نشریه علمی "پدافند الکترونیکی و سایبری"
سال نهم، شماره ۱، بهار ۱۴۰۰، ص ۱۴۷-۱۳۷

علمی – پژوهشی

کاوش: ارائه روش تحلیل بات‌نت و تأثیر ترافیک عادی شبکه بر مرحله انتخاب و استخراج ویژگی مبتنی بر فاصله مینکوفسکی

محمدجواد فقیه‌نیا[۱]، رضا جلایی[۲*]، حامد شجاعی یاس[۳]

۱ و ۳- کارشناسی ارشد، مرکز تحقیقات صدر، ۲- استادیار، دانشگاه جامع امام حسین(ع)، تهران، ایران

(دریافت: ۱۳۹۹/۰۳/۲۱، پذیرش: ۱۳۹۹/۰۸/۰۵)

چکیده

گسترش روزافزون تهدید بات‌نت و توسعه بسترهای جدید استقرار بات‌نت مانند اینترنت اشیا، لزوم مقابله را نشان می‌دهد. پژوهش‌هایی که در حوزه تشخیص بات‌نت مبتنی بر روش‌های یادگیری ماشین انجام شده است؛ نشان می‌دهد این روش‌ها کارایی لازم را جهت تشخیص بات‌نت دارند. این درحالی است که عدم وجود یک مجموعه دادگان استاندارد در این حوزه، یکی از چالش‌ها در سامانه‌های تشخیص بات‌نت است که موجب افزایش نرخ خطا و کاهش نرخ تشخیص در محیط واقعی می‌شود. در این مقاله، ترافیک عادی و بات‌نت با ارائه روشی مبتنی بر بردار فاصله مینکوفسکی تحلیل شده است. نتایج مقاله نشان می‌دهد که جریان ترافیک عادی، مرحله انتخاب و استخراج ویژگی را با تغییر در اهمیت ویژگی‌ها مؤثر می‌کند. این روش به ویژگی‌ها بر اساس نزدیک نمودن بردارهای رفتاری بات-بات و دور نمودن بردارهای رفتاری بات‌عادی امتیاز می‌دهد. نتایج این آزمایش‌ها بر روی ده مجموعه دادگان عادی و سه مجموعه دادگان بات، نشان داد امتیاز یک ویژگی در محیط‌هایی با ترافیک عادی متفاوت بیش از ۵۰٪ افزایش یا کاهش دارد.

کلیدواژه‌ها: بات‌نت، جریان شبکه، استخراج ویژگی، فاصله مینکوفسکی، فاصله رفتاری

۱- مقدمه

اولین بدافزار رایانه‌ای، به نام ویروس برین[۱] در سال ۱۹۸۶ نوشته شد [۱]. دو سال بعد و به‌طور تصادفی اولین کرم توسط رابرت موریس[۲] در دانشگاه MIT نوشته شد. اما کدرد[۳] اولین کرمی بود که در سال ۲۰۰۰ در سطح جهان گسترش یافت. از همان سال‌های اولیه ایجاد بدافزارها، چالش‌های مقابله با آن‌ها نیز مورد توجه محققان قرار گرفت. در سال‌های اخیر نوعی از بدافزار به نام بات‌نت وارد عرصه شده است که به‌دلیل بهره‌مندی از کانال فرمان و کنترل[۴] خطرناک‌تر از انواع قبلی خود هستند. آن‌ها با استفاده از این کانال، امکان مدیریت و ارسال فرمان مخرب را در میزبان‌های مختلف فراهم می‌کنند و از این‌رو تهدید بزرگ‌تری هستند.

تفاوت اصلی بات‌نت و سایر بدافزارها[۵] زیرساخت فرمان و کنترل آن‌ها است [۲]. بات‌ها ماشین‌های آلوده‌ای هستند که یک فایل اجرایی از بدافزار را اجرا می‌کنند [۳] و هر بات‌نت مجموعه‌ای از بات‌ها است که اقدام به دریافت و ارسال دستورات از طریق کانال فرمان و کنترل خود می‌کند. این کانال، سازوکاری است که توسط مدیر بات[۶] ایجاد شده است. به‌طور کلی یک بات‌نت از سه جزء اصلی تشکیل می‌شود که شامل مدیربات، سرویس‌دهنده کانال فرمان و کنترل و قربانی است [۳].

تشخیص ترافیک تولیدشده در شبکه با استفاده از تجزیه و تحلیل و شناسایی تولیدکننده آن ترافیک از مسائل مورد بحث محققان است. آن‌ها به دنبال روشی هستند که بتواند ترافیک رمز شده در شبکه را به شکل بلادرنگ مورد تحلیل قرار دهد. در حال حاضر اکثر روش‌های مبتنی بر تشخیص رفتار شبکه از دو نوع داده ورودی برای الگوریتم تشخیص استفاده می‌کنند که شامل بسته‌ها[۷] و جریان شبکه[۸] است. مزیت اصلی در روش دریافت بسته‌های شبکه در اختیار داشتن کامل بدنه[۹] و اطلاع از تمامی بیت‌های[۱۰] ارسالی است؛ اما این روش دارای کاستی‌هایی است که

*رایانامه نویسنده مسئول: Rjalaei@ihu.ac.ir

[1] Brain
[2] Robert Tappan Moris
[3] Code Red
[4] Command and Control
[5] Malware
[6] Botmaster
[7] Packets
[8] Flow
[9] Payload
[10] Bits

مهم‌ترین آن‌ها نیاز به فضای ذخیره‌سازی زیاد و نقض حریم خصوصی است [۲].

تحلیل رفتار بات‌نت‌ها مبتنی بر جریان شبکه، روشی جهت تشخیص بات‌نت است که از فراداده[1] استفاده می‌کند. در این مقاله روشی پیشنهادی می‌شود که علاوه بر در نظر گرفتن رفتار ترافیکی بات-بات، مدلی از امتیازدهی پیشنهاد می‌دهد که رفتار ترافیکی بات- عادی را نیز درنظر می‌گیرد. این در حالی است که در روش‌های قبلی تمرکز بر روی رفتار بات-بات است. روشی که در این مقاله ارائه می‌شود اهمیت ویژگی را هم بر اساس مشابهت رفتاری بات-بات و هم عدم مشابهت رفتاری بات-عادی درنظر می‌گیرد. فاصله برداری مینکوفسکی و معکوس آن نیز ارزش ویژگی‌ها را کمی می‌کند تا تأثیر ترافیک عادی به‌طور دقیق تحلیل شود.

در ادامه نوآوری‌های روش پیشنهادی ارائه می‌شوند.

۱- ارائه روشی جهت بررسی میزان اثرگذاری ترافیک عادی بر اهمیت ویژگی‌ها
۲- ارائه روش تحلیلی بردار جریان مبتنی بر فاصله رفتاری. در این مقاله با استفاده از تشکیل نمودار جریان-ویژگی بردار کلی رفتار بات و یا ترافیک عادی استخراج می‌شود.
۳- تحلیل مینکوفسکی یک بعدی، دوبعدی و سه بعدی با هدف کمی‌سازی فاصله رفتاری ترافیک بات-بات و بات- عادی.
۴- رده‌بندی و امتیازدهی ویژگی‌های منتخب با در نظر گرفتن ترافیک عادی و بدون آن.
۵- به‌کارگیری مفهوم فاصله مینکوفسکی معکوس جهت امتیازدهی به ویژگی‌ها در ترافیک بات-بات

ساختار مقاله به شرح زیر تنظیم و نگارش شده است، در ادامه و در بخش ۲ مفاهیم پایه مرتبط با این پژوهش مرور خواهند شد. بخش ۳ پژوهش‌های مرتبط را مرور می‌کند. بخش ۴ روش پیشنهادی و در نهایت در بخش ۵ نتایج حاصل از آزمایش‌های انجام شده ارائه می‌شود.

۲- مفاهیم پایه

در این بخش مفاهیم پایه مرتبط با این پژوهش مورد بررسی قرار می‌گیرد.

[1] Meta Data

۲-۱- جریان شبکه

بر طبق [۲۰] هر نرم‌افزار کاربردی و یا پروتکل ارتباطی، الگوی رفتاری منحصربه‌فرد از خود نشان می‌دهد، به‌طوری‌که می‌توان با بهره‌گیری از این الگوهای رفتاری، نرم‌افزار کاربردی مولد را تشخیص داد. یکی از روش‌های استخراج این الگوهای رفتاری استفاده از جریان‌های شبکه است. سامانه‌های تشخیص نفوذ مبتنی بر جریان شبکه نیز از داده‌های جریان شبکه جهت تشخیص نفوذ استفاده می‌کنند [۲۱].

جریان شبکه یا نت‌فلو[2] [۲]، اولین بار به‌عنوان یک مشخصه جدید در مسیریاب‌های سیسکو[3] در سال ۱۹۹۶ مطرح شد [۲۲]. سازمان استاندارد IETF[4] نسخه ۹ از نت‌فلو را تحت عنوان پروتکل استاندارد IPFIX[5] [۴] منتشر کرده است. این استاندارد شامل پنج ویژگی آدرس منبع[6]، پورت منبع[7]، آدرس مقصد[8]، پورت مقصد[9] و پروتکل[10] است.

۲-۲- مفاهیم آماری

در این مقاله جهت بررسی فاصله رفتاری، از فاصله مینکوفسکی[11] استفاده شده است که یکی از رایج‌ترین سنجه‌های محاسبه فاصله است. فاصله اقلیدسی مینکوفسکی از رابطه (۱) محاسبه می‌شود[۵].

$$d(i.j) = \sqrt{(x_{i1} - x_{j1})^2 + \cdots + (x_{ip} - x_{jp})^2} \qquad (۱)$$

i و j بیانگر منبع تولید جریان شبکه و p شماره ترتیب جریان‌های شبکه در بردار رفتاری است. برای مثال، نت‌اسکای=i ، $i = SSH$ و $p = 3$، جریان سوم تولیدشده توسط بات نت‌اسکای و SSH است.

همچنین در روش پیشنهادی از روش Min-Max برای نرمال‌سازی استفاده می‌شود:

$$A = \left(\frac{A - \min \text{value } of\ A}{\max value\ of\ A - \min value of\ A}\right) \times (D - C) + C \qquad (۲)$$

A مقادیر بازه موجود و [C, D] بازه‌ای است که مقادیر A در آن نگاشت می‌شوند. و در اینجا [۱۰, ۰] در نظر گرفته می‌شود.

[2] Netflow
[3] Cisco
[4] Internet Engineering Task Force
[5] Ip Flow Information Export
[6] Source IP
[7] Sourcr Port
[8] Destination IP
[9] Destination Port
[10] Protocol
[11] Minkowski

۳- مروری بر پژوهش‌های مرتبط

تشخیص کانال فرمان و کنترل بات‌نت با استفاده از روش‌های یادگیری ماشین[1] تا حد زیادی ضعف‌های سامانه‌های مبتنی بر امضا[2] را پوشش می‌دهد. از ۳۶ روش مورد بررسی در زمینه تشخیص بات‌نت، ۳۰ روش از روش‌های یادگیری ماشین استفاده می‌کنند، این فراوانی نشان می‌دهد که روش‌های یادگیری ماشین در بین محققان نیز از محبوبیت برخوردار است [۶]. در مقابل به دلیل شباهت رفتاری ترافیکی بات و ترافیک عادی[3] در شبکه، وجود نرخ مثبت غلط[4] در این سامانه‌ها اجتناب‌ناپذیر است. روش‌های پیشنهادی اخیر در حوزه تشخیص کانال فرمان و کنترل توانسته‌اند نرخ مثبت غلط را تا ۱٪ کاهش دهند [۷ و ۸]. اما این محاسبات وابسته به مجموعه دادگانی[5] است که جهت آموزش[6] و آزمون[7] از آن استفاده شده است. به بیان دیگر ارزیابی یک روش یادگیری ماشین با مجموعه دادگان مختلف ممکن است نتایج متفاوتی داشته باشد. بشرا[8] و همکاران [۷] نشان می‌دهند که تعداد مثبت غلط وابسته به ترافیک عادی، متفاوت است. در واقع ترافیک عادی متفاوت، نرخ مثبت غلط متفاوتی ایجاد می‌کند.

این مقاله، نشان می‌دهد که در نظر گرفتن ترافیک عادی در مرحله استخراج و انتخاب ویژگی[9] باعث افزایش دقت روش‌های مبتنی بر یادگیری ماشین می‌شود.

در ادامه، کارهای مرتبط با این مقاله در دو حوزه بررسی روش‌های تشخیص و مجموعه دادگان موجود مورد بررسی قرار می‌گیرد.

۳-۱- روش‌های تشخیص بات

گارسیا[10] و همکاران [۹] با بررسی رفتار یک بات طی ۵۷ روز، اقدام به انتخاب ویژگی‌های اندازه‌ی جریان شبکه[11] و مدت زمان جریان شبکه[12] و استخراج ویژگی تناوبی[13] کردند. روش آن‌ها با مدل‌سازی رفتار بات مبتنی بر زنجیره مارکوف[14] تشخیص با دقت ۹۶٪ را به‌دست آورد. باتکشن[15] [۷] روشی است که از طریق مدل‌سازی رفتاری بات مبتنی بر زنجیره مارکوف اقدام به تشخیص بات‌ها در شبکه می‌کند. در این روش با ارزیابی بات‌های مختلف سه ویژگی پروتکل[16]، سرویس[17] و حالت اتصال[18] از ابزار زیک[19] انتخاب شده است. باتکشن در مرحله تشخیص، ترافیک عادی را نیز در نظر می‌گیرد. عبدالرحمن[20] و همکاران [۸] روش نوینی مبتنی بر ویژگی‌های جریان شبکه با استفاده از شبکه عصبی عمیق ارائه داده‌اند که شامل سه مرحله استخراج ویژگی، ساخت مدل رده‌بندی و ارزیابی است. نوآوری اصلی این روش، ایجاد گراف ارتباط[21] در مرحله استخراج ویژگی است. آن‌ها با تحلیل بات نرسیز[22] ویژگی‌های اولیه شامل، مدت زمان، اندازه، تعداد بسته‌ها، تناوبی بودن، حالت و تعداد ارتباط TCP را بررسی کردند. در سال ۲۰۲۰ نیز سوزان[23] و همکاران [۱۰] یک روش ترکیبی مبتنی بر تحلیل میزبان و شبکه جهت تشخیص بات پیشنهاد دادند. در این روش نیز با بررسی بات‌های مختلف اقدام به استخراج و انتخاب ۱۱ ویژگی شده است.

در تمامی روش‌های تشخیص بررسی شده که دقت تشخیص آن‌ها بیش از ۹۶٪ بوده است، می‌توان عدم توجه به ترافیک عادی را در مرحله انتخاب و استخراج ویژگی مشاهده کرد. در این روش‌ها، ویژگی مدنظر با تحلیل رفتار بات انتخاب می‌شود. این مقاله با ارائه روشی مبتنی بر فاصله مینکوفسکی نشان می‌دهد که ویژگی‌هایی در تشخیص بات موثرتر هستند که علاوه بر نزدیک نمودن بردار رفتاری بات‌ها در شبکه، بردار رفتاری عادی و بات را نیز از یکدیگر دور می‌کنند.

۳-۲- مجموعه دادگان موجود

در روش‌های مبتنی بر یادگیری ماشین، مجموعه دادگان اهمیت ویژه‌ای دارد. در صورتی‌که مجموعه دادگان منتخب نتواند محیط واقعی را شبیه‌سازی کند، سیستم حاصل از آموزش با آن مجموعه دادگان، برای به‌کارگیری در محیط واقعی مناسب نیست. در انتخاب مجموعه داده بات‌نت، شرایط زیر در نظر گرفته می‌شود [۶].

۱- ایجاد ترافیک پس‌زمینه[24] با بیش از نیمی از میزبان‌های عادی در هر سناریو

[1] Machine Learning
[2] Signiture
[3] Normal
[4] False Positive Rate
[5] Dataset
[6] Train
[7] Test
[8] Bushra
[9] Feature Extraction and Feature Selection
[10] Garcia
[11] Flow Size
[12] Flow Duration
[13] Periodic
[14] Markov Chain
[15] BOTection
[16] Protocol
[17] Service
[18] Connetion State
[19] Zeek
[20] Abdurrahman
[21] Commiunication Graph
[22] Nersis
[23] Suzan
[24] Background

۲- اجرای طولانی باتها به‌طوری‌که یک دور کامل از رفتار خود را نشان دهند.

۳- برچسب‌گذاری ترافیک‌های آلوده به بات و عادی و مشخص کردن کانال‌های فرمان و کنترل

۴- اطمینان از صحت عملکرد کانال فرمان و کنترل

۵- به‌کارگیری باتها با چرخه حیات متفاوت

مجموعه دادگان ISOT توسط شریف و همکاران [۱۱] ایجاد شده است که ترکیبی از مجموعه دادگان موجود قبلی است. ترافیک باتهای زئوس[1] و ولداک[2] در این مجموعه دادگان وجود دارد و برچسب‌گذاری آن مبتنی بر بسته[3] است.

مجموعه دادگان ISOT HTTP توسط النازی[4] و همکاران[۱۲] ایجاد شده است. در این مجموعه دادگان تمرکز بر ایجاد ترافیک DNS بات و عادی است. نقطه قوت این مجموعه دادگان، وجود تنوع در نرم‌افزارهای کاربردی عادی است (جدول (۱)).

جدول (۱): مقایسه مجموعه دادگان

نام مجموعه دادگان	ترافیک عادی	زمان اجرای طولانی	برچسب جریان	اطمینان از برقراری فرمان و کنترل	تحلیل رفتاری
ISOT Botnet[11]	✓	✓	✗	✗	✗
ISOT HTTP Botnet[12]	✓	✓	✗	✗	✗
CTU-13 Dataset[13]	✓	✓	✓	✓	✗
Botnet Dataset[14]	✓	✓	✗	✗	✗
Android Botnet[15]	✓	✓	✗	✗	✗
Bot-IOT Dataset[16]	✓	✓	✓	✓	✓

مجموعه دادگان CTU-13 [۱۳] به دلیل داشتن ویژگی‌هایی همچون اجرای طولانی، برچسب‌گذاری جریان شبکه، اطمینان از عملکرد بات و به‌کارگیری ترافیک پس‌زمینه، در کنار ترافیک عادی یکی از مجموعه دادگان محبوب در بین محققان است. در این مجموعه دادگان از چندین خانواده باتنت مانند نرسیز، آربات[5]، ویروت[6]، منتی[7]، سوگو[8]، انسیز[9] و مورلو[10] استفاده شده است. جریان ترافیک عادی نیز در این مجموعه دادگان برچسب‌گذاری شده است.

بیگی[11] و همکاران [۱۴] با ترکیب سه مجموعه دادگان قبل از خود [۱۱، ۱۳ و ۱۷] مجموعه دادگانی ارائه کرده است که طیف وسیعی از خانواده‌های بات را دارد. برچسب‌گذاری در این مجموعه دادگان مبتنی بر IP است.

حبیبی و همکاران [۱۵] مجموعه دادگانی متشکل از ۱۴ خانواده مختلف از باتهای اندرویدی ایجاد نمودند. در مرجع [۱۶] مجموعه دادگان جدیدی با تمرکز بر باتهای IOT[12] ارائه شده است. این مجموعه داده به جای استفاده از خانواده‌های مختلف باتنت از حملات رایج در این نوع باتها استفاده می‌کند. حملات استخراج داده[13]، کی لاگینگ[14]، پویش سیستم عامل[15] و پویش سرویس[16] در این مجموعه دادگان جهت ایجاد رفتار باتنت اجرا شده است.

بررسی مجموعه دادگان موجود نشان می‌دهد که تلاش‌های صورت گرفته جهت ایجاد مجموعه دادگان بیشتر جهت تولید رفتار باتها است و به حضور هم‌زمان کاربران و نرم‌افزارهای کاربردی قانونی در شبکه آلوده، توجه کمتری شده است. این موضوع ناشی از گستردگی ترافیک عادی در شبکه است. به‌بیان‌دیگر، هزاران نرم‌افزار کاربردی قانونی، وب‌سایت و کاربری شبکه وجود دارد که گنجاندن همه‌ی آن‌ها در مجموعه دادگان، عملی نیست. این مقاله به دلایل زیر، مجموعه دادگان جدیدی تولید کرده است،

۱- در اکثر مجموعه دادگان از برچسب عادی جهت نشان دادن ترافیک عادی استفاده می‌شود. این در حالی است که در آزمایش‌های این مقاله لازم است ترافیک هر نرم‌افزار کاربردی، متناظر با هر بردار رفتاری، مقایسه شود.

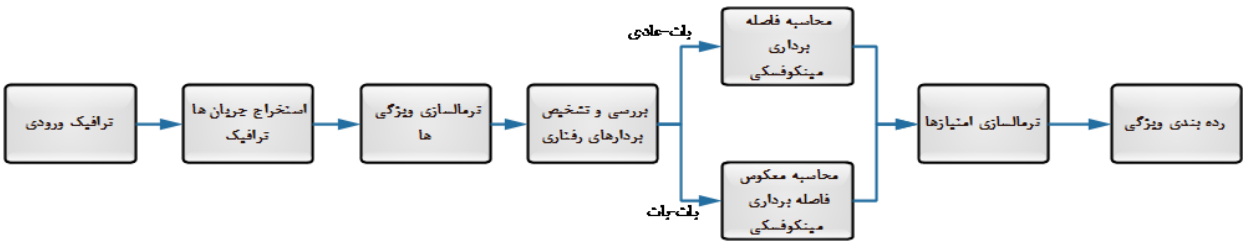
۲- اجرای هم‌زمان ترافیک باتنت و عادی در یک محیط جهت ارزیابی اثر آن‌ها بر هم.

بنابراین با ایجاد آزمایشگاهی متشکل از میزبان‌های آلوده به بات و کاربران عادی، مجموعه دادگان جدیدی تولید شد.

[1] Zeus
[2] Waledac
[3] Packet
[4] Alenazi
[5] Rbot
[6] Virut
[7] Menti
[8] Sogou
[9] NSIS
[10] Mulrlo
[11] Beigi
[12] Internet of Things
[13] Data Exfiltration
[14] Keyloging
[15] OS Scan
[16] Service Scan

۴- روش پیشنهادی

برای بررسی و تحلیل میزان اثرگذاری ترافیک عادی شبکه بر مرحله انتخاب و استخراج ویژگی، ترافیک تولیدشده توسط هریک از این منابع، مدل‌سازی شده و از نظر میزان فاصله با یکدیگر مقایسه می‌شوند. این کار در دو مرحله انجام می‌شود، در مرحله اول جریان‌های ترافیک تولیدشده توسط بات و منابع عادی، با حذف درگاه منبع[1] در دسته‌هایی به نام "چهارتایی[2]" دسته‌بندی می‌شوند. و در مرحله دوم، با اضافه کردن توالی زمانی از مقادیر هر ویژگی، بردار رفتاری آن منبع و مبتنی بر ویژگی منتخب، مدل می‌شود. به دلیل تفاوت در مقادیر ویژگی‌ها قبل از محاسبه فاصله بردارهای رفتاری، تمامی آن‌ها نرمال‌سازی[3] می‌شوند. در ادامه فاصله برداری محاسبه می‌گردد. در این روش مؤثرترین ویژگی، باید کم‌ترین فاصله بین بات‌ها و بیشترین فاصله را از بردار رفتاری عادی ایجاد کند. به بیان دیگر هر ویژگی، بالاترین امتیاز مثبت صحیح بیشتر و مثبت غلط کم‌تری ایجاد می‌کند. با توجه به آنچه گفته شد، در این مرحله از بردار فاصله مینکوفسکی[4] یک بعدی، دوبعدی و سه بعدی جهت محاسبه فاصله برداری ترافیک بات از ترافیک عادی استفاده می‌شود. از طرفی در محاسبه فاصله رفتاری ترافیک بات‌ها به دلیل اینکه کم‌ترین فاصله باید بیشترین امتیاز را دریافت کند، از فاصله مینکوفسکی معکوس استفاده می‌شود. برای رده‌بندی ویژگی‌های مختلف، میانگین امتیازهای کسب‌شده در مرحله تحلیل فاصله رفتاری استفاده شده است. هر ویژگی که در مجموع مقایسه‌ها، میانگین امتیاز بیشتری کسب کند در رده‌بندی، جایگاه بهتری خواهد داشت. مراحل روش امتیازدهی پیشنهادی در شکل (۱) نشان داده شده است.



شکل (۱): روش پیشنهادی امتیازدهی به ویژگی‌ها

۴-۱- تحلیل آماری

پس از استخراج بردار رفتاری هر یک از منابع تولید ترافیک منتخب، فاصله‌ی مینکوفسکی بین این بردارها بررسی می‌شود. معیار فاصله مینکوفسکی برای ویژگی‌های مختلف، مقادیر متفاوتی دارد. هفت ویژگی اندازه (s)، مدت زمان (d)، تناوب (p)، اندازه- مدت زمان (sd)، اندازه- تناوب (sp)، مدت زمان- تناوب (dp) و اندازه- مدت زمان- تناوب (sdp) برای ارزیابی فاصله رفتاری ترافیک عادی با بات‌نت‌ها انتخاب شده است. با توجه به اینکه معیارها با چندین متغیر در نظر گرفته می‌شود، برای بررسی حالات ترکیبی (sd، sp، dp)، از فاصله اقلیدسی در فضای دوبعدی برای هر متغیر استفاده می‌شود (رابطه (۳).

$$d = \sqrt{(s_2 - s_1)^2 + (p_2 - p_1)^2} \qquad (۳)$$

برای بعد dp نیز از رابطه (۴) استفاده می‌شود.

$$d = \sqrt{(d_2 - d_1)^2 + (p_2 - p_1)^2} \qquad (۴)$$

همچنین از رابطه (۵) برای محاسبه هر نقطه در ابعاد sd استفاده می‌شود.

$$d = \sqrt{(s_2 - s_1)^2 + (d_2 - d_1)^2} \qquad (۵)$$

در نهایت نیز از رابطه (۶) مینکوفسکی برای بردار با نقاط دوبعدی sp استفاده می‌شود.

$$d(i.j) = \sqrt{\begin{matrix}(s_{i1} - s_{j1})^2 + (p_{i1} - p_{j1})^2 + \cdots \\ \cdots + (s_{ip} - s_{jp})^2 + (p_{ip} - p_{jp})^2\end{matrix}} \qquad (۶)$$

در فضای سه‌بعدی نیز، فاصله با استفاده از رابطه (۷) محاسبه می‌شود.

$$d = \sqrt{(s_2 - s_1)^2 + (p_2 - p_1)^2 + (d_2 - d_1)^2} \qquad (۷)$$

که با جایگذاری در رابطه فاصله اقلیدسی مینکوفسکی، رابطه (۸) حاصل می‌شود.

$$d(i.j) = \sqrt{\begin{matrix}(s_{i1} - s_{j1})^2 + (p_{i1} - p_{j1})^2 + (d_{i1} - d_{j1})^2 + \cdots \\ \cdots + (s_2 - s_1)^2 + (p_2 - p_1)^2 + (d_2 - d_1)^2\end{matrix}} \qquad (۸)$$

[1] Sorce Port
[2] 4-Tuple
[3] Normalize
[4] Minkowski

جدول (۲): مجموعه دادگان بات

ردیف	وضعیت‌های مختلف شبکه بات	مدت اجرا (ساعت)	تعداد جریان
۱	۱- راه‌اندازی شبکه محلی با ۱۰ میزبان ۲- آلوده سازی میزبان اول با زئوس ۳- آلوده سازی میزبان دوم با نکرس	۹۰	۲۴۷۷۹۶
۲	۱- راه‌اندازی شبکه محلی با ۱۰ میزبان ۲- آلوده سازی میزبان اول با نت‌اسکای	۸۹	۶۲۵۱۹۷
۳	۱- راه‌اندازی شبکه محلی با ۱۰ میزبان ۲- آلوده سازی میزبان اول با ولداک	۶۷	۱۰۱۳۴۷

۵- سکوی آزمایش

در این قسمت مراحل ایجاد مجموعه دادگان و انتخاب ویژگی‌ها و نحوه‌ی اجرای آزمایش‌ها ارائه می‌شود.

۵-۱- نحوه اجرای آزمایش

در این بخش، مراحل تولید ترافیک کانال فرمان و کنترل، تولید ترافیک عادی، برچسب‌زنی بر روی ترافیک فرمان و کنترل، انتخاب ویژگی جریان شبکه و نحوه اجرای آزمایش ارائه می‌شود.

- **تولید ترافیک کانال فرمان و کنترل**

در این بخش نحوه اجرای بات‌های زئوس، ولداک و نت‌اسکای در محیط آزمایشگاهی ارائه می‌شود. نکته مهم در بررسی ترافیک فرمان و کنترل، اطمینان از صحت عملکرد بات و ایجاد صحیح کانال فرمان و کنترل است. مجموعه دادگان ایجاد شده در این مرحله، مطابق با شرایط مرجع [۱۸] است.

- **تولید ترافیک عادی**

مجموعه دادگان عادی شبکه به ترافیکی اشاره دارد که توسط کاربر عادی و همچنین نرم‌افزارهای کاربردی قانونی ایجاد شده است. یک کاربر عادی می‌تواند از طریق جستجو و وب‌گردی بر روی پورت ۴۴۳ ایجاد ترافیک کند. عملیات دانلود، خواندن ایمیل، دسترسی راه دور به میزبانی دیگر از طریق SSH[1] از دیگر مواردی هستند که توسط کاربر ایجاد ترافیک می‌کنند. این ترافیک‌ها به دلیل اینکه توسط یک کاربر انسانی ایجاد شده‌اند در اندازه، مدت زمان و تناوبی بودن جریان‌ها، رفتاری تصادفی دارند.

مجموعه دادگان N-1، برای بررسی رفتار عادی شبکه ایجاد شده است. این مجموعه دادگان شامل دو میزبان لینوکس و ویندوز است که پس از شش ساعت اتصال و تعامل با اینترنت ایجاد شده است. در مجموع ۱۹۵۳۰ جریان شبکه در این مجموعه دادگان به‌دست آمد که برای استفاده در سناریوهای مختلف به ۱۰ مجموعه داده متفاوت تقسیم می‌شود. این مجموعه دادگان یا به‌طور مستقیم از یک منبع تولید ترافیک عادی تشکیل شده‌اند یا ترکیبی تصادفی از دو یا چند منبع ترافیک عادی هستند (جدول (۳)).

جدول (۳): مجموعه دادگان ترافیک عادی شبکه

نام	فعالیت‌ها	تعداد جریان
N-1	۱- جستجو گوگل(GS۲) ۲- ارتباط با سرویس‌دهنده جانگو (DS[3]) ۳- SSH ۴- سایت خبرگزاری فارس(Farsnews) ۵- دانلود سیستم‌عامل لینوکس(FD[4])	۱۹۵۳۰

- **انتخاب ویژگی جریان شبکه**

توسعه‌دهندگان روش‌های مبتنی بر یادگیری ماشین از فرآیند انتخاب ویژگی جهت کاهش زمان محاسبات، بهبود پیش‌بینی و فهم بهتر الگوهای رفتاری استفاده می‌کنند [۱۹]؛ بنابراین، تحلیل رفتار مبتنی بر ویژگی‌ها اصلی، فهم بهتری از رفتار ترافیکی بات در اختیار قرار می‌دهد. مطابق پژوهش‌های انجام‌شده [۱۸، ۲۰ و ۲۲] سه ویژگی اندازه[5]، مدت زمان[6] و تناوبی بودن[7] جریان ترافیک برای ارائه تحلیل رفتاری انتخاب شدند. این سه ویژگی در تشخیص ناهنجاری، دارای خاصیت‌های عدم وابستگی بین ویژگی‌ها و سادگی استخراج از جریان‌های شبکه هستند [۲۳].

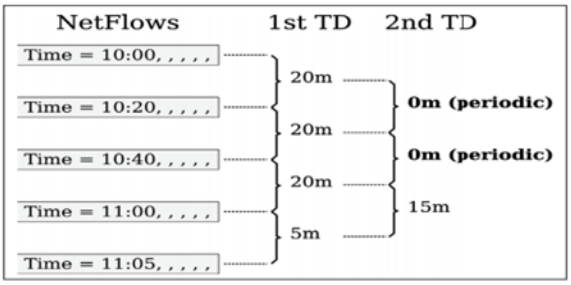
ویژگی‌های اندازه و مدت زمان از ویژگی‌های ذاتی جریان شبکه هستند اما ویژگی تناوبی بودن به‌طور مستقیم از جریان شبکه قابل استخراج نیست و باید محاسبات جداگانه‌ای داشته باشد. هر بات برای تعامل با سرویس‌دهنده فرماندهی خود از یک الگوی ثابت پیروی می‌کند. به بیان دیگر، میزان وقفه بین هر اتصال دارای الگوی تقریباً پایداری است. این در حالی است که

[1] Secure Shell

[2] Google Search
[3] Django Server
[4] File Download
[5] Size
[6] Duration
[7] Periodic

رفتار کاربر عادی دارای وقفه‌های تصادفی است [۶۰]؛ بنابراین، هرچه جریان‌ها تناوبی‌تر تولید شوند، رفتار آن‌ها به رفتار بات شبیه‌تر خواهد بود.

این ویژگی در مرجع [۵۹] تشریح شده است. بر این مبنا، درصورتی‌که اختلاف بین ایجاد سه جریان در شبکه با مقدار ثابتی برابر شود، آن جریان‌ها به‌صورت تناوبی تولید شده‌اند. این مفهوم در شکل (۲) نشان داده شده است.



شکل (۲): نحوه محاسبه تناوبی بودن در جریان شبکه. TD بیانگر اختلاف زمان بین دو جریان شبکه است [۹].

- برچسب‌زنی بر روی ترافیک فرمان و کنترل

در این قسمت نتایج حاصل از بررسی جریان‌های عادی شبکه و بات‌ها ارائه شده است. با حذف درگاه مبدأ، چندین جریان شبکه در یک چهارتایی قرار می‌گیرند. چهارتایی‌های استخراج‌شده از مجموعه دادگان بات‌نت به ترتیب در جدول (۴) ذکر شده است.

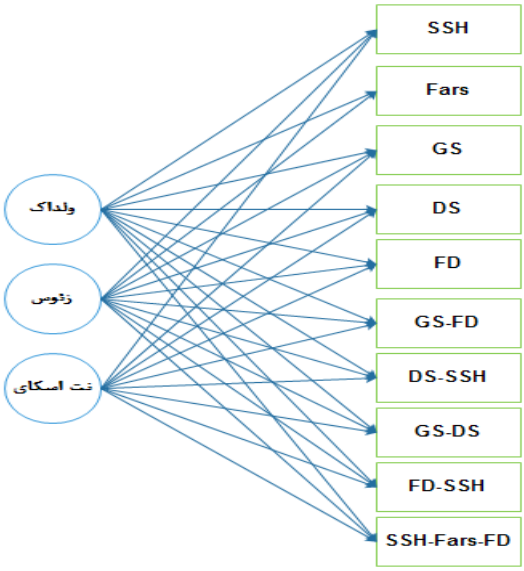
جدول (۴): چهارتایی‌های منتخب از بات‌نت‌ها

نام	ویژگی‌ها			تعداد C&C
	C&C منتخب	درگاه مقصد	پروتکل	
زئوس	162.206.16.208	۱۳۰۹۹	TCP	۱
نت‌اسکای	82.166.16.73	۲۵	TCP	بین ۱ تا ۱۰
ولداک	91.221.219.119	۸۰	TCP	بیش از ۵۰

- **نحوه اجرا**

برای نشان دادن اهمیت تأثیر ترافیک عادی شبکه بر ویژگی‌های جریان شبکه، ۱۰ مجموعه داده ترافیک عادی شبکه و سه مجموعه داده بات مورد استفاده قرار گرفته است. در ابتدا فاصله مینکوفسکی بات و ترافیک عادی مانند شکل (۳) به‌صورت یک به یک محاسبه می‌شود. همین فرآیند برای محاسبه فاصله رفتاری بات‌ها با یکدیگر نیز انجام می‌شود. پس از محاسبه فاصله رفتاری مختلف و نرمال‌سازی، هر یک از ویژگی‌ها با سناریوهای متفاوت امتیازدهی می‌شوند. در تمام سناریوها هر سه بات حضور دارند و تنها نوع و تعداد منابع تولید ترافیک عادی تغییر می‌کند.

با ترکیب‌های مختلف فاصله‌های رفتاری عادی-بات در مجموع برای هر ویژگی ۶۰ فاصله رفتاری محاسبه می‌شود که با احتساب هفت ویژگی متفاوت و سه فاصله رفتاری بات-بات در این آزمایش‌ها، ۴۲۳ فاصله رفتاری بردارهای مختلف تحت چهار سناریو آزمون محاسبه و تحلیل می‌شود. برای رده‌بندی در هر یک از این آزمون‌ها از میانگین کل امتیاز اکتسابی هر ویژگی استفاده می‌شود. همچنین با استفاده از مفهوم انحراف معیار، میزان قابل اتکا بودن هر یک از ویژگی‌ها نشان داده می‌شوند. در این مفهوم، هر چه انحراف معیار کم‌تر باشد آن ویژگی در محیط‌های مختلف، رفتار باثبات‌تری از خود نشان می‌دهد.



شکل (۳): محاسبه ۳۰ فاصله رفتاری بات و ترافیک عادی

۶- نتایج

در ادامه حاصل محاسبه فاصله در حالات مختلف در جدول‌های (۵)، (۶) و (۷) آورده شده است. برای اختصار، تنها نتایج ارزیابی سناریو اول نشان داده شده است. جدول (۵) فاصله رفتاری هر یک از موجودیت‌های منتخب را در فضای یک بعدی نشان می‌دهد. در این جدول از مقایسه یک به یک الگوهای رفتاری استفاده شده است. برای مثال در جدول (۵) فاصله مینکوفسکی بردار رفتاری بات زئوس نسبت به بردار رفتاری SSH از نظر ویژگی تناوبی بودن، ۸۳/۸۵۲ است.

این اعداد برای فضای دوبعدی در جدول (۶) آورده شده است. در جدول (۷) بررسی فاصله رفتاری با در نظر گرفتن سه ویژگی تناوبی- اندازه- مدت زمان به‌صورت هم‌زمان انجام شده است.

نتایج کلی اجرای چهار سناریوی آزمایش مختلف در

جدول (۸) ذکر شده است. سناریوی یک به ارزیابی اهمیت ویژگی‌ها در زمان حضور دو منبع ترافیک عادی می‌پردازد به همین شکل سناریوهای ۲، ۳ و ۴ به ترتیب ۳، ۵ و ۱۰ منبع عادی شبکه را مورد بررسی قرار می‌دهند.

در ادامه تحلیل نتایج از سه منظر ترتیب ویژگی‌ها، میانگین و انحراف معیار امتیازهای کسب‌شده تحلیل می‌شوند:

- ترتیب اهمیت ویژگی‌ها: ترتیب اهمیت ویژگی‌ها در حضور ترافیک‌های عادی مختلف تغییر می‌کند. این تفاوت در رده‌بندی نیز مشهود است برای مثال ویژگی SD در سناریوهای ۱و ۲ رتبه اول را ازنظر میانگین امتیاز کسب کرده است اما در سناریوهای ۳و ۴ کاهش کارایی زیادی دارد؛ به‌طوری‌که از جایگاه اول به رتبه ۴ تنزل پیدا می‌کند. این مسئله در دیگر ویژگی‌ها نیز مشهود است. از طرف دیگر هر چه حجم ترافیک عادی، بیشتر بررسی می‌شود، ویژگی‌های ذاتی جریان شبکه مانند اندازه و مدت زمان، کارایی خود را از دست می‌دهند. در مقابل ویژگی‌های ترکیبی و محاسباتی مانند تناوبی بودن و SDP و غیره با افزایش کارایی روبرو می‌شوند. در نهایت با بررسی ترتیب اهمیت می‌توان دریافت که برخی از ویژگی‌ها اضافی هستند. برای مثال ویژگی اندازه و مدت زمان تولید در تمامی سناریوها کارایی یکسانی داشته‌اند و می‌توان یکی از آن‌ها را حذف نمود. این مسئله برای ویژگی‌های SDP و SP نیز درست است.
- میانگین: امتیازهای کسب‌شده در سناریوهای مختلف توسط ویژگی‌ها نشان می‌دهد هر چه حجم ترافیک عادی در شبکه بیشتر باشد اختلاف اهمیت ویژگی‌ها جهت دسته‌بندی افزایش می‌یابد. به‌عنوان نمونه در این آزمایش میانگین امتیاز بهترین و بدترین ویژگی در سناریو ۱ به ترتیب برابر ۶/۲۱ و ۵/۲۸ است اما در سناریوی ۴ این مقادیر به ترتیب ۵/۸۴ و ۲/۳ است. این مسئله نشان می‌دهد در محیط‌هایی با تنوع ترافیکی عادی بالا، انتخاب دقیق ویژگی و حذف ویژگی‌های ناکارامد اهمیت ویژه‌ای دارد.
- انحراف معیار: بررسی انحراف معیار ویژگی‌ها نشان می‌دهد هر چه انحراف معیار ویژگی کم‌تر باشد ثبات بیشتری در سناریوهای مختلف به وجود می‌آید. برای مثال ویژگی اندازه دارای اختلاف انحراف معیار ۲۵/۲٪ است و اهمیت این ویژگی در سناریوهای مختلف از اول به چهارم تغییر پیدا می‌کند این در حالی است که ویژگی SDP با اختلاف انحراف معیار ۲۱/۶٪ در سناریوهای مختلف، بین رتبه اول و دوم از نظر اهمیت جا به جا می‌شود. بنابراین زمانی‌که از محیط استقرار شناخت کافی وجود ندارد و یا تحول ترافیکی در آن محیط زیاد است بهترین ویژگی جهت سامانه‌های تشخیص بات ویژگی‌هایی با انحراف معیار کم‌تر هستند.

جدول (۵): فاصله بردارهای رفتاری بات-بات و بات -عادی در سناریوی ۱ (ویژگی‌های تناوب، مدت زمان و اندازه)

	نت‌اسکای			ولداک			زئوس		
	S	D	P	S	D	P	S	D	P
SSH	۶/۷۴۴	۱/۰۷۵	۶۷/۲۰۱	۶/۷۴۵	۱/۰۷۶	۵۰/۱۲۰	۶/۷۴۵	۱/۰۷۶	۵۳/۸۵۲
GS	۶/۵۰۸	۰/۱۶۹	۱۴/۲۸۳	۶/۵۱	۰/۱۶۹	۳۳/۲۲۶	۶/۵۱	۰/۱۶۹	۴۹/۶۷۹
زئوس	۰/۰۱	۰/۰۰۱	۴۹/۱۵۳	۰	۰	۵۲/۶۵	۰	۰	۰
ولداک	۰/۰۱	۰/۰۰۱	۳۷/۷۸۹	۰	۰	۰	۰	۰	۵۲/۶۵
نت اسکای	۰	۰	۰	۰/۰۱	۰/۰۰۱	۳۷/۷۸۹	۰/۰۱	۰/۰۰۱	۴۹/۱۵۳

جدول (۶): فاصله بردارهای رفتاری بات-بات و بات-عادی در سناریوی ۱ (ویژگی‌های PD، PS و SD)

	نت‌اسکای			ولداک			زئوس		
	SD	PS	PD	SD	PS	PD	SD	PS	PD
SSH	۶/۸۳۰	۶۷/۵۳۹	۶۷/۲۱	۶/۸۳۱	۵۰/۵۷۲	۵۰/۱۳۱	۶/۸۳۱	۵۴/۲۷۲	۵۳/۸۶۲
GS	۶/۵۱۱	۱۵/۶۹۶	۱۴/۲۸۴	۶/۵۱۲	۳۳/۸۵۸	۳۳/۲۲۷	۶/۵۱۲	۵۰/۱۰۴	۴۹/۶۷۹
زئوس	۰/۰۱	۴۹/۱۵۳	۴۹/۱۵۳	۰	۵۲/۶۵	۵۲/۶۵	۰	۰	۰
ولداک	۰/۰۱	۳۷/۷۸۹	۳۷/۷۸۹	۰	۰	۰	۰	۵۲/۶۵	۵۲/۶۵
نت‌اسکای	۰	۰	۰	۰/۰۱	۳۷/۷۸۹	۳۷/۷۸۹	۰/۰۱	۴۹/۱۵۳	۴۹/۱۵۳

جدول (۷): فاصله بردارهای رفتاری بات-بات و بات-عادی در سناریوی اول (ویژگی SDP)

	نت‌اسکای	ولداک	زئوس
	SDP	SDP	SDP
SSH	۱۴۳۹۶۱/۵۸۱	۱۴۳۹۸۴/۰۲۱	۱۴۳۹۸۴/۰۰۶
GS	۱۳۷۲۳۴/۳۲۰	۱۳۷۲۷۴/۱۷۵	۱۳۷۲۷۴/۱۷۷
زئوس	۲۱۰/۰۵۲	۲۶/۴۲۰	۰
ولداک	۲۰۹/۴۵۱	۰	۲۶/۴۲۰
نت اسکای	۰	۲۰۹/۴۵۱	۲۱۰/۰۵۲

جدول (۸): رده‌بندی ویژگی‌ها در سناریوهای مختلف

آزمون	سناریو	ترتیب اهمیت ویژگی‌ها	میانگین	انحراف معیار
۱	ترافیک بات های نت‌اسکای، زئوس و ولداک ترافیک عادی: SSH و GS	SD	۶/۲۱	۴/۷۹
		مدت زمان و اندازه	۶	۵/۷
		SDP و SP	۵/۳۷	۲/۹۵
		PD	۵/۳۵	۲/۹۶
		تناوبی	۵/۲۸	۳/۳
۲	ترافیک بات های نت‌اسکای، زئوس و ولداک ترافیک عادی: DS و FD و Fars	SD	۵/۴۵	۴/۶۹
		SP و SDP	۵/۲۹	۲/۸
		PD	۵/۲۲	۲/۸۵
		تناوبی بودن	۵/۱۴	۲/۷۳
		اندازه و مدت زمان	۵	۵/۱۴
۳	ترافیک بات های نت‌اسکای، زئوس و ولداک ترافیک عادی: SSH و GS و Fars و DS و FD	تناوبی	۵/۷۷	۳/۲
		SP و SDP	۵/۶۵	۲/۷۸
		PD	۵/۵۸	۲/۸۴
		SD	۴/۳۳	۴/۴۹
		S و D	۳/۷۵	۴/۹۴
۴	ترافیک بات های نت‌اسکای، زئوس و ولداک ترافیک عادی: تمامی ۱۰ دیتاست عادی	SP و SDP	۵/۸۴	۲/۳۱
		PD	۵/۷۹	۲/۳۷
		P	۵/۱۵	۲/۶۹
		SD	۳/۰۷	۳/۸۶
		S و D	۲/۳	۴/۲۶

۷- نتیجه‌گیری

این مقاله تحلیل ترافیک باتنت و مقایسه آن با رفتار عادی شبکه را ارائه کرد. در این بخش‌ها مهم‌ترین ویژگی‌های یک مجموعه دادگان استاندارد بررسی شد و همچنین برای اطمینان از عملکرد کانال فرمان و کنترل و تکمیل چرخه حیات باتنت مجموعه دادگان اختصاصی مورد بررسی قرار گرفت. لذا با توجه به نیاز به هم‌زمانی حضور ترافیک عادی و باتنت و عدم تنوع کافی ترافیک عادی در مجموعه دادگان موجود، مجموعه دادگان ایجاد شد. در نهایت با ارائه روشی مبتنی بر فاصله مینکوفسکی، تأثیر ترافیک عادی بر مرحله انتخاب و استخراج ویژگی مورد بررسی قرار گرفت. نتایج حاصل از این تحقیق در مورد اهمیت ترافیک عادی محیط استقرار شامل موارد زیر هستند.

- ترتیب اهمیت ویژگی‌ها در حضور ترافیک‌های عادی مختلف دچار تغییر می‌شود.
- در محیط‌هایی با تنوع ترافیکی بالا انتخاب دقیق ویژگی و حذف ویژگی‌های ناکارامد اهمیت ویژه‌ای دارد.
- زمانی‌که از محیط استقرار شناخت کافی وجود نـدارد و یا تحول ترافیک عادی در آن محیط زیاد است؛ بهترین ویژگی جهت سـامانه‌هـای تشـخیص، ویژگـی‌هـایی بـا انحراف معیار کم‌تر و میانگین بهره‌وری بالاتر هستند.
- هر چه تنوع ترافیک عادی در محیط استقرار بیشتر باشد ویژگی‌های ترکیبی و محاسباتی، پیچیدگی بیشتر مانند تناوب یا SP اهمیت بالاتر دارند.
- ویژگی‌های ترکیبی و محاسباتی با پیچیدگی بالاتر، ثبات و پایداری بیشتری در محیط‌های مختلف دارند.
- هرچه ترافیک عادی محیط استقرار تنوع کم‌تری داشته باشد کارایی ویژگی‌های ذاتی جریان شبکه مانند انـدازه و مدت زمان افزایش پیدا می‌کند.

در نتیجه اهمیت ترافیک عادی در زمان حضور بات قابل توجه است به‌طوری‌که می‌تواند مرحله انتخاب و استخراج ویژگی را با تغییر زیادی مواجه کند و بهره‌وری یک ویژگی را بیش از ٪۵۰ افزایش و یا کاهش دهد. بنابراین، هر چه سیستم تشخیص باتنت با ترافیک عادی محیط استقرار نهایی خود آشنایی بیشتری داشته باشد؛ نتیجه بهتری را به عنوان یک سیستم نظارتی ارائه می‌دهد.

۸- مراجع

[1] N. Milošević, "History of malware," arXiv preprint arXiv:1302.5392, 2013.

[2] R. Jalaei and M R Hasani Ahangar, "A Analytical Survey on Botnet and Detection Methods," Electronical & Cyber Defence, vol. 4, pp. 25-46, 2017. (In Persian)

[3] Cisco, "Netflow " [Online]Available: https://www.cisco.com/c/en/us/tech/quality-of-service-qos/netflow/index.html, 2020.

[4] B. Claise, S. Bryant, G. Sadasivan, S. Leinen, T. Dietz, and B. Trammell, "Specification of the ip flow information export (ipfix) protocol for the exchange of ip traffic flow information (rfc 5101)," Technical report, The Internet Engineering Task Force (IETF), 2008.

[5] J. Han, J. Pei, and M. Kamber, "Data mining: concepts and techniques," Elsevier, 2011.

[6] A. H. Lashkari, G. D. Gil, J. E. Keenan, K. F. Mbah, and A. A. Ghorbani, "A survey leading to a new evaluation framework for network-based botnet detection," in Proceedings of the 2017 the 7th International Conference on Communication and Network Security, pp. 59-66, 2017.

[7] B. A. Alahmadi, E. Mariconti, R. Spolaor, G. Stringhini, and I. Martinovic, "BOTection: Bot Detection by Building Markov Chain Models of Bots Network Behavior," 2019.

[8] A. Pektaş and T. Acarman, "Deep learning to detect botnet via network flow summaries," Neural Computing and Applications, vol. 31, pp. 8021-8033, 2019.

[9] S. García, "Identifying, Modeling and Detecting Botnet Behaviors in the Network," 2014.

[10] S. Almutairi, S. Mahfoudh, S. Almutairi, and J. S. Alowibdi, "Hybrid Botnet Detection Based on Host and Network Analysis," Journal of Computer Networks and Communications, vol. 2020, 2020.

[11] I. T. Sherif Saad, A. A. Ghorbani, Bassam Sayed, D. Zhao, Wei Lu, John and P. H. Felix, "Detecting P2P botnets through network behavior analysis and machine learning," Presented at the Proceedings of 9th Annual Conference on Privacy, Security and Trust (PST2011), Montreal, Quebec, Canada, 2011.

[12] A. Alenazi, I. Traore, K. Ganame, and I. Woungang, "Holistic Model for HTTP Botnet Detection Based on DNS Traffic Analysis," in International Conference on Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments, pp. 1-18, 2017.

[13] S. Garcia, "The CTU-13 Dataset. A Labeled Dataset with Botnet, Normal and Background Traffic," S. Lab, Ed., ed, 2014.

[14] E. B. Beigi, H. H. Jazi, N. Stakhanova, and A. A. Ghorbani, "Towards effective feature selection in machine learning-based botnet detection approaches," in 2014 IEEE Conference on Communications and Network Security, pp. 247-255, 2014.

[15] A. F. A. Kadir, N. Stakhanova, and A. A. Ghorbani, "Android botnets: What urls are telling us," in International Conference on Network and System Security, pp. 78-91, 2015.

[16] N. Koroniotis, N. Moustafa, E. Sitnikova, and B. Turnbull, "Towards the development of realistic botnet dataset in the internet of things for network forensic analytics: Bot-iot dataset," Future Generation Computer Systems, vol. 100, pp. 779-796, 2019.

[17] A. Shiravi, H. Shiravi, M. Tavallaee, and A. A. Ghorbani, "Toward developing a systematic approach to generate benchmark datasets for intrusion detection," computers & security, vol. 31, pp. 357-374, 2012.

[18] A. H. Lashkari, G. D. Gil, J. E. Keenan, K. Mbah, and A. A. Ghorbani, "A survey leading to a new evaluation framework for network-based botnet detection," in Proceedings of the 2017 the 7th International Conference on Communication and Network Security, pp. 59-66, 2017.

[19] G. Chandrashekar and F. Sahin, "A survey on feature selection methods," Computers & Electrical Engineering, vol. 40, pp. 16-28, 2014.

[20] S. García, V. Uhlíř, and M. Rehak, "Identifying and modeling botnet C&C behaviors," in Proceedings of the 1st International Workshop on Agents and CyberSecurity, 2014.

[21] S. Garcia, "Modelling the network behaviour of malware to block malicious patterns. the stratosphere project: a behavioural ips," Virus Bulletin, pp. 1-8, 2015.

[22] D. Acarali, M. Rajarajan, N. Komninos, and I. Herwono, "Survey of approaches and features for the identification of HTTP-based botnet traffic," Journal of Network and Computer Applications, vol. 76, pp. 1-15, 2016.

[23] F. Iglesias and T. Zseby, "Analysis of network traffic features for anomaly detection," Machine Learning, vol. 101, pp. 59-84, 2015.

Journal of Electronical & Cyber Defence
Vol. 9, No. 1, 2021, Serial No. 33

Kavosh: Offering an Analysis Method and the Impact of Normal Network Traffic on Selection and Extraction Based on the Minkowski Distance

M. J. Faghihniya, R. Jalaei*, H. Shojaee Yas
*Imam Hossein Comprehensive University
(Received: 03/05/2020, Accepted: 26/10/2020)

ABSTRACT

The growing spread of botnet threats and the development of new platforms for deploying botnets such as the Internet of Things urges the need for confrontation. Research in the field of botnet detection based on machine learning methods, shows that these methods have the necessary efficiency for botnet detection. In this paper, normal and botnet traffic are analyzed by the proposed method based on the Minkowski distance vector. The results of the article show that normal traffic flow affects the feature selection and extraction stage by changing the importance of features. This method scores the features based on near bot-bot behavioral vectors and far bot-normal behavioral vectors. The results of these experiments on ten sets of normal data and three sets of bot data showed that the score of a feature increases or decreases by more than 50% in environments with various normal traffic.

Keywords: Botnet, Network Flow, Feature Extraction, Minkowski Distance, Behavioral Distance

* Corresponding Author Email: rjalaei@ihu.ac.ir