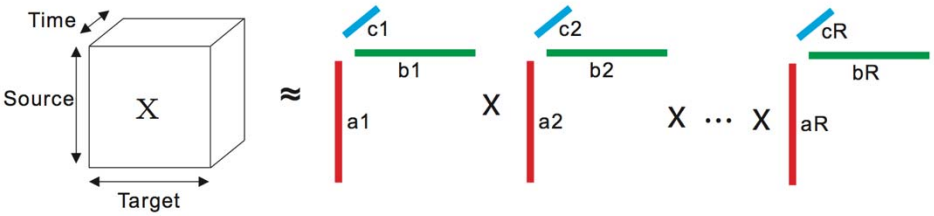
- 自我檢測(Self-inspection)元件

  該元件重要目的在於對於所蒐集的資安事件或日誌資訊進行資料蒐集以及服務配置上的檢視與檢測，其目的包含：未完善的配置、非技術面的配置議題及評估何種環境導向的屬性需求。

  自我檢測元件將以資料導向的方式(data-driven)的重點在於找出在統計特徵具有顯著異常的特徵，所採用的是以統計與樞紐分析的機制為主，此外會採用奇異值分解(Singular Value Decomposition)或是採用張量分解(Tensor Decomposition)等技術。

  在產生異質資料多維陣列後，我們需對於該多維陣列進行分解，本計畫採用張量的概念，可對於多維的資料進行分解，其分解採用 PARAFAC 分解法，可將張量分解為不同維度的向量，如圖三所示，可將原本的多維度矩陣，參考每個維度異質性資料的共相依程度進行分解成 a、b 與 c 向量，藉此分解出各個屬性的特性。

  分解後依照不同資料面向進行資料群聚以及異常偵測，本計畫預計採用 Apache Mahout 巨量資料分析套件中的 K-means 對於資料進行分群，並透過距離演算法找出偏離於原點的群，其中會採用共變異數的方式評估每一群的重要程度，進而排序出優先分析的群聚類型，每一個群聚皆代表潛在的惡意程式樣態。



圖三：PARAFAC 張量分解示意　(資料來源:本計畫自行整理)

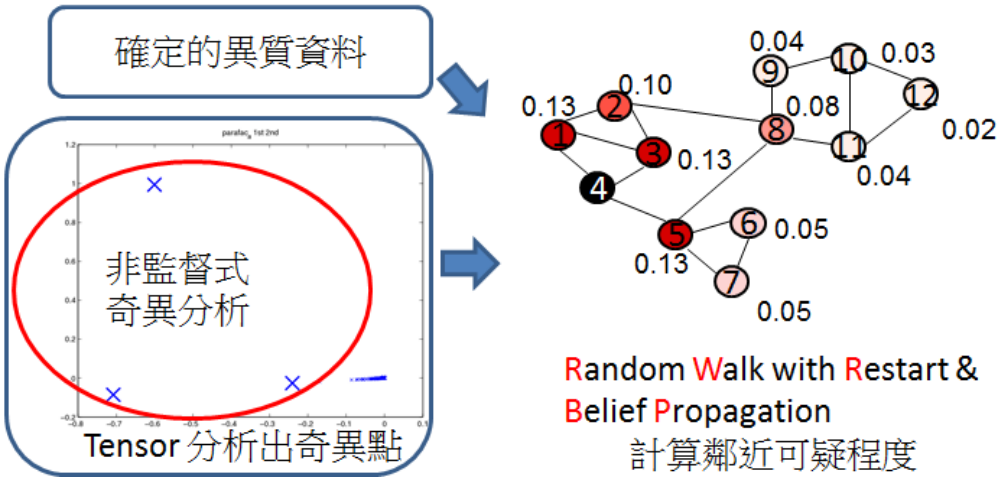
- 威脅感知(Thread Recognition)元件

  威脅感知元件是以資料導向的方式協助建立已知或是潛在威脅模型，其中需透過模式參數估計找出屬於不同環境的基礎參數，並且依照不同威脅形態產生模型群組，該元件的目的在於找出多步驟或是複合式行為的威脅行為。

- 威脅評估與預測(Prioritization and Prediction)元件

  該元件基於威脅感知元件所產生的不同模型，透過風險評估與大規模擴散關聯分析的演算，在部分已知的專家分析結果下，對於大量未知的事件或是行為進行評估，其中有兩種模式，一為找出予以標記行為相似特徵的事件，進而推論出潛在惡意的威脅行為，另一種為找出與正常行為變異過大的行為，找出潛在異常的行為。透過威脅評估與預測，可找出潛在異常事件。

- 採用 RWR 演算法，以相鄰矩陣為圖形模型的呈現作為輸入的參數，我們可以把攻擊者的 IP 當作是圖中(Graph)的節點，進而類比攻擊的行為樣態，例如：其關聯可產生類似的攻擊時間差和觸發的事件，以連結進行關聯。在 RWR 演算法中，另一個參數是奇異點，奇異點也是圖中的點，RWR 的演算從奇異點出發，計算出鄰近的點的可疑程度，其中，演算法的有效性驗證方面，以往的黑名單一直有人工判斷與持續更新，有一定信度，故可以利用以往的黑名單來驗證 RWR 演算法之效度。

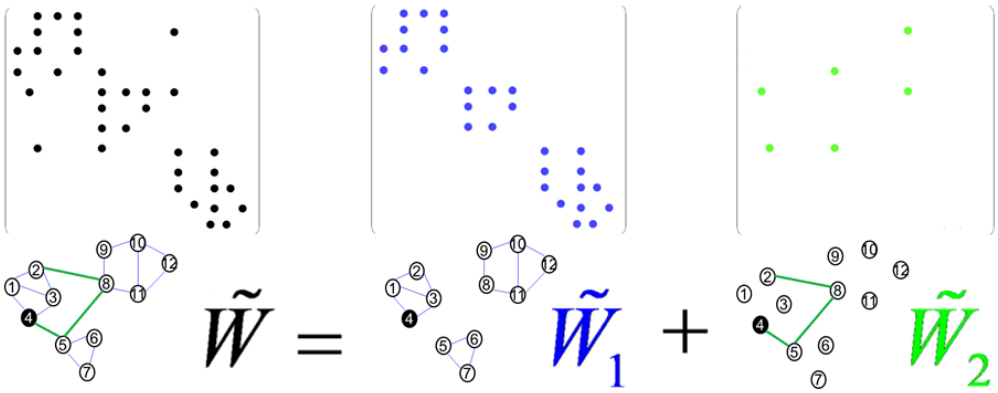


圖四：跨時域樣態異常偵測與分析機制 (資料來源:本計畫自行整理)

詳細的 RWR 演算法的描述如下，中間是相鄰矩陣，欲解向量 $\vec{r}_i$ 使之滿足以下公式

$$\vec{r}_i = c\tilde{W}\vec{r} + (1-c)\vec{e}_i \quad (1)$$

可疑程度向量　鉅量節點相鄰矩陣　奇異點



圖五：可規模化鉅量節點相鄰矩陣分解示意圖

在這個演算法中，$\tilde{W}$ 是巨量資料參考[17]Tensor 分析結果所產生的矩陣，所以有比較明顯的群聚。因為 $\tilde{W}$ 屬於巨量資料，我們需要把它分解成如 $\tilde{W}_1$ 的格式，形成對角線上的小矩陣，另外會有一些稀疏的群，形成的矩陣如 $\tilde{W}_2$。$\tilde{W}_1$ 對角線上的每一個小矩陣都可以視為有某種類似特性的資料，而 $\tilde{W}_2$ 上的每一個點所代表的資料都可以視為奇異資料，可能具有特殊的行為樣態。但是奇異點也有可能是 $\tilde{W}_1$ 對角線上的一個小矩陣如公式(2)(3)。因為 Tensor 分析是非監督式的群聚分析演算法，所以我們可以參考我們從過去分析出來的確定的異質資料來挑選奇異點。

$$Q_{1,i}^{-1} = \left(I - \tilde{W}_{1,i}\right)^{-1} \quad (2)$$

$$Q_{1,i}^{-1} = \left(I - \tilde{W}_{1,i}\right)^{-1}，\ \tilde{W}_2 \approx USV，\ \tilde{\Lambda} = \left(S^{-1} - cVQ_1^{-1}U\right)^{-1} \quad (3)$$

RWR 的演算法最耗時的地方是解 Q_1^{-1}，由於對角矩陣的特性，我們可以利用分散式計算去解 $\tilde{W}_1$ 對角線上的每一個小矩陣的 $Q_{1,i}^{-1}$，然後再合併成 Q_1^{-1}，最後求出可以程度向量，如公式(4)(5)。

$$Q^{-1} \approx Q_1^{-1} + cQ_1^{-1}U\tilde{\Lambda}VQ_1^{-1} \quad (4)$$

$$\bar{r}_i = \left(1-c\right)\left(Q_1^{-1}\bar{e}_i + cQ_1^{-1}U\tilde{\Lambda}VQ_1^{-1}\bar{e}_i\ \right) \quad (5)$$

重點在於如何讓資料盡可能集中在對角線的每個矩陣上面，以及如何分配各個 Hadoop 運算單位適當的工作量，都會影響計算的效率以及結果。雖然 Tensor 分析可以對資料進行分群，目前為止尚未有研究對於 Tensor 分析和 RWR 演算法做很好的整合。未來預期對於 Tensor 與圖形探勘演算法進行整合工作，可簡化矩陣 $\tilde{W}_1$ 的呈現，進而讓矩陣的分解更有效率。

肆、G-SOC Big Data Analytics

本章我們將以本國政府 G-OSC 巨量資料分析架構進行說明，首先說明在實務上 G-SOC 介接相關資料在技術面上的議題，接著說明 SOC 資料倉儲架構，最後說明資料分析的實證結果。

4.1 SOC 資料介接與倉儲

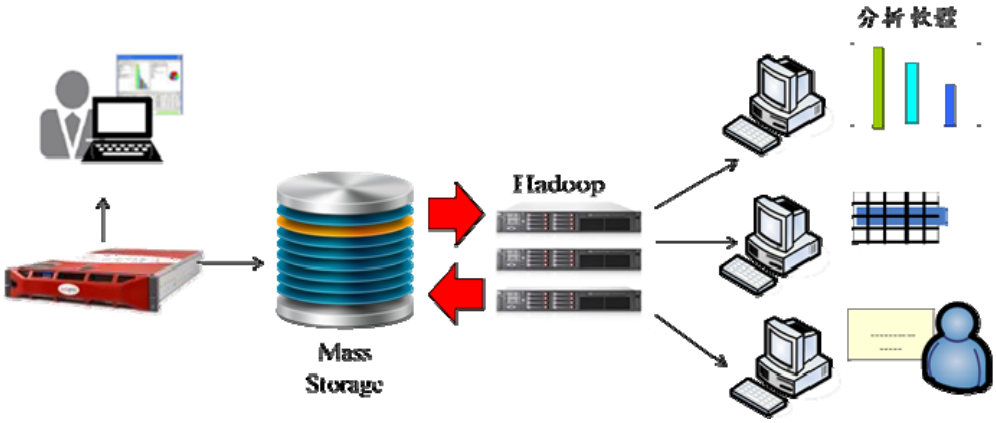
不同機關各有相異的防禦配置，欲接收多樣的異質設備資料，以 ArcSight 平台為例，是透過 SmartConnector 將已知的資料型態設備以 Syslog 進行直接收取，如：Firewall、IPS 等網路設備；防毒類型設備則透過專屬的 SmartConnector 直接撈取資料庫內容，轉換為 CEF 格式讀入資料。未知格式之設備系統 Log 資訊，則人工分析資料結構，透過 FlexConnector 以自訂 Parser 轉換資料為相對應欄位的 CEF 格式讀入系統。若資料為不定內容，不具固定格式，則使用預設的欄位對應轉換表，協助機關將既有資訊對應至相對欄位中，透過關鍵值提取的方式進行 Parser，亦將資料轉換為 CEF 格式讀入系統。

因應各家業者來源資料格式與內容定義的不同所產生的資料認知落差，前置處理方式透過自定義的通用類型名稱結構表做為正規化的依據，依照來源事件的關鍵字或詞意，運用事件對照(lookup)的方式進行初步分類，若難以判斷之部分再施以人工分類。對於相似類型事件賦予該筆事件可能所屬類別型態的代碼，再使用代碼結構組合進行細部分類，正規化整理多樣來源名稱等資訊，並於事件分類後依需求或例外狀況再進行微調，整合來源事件與降低資料認知落差。

4.2 SOC 資料倉儲架構

G-SOC 二線監控平台規劃與設計，首先曾針對 100M 頻寬單位進行測試，單日約可蒐集約 50 萬件事件(RAW event=500~1800 bytes)，同時間僅能傳送 10~20 件事件，乍看之下似乎可完全傳送單位防火牆與入侵偵測設備之事件量(每日約 16000 件，防火牆 12000 件、入侵偵測設備 4000 件)，但這是所有頻寬在進行事件傳送產生之數據值，所以兼顧現實環境與頻寬考量，分散式架構僅能蒐集關聯後之事件 EPS(Event Per Second，1 EPS 約占 1765 bytes)。接著，壓縮比平均在 10:1，加密傳輸的話，會須要 30% header，1765 / 10 * 1.3 就是每秒 1 EPS 的資料量，1EPS 每日資料量約在 34MB，10EPS 約在 340MB，以此類推，一般類似政府機關資安等級之 A 級機關，一天的 EPS 約在 100~300 件。

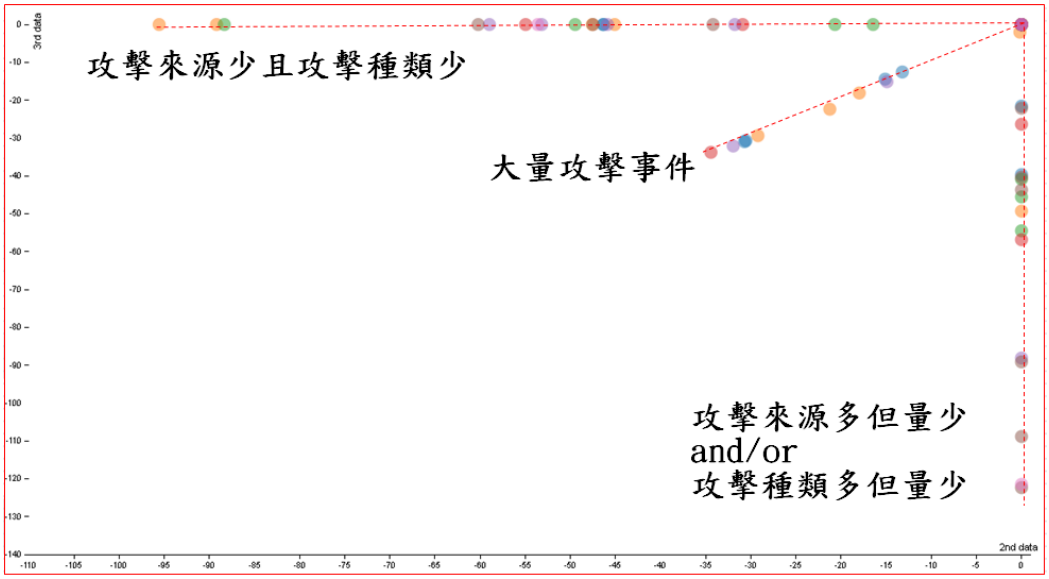
所以為解決 G-SOC 收容之 SOC 業者與政府機關 SOC 監控日益龐大之監控資料量以及未來可能收容之非結構化異質資料，G-SOC 二線監控平台保留三個月資料進行線上即時偵測，超過三個月的資料全部儲存至 archive 至 Hadoop 進行永久保存以及分析。



圖六：資安營運中心巨量資料分析系統架構圖

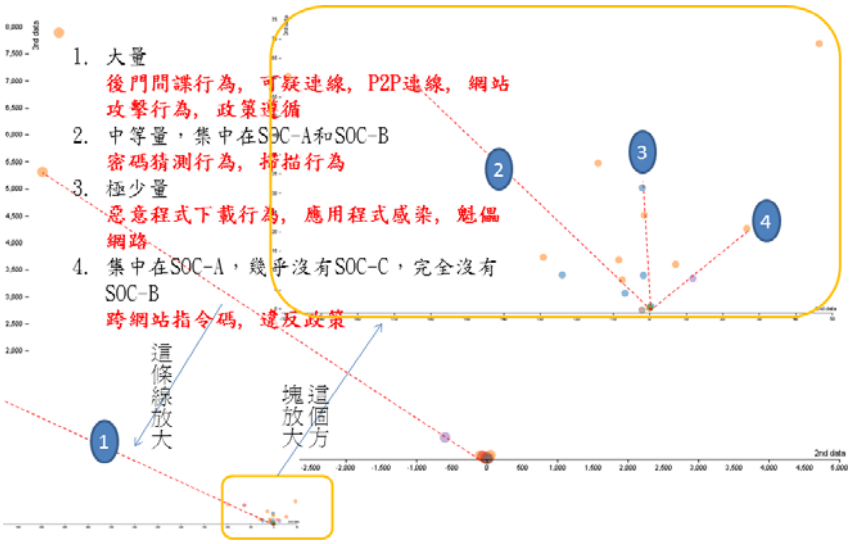
4.3 SOC 資料分析結果

首先我們先對於事件單的機關名稱、攻擊種類、攻擊來源透過奇異值分解(Singular Value Decomposition)和張量分解(Tensor Decomposition)等技術[17,18]得到的分析圖七所示。其中可發現透過分析機制，可將不同特性的機關依事件觸發行為對照其威脅特性，分為不同群組，此方式有別於以規則或是統計方式，可隨著資料特性的變化，而持續進行分析與運算，所採取的分析機制皆可運作於 Hadoop 平台上。



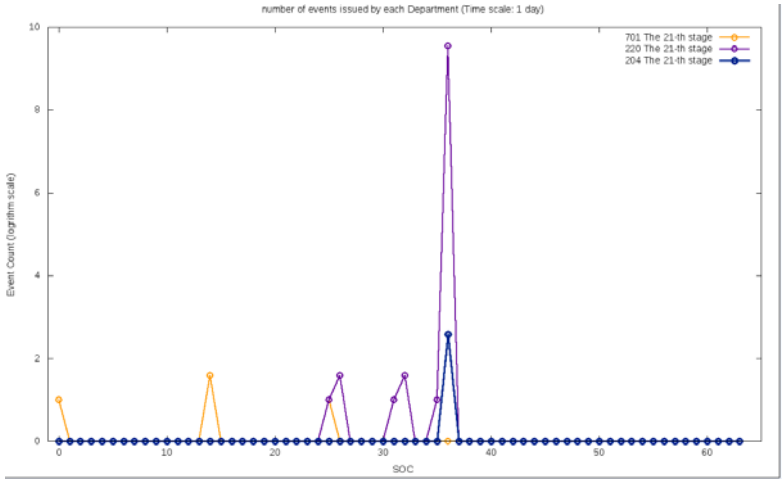
圖七：對於機關名稱、攻擊種類、攻擊來源進行張量分解(Tensor Decomposition)

根據時間、SOC 廠商、攻擊種類透過奇異值分解(Singular Value Decomposition)和張量分解(Tensor Decomposition)等技術得到的分析如圖八所示。其中可發現透過分析機制，可將不同的攻擊種類在事件觸發行為上依照其期觸發量和紀錄事件的 SOC 廠商分布，分為不同群組。機關可以根據此圖所顯示的威脅趨勢，加強防禦或督導，或者了解是否有誤報的可能。也可以了解各個 SOC 廠商的強項與短缺。此分析與運算的持續進行，有助於觀察此現象是否有改善。



圖八：透過張量分解(Tensor Decomposition)分析事件單事件類型特性

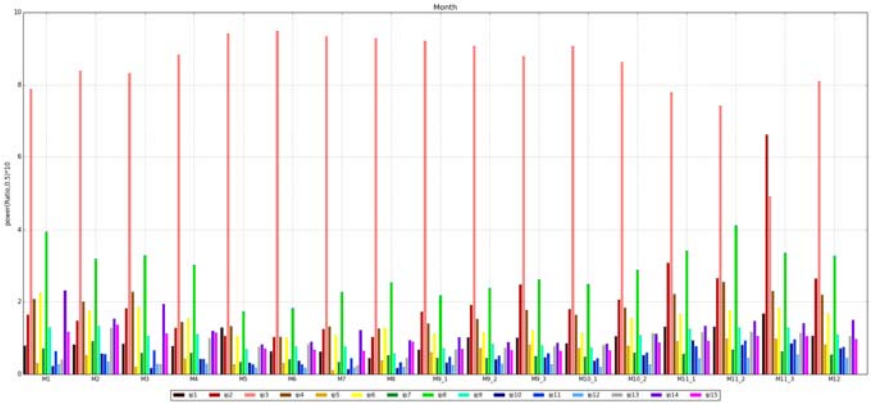
然而上述的圖不易看出事件隨時間的變化，如果使用二維的圖來呈現事件隨時間的變化，時間將會占據一維的空間。所以我們以動態並多種顏色線條的呈現方式，在使用者可以接受的複雜度之內，盡可能的豐富呈現的資訊，讓比較明顯的關聯可以透過肉眼判斷出來。根據 SOC 廠商、各種攻擊種類的頻率隨著時間演進的動態如圖九所示，不同顏色表示不同的 SOC 廠商，由左到右共顯示六十多種攻擊種類，在圖中可看出紫色所代表的 SOC 廠商在某時間點的 36 號攻擊特別的高，而同一時間點這樣的現象也被藍色 SOC 廠商偵測到了。



圖九：根據 SOC 廠商、各種攻擊種類的頻率隨著時間演進的動態圖

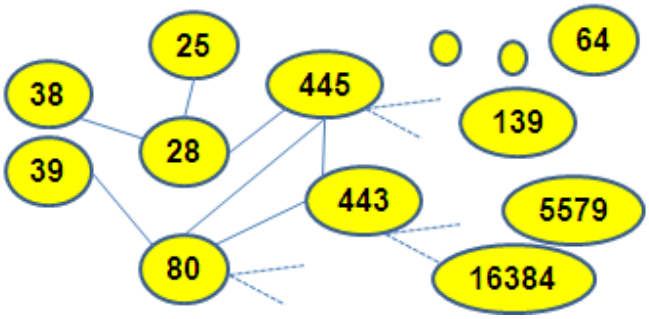
如果我們想要進一步觀察事件觸發的細節，了解來源 IP 和目標 IP 的關係，然而面對如此複雜的資料，我們可以從比較抽象的概念來了解攻擊的樣態圖。圖十是透過 IP 之間互動的關係，把 IP 分成 15 類左右，分別用 15 種左右的顏色表示，再把所有的 Log

切成幾個時間區段，每個時間區段分別去統計各類 IP 出現次數的比例。圖中可看出平常棕色線都明顯低於粉紅色線，但在倒數第二個時間區段棕色線卻異常高於粉紅色線，可推測該時間區段正是單位內部在測試弱點掃描的時間區段。由此可見抽象的樣態圖可以把複雜資料(IP 的 Dimension 很大)中的幾種異常現象，抽象成簡單的樣態(15 種顏色)，以肉眼所能觀察出的呈現方式呈現給使用者觀看。



圖十：入侵偵測系統資料的每月 IP 樣態圖。

此外，如果我們想要知道攻擊擴散的風險，可以透過 belief propagation 算法，例如，圖十一是從 Port 的角度看風險擴散程度，使用 Random Walk with Restart 演算法所得到的結果：



1	445
2	443
3	28
4	139
5	5579
6	38

圖十一：入侵偵測系統資料的每月 IP 樣態圖。

伍、結論

本研究對於資安監控與巨量資料分析進行探討，由於現有資安監控與事件管理平台，缺乏橫向擴充的能力，因此無法進行長時間的事件關聯，而異質事件的樣態與關聯規則亦缺乏有效發掘的方法。在考量到監控的成本下，如：資料倉儲資源的需求、分析運算平台的需求以及跨網路資料傳輸需求等，本研究提出資安監控與巨量資料架構，由資安監控切入，說明資安分析即服務之監控平台(Security Analytics as a Service)架構，並著重於技術分析的三個核心模組：Self-Inspection Modules、Threat Recognition Modules 與 Prioritization, Predictions Decisions Modules。本研究透過政府資安 G-SOC 二線監控平台，收容各個不同來源 SOC 資料，作為分析的案例進行探討。

參考文獻

[1] Driven Discovery of Temporal and Structural Information for Activity Recognition. IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[2] Dempster, A. P., Laird, N. M., and Rubin, D. B. 1977. Maximum Likelihood from Incomplete Data via the EM Algorithm. Journal of the Royal Statistical Society. Series B (Methodological), 39(1):1-38.

[3] Dietterich, T. G. 2000. Ensemble Methods in Machine Learning. In J. Kittler and F. Roli (Ed.) First International Workshop on Multiple Classifier Systems, Lecture Notes in Computer Science (pp. 1-15). New York: Springer Verlag.

[4] Green, O., Bader, D.A. 2013. Faster Betweenness Centrality Based on Data Structure Experimentation. 13th International Conference on Computational Science (ICCS).

[5] Green, O., McColl, R., and Bader, D.A. 2012. A Fast Algorithm for Streaming Betweenness Centrality. ASE/IEEE International Conference on Social Computing (SocialCom)

[6] Hamid, R., et al. 2009. A Novel Sequence Representation for Unsupervised Analysis of Human Activities, Artificial Intelligence.

[7] Hastie, T., Tibshirani, R., and Friedman, J. 2008, The Elements of Statistical Learning (2nd edition). Springer-Verlag.

[8] Hempstalk, K., Frank, E., and Witten, I.H. 2008. One-class classification by combining density and class probability estimation. In: W. Daelemans et al. (Eds.), ECML PKDD 2008, Part I, LNAI 5211, pp. 505–519.

[9] Mappus, R. and Briscoe, E. 2013. Layered Behavioral Trace Modeling for Threat Detection. IEEE Intelligence and Security Informatics.

[10] Neal, R. and Hinton, G. E. 1998. A view of the EM algorithm that justifies incremental, sparse, and other variants. In Learning in Graphical Models, (pp. 355-368), Kluwer Academic Publishers.

[11] Riedy, J. and Bader, D.A. 2013. Multithreaded Community Monitoring for Massive Streaming Graph Data. 7th Workshop on Multithreaded Architectures and Applications (MTAAP), Boston, MA, May 24, 2013.

[12] Riedy, J., Meyerhenke, H., and Bader, D.A. 2012. Scalable Multi- threaded Community Detection in Social Networks. 6th Workshop on Multithreaded Architectures and Applications (MTAAP), Shanghai, China, May 25, 2012.

[13] Zhou, Z-H. 2012. Ensemble Methods: Foundations and Algorithms, Chapman & Hall/CR

[14] Yen, T. F., Oprea, A., Onarlioglu, K., Leetham, T., Robertson, W., Juels, A., & Kirda, E. (2013, December). Beehive: large-scale log analysis for detecting suspicious activity in enterprise networks. In Proceedings of the 29th Annual Computer Security Applications Conference (pp. 199-208). ACM.

[15] Analytics, Big Data. "Big Data Analytics for Security." (2013).

[16] West, A. G., & Mohaisen, A. (2014). Metadata-driven Threat Classification of Network Endpoints Appearing in Malware. In DIMVA'14: Proceedings of the 11th Conference on Detection of Intrusions and Malware & Vulnerability Assessment (to appear).

[17] Jimeng Sun, Dacheng Tao, Spiros Papadimitrious, Philip S. Yu, Christos Faloutsos, Incremental tensor analysis: theory and applications, ACM Transaction on Knowledge Discovery from Data, Vol. 2, Issue 3, 2008.

[18] E. E. Papalexakis, C. Faloutsos, and N. D. Sidiropoulos, “Parcube: sparse parallelizable tensor decompositions,” in Machine Learning and Knowledge Discovery in Databases. Springer, 2012, pp. 521–536.