國立臺北大學　資訊工程學系

專題報告

主動式惡意程式自動化蒐集系統與樣本搜尋分析服務

指導老師：曾俊元 老師

專題組員：李建瑩、李思賢

執行期間：中華民國 100 年 02 月至 10 月

國立台北大學資訊工程學系專題報告

主動式惡意程式自動化蒐集系統與樣本搜尋分析服務

專題組員:李建瑩、李思賢
指導老師:曾俊元老師
執行期間:100 年 02 月 至 100 年 10 月

1. 摘要

本專題承接交通大學的病毒蒐集系統,嘗試在眾多可疑檔案中,使用市面上不同的防毒軟體逐一掃描並分類。以掃描結果為基礎,發展系統性分類及針對病毒進行交叉比對的分析,以期能藉由此系統輔助惡意程式分析之研究人員快速掌握最新病毒資訊。

本研究的目的是建立一套完整的惡意程式分析系統,從一開始的惡意程式蒐集、防毒軟體掃描、資料庫建置,最後以圖表化的方式產生分析結果。隨著病毒蒐集時間的累積,希望建構一個龐大的資料庫樣本提供相關人員作為研究的基礎。

關鍵字:virus collection, virus management, contrast

2. 簡介

1.背景與動機

現代惡意程式種類繁多、傳染速度十分迅速,但市面的防毒軟體對其惡意程式定義皆不盡相同,且分析資料也過於分散、粗糙。

此專題想法為建構一個親和的使用者介面,將蒐集到的惡意程式樣本儲存至資料庫做為基礎,讓使用者可透過本專題設計之介面,查詢自己感興趣的資料進行分析,不再受限於防毒軟體公司提供的報告內容。

2.主要預期效果

(1)依據系統蒐集的惡意程式,將其有關的資訊如:系統取得時間、檔案大小、檔案類型、各家掃毒軟體檢查後病毒類型統合成一個完整的病毒資料庫。

(2) 利用此資料庫儲存的資訊,可以進行更多樣化的分析功能,例如:市面上防毒軟體偵測率比較、惡意程式類型所佔比率之分析…等。

3. 專題進行方式

3.1 重點與方向

我們研究的重點在於為使用者提供一個簡潔、有架構的環境,並搭配惡意程式蒐集模組,使資料庫能持續搜集到最新的惡意程式樣本。

起初開始的方向是將所有收集的檔案用四家防毒軟體在系統內進行掃描。後來在網路上發現了一個名為 VirusTotal 的網站。這個網站提供一個擁有四十三家掃毒軟體的平台資源,無論在世界上任何角落都可以自行將檔案上傳至 VirusTotal,享有平台內四十三家掃毒軟體的掃描資訊。

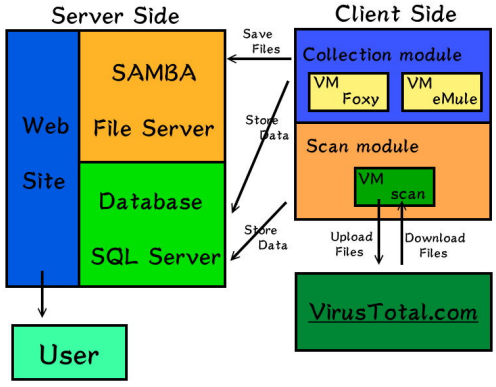
在權衡比較之後便決定改採用 VirusTotal 這個網站所提供的服務。

優點如下

1. 可以使用 43 家防毒軟體對此檔案進行掃描後的結果(定義時間、病毒類型..等等)。
2. 網站提供上傳檔案和下載檔案報告之 API，透過程式的撰寫，可自動化取得相關資訊。

(圖一、VirusTotal結果頁畫面)

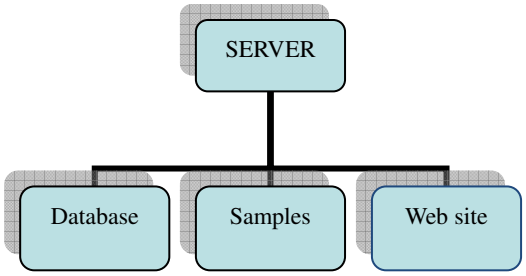
3.2 系統架構



(圖二、系統架構圖)

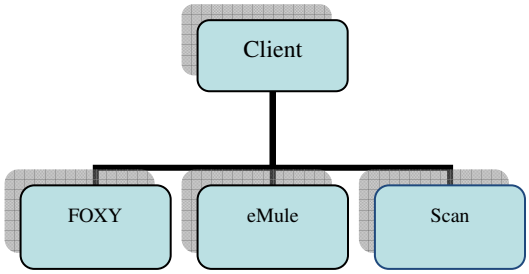
我們將此實驗系統大致分為兩個架構：Server 端和 Client 端，以下為架構說明。

Server 端架構如下



Server 端使用 Ubuntu Server 10.04 作為主要作業系統，在此 Server 上提供 SAMBA 服務，儲存來自 Client 端上傳的疑似惡意病毒樣本；建立 MySQL Database，惡意程式的樣本資訊；最後一部份是設計網頁介面，讓使用者分析統計本系統所蒐集的樣本資訊。

Client 端架構如下



Client 端下大致可分為「病毒蒐集模組」和「病毒掃描模組」。病毒蒐集模組使用了人們使用已久的 p2p 軟體 Foxy 及 eMule 作為蒐集工具。為了避免 Client 端系統因為惡意程式的破壞，而需要花費大量時間精力重新恢復環境，本專題將蒐集模組建置在 VirtualBox 底下，萬一環境遭破壞而無法進行蒐集時，可以快速還原至原先未遭破壞的環境。病毒掃描模組則是利用 VirusTotal 所提供的平台，將檔案上傳至 VirusTotal 掃描並獲取掃描結果後，再將資訊存入 Server 上的 Database。

3.3 實驗設計與流程

Client 端

1. FOXY && eMule

Step1

自Server上取得：keygen、破解、game、crack、hack、遊戲等關鍵字，使用這些字彙做為主要搜尋的關鍵字，並自動下載檔案大小小於5MB，且副檔名是如.exe、.rar、.zip...等常見於惡意程式的檔案。

若下載的檔案副檔名是如rar、zip、7z等等的壓縮檔，會將之解壓縮後刪去副檔名上述關鍵字的檔案。

Step2

下載下來的檔案會上傳至Server，同時將檔案做 sha1 編碼後做為檔案的檔名，以避免檔名重複的情況發生。並將有關檔案的所有資訊存入 Server 的 SQL Database 上後刪除該檔案。

2.SCAN

Step1

以先前檔案的 sha1 碼至 Virustotal 上做搜尋，若是發現檔案在 VirusTotal 上已有報告(表示已有使用者上傳掃描過相同的檔案)，則獲取報告並將報告資訊存入 Server 資料庫內並標記為已取得報告。

Step2

若檔案在 VirusTotal 上無報告，則上傳檔案至 VirusTotal 上，讓伺服器上的四十三家掃毒軟體掃描該上傳檔案。

此一程式是一個迴圈，重覆執行 Step1(獲取報告並存入Server 資料庫)及 Step2(上傳檔案)。

實作方法為先在資料庫內搜尋(query)尚未至VirusTotal檢查是否有報告的檔案，將得到的結果做成一個佇列，至 VirusTotal 一個一個將 sha1 碼做搜尋，直到佇列為空(step1 結束)。接下來再至資料庫搜尋(query)尚未獲取報告的檔案。同樣的，將結果製成一個佇列，一個一個將檔案上傳至 VirusTotal 伺服器上，直至佇列為空。程式會一直重覆此二步驟，不斷產生更新的資料。

Server 端網站

網站主要目的為提供一個友善的介面讓使用者或研究人員可以更容易利用此系統，取得有用的資訊。由於此系統主要儲存惡意程式，對於電腦有害的資料，不能隨意提供給一般人使用，在帳號控管上本研究設計三個層次的權限。網站功能大略分為以下三項：

(一)依條件查詢

依檔名、sha1 碼、系統發現時間、VirusTotal 最早發現時間、病毒類型、下載來源...等等關鍵詞做為資料查詢的條件做查詢。

- 提供查詢的項目有：

1. **sha1 碼**
2. **檔案名稱**
3. **檔案大小**
4. **下載來源(Foxy or eMule)**
5. **系統發現時間**
6. **VirusTotal 最早的發現時間**

病毒敘述(如攻擊方式、行為，目前僅提供 Kaspersky 官方撰寫的病毒敘述)

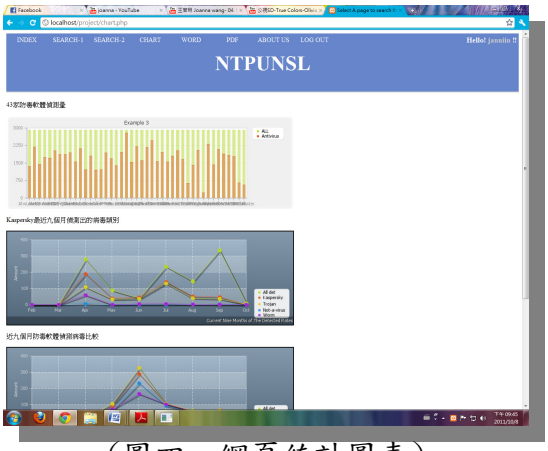
(二)樣本下載

提供有權限的使用者下載檔案樣本做為研究使用。

（圖三、系統搜尋結果網頁頁面）

(三)圖表分析

提供近期系統資料庫得到的資訊，如：各家掃毒軟體回報系統樣本為有危險的比率、回報為病毒的檔案中各類型病毒所佔比例、近期各家防毒軟體偵測病毒比率...等。(如圖四)

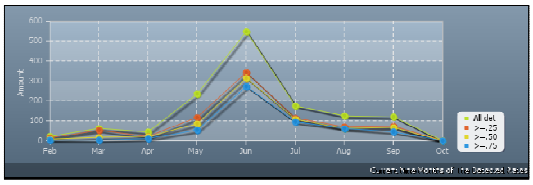


（圖四、網頁統計圖表）

4. 主要成果與評估

一般使用者大多只能使用單一家掃毒軟體，得到的病毒資料和敘述面向難免通常較為單一。而 VirusTotal 提供了一個可以得到不只一家掃毒軟體掃毒結果的搜尋平台。但在分析上 VirusTotal 較沒有辦法提供更多面向的分析統計資料(如：某個月份下載得到的檔案中，各類型惡意程式所占的比例)。

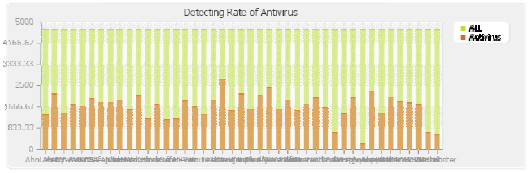
而我們利用了本身自己的蒐集系統再加上 VirusTotal 的防毒軟體資源，將資料彙整起來，其中成果最為豐碩的是使用者可以選擇自己關注的項目來做資料的比對。



（圖五、回報率與數量比較）

圖五是在各月搜集到的惡意程式的防毒軟體的回報數量比較。綠色線(All detected)表示的是所有有被任一防毒軟體回報為有問題的檔案數量。橙色線(>=0.25)表示的是所有有被大於25%以上防毒軟體回報為惡意程式的檔案數量；同樣的，黃色線和淺藍色線代表的則是被 50%以及 75%以上防毒軟體回報為惡意程式的檔案數量。偵測比率分別是：25%以上平均比率為 61.55%、50%以上平均比率為 46.55%、75%以上平均比率為 31.86%

可以發現對於同一檔案，回報該檔含有惡意程式的防毒軟體每增加 25%，則平均偵測數量比就會減少約 15%。在進行此分析之前，其實我們原本認為下降的比率會大於 20%。但後來思考其原因，很有可能是市面上防毒軟體之間對於惡意程式的偵測能力事實上不會有明顯或甚至是過分誇張的差別，對於各等級惡意程式的反應，防毒軟體基本上都能做到。



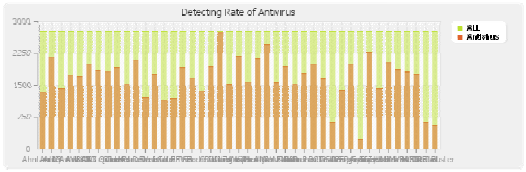
（圖六、各家防毒軟體對系統樣本偵測率）

圖六表示了對於有被任一防毒軟體回報為病毒的檔案中，43 家防毒軟體

對於這些檔案的偵測率。

由此圖可以發現對於有被回報為病毒或是有問題的檔案，大部分防毒軟體的回報率大約是四成到五成之間。其中回報率最高的是 Ikarus，回報率高達 58.32%(2740)。而相反的回報率最低的是 SUPERAntiSpyware，回報率大約是 4.53%(213)。平均回報率約為 34.07%(1600)。

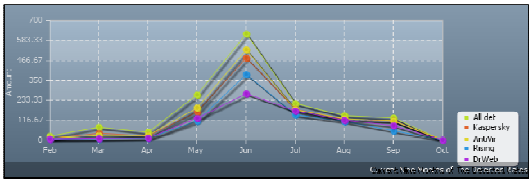
會有這麼低的回報率除了肇因於防毒軟體本身的掃毒能力之外，可能的原因還有掃毒軟體的誤判。但這個方面我們目前系統還沒辦法認定是否為誤判。



（圖七、修改圖六數值後）

考慮到"只要有任一防毒軟體"回報為有問題，則誤判比率可能太大，則此回報率的數值可能變得沒有價值。我們將總數改為若"該檔有超過 12.5%(大約是大於 5 家)的防毒軟體回報為有問題"則取用之。調整之後得到了圖七。為最高回報率為 99.38%，最低回報率為 7.73%，平均回報率約為 58.05%

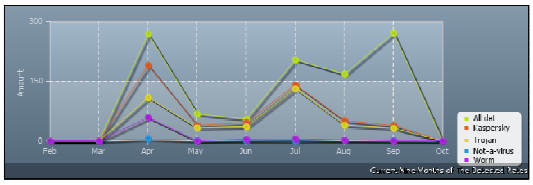
我們在後續會以 **12.5%**做為一個判定檔案為惡意程式可能的標準。換句話說，只要有超過 5 家防毒軟體回報該檔為惡意程式，我們便會判定該檔為惡意程式。



（圖八、防毒軟體各月回報率比較）

圖八為近幾個月 12.5%以上防毒軟體回報有問題(由前圖的定義，這些檔案為惡意程式)、Kasperky、AntiVir、Rising 以及 DrWeb 等防毒軟體回報有問題的檔案數量比較。這四家防毒軟體中，偵測率最高的是 AntiVir 軟體，平均偵測率為 69.53%。最低的是 Rising 軟體，平均偵測率為 43.71%。

在比較各家軟體掃毒能力時，我們由 VirusTotal 上提供的 43 家軟體挑了許多家防毒軟體組合比較，發現到，事實上市面上許多較為知名的軟體他們的偵測率結果差別差不到 5%，以此結果畫出的圖表可以看出，使用這些較為著名的防毒軟體安全性似乎都是非常不錯的。



（圖九、針對 Kaspersky 所偵測病毒類型比較）

圖九為掃毒軟體卡巴斯基各月回報各種類型的病毒類型的數量，可以發現木馬程式所佔的比例最少都有五成以上的比例，最大所佔比率為 92.91%，最低所佔比率為 57.37%，平均比率約為 78.94%，可說是病毒的最大宗。其次為網路蠕蟲，最大所佔比率為 31.05%，平均比率為 7.78%。

由此圖可以發現目前最流行的病毒為木馬程式，由我們自行建立的網站和資料庫中也可發現，若是隨意搜尋且不刻意避開 Troj、Tro、Trojan...等等關鍵字所得到的資料，幾乎看到的結果都會是木馬程式。可以顯現木馬程式的主流性、危害性以及各家防毒軟體公司對於木馬程式防治的迫切性絕對都是非常大的。

由各類型病毒所佔比率的結果也

可以推論導致這個結果的最可能原因會是"網路的普及性". 在現代社會，無論是生活、工作...等大小事，人們已經難以和網路作分隔。有鑑於此，駭客等惡意程式撰寫者為了增加自己所製作的惡意程式的影響力，必定走向大量編寫產生與網路最有關係的木馬程式和網路蠕蟲等程式的選擇。

5. 結語與展望

在病毒蒐集模組方面，原先認為隨著時間推進會因為系統重複檔案愈來愈多造成每日樣本下載量逐日下降。但出乎預料的是每日樣本的下載量波動似乎並不大，每日樣本下載數量至少還是有五十個。推測原因是由於 Foxy 及 eMule 仍為一般民眾使用的主流下載器，因此每日檔案更新速度仍是非常快速。

在分析樣本中惡意程式、防毒軟體偵測率以及惡意程式等等關係時，發現了很多有趣的現象，如：主流防毒軟體對於惡意程式偵測率的比較、主流病毒的類型分佈比率和日常生活的關係等等，皆是在分析的過程中探索到的有趣例子。藉由這些例子，也可以自行做一些猜想，如：惡意程式編寫者編寫的惡意程式會盡量貼近大部分人類的行為以提高自己惡意程式的影響能力。因此我們可以由這個現象猜想下一個惡意程式編寫者的目標極有可能是平板電腦或智慧型手機等移動式平台！

在分析的過程中我們也體會到由於時間的限制，我們無法讓系統能得到更多的樣本數以及更長的時間去分析時間與下載得到的惡意程式之間的關係，以及設計更多的分析方法分析其他更深入的統計結果。

我們非常肯定這個系統的樣本蒐集能力以及可能的分析潛力，此系統若能持續運作，所蒐集的樣本數量會大幅提升。對於研究惡意程式的相關人員此系統必能提供強而有力的幫助！

6. 銘謝

感謝曾俊元老師在專題整個過程中的指導。在每一步的想法上都會與我們討論，使我們在整個專題想法上更加周全、成熟。當遇到瓶頸時，老師更會引導我們新的方向、刺激我們有新的想法出來。在實作方面，也提供經驗，使我們整個專題能順利完成，解決難關。感謝老師的教導，才能有此實驗結果。

也要感謝學長們，剛進實驗室時，對環境、使用系統相當不熟悉。在他們熱心的幫助下，漸漸的步上軌道。在實驗室的時光，除了互相交換在專題上的意見與想法，使此專題能更周全。在經驗的傳授上，更使我們受益良多。

參考文獻

[1]李蔚澤 著·2007·Ubuntu7.10 Linux 全方位學習・碁峰

[2]吳弘凱 著・2006・PHP+MySQL 快速入門・松崗

[3]吳明哲、何珮琪、張志成、何嘉益、周家旬 著・2009・Visual C# 2008 學習範本・松崗

[4]河內正紀 著·2007·CSS LAYOUT 達人の階梯・旗標出版股份有限公司

[5]Silberschatz & Korth & Sudarshan 著·2011·Database System Concepts・McGraw Hill

[6]Musciano & Kennedy 著・2001・HTML&XHTML 大全・O'REILLY