İnformasiya Təhlükəsizliyi Üzrə Tədqiqatlar üçün Böyük və Kiçik Verilənlər

Yadigar İmamverdiyev
AMEA İnformasiya Texnologiyaları İnstitutu, Bakı, Azərbaycan
yadigar@lan.ab.az

***Xülasə* — Big Data texnologiyaları informasiya təhlükəsizliyi üzrə tədqiqatlarda böyük verilənlər toplularından istifadə edilməsini tələb edir. Lakin belə verilənlərin toplanması, tədqiqatçıların geniş dairəsi üçün əlyetər edilməsi sahəsində bir sıra hüquqi, təhlükəsizlik, etik, intellektual mülkiyyət, rəqabət və digər problemlər mövcuddur. Məqalədə informasiya təhlükəsizliyi sahəsində aktual tədqiqat istiqamətləri üzrə mövcud verilənlər topluları haqqında məlumat verilir, bu toplulara müraciət modelləri analiz edilir.**

Açar sözlər — informasiya təhlükəsizliyi; Big Data; verilənlərin paylaşılması; verilənlər toplusu; botnet

I. GİRİŞ

Big data əsrində informasiya təhlükəsizliyi üzrə tədqiqatların aparılması üçün verilənlər bazis rolunu oynayır. Zəngin və etibarlı verilənlər toplularının əlyetər olması elmi-tədqiqat işinin aparılması üçün yaxşı bir başlanğıcdır.

Lakin informasiya təhlükəsizliyi üzrə elmi-tədqiqatlarda istifadə edilən verilənlərin toplanması kifayət qədər çətin məsələdir. Bu çətinlik belə verilənlərin tədqiqatçılar arasında paylaşılması üçün səbəblərdən biridir. Bununla yanaşı, burada bəlkə də daha vacib olan digər bir səbəb də var. Eksperimentlərdə istifadə edilən verilənlərin paylaşılması nəticələri yoxlamağa və metodları müqayisə etməyə imkan verir. Bu yaxşı əsaslandırılmış elmi iş üçün fundamentlərdən biridir.

İnformasiya təhlükəsizliyi üzrə elmi-tədqiqatlarda istifadə edilən verilənlərin paylaşılmasında bir sıra etik, hüquqi və texnoloji problemlər meydana çıxır, bu problemlər [1]-də araşdırılır.

Bu işdə informasiya təhlükəsizliyi üzrə tədqiqatlarda istifadə edilən verilənlər topluları analiz edilir, onların toplanması, saxlanması və istifadəsi üzrə hüquqi, texnoloji və təhlükəsizlik problemləri araşdırılır. Verilənlər AMEA İnformasiya Texnologiyaları İnstitutunda informasiya təhlükəsizliyi üzrə aparılan və aparılması planlaşdırılan aşağıdakı elmi-tədqiqat istiqamətləri üzrə təsnif edilir:

- kiberhücumların aşkarlanması;
- DDoS hücumların aşkarlanması;
- botnetlərin aşkarlanması;
- spamın aşkarlanması;
- zərərli proqramların aşkarlanması;
- veb-texnologiyaların təhlükəsizliyi;
- informasiya təhlükəsizliyi risklərinin idarə edilməsi;
- informasiya təhlükəsizliyinin qiymətləndirilməsi;
- informasiya təhlükəsizliyi insidentlərinin idarə edilməsi;
- simsiz şəbəkələrdə informasiya təhlükəsizliyi problemləri.

Qeyd edək ki, verilənlərlə bağlı oxşar problem informasiya təhlükəsizliyi sahəsində tədqiqat aparan digər tədqiqatçıları da narahat edir. Aparıcı tədqiqatçıların bir qrupu 2010-cu ildə informasiya təhlükəsizliyi elminin reprezentativ verilənlərə ehtiyaclarını vurğulayan “arzular siyahısı”nı tərtib etmişdi [2]. Siyahı müəllifləri hazırda əlyetər olmayan spesifik verilənlər üçün metaverilənlərin toplanmasını asanlaşdıran və hüquqi və gizlilik problemlərinin həllinə kömək edən məlumatların paylaşılması prosesinə tələbləri də müəyyən etməyə cəhd edirlər.

İşin strukturu aşağıdakı kimidir. İkinci bölmədə müdaxilələrin aşkarlanması sistemlərinin (ing. Intrusion Detection System, IDS) test edilməsi üçün mövcud verilənlər topluları analiz edilir. Üçüncü bölmədə verilənlərin paylaşımına yanaşmalardan biri kimi PREDICT modeli və dördüncü bölmədə WINE verilənlər toplusu araşdırılır. Beşinci bölmədə botnetlərin aşkarlanması üçün yeni işlənmiş verilənlər topluları və onlara yürüdülən tələblər təhlil olunur.

II. IDS SİSTEMLƏRİNİN TEST EDİLMƏSİ ÜÇÜN VERİLƏNLƏR

IDS sistemləri tarixən informasiya təhlükəsizliyi sahəsində verilənlərin intellektual analizinə əsaslanan ilk sistemlərdir. 1990-cı illərin ortalarından belə sistemlərin test edilməsi üçün etalon test verilənlərinin vacibliyi etiraf edildi və DARPA verilənləri topluları meydana çıxdı [3]. Ümumiyyətlə, IDS sistemləri üçün test toplularının kritik analizinə həsr olunmuş tədqiqat işləri yerinə yetirilmiş və yeni topluların yaradılması ilə bağlı xeyli təcrübə toplanmışdır. Bunu nəzərə alaraq bu bölmədə belə toplulların bir qədər ətraflı analizi aparılır.

IDS sistemlərinin test edilməsi üçün tədqiqatçılara bir çox verilənlər toplusu əlyetərdir: DARPA [3], KDD’99 [4], Internet Trafik Arxivi [5], LBNL [6], CAIDA [7], DEFCON [8,9], PREDICT [10], və ISCX 2012 IDS verilənlər topluları [11].

DARPA verilənlər topluları şəbəkə təhlükəsizliyi məqsədləri üçün orta ölçülü ABŞ aviabazasında müşühidə edilən şəbəkə trafiki imitasiya edilməklə yaradılmışdı. DARPA’98 və ’99 verilənlərinin diqqətli analizi nəticəsində bir sıra nöqsanlar aşkarlandı [12-14]: real trafikə tam oxşamaması,

verilənlər toplusuna süni yaradılmış verilənlərin daxil edilməsi nəticəsində meydana çıxan çoxsaylı uyğunsuzluqlar və s.

DARPA′98-də 7 həftəlik şəbəkə trafikinin təxminən 5 Qbaytlıq tcpdump verilənləri var. Onları hərəsi 100 bayt olmaqla 5 milyon bağlantı haqqında məlumatlara bölmək olar. Iki həftəlik test verilənlərində təxminən 2 milyon bağlantı məlumatı var. KDD təlim verilənləri toplusunda təxminən 4.9 milyon bağlantı vektoru var, hər bir vektorda 41 əlamət var və normal və ya müəyyən bir hücum növü ilə işarə edilib. İmitasiya edilmiş hücumlar dörd kateqoriyadan birinə aiddir:

1) Xidmətdən imtina hücumu (Denial of Service, DoS) – hücum edən müəyyən servisləri həddindən artıq yükləyir ki, qanuni istifadəçilərə xidmətdən imtina edilsin.
2) User to Root Attack (U2R): Hücum edən müəyyən eksploytlardan istifadə edərək sistemdə normal istifadəçi hesabından administrator hesabına yüksəlməyə cəhd edir.
3) Remote to Local Attack (R2L): hücum edən kompüterdə müəyyən boşluqlardan istifadə edərək həmin kompüterdə lokal istifadəçi hesabına giriş əldə etməyə cəhd edir.
4) Zondlama hücumu (ing. probing attack): informasiya təhlükəsizliyini pozmaq məqsədilə kompüter şəbəkəsi haqqında informasiya toplamağa cəhd edilir.

KDD′99-un test verilənlərinə təlim verilənlərində olmayan spesifik hücum növləri daxildir, bu aşkarlamanı reallığa daha yaxın edir. Təlim verilənlərində 24 hücum növü vardır ki, onlardan 14-ü test verilənlərində yoxdur. Təlim verilənlərindəki hücum növlərinin adları və ətraflı təsvirləri [4]-də verilir.

DARPA və KDD verilənlər topluları köhnəlmiş hesab edilir, bir sıra nöqsanlara malikdirlər, lakin hələ də istifadə edilirlər. Onların aşağıdakı nöqsanlarını göstərmək olar [12–14]:

1) Gizlilik naminə həm fon, həm də hücum verilənləri bir neçə ay ərzində ABŞ hərbi aviasiya bazalarında müşahidə edilən verilənlərə oxşar olan sintetik verilənlər generasiya edilmişdir. Lakin səhv aşkarlama xarakteristikalarının analitik və ya eksperimental yoxlamasına cəhd edilməmişdir. Eyni zamanda, sintez edilmiş verilənlər real şəbəkədəki trafikə oxşar deyil.

2) DARPA′98-də istifadə edilən tcpdump kimi trafik kollektorları ağır trafik yüklənməsi zamanı paketləri atırlar. Lakin atılmış paketlərin mümkünlüyü yoxlanılmamışdır.

3) Hücumların dəqiq tərifləri verilmir. Məsələn, əgər iterasiyaların sayı müəyyən həddi keçmirsə, onda zondlama əməliyyatı hücum hesab edilmir. Oxşar olaraq, bufer daşması yaradan paket də həmişə hücumu təmsil etmir. Buna görə də, bu şərtlər altında qiymətləndirənlə qiymətləndirilən arasında tərifləri barədə razılaşma olmalıdır. Lakin DARPA'98-də şəbəkə hücumlarının aydın tərifləri yoxdur.

4) İzafi məlumatların sayı olduqca çoxdur. Təlim və test verilənlərində məlumatların uyğun olaraq 78%-i və 75%-i təkrarlanır. Təlim verilənlərində böyük sayda təkrarlanma öyrənmə alqoritmlərində tezliyi çox olan yazılara doğru meyletməyə səbəb olur və daha zərərli ola bilən az tezlikli məlumatlardan öyrənmənin qarşısını alır. Test verilənlərində təkrarlanma isə daha yüksək aşkarlanma faizi olan yüksək tezlikli məlumatlara doğru meyletməyə səbəb olacaq.

5) Məlumatların çətinlik səviyyəsi aşağıdır. Ədəbiyyatda məlumat verilənlərin aşkarlanma metodlarının dəqiqliyi 98%, səhv aşkarlanma faizi isə 1% səviyyəsindədir. Lakin praktiki İDS sistemlərində bu metodlar istifadə edilmir, hesab edilir ki, onlar hələlik mükəmməl texnologiyalar deyil.

KDD’99 verilənlər toplusu da DARPA’98-də toplanmış verilənlərdən əldə edilmişdir və eyni problemlər qalır. Bu problemləri aradan qaldırmaq məqsədilə NSL_KDD verilənlər toplusu KDD`99-dan seçilmiş verilənlərdən təşkil edilib [14]:

- Təlim toplusunda izafi yazılar yoxdur.
- Test toplusunda təkrar yazılar yoxdur.
- Hər bir qrupda seçilmiş yazıların sayı ilkin KDD toplusundakı yazıların faiz payına tərs mütənasibdir.

CAIDA (Center for Applied Internet Data Analysis) [7] müxtəlif növ verilənlər toplayır və onları elmi-tədqiqatlar üçün paylaşır. CAIDA verilənlər toplularının əksəriyyəti xüsusi hadisələrlə və hücumlarla əlaqəlidir. Bu verilənlərin bir çoxu anonim magistral trafikidir və onlardan faydalı yük, bəzən protokol məlumatları, gediş ünvanı və s. tamamilə silinib. Verilənlər toplularının bu nöqsanları onların etalon test verilənləri kimi effektiv istifadəsinə təsir edir.

DEFCON verilənlər toplusu CTF (Capture The Flag) yarışları zamanı yığılmış trafikdir, bu verilənlər real şəbəkə verilənlərindən çox fərqlidir, çünki normal fon trafikindən fərqli olaraq əsasən müdaxilə trafikindən ibarətdir. Bu nöqsanına görə DEFCON verilənləri adətən həyəcan siqnallarının korrelyasiyası üsullarının qiymətləndirilməsi üçün istifadə edilir [8].

ISCX 2012 IDS Data Set – istifadəçilərin davranışlarını imitasiya edən real trafik (SSH, HTTP, SMTP) generasiya edən real qurğulardan istifadə edilməklə toplanmışdır [11].

Profil anlayışına müdaxilələrin ətraflı təsviri və tətbiqi proqramların, protokolların və aşağı səviyyə şəbəkə subyektlərinin abstrakt paylanma modelləri daxildir. Real trafik analiz edilərək trafik generasiya edən agentlərin profilləri yaradılmışdır. Real verilənlər generasiya etmək üçün qaydalar müəyyən edilmiş və profillə birlikdə istifadə edilmişdir. Bu qaydalar verilənlər toplusunun reallığı, qiymətləndirmə imkanları, tamlıq, ümumi toplama və zərərli fəaliyyət baxımdan effektivliyi üçün olduqca vacibdir. Daha sonra profillərdən istifadə edilərək verilənlər generasiya edilmişdir. Verilənlər toplusunda anomal fəaliyyətə aid verilənləri generasiya etmək üçün müxtəlif çoxmərhələli hücum ssenariləri yerinə yetirilmişdir.

III. PREDICT VERİLƏNLƏR TOPLUSU

PREDICT (Protected Repository for the Defense of Infrastructure Against Cyber Threats) ABŞ Milli Təhlükəsizlik Nazirliyi tərəfindən dəstəklənir, 2008-ci ildən fəaliyyətdədir və şəbəkə təhlükəsizliyi üzrə tədqiqatları inkişaf etdirmək üçün real verilənlərin paylaşımını həyata keçirir.

PREDICT ölkənin informasiya infrastrukturuna olan kiber-təhdidləri qiymətləndirmək və ölkənin kiber təhlükəsizlik potensialını artırmaq üçün yeni modellərin, texnologiyaların və məhsulların yaradılması üçün tədqiqat verilənləri bazasıdır (repozitarisidir).

Verilənlər provayderi (Data Provider) və verilənlər hostu (Data Host) anlayışları istifadə olunur. Verilənlər provayderi özlərinin sahibi olduqları və ya nəzarət etdikləri verilənləri PREDICT vasitəsilə icazə verilmiş tədqiqatçılara əlyetər edir. Verilənlər hostu PREDICT verilənlər toplularını saxlamaq və onların icazə verilmiş tədqiqatçılara ötürülməsini koordinasiya etmək üçün hesablama infrastrukturu təqdim edən təşkilatdır.

PREDICT verilənlər toplularının provayderləri Kolorado Dövlət Universiteti, Miçiqan Universiteti, Viskonzin Universiteti, Cənubi Kaliforniya Universiteti – İnformasiya Elmləri İnstitutu (USC–ISI), Kaliforniya Universiteti San-Diyeqo (UCSD), Corciya Texnologiya İnstitutu (Georgia Tech), Merit Networks, SKAION və Packet Clearing House-dur.

PREDICT repozitarisində hazırda 526 verilənlər toplusu var. Toplama müddətləri bir neçə saatdan günlərə və aylara, verilənlərin həcmi isə baytlardan terabaytlara (TB) çatır. Cədvəl 1-də toplanmış verilənlərin kateqoriyalarına və həcmlərinə aid məlumatlar verilir.

CƏDVƏL 1. PREDICT TOPLUSUNA DAXİL OLAN BƏZİ VERİLƏNLƏR

Verilənlər Hostu/Provayderi	Verilənlərin kateqoriyası	Verilənlərin həcmi
UCSD/CAIDA	Internet Topology Data	1612.1 TB
UCSD/CAIDA	Blackhole ünvan fəzası verilənləri	2113.1 TB
USC-ISI	Trafik axını, Internet topologiya, Ünvan paylanması verilənləri	42 TB
Kolorado Dövlət Universiteti	Trafik axını verilənləri, Spam loqları, IP Reputasiya siyahıları	90 TB
Miçiqan Universiteti	Tətbiqi proqramlar səviyyəsi təhlükəsizlik verilənləri	120.0 TB
Merit Networks	Trafik axını verilənləri	1400.0 TB
Georgia Tech	Sinkhoul verilənləri	0.01 TB
Viskonzin Universiteti	IDS və şəbəkə ekranı verilənləri	2.0 TB
Packet Clearing House	Sintetik verilənlər	7083.4 TB
SKAION	Sintetik verilənlər	119.2 TB

PREDICT paylanmış verilənlər repozitarisidir, onun verilənləri Verilənlər provayderləri tərəfindən təqdim olunur və verilənlər hostlarında saxlanır və müraciət edilir. PREDICT Koordinasiya Mərkəzi (PKM) verilənlərin repozitariyə daxil edilməsini idarə edir, bu verilənlərə müraciət xahişlərini nəzərdən keçirir və icazə verir. PKM verilənlərin tamlığını və konfidensiallığını, eləcə də müvafiq məqsədlərdə istifadəsini təmin edəcək protokollarla işləyir.

Hazırda toplanmış verilənlər kateqoriyasına aşağıdakılar daxildir: ünvan fəzasının paylanması, BGP marşrutlama verilənləri, Blackhole ünvan fəzası, DNS, IDS və şəbəkələrarası ekran verilənləri, infrastruktur, Internet topologiyası, IP paket başlıqları, sintetik genereasiya edilmiş verilənlər, trafik axını, e-poçt spamı. PREDICT-ə təqdim edilən hər bir verilənlər toplusu korporativ siyasətə və öhdəliklərə uyğunluğu təmin etmək üçün verilənlər provayderinin hüquqi və etik şurası tərəfindən baxılmalı və icazə verilməlidir.

Repozitaridəki verilənlər toplusu üçün qeyri-məhdud, kvazi-məhdud, və məhdud giriş hüquqları müəyyən olunur (qeyri-kommersiya məqsədləri üçün qeyri-məhdud və ya kvazi-məhdud hüquqları da vardır). Məhdud giriş sinfinə aid verilənlərə giriş əldə etmək üçün yazılı razılaşma memorandumu tələb edilir:

PREDICT portalı (www.predict.org) bu repozitariyə giriş nöqtəsidir və əlyetər verilənlər toplularını təsvir edən kataloqa malikdir. Lakin verilənlərə müraciət xahişi göndərmək üçün portalda istifadəçi hesabı yaratmaq tələb edilir.

İstifadəçi hesabı yaradılması üçün bütün müraciətlərə PKM tərəfindən baxılır və icazə verilir; lakin digər ölkələrdən olan tədqiqatçılara hesab açılmasına icazə verilməsi prosesində müraciətin edildiyi ölkədə olan PREDICT koordinatorunun rəyi və icazəsi tələb edilir. Verilənlər toplusundan istifadə etmək üçün kibertəhlükəsizlik sahəsində tədqiqat aparan ABŞ və bir neçə ölkə (Avstraliya, Birləşmiş Krallıq, Kanada, İsrail və Yaponiya) tədqiqatçıları müraciət edə bilərlər.

IV.WINE VERİLƏNLƏR TOPLUSU

Zərərli proqramların aşkarlanması sahəsində tədqiqatlar üçün yararlı olan real iş mühitində toplanmış verilənlər toplusuna nümunə Symantec Research Labs tərəfindən nəzarət edilən WINE toplusudur [15]. WINE verilənlər toplusuna zərərli proqram nümunələri, binar reputasiya, URL-reputasiya, e-poçt spamı, anti-virus telemetriya məlumatları daxildir (Cədvəl 2). Zərərli proqramlar kolleksiyasına siqnaturalarının yaradılmasında istifadə edilmiş zərərli proqram nümunələri (virus, soxulcan, bot və s.) daxildir. Binar reputasiya verilənləri naməlum binar fayllar haqqında məlumat təmin edir (məsələn, antivirus siqnatutası hələ yaradılmamış fayllar), bu fayllar Symantec reputasiya əsasında təhlükəsizlik proqramında iştirak edən istifadəçilər tərəfindən göndərilir. Anti-virus telemetriyası Symantec tərəfindən siqnaturası yaradılmış və antivirusların aşkarladığı məlum təhdidlərin baş verməsini də qeydə alır. WINE verilənlər toplusu informasiyanın nə zaman, harada və necə toplandığını göstərən metaverilənləri də müəyyən edir.

CƏDVƏL 2: WINE VERİLƏNLƏRİ

Verilənlər toplusu	Mənbələr
Binar reputasiya	50 milyon kompüter
Anti-virus telemetriyası	130 milyon kompüter
E-poçt spamı	2.5 milyon tələ-hesab
URL reputasiya	10 milyon domen
Zərərli proqram nümunələri	200 ölkə

WINE giriş modelində tədqiqatçı verilənlər toplusu kataloquna baxır və verilənlərə giriş üçün sorğu hazırlayır. Sorğuların doğruluğu və soruşulan verilənlərin əlyetərliyi daxili və xarici tədqiqatçıların məsləhət şurası tərəfindən qiymətləndirilir. Verilənlərə giriş icazəsi verilmiş tədqiqatçılar WINE infrastrukturu olan laboratoriyaya gəlməli və tədqiqatlarını orada aparmalıdır. WINE istifadə edilməklə yaradılan intellektual mülkiyyət tədqiqatçıların özlərinə məxsusdur və onlar öz nəticələrini məsuliyyətlə çap etmələri təşviq edilir.

V. BOTNET TRAFİKLƏRİ ÜZRƏ VERİLƏNLƏR

Botnetlərin aşkarlama metodlarının qiymətləndirilməsi real trafiki qəbul edilən səviyyədə imitasiya edən kifayət qədər qeyri-bircins olan verilənlərlə eksperimentlərin aparılmasını tələb edir. Botnet detektorlarının praktikada effektivliyi onların qiymətləndirilməsi zamanı istifadə edilən reallığa yaxın botnet trafikindən çox asılıdır. Belə verilənlərin yoxluğu botnetlərin aşkarlanmasını qiymətləndirmək üçün ədəbiyyatda dəfələrlə vurğulanmış problemlər səbəbindəndir [16,17].

Botnetlərin qiymətləndirilməsi üçün istifadə edilən verilənlər toplusu aşağıdakı problemlər nəzərə alınmaqla yaradılmalıdır [17]:

Ümumilik. Mövcud botnet verilənləri toplularında yalnız bir neçə (adətən 2-3) botnetdən götürülmüş verilənlər olur. Təbiətcə məhdud olan bu yanaşmalar qeyri-pratik və yeni təhdidlərə qarşı qeyri-dəqiqdir (belə verilənlərlə yaradılmış detektorlar çox spesifik olan botnet davranışlarının çox az sayda xarakteristikasını əks etdirir).

Realizm. Botnet trafiki nəzarət edilən mühitdə generasiya edilir və ya toplanır. Botnetin nəzərdə tutulmuş bütün zərərli fəaliyyətini yerinə yetirməsi üçün yararlı mühit (botnet tərəfindən aşkarlanmayan) təmin etmək asan məsələ deyil. Eyni zamanda, toplama müddəti kifayət qədər uzun olmalıdır ki, “yuxuya getmiş” botnetlər də öz funksiyalarını nümayiş etdirsinlər.

Reprezentivlik. Botnet verilənlərini toplanması zamanı daha bir problem detektorun praktiki istismarı zamanı qarşılaşacağı real mühiti əks etdirən şəbəkə trafikini toplamaq imkanıdır. Gizlilik tələbləri səbəbindən real istismar mühitində olan fon trafikini toplamaq bir çox halda mümkün deyil, nəticədə trafik ya imitasiya edilir, ya da nəzarət edilən mühitdə toplanılır.

UNB ISCX Botnet verilənləri toplusu yuxarıdakı meyarlar nəzərə alınmaqla yaradılıb və topluya aşağıdakılar daxildir:

- ISOT verilənlər toplusu müxtəlif verilənlər topluları birləşdirməklə yaradılıb [17]. Topluda həm zərərli (Storm və Zeus botnetlərindən), həm də zərərli olmayan trafik (HTTP trafiki, bittorrent) vardır.
- Malware Capture Facility Project tərəfindən generasiya edilmiş botnet trafiki [18] – uzunmüddətli botnet trafikinin generasiyası və toplanması üzrə tədqiqat layihəsidir. Bu topludan təlim verilənləri üçün 4 botnet izi (Neris, Rbot, Virut və NSIS) və test verilənləri üçün (Neris, Rbot, Virut, NSIS, Menti, Sogon və Murla) doqquz botnet izi seçilmişdir.

Bu botnet izlərini vahid topluda birləşdirmək üçün overley metodologiyası istifadə edilmişdir, sintetik verilənlər toplusunun yaradılması üçün ən populyar metoddur [17]. Zərərli verilənlər adətən honeypotlar və ya botnet binar kodu ilə yoluxmuş kompüterlər vasitəsilə nəzarət edilən mühitdə toplanılır. Botnet izlərini normal verilənlərlə birləşdirmək üçün zərərli verilənlər ya daxili, ya da xarici şəbəkədə mövcud olan kompüterlərə inikas etdirilir. Sonra TCPReplay istifadə edilməklə zərərli və normal trafik qarışdırılıb və TCPdump vasitəsilə vahid topluda toplanılıb.

NƏTİCƏ

Bu işdə informasiya təhlükəsizliyi üzrə tədqiqatlarda istifadə edilən verilənlər topluları analiz edilmişdir. Analizin nəticələrinə görə, bəzi topluların köhnəlmiş olmasına və eksperimentlərin nəticələrinə təsir edən nöqsanlara malik olmalarına baxmayaraq, yenə də istifadə edilir. Bir sıra mühüm verilənlər toplusuna isə yalnız müəyyən ölkə tədqiqatçılarının giriş imkanları vardır. Daha bir müşahidəyə görə, qabaqcıl tədqiqat qrupları bir qayda olaraq, verilənlərin toplanmasına və nişanlanmasına da yetərli resurslar ayırırlar.

ƏDƏBİYYAT

[1] S.E. Coull, E. Kenneally, “A qualitative risk assessment framework for sharing computer network data,” 2012 TRPC. http://dx.doi.org/10.2139/ssrn.2032315

[2] J.Camp et al. Data for Cybersecurity Research: Process and Wish List. NSF Workshop on Cyber Security Data for Experimentation, 2010. http://www.cc.gatech.edu/fac/feamster/papers/data-wishlist.pdf

[3] Lincoln Laboratory, M.I.T., 2011. DARPA Intrusion Detection Evaluation. http://www.ll.mit.edu/mission/communications/ist/corpora/ideval/index.html.

[4] University of California, 2011. KDD Cup 1999 Data. http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html.

[5] Lawrence Berkeley National Laboratory, 2010. The Internet Traffic Archive. http://ita.ee.lbl.gov/index.html..

[6] Lawrence Berkeley National Laboratory and ICSI, LBNL/ICSI Enterprise Tracing Project. www.icir.org/enterprise-tracing/.

[7] CAIDA Data. http://www.caida.org/data/

[8] The Shmoo Group, 2011. Defcon. http://cctf.shmoo.com/.

[9] B. Sangster, T.J. O'Connor, T. Cook, R. Fanelli, E. Dean, J. Adams, C. Morrell, and G. Conti, "Toward Instrumenting Network Warfare Competitions to Generate Labeled Datasets;" USENIX Security's Workshop on Cyber Security Experimentation and Test (CSET), 2009.

[10] PREDICT, 2011. http://www.predict.org/.

[11] A. Shiravi, H. Shiravi, M. Tavallaee, and A. A. Ghorbani, “Towards developing a systematic approach to generate benchmark datasets for intrusion detection,” Computers & Security, Vol. 31, No. 3, pp. 357–374, 2012.

[12] J. McHugh, “Testing intrusion detection systems: A critique of the 1998 and 1999 DARPA intrusion detection system valuations as performed by Lincoln Laboratory,”ACM Transactions on Information and System Security, Vol. 3, No.4, pp. 262–294, November 2000.

[13] M. Tavallaee, N. Stakhanova, and A. A. Ghorbani, “Toward credible evaluation of anomaly-based intrusion-detection methods,” Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on, Vol. 40, No. 5, pp. 516–524, 2010.

[14] S. Revathi, A. Malathi “A detailed analysis on NSL-KDD dataset using various machine learning techniques for intrusion detection,” International Journal of Engineering Research & Technology (IJERT), Vol. 2, Issue 12, pp. 1848-1853, December 2013.

[15] T. Dumitras, D. Shou, “Toward a standard benchmark for computer security research: The Worldwide Intelligence Network Environment (WINE),” Proc. of the First Workshop on Building Analysis Datasets and Gathering Experience Returns for Security, pp. 89-96, 2011.

[16] A. J. Aviv and A. Haeberlen, “Challenges in experimenting with botnet detection systems,” in USENIX 4th CSET Workshop, 2011.

[17] D. Zhao, I. Traore, B. Sayed, W. Lu, S. Saad, A. Ghorbani, and D. Garant,“Botnet detection based on traffic behavior analysis and flow intervals,”Computers & Security, 2013.

[18] S.García, (2013). Malware Capture Facility Project. CVUT University. Retrieved February 03, 2013, from https://agents.fel.cvut.cz/malware-capture-facility.