نشریه علمی "پدافند الکترونیکی و سایبری"
سال هفتم، شماره ۳، پاییز ۱۳۹۸، ص ۸۱-۷۱

بهبود روش‌های انتساب بار داده در فرآیند جرم‌شناسی شبکه‌های کامپیوتری به کمک فیلتر بلوم سلسله مراتبی در زمان

زینب ساسان[۱]، مهدی خرازی[۲*]
۱- کارشناسی ارشد، ۲- استادیار، دانشگاه صنعتی شریف
دریافت: ۹۷/۰۶/۱۹، پذیرش: ۹۷/۱۲/۱۴)

چکیده

انتساب حملات سایبری در سطح شبکه‌های کامپیوتری به عوامل آن، یکی از مهم‌ترین مراحل جرم‌شناسی شبکه محسوب می‌شوند. در فرآیند انتساب در برخی موارد تنها به بار داده بسته‌های تبادل‌شده در شبکه دسترسی وجود دارد و از این رو روش‌های انتساب بار داده معرفی شده‌اند. در روش‌های انتساب بار داده باید کل ترافیک در قالب خلاصه ذخیره شده و حریم خصوصی کاربران حفظ شود که برای این منظور از ساختار داده تصادفی فیلتر بلوم استفاده می‌شود. پژوهش‌هایی که تاکنون در این حوزه انجام شده تلاش می‌کنند تا خطای مثبت- نادرست فیلترهای بلوم را کاهش داده و نسبت کاهش حجم داده را بهبود دهند ولی تاکنون پژوهش قابل توجهی در خصوص عملیاتی کردن این روش‌ها در سطح شبکه‌های کامپیوتری انجام نشده است. خروجی یک روش انتساب بار داده، باید شناسه‌های جریانی باشد که مشکوک به انتقال نمونه ترافیک مخرب هستند. چالشی که در راستای عملیاتی کردن این روش‌ها در این پژوهش به آن پرداخته شده، زیاد بودن تعداد پرس‌وجوها در فرآیند یک انتساب است. زیاد بودن پرس‌وجوها از زیاد بودن شناسه‌های جریان و فیلترهای بلوم در بازه‌های زمانی طولانی مدت ناشی می‌شود. در این پژوهش راه‌کاری مبتنی‌بر سلسله‌مراتب زمان ارائه‌شده که فضای پرس‌وجو را کاهش داده و سعی می‌کند تعداد شناسه‌های جریان که به اشتباه گزارش شده‌اند را کاهش دهد. ارزیابی‌ها نشان می‌دهد در رویکرد مبتنی بر سلسله‌مراتب زمان، احتمال رخ ندادن خطا در برخی از شاخه‌های سلسله‌مراتب وجود داشته و از شناسه‌های جریان مربوط به آن شاخه برای پرس‌وجو صرف‌نظر می‌شود. این موضوع در نهایت می‌تواند به کاهش خطای نهایی سامانه انتساب بار داده منجر شود به‌طوری که مقدار خطای سامانه در سناریوی طراحی‌شده، در روش قبلی برابر با ۵/۶۶ درصد بوده و این مقدار به ۳/۹۸ درصد کاهش پیدا کرده و ۸۴۰۰ شناسه جریان کمتری به اشتباه گزارش می‌شود.

کلیدواژه‌ها: جرم‌شناسی دیجیتال، جرم‌شناسی شبکه، روش‌های انتساب بار داده، روش‌های انتساب بار داده عملیاتی

۱- مقدمه

با توجه به توسعه شبکه‌های کامپیوتری به‌عنوان مهم‌ترین زیرساخت تبادل اطلاعات، وابستگی هر روزه جامعه اعم از شرکت‌ها، سازمان‌ها و کاربران به آن به‌طور چشم‌گیری افزایش یافته است. در این حین تعداد حملاتی که به‌واسطه این زیرساخت ارتباطی انجام می‌شوند نیز هم در تعداد و هم در درجه‌ی تاثیرگذاری، بیشتر شده است.

روش‌های پیشگیرانه و دفاعی متعدد برای جلوگیری و یا کاهش تاثیر حملات پیشنهاد شده و در شبکه استقرار یافته‌اند. البته بسیاری از این روش‌ها، یا به‌خاطر ضعف در نحوه به‌کارگیری و یا کشف و به‌کارگیری آسیب‌پذیری‌های جدیدتر، به درستی نمی‌توانند از وقوع حملات جلوگیری کنند. بعد از وقوع حمله، یکی از مهم‌ترین اقدامات بررسی شواهد موجود برای کشف نحوه انجام حمله، آسیب‌پذیری استفاده شده و تحلیل دقیق از تاثیر نهایی آن است. در این راستا، هدف روش‌های جرم‌شناسی دیجیتال[۲]، بررسی شواهد موجود بعد از یک حمله با هدف کسب اطلاعات ذکرشده می‌باشد.

در شبکه‌های کامپیوتری، با توجه به نرخ بالای ارسال داده، به‌صورت پیش‌فرض داده‌های منتقل‌شده ذخیره نمی‌گردند و در عمل این کار پرهزینه بوده و در بازه‌های زمانی طولانی مدت امکان‌پذیر نمی‌باشد. لذا معمولا مدیران شبکه در مواردی فقط

* رایانامه نویسنده مسئول: kharrazi@sharif.edu

[2] Digital Forensics

سرآیند[1] بسته‌های منتقل‌شده را برای بازه زمانی محدودی ذخیره کرده و محتویات بسته را ذخیره نمی‌کنند. از طرفی دیگر نیاز به جرم‌شناسی رایانه‌ای بسیار حائز اهمیت می‌باشد تا بعد از وقوع حمله بتوان به آن استناد کرده و حمله کننده و یا وقایع مرتبط را به آن منتسب نمود.

به‌طور دقیق‌تر، مرحله بررسی و انتساب[2] یکی از مهم‌ترین مراحل فرآیند جرم‌شناسی شبکه[3] محسوب می‌شود. در پژوهش [۲-۱] نیز انتساب در حوزه شبکه‌های کامپیوتری، مشخص کردن مبدأ یا مقصد نمونه‌ای از ترافیک شبکه تعریف شده است. با توجه به ماهیت حملات، روش‌های مختلفی برای انتساب ارائه شده است. به‌عبارت دیگر با توجه به ماهیت و نوع حمله باید روش انتساب متناسب با آن انتخاب شود [۳]. در حالت کلی، انتساب حملات در سطح شبکه‌های کامپیوتری براساس سرآیند و یا بار داده[4] بسته‌ها انجام می‌شود که در بخش‌های بعدی به تفضیل توضیح داده شده است.

در برخی موارد تنها شواهدی که برای انجام فرآیند جرم‌شناسی شبکه وجود دارد، بار داده بسته‌های تبادل‌شده در شبکه است. روش‌های انتساب بار داده[5] برای این منظور معرفی شده‌اند که خروجی این روش‌ها باید آدرس‌های مبدأ و مقصدی باشد که مشکوک به حمل بار داده مخرب و یا آلوده هستند.

به‌عنوان کاربردهایی از روش‌های انتساب بار داده می‌توان شرایطی را در نظر گرفت که شبکه سازمانی به کرم[6] و بدافزار[7] آلوده شده است و تنها بخشی که برای انجام جرم‌شناسی وجود دارد، امضای بدافزار[8] است. در مثالی دیگر یکی از کارکنان ناراضی شرکت، پرونده‌های حساس و مهم را به خارج از سازمان ایمیل کرده است. دریافت ایمیل‌های فیشینگ[9] نیز یکی از تهدیدهای مهمی است که سازمان‌ها با آن روبرو هستند و نقطه مشترک در تمامی این مثال‌ها، شواهدی است که برای اجرای جرم‌شناسی وجود دارد و آن بار داده بسته‌های شبکه است.

در فرآیند جرم‌شناسی و در مرحله انتساب، نمونه‌ای از ترافیک مشکوک وجود دارد که باید مشخص شود توسط چه موجودیت‌هایی و با چه آدرس مبدأ و مقصدی ارسال شده‌اند. به‌طور کلی به ترکیب دوتایی آدرس مبدأ و مقصد، شناسه جریان[10] گفته می‌شود. برای اجرای این فرآیند باید تمامی ترافیک شبکه در بازه‌های زمانی مشخصی ذخیره شود تا در فرآیند جرم‌شناسی و انتساب بتوان نمونه ترافیک مشکوک را جستجو کرد. در بسیاری از سازمان‌ها برای انجام فرآیند جرم‌شناسی شبکه لازم است که ترافیک برای بازه زمانی طولانی ۳ الی ۶ ماهه ذخیره شود، تا بعد از کشف حادثه، بتوان به گذشته برگشت و داده‌ها را مورد بررسی قرار داد.

ذخیره‌سازی ترافیک شبکه برای انجام فرآیند انتساب با چالش‌هایی مواجه است. در طولانی مدت به‌منظور جمع‌آوری کل ترافیک شبکه به فضای ذخیره‌سازی زیادی نیاز است. همچنین پرس‌وجوی نمونه ترافیک مشکوک در این حجم عظیم از ترافیک، زمان‌بر خواهد بود. نگهداری ترافیک شبکه از دیدگاه کاربران مشکلات حریم خصوصی[11] را نیز به دنبال خواهد داشت. برای رفع این ۳ چالش باید فرآیند ذخیره‌سازی ترافیک شبکه به‌صورت کارایی انجام شود که برای این منظور اولین بار پیشنهاد استفاده از ساختار داده تصادفی فیلتر بلوم[12] [۴] توسط شانموگاساندارام و همکارانش [۱] ارایه گردید.

در ساختار داده فیلتر بلوم، مقادیر درهم‌سازی[13] از بسته‌های شبکه محاسبه می‌شود که این مقادیر متناسب با اندازه بردار بیتی در فیلتر بلوم بوده و خانه‌های مربوط به آن را یک می‌کند. به عبارت دیگر در این حالت چکیده‌ای از ترافیک ذخیره شده و پرس‌وجو بسیار سریع‌تر انجام می‌شود. همچنین در این نوع از ذخیره‌سازی، می‌توان در مورد حضور و یا عدم حضور یک بسته در ترافیک شبکه از فلیتر بلوم پاسخ مثبت یا منفی دریافت کرد ولی نمی‌توان محتوای بسته‌ها و ترافیک را مشاهده کرد که این موضوع، چالش مربوط به حریم خصوصی را در روش‌های انتساب بار داده برطرف می‌کند.

ساختار داده تصادفی فیلتر بلوم که در بخش‌های بعدی بیشتر توضیح داده خواهد شد. در پرس‌وجوی یک نمونه، دارای خطای مثبت- نادرست[14] است. همچنین با نسبت کاهش حجم داده[15] مشخصی می‌توان ترافیک را در فیلتر بلوم ذخیره کرد. در پژوهش‌هایی که تاکنون در حوزه روش‌های انتساب بار داده انجام شده، تمرکز اصلی بر روی کاهش خطای مثبت- نادرست و افزایش نسبت کاهش حجم داده بوده است. به‌عبارت دیگر تلاش شده تا فیلتر بلوم خطای کمتری داشته و داده‌ها به‌صورت

[1] Header
[2] Attribution
[3] Network Forensics
[4] Payload
[5] Payload Attribution methods
[6] Worm
[7] Malware
[8] Malware Signiture
[9] Phishing

[10] Flow ID
[11] Privacy
[12] Bloom Filter
[13] Hash
[14] False-Positive
[15] Data reduction ratio

فشرده‌تری در یک فیلر بلوم ذخیره شوند. در این پژوهش مسئله مهمی که به آن پرداخته شده، بررسی چالش‌های پیاده‌سازی و عملیاتی کردن روش‌های انتساب بار داده و ارائه راه‌کاری برای این مسأله است.

یکی از چالش‌های اساسی بر سر راه پیاده‌سازی روش‌های انتساب بار داده این است که بررسی و انتساب باید روی حجم زیادی از ترافیک که در بازه‌های زمانی طولانی ۳ الی ۶ ماهه جمع‌آوری شده است، انجام شود. در این بازه زمانی طولانی، تعداد شناسه‌های جریان بسیار زیادی ذخیره شده و در مرحله انتساب باید تک‌تک این شناسه‌ها بر روی فیلتر بلوم پرس‌وجو شود. همچنین طولانی شدن بازه جمع‌آوری ترافیک و نداشتن یک حافظه بزرگ و یکپارچه، باعث می‌شود تا ظرفیت یک فیلتر بلوم پر شده و مقادیر آن در دیسک ذخیره شود. به‌عبارت دیگر باید از یک فیلتر بلوم به فیلتر بلوم دیگر تعویض صورت گیرد.

پس در حالت کلی، در بازه‌های زمانی طولانی، تعداد شناسه‌های جریان و همچنین تعداد فیلترهای بلوم افزایش می‌یابد. همچنین اشاره شد که فیلتر بلوم یک ساختار داده تصادفی است که دارای خطای مثبت نادرست است. به‌عبارت دیگر هرچه تعداد پرس‌وجوهایی که بر روی این ساختار داده تصادفی صورت می‌گیرد بیشتر باشد، احتمال وقوع خطا نیز افزایش خواهد یافت. پس یکی از اساسی‌ترین چالش‌ها در عملیاتی کردن روش‌های انتساب بار داده، بزرگ بودن فضای پرس‌وجو است که از زیاد بودن تعداد شناسه‌های جریان و تعداد فیلترهای بلوم ناشی می‌شود.

در این پژوهش، راه‌کاری مبتنی بر سلسله مراتب زمان[1] ارائه شده تا فضای پرس‌وجو بر روی فیلترهای بلوم در بازه زمانی طولانی، کوچک‌تر شود. در این راه‌کار تلاش می‌شود تا از حداکثر حافظه یکپارچه‌ای که وجود دارد، به‌عنوان فیلتر بلوم استفاده شده و تعداد تعویض‌ها از یک فیلتر بلوم به فیلتر بلوم دیگر به حداقل برسد. همچنین در ساختار سلسله‌مراتبی در برخی از شاخه‌های این سلسله‌مراتب پرس‌وجو صورت نخواهد گرفت و تعداد شناسه‌های جریان برای پرس‌وجو کمتر می‌شود. به‌عبارت دیگر راه‌کاری که ارائه شده، سعی دارد با کاهش فضای پرس‌وجو، خطایی نهایی روش انتساب بار داده را کاهش داده و پیاده‌سازی آن در بازه‌های زمانی طولانی‌تر را ممکن سازد.

در ادامه این پژوهش، در بخش بعدی به معرفی و بررسی کارهای پیشین در رابطه با روش‌های انتساب و به‌ویژه انتساب بار داده پرداخته شده است. در بخش ۳ با هدف تعریف مسأله، چالش‌های موجود در عملیاتی کردن روش‌های انتساب بار داده مطرح و در بخش ۴ راه‌کار پیشنهادی ارائه شده است. در بخش ۵ روند پیاده‌سازی و نتایج حاصل از راه‌کار پیشنهادی مورد ارزیابی قرار گرفته است. در نهایت در بخش ۶، مطالعات و نتایج حاصل از این پژوهش جمع‌بندی شده است.

۲- کارهای پیشین

همان‌طور که در بخش قبل اشاره شد، با توجه به ماهیت حمله و همچنین شواهدی که برای انتساب حمله وجود دارد، روش‌های انتساب مختلفی ارائه شده است. در یک طبقه‌بندی کلی می‌توان روش‌های انتساب بار داده را به دو دسته مبتنی‌بر سرآیند و بار داده بسته‌های شبکه تقسیم کرد.

در روش‌های مبتنی بر سرآیند بسته‌ها، می‌توان به روش‌های انتساب حملات گام صخره‌ای[2] [۵-۶] و انتساب با استفاده از سامانه‌های تله عسل[3] [۷] اشاره کرد. همچنین یکی از حوزه‌های پژوهشی مهم، روش‌های ردیابی به سمت عقب[4] [۸-۱۱] است. در این روش‌ها، ردیابی از رایانه قربانی شروع شده و به صورت بازگشتی در مسیر حمله به سمت مهاجم (یا واسط حمله) به عقب گام برداشته می‌شود. گستره وسیعی از پژوهش‌ها برای انتساب حملات در محافل علمی روی این دسته متمرکز شده‌اند. شایان ذکر است که رویکردهای مختلف ذکر شده خارج از حوزه این پژوهش می باشند و به جزئیات آنها در این مقاله پرداخته نمی‌شود.

همان‌طور که در بخش قبل اشاره شد، در بسیاری از موارد تنها بار داده بسته در دسترس است و روش‌های مبتنی بر سرآیند کارایی نداشته و باید از روش‌های انتساب بار داده استفاده کرد. در روش‌های انتساب بار داده ضروری است کل ترافیک شبکه برای بررسی‌های پیش از وقوع حمله ذخیره شود. این ذخیره‌سازی همان‌طور که در بخش قبل به آن اشاره شد، با چالش‌هایی مواجه بود. برای برطرف کردن چالش‌های ذخیره‌سازی ترافیک در روش‌های انتساب بار داده از ساختار داده فیلتر بلوم استفاده شده است.

فیلتر بلوم یک ساختار داده تصادفی برای نمایش مجموعه داده‌ها است که می‌تواند عمل درج و پرس‌وجو برای عضویت را انجام دهد. فیلتر بلوم فضای ذخیره‌سازی در قالب فشرده را با

[1] Time based Hierarchical

[2] Stepping Stone
[3] Honeypot
[4] Traceback

هزینه اضافه شدن مقدار مشخصی خطای مثبت-نادرست ممکن می‌سازد. به‌عبارت دیگر در این ساختار داده امکان دارد برای پرس‌وجو در مورد حضور عنصری که در آن درج نشده است، به اشتباه پاسخ مثبت برگردانده شود. همچنین این ساختار داده دارای خطای منفی- نادرست[1] نبوده و زمانی که به یک پرس‌وجو پاسخ منفی برگردانده می‌شود، این پاسخ کاملا درست است. مقدار خطای مثبت- نادرست از رابطه (۱) زیر به‌دست می‌آید:

$$FP = \left(1 - \left(1 - \frac{1}{m}\right)^{nk}\right)^{k} = \left(1 - e^{\frac{-kn}{m}}\right)^{k} \qquad (۱)$$

در این رابطه، FP مقدار خطای مثبت- نادرست، m اندازه فیلتر بلوم، k تعداد توابع چکیده‌ساز و n تعداد کل عناصر درج شده در فیلتر بلوم است. فیلترهای بلوم در حوزه‌های مختلف علوم کامپیوتر به‌ویژه در زمینه شبکه‌های رایانه‌ای بسیار مورد توجه قرار گرفته‌اند [۱۴-۱۲].

اولین کار در زمینه روش‌های انتساب بار داده در پژوهش [۱۵] توسط شانموگاساندارام و همکارانش در سال ۲۰۰۳ انجام شد. آن‌ها ایده و معماری ForNet برای جرم‌شناسی شبکه را معرفی کردند که یکی از مؤلفه‌های اصلی آن روش انتساب بار داده بود. وی و همکارانش ایده استفاده از ساختمان داده فیلتر بلوم در فرآیند انتساب بار داده را مطرح کردند و روش‌های فیلتر بلوم مبتنی بر قطعه[2] و فیلتر بلوم سلسله مراتبی[3] [۲] را ارائه دادند.

در روش فیلتر بلوم مبتنی بر قطعه، بار داده یک بسته برای این‌که قابلیت پرس‌وجو بر روی زیررشته نیز وجود داشته باشد، به قطعه‌هایی تقسیم می‌شود. در ادامه قطعه‌ها در فیلتر بلوم درج می‌شوند. در این روش مشکلات ترازبندی[4] و تصادم شماره قطعه‌ها[5] وجود دارد که در ادامه پونک و همکارانش در پژوهش [۱۶] روش‌های موجود را بهبود داده و به ارائه روش‌های جدید پرداختند.

روش‌های انتساب بار داده معرفی شده با تقسیم بار داده بسته به قطعه‌هایی با طول ثابت و یا متغیر سعی در انتساب زیررشته‌ها داشتند. در حالتی که طول قطعه ثابت بود مشکل ترازبندی وجود داشت و باید تمامی حالت‌های ترازبندی پرس‌وجو می‌شد. راه‌حل استفاده از الگوریتم‌های اثرانگشت‌گیری[6] و تقسیم بسته‌ها به قطعه‌های با اندازه متغیر به این منظور ارائه شد. مشکل دیگر تصادم شماره قطعه‌ها در تشخیص قطعه‌های متوالی بود. برای این مشکل نیز استفاده از رویکرد سلسله‌مراتب و هم‌پوشانی[7] پیشنهاد شده است. ترکیبی از این رویکردها برای حل دو مشکل ذکر شده، سیر تکاملی روش‌های مختلف انتساب بار داده را به وجود آورده است. از دیگر روش‌های انتساب بار داده که سعی در کاهش خطای مثبت- نادرست و بهبود نسبت کاهش حجم داده دارند، می‌توان به پژوهش‌های [۱۹-۱۷] اشاره کرد.

مسئله دیگری که همچنان وجود داشت و بررسی روی آن انجام نشده بود، پرس‌وجو بر روی رشته دارای الگوی مبهم[8] بود. برخی از بدافزارها دارای ماهیت چندریختی[9] هستند و در حین توزیع در سطح شبکه سازمان، بار داده آن‌ها تغییر می‌کند. با این حال بخش‌هایی از این بار داده ثابت بوده و می‌توان عملیات انتساب را انجام داد ولی بخش‌هایی از آن نیز مبهم است. در پژوهش‌های قبلی باید تمامی حالات برای بخش مبهم در نظر گرفته می‌شد و این موضوع زمان اجرای انتساب را افزایش داده و دقت را کاهش می‌داد. برای حل این مشکل حقیقت و همکارانش روش چند فیلتر بلوم وابسته به نویسه[10] [۲۰] را پیشنهاد دادند.

همان‌طور که در بخش قبل اشاره شد، در روش‌های انتساب بار داده که تاکنون ارائه شده‌اند، تلاش شده خطای مثبت-نادرست در فیلتر بلوم را کاهش داده و نسبت کاهش حجم داده را بهبود دهند و بیشتر ارزیابی‌ها مربوط به گرفتن جواب بله یا خیر به حضور یک رشته پرس‌وجو در فیلتر بلوم است. در حالی‌که خروجی یک روش انتساب بار داده باید ارائه شناسه‌های جریانی باشد که مشکوک به انتقال نمونه ترافیک مخرب هستند.

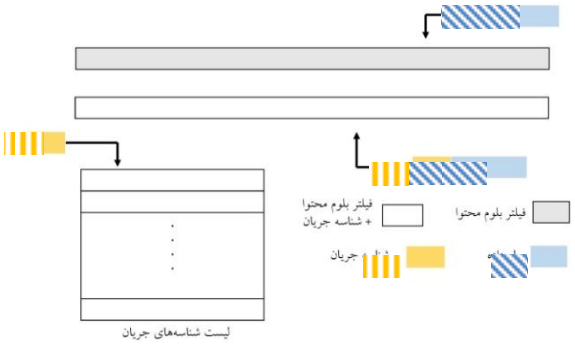
راه‌کاری که در کار شانموگاساندارام و همکارانش در روش فیلتر بلوم سلسله‌مراتبی [۲] برای تکمیل این سامانه ارائه شده است، به مشخص کردن جریان‌های حامل رشته پرس‌وجو نیز می‌پردازد. همچنین دو پژوهش انجام شده در زمینه انتساب بار داده عملیاتی، پروژه‌های ForNet [۱۵] و TOPO [۲۱] هستند، که در آن‌ها به استفاده از روش‌های انتساب بار داده اشاره شده است. نکته‌ای که باید به آن توجه شود این است که در این پژوهش‌ها ارزیابی کافی برای انتساب جریان‌های حامل رشته پرس‌وجو انجام نشده تا رشته مورد نظر را به آدرس‌های مبدأ و مقصد مشخصی انتساب دهد. همچنین نتایجی که در این زمینه

[1] False-Negative
[2] Block-Based Bloom Filter
[3] Hierarchical Bloom Filters
[4] Alignment
[5] Offset Collision
[6] Fingerprint algorithms
[7] Shingling
[8] Wildcard
[9] Polymorphism
[10] Character Dependent Multi-Bloom Filters

ارائه شده جامع نبوده و براساس آن ارزیابی کاملی نمی‌توان انجام داد.

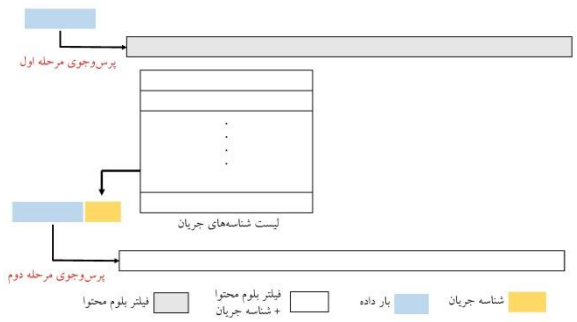
۳- تعریف مسئله

پیش از پرداختن به چالش‌های عملیاتی کردن روش‌های انتساب بار داده، ضروری است تا روند درج ترافیک در فیلتر بلوم و انجام پرس‌وجو مورد بررسی قرار بگیرد. برای انتساب جریان به رشته پرس‌وجو، نیاز است تا مانند رویکرد روش فیلتر بلوم سلسله‌مراتبی [۲] از دو عدد فیلتر بلوم استفاده شود. در یکی از فیلترها فقط محتوای بار داده بسته‌ها ذخیره شده و در دیگری، شناسه جریان همراه با محتوای بسته الحاق شده و درج شود. همچنین نیاز است تا برای هر بازه زمانی، فهرستی از شناسه‌های جریان نگهداری شود. این روند درج در شکل (۱) نشان داده شده است.



شکل (۱): درج بار داده در فیلترهای بلوم همراه با شناسه جریان

در فاز انجام پرس‌وجو، پس از این‌که از حضور رشته پرس‌وجو در فیلتر بلوم محتوا (پرس‌وجوی مرحله اول) اطمینان حاصل شد، رشته پرس‌وجو با تمامی شناسه‌های جریان موجود در فهرست مربوط به آن فیلتر بلوم، الحاق شده و پرس‌وجو مرحله دوم انجام می‌شود. اگر پاسخ به پرس‌وجو مربوط به یک شناسه مثبت باشد، آن شناسه جریان، به‌عنوان مبدأ و مقصد احتمالی رشته پرس‌وجو، گزارش می‌شود. فرآیند پرس‌وجو در شکل (۲) نشان داده شده است.



شکل (۲): پرس‌وجو در فیلترهای بلوم برای انتساب جریان

همان‌طور که در بخش‌های قبل نیز اشاره گردید، یکی از چالش‌های مهم سامانه‌های انتساب بار داده مسئله عملیاتی شدن و استفاده از این سامانه‌ها در کاربردهای مرتبط است. این چالش از دو منظر قابل بررسی است. ابتدا باید خروجی مطلوب سامانه‌های انتساب بار داده تولید شود. این خروجی شناسه‌های جریانی است که مشکوک به ارسال نمونه ترافیک مخرب هستند. از منظر دیگر، عملیاتی بودن این سامانه‌ها در بازه‌های طولانی ۳ الی ۶ ماهه مطرح می‌شود. در هر دو چالش، مشکل اساسی که وجود دارد، تعداد زیاد شناسه‌های جریان و یا فیلترهای بلوم است که باید پرس‌وجو شوند.

همان‌گونه که اشاره شد، ساختار داده فیلتر بلوم یک ساختار داده تصادفی است که در آن احتمال وقوع خطای مثبت-نادرست در پاسخ به یک پرس‌وجو وجود دارد. هرچه در یک فرآیند انتساب، تعداد پرس‌وجوهایی که انجام می‌شود بیشتر باشد، احتمال خطای نهایی سامانه نیز افزایش می‌یابد. اگر خطای هر یک از فیلترهای بلوم در سامانه انتساب بار داده p در نظر گرفته شود، هر پرس‌وجو در سامانه انتساب بار داده، معادل با یک آزمایش برنولی است که در آن احتمال موفقیت (بروز خطای مثبت- نادرست) برابر با p و احتمال شکست برابر با $q = 1-p$ خواهد بود. به عبارت دیگر هر پرس‌وجود در سامانه انتساب بار داده، یک آزمایش برنولی محسوب می‌شود.

حال اگر فرض شود پرس‌وجوهایی که در یک فرآیند در سامانه انتساب بار داده انجام می‌شود برابر با T_q آزمایش مستقل برنولی باشد، احتمال این‌که در K بار آزمایش، موفقیت حاصل شود، دارای توزیع دوجمله‌ای است و احتمال آن برابر با:

$$\Pr(X = K) = \binom{\mathrm{Tq}}{K} p^K q^{\mathrm{Tq}-K} \qquad (۲)$$

خواهد بود. به‌عبارت دیگر احتمال این‌که در یک فرآیند، K بار خطای مثبت- نادرست رخ دهد محاسبه شده است. امید ریاضی این متغیر تصادفی در توزیع دوجمله‌ای برابر با $E(x)= T_q$ p خواهد بود. همان‌طور که مشاهده می‌شود در یک انتساب، میانگین پاسخ‌های مثبت- نادرستی که مشاهده می‌شود با تعداد پرس‌وجوها در سامانه رابطه مستقیم دارد.

زیاد بودن تعداد پرس‌وجو بر روی فیلترهای بلوم در یک انتساب، از ۲ دیدگاه قابل بررسی است. از یک دیدگاه در بازه‌های زمانی مختلف، تعداد شناسه‌های جریان که در فهرستی نگهداری می‌شود بسیار زیاد است. همان‌طور که اشاره شد، اگر رشته پرس‌وجو در فیلتر بلوم اولیه وجود داشته باشد، رشته پرس‌وجو باید با تمامی شناسه‌های جریان موجود در آن فهرست الحاق

شده و در فیلتر دوم پرس‌وجو شود. تعداد زیاد شناسه‌های جریان، تعداد پرس‌وجوها را افزایش می‌دهد.

از دیدگاه دیگر در بسیاری از سازمان‌ها و نهادها نیاز است تا داده‌های شبکه برای بررسی‌های جرم‌شناسی برای مدت طولانی ۳ الی ۶ ماه نگهداری شود. باید به این نکته توجه داشت که به دلیل محدودیت در حافظه، در بررسی‌های طولانی‌مدت به ناچار پس از پر شدن فیلتر بلوم، داده‌ها در حافظه دائمی ذخیره شده و از فیلتر بلوم تازه‌ای برای ادامه کار استفاده می‌گردد. در این شرایط برای انجام پرس‌وجو در کل بازه زمانی، باید تمامی فیلترهای بلوم مورد پرس‌وجو قرار گیرند. افزایش تعداد فیلترهای بلوم نیز در نهایت تعداد پرس‌وجوها در یک انتساب را افزایش می‌دهد.

به‌عنوان مثال در ترافیکی که در این پژوهش مورد استفاده قرار گرفته و بر روی سوئیچ هسته در دانشکده‌ی کامپیوتر دانشگاه صنعتی شریف ضبط شده است، در بازه زمانی تقریبا ۷۸۴ ثانیه، ۵۰۰۴۰۲ جریان منحصربه‌فرد مشاهده شده است. حال اگر با همین نسبت، بازه زمانی ۳ ماهه در نظر گرفته شود، تعداد شناسه‌های جریان برای پرس‌وجوی مرحله دوم بسیار زیاد خواهد بود.

همچنین اشاره به این نکته نیز ضروری است که در روش‌های انتساب بار داده، برای پشتیبانی از پرس‌وجو بر روی زیررشته‌ها، بسته‌ها به قطعاتی تقسیم می‌شوند. لذا در زمان پرس‌وجو نیز باید رشته مورد نظر قطعه‌بندی شده و به ازای هر شناسه جریان، چندین پرس‌وجو بر روی فیلتر بلوم انجام می‌شود و این مسأله تعداد پرس‌وجوها را چندین برابر می‌کند. با راه‌کاری که در بخش بعدی توضیح داده شده، تلاش می‌شود بازه پرس‌وجو کاهش یافته و خطای نهایی سامانه کمتر شود.

۴- راه‌کار پیشنهادی: سلسله‌مراتب زمان

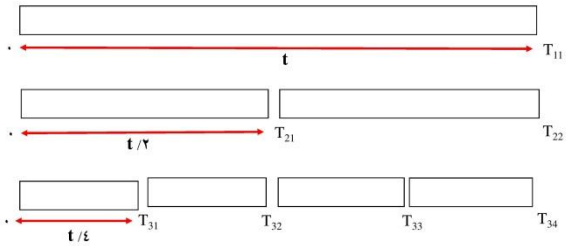
همان‌طور که در بخش قبل مطرح شد، چالش‌های مهم موجود در حوزه انتساب بار داده، تعداد زیاد شناسه‌های جریان و فیلترهای بلوم برای پرس‌وجو و عملیاتی نبودن این روش‌ها در بررسی‌های طولانی‌مدت است. راه‌کار پیشنهادی تلاش می‌کند فضای پرس‌وجو (اعم از تعداد شناسه‌های جریان و تعداد فیلترهای بلوم برای پرس‌وجو) را کاهش داده و در نهایت خطای کلی سامانه را کمتر کند. همچنین برای جلوگیری از تعویض‌های مکرر بین فیلترهای بلوم، سعی شده است از حداکثر فضای حافظه برای داشتن یک فیلتر بلوم یکپارچه استفاده شود و تعداد فیلترهای بلوم برای پرس‌وجو کاهش یابد.

در راستای کاهش تعداد جریان‌های مشکوک برای پرس‌وجو، باید روند از بالا به پایینی برای پرس‌وجو وجود داشته باشد تا در هر مرحله تعدادی از جریان‌ها کاهش یابند. این روند از بالا به پایین در ساختار سلسله‌مراتبی وجود دارد. همان‌طور که در پروژه‌ی TOPO [۲۱] در ساختار سلسله‌مراتبی شبکه، تعداد شاخه‌های سلسله‌مراتب رفته‌رفته هرس می‌شود و برای پرس‌وجو تعداد محدودی مسیریاب بالادستی باقی می‌ماند، برای راه‌حل پیشنهادی نیز با این دیدگاه ساختار سلسله‌مراتب زمان در نظر گرفته شده است.

تاکنون سلسله‌مراتب بر روی محتوا و مکان قرارگیری فیلترهای بلوم در ساختار شبکه ارائه شده است. شانموگاساندارام و همکارانش در روش فیلتر بلوم سلسله‌مراتبی[۲]، سلسله‌مراتب بر روی محتوای بسته را اعمال نمودند که باعث کاهش خطا و رفع مشکل تصادم شماره قطعه‌ها شده است. در پژوهش مربوط به پروژه‌ی TOPO نیز از ساختار سلسله‌مراتبی شبکه برای جایگذاری فیلترهای بلوم در ساختار شبکه و کاهش تعداد مسیریاب‌های بالادستی برای انجام پرس‌وجو، استفاده شده است. در این پژوهش برای رفع چالش‌های موجود، روش سلسله‌مراتب زمان ارائه شده که جزئیات این ساختار و روند درج و پرس‌وجو در ادامه توضیح داده شده است.

۴-۱- ساختار سلسله‌مراتب زمان

در ساختار سلسله‌مراتب زمان، سطوح مختلفی از فیلترهای بلوم در نظر گرفته شده است. در هر سطح تعداد فیلترهای بلوم و تعویض از یک فیلتر بلوم به فیلتر بلوم دیگر متفاوت خواهد بود. در سطوح پایین سلسله‌مراتب، بازه زمانی که هر فیلتر بلوم ترافیک را نگهداری می‌کند، کوچک‌تر می‌شود. نمونه‌ای از ساختار سلسله‌مراتب زمان در شکل (۳) نشان داده شده است.



شکل (۳): نمونه‌ای از ساختار سلسله‌مراتب زمان

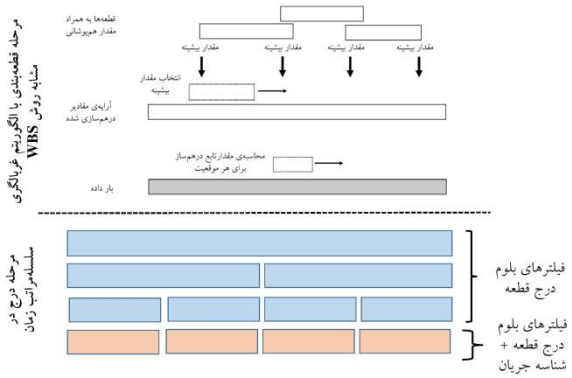
در این ساختار سلسله‌مراتبی ترافیک در هر لحظه در تمامی سطوح درج می‌شود. به‌طور مثال سلسله‌مراتب زمان در شکل (۳) از ۳ سطح تشکیل شده است. اگر فرض شود که فیلتر بلوم بالاترین سطح، ترافیک یک شبانه‌روز را ذخیره کند و هریک از فیلترهای بلوم سطح ۲ و ۳ به‌ترتیب ترافیک ۱۲ و ۶ ساعت را

ذخیره کنند، ترافیک ساعت اول در فیلترهای بلوم با بازه زمانی 0-T11، 0-T21 و 0-T31 ذخیره خواهد شد. ترافیک مربوط به ساعت ۲۴-۲۳ نیز در فیلترهای بلوم بازه 0-T11، T21-T22 و T33-T34 ذخیره می‌شود.

۴-۱-۱- درج بسته‌ها در سلسله‌مراتب زمان

براساس روش‌های انتساب بار داده، یک بسته برای درج در این ساختار، به قطعه‌هایی تقسیم شده و سپس این قطعه‌ها در تمامی سطوح سلسله‌مراتب در فیلتر بلوم درج می‌شود. باید به این نکته اشاره شود که در این راه‌کار، هر فیلتر بلوم در هر یک از سطوح، دارای بازه زمانی آغاز و پایان است. بسته مورد نظر و قطعات متناظر با آن، در هر سطح در آن فیلتر بلوم درج خواهد شد که بازه زمانی آن متناسب با زمان رسیدن بسته باشد.

همچنین برای این که راه‌حل ارائه شده، انتساب جریان را پشتیبانی نماید، در پایین‌ترین سطح سلسله‌مراتب زمان و متناظر با برگ‌های سلسله‌مراتب، فیلترهای بلوم دیگری در نظر گفته شده است. محتوای قطعه‌های بسته با شناسه جریان، الحاق‌شده و در این فیلترهای بلوم درج می‌شوند. در کنار هر یک از فیلترهای بلوم پایین‌ترین سطح، فهرستی وجود دارد که شناسه جریان نیز به این فهرست اضافه می‌شود. روند درج بار داده بسته‌ها در شکل (۴) نشان داده شده است.



شکل (۴): روند درج بسته‌ها در روش سلسله‌مراتب زمان

۴-۱-۲- پرس‌وجو در سلسله‌مراتب زمان

پیش از شروع توضیح در مورد روند پرس‌وجو در سلسله‌مراتب پیشنهادی، لازم است ابتدا رابطه فرزند- والد بین فیلترهای بلوم سلسله‌مراتب زمان توضیح داده شود. هر فیلتر بلوم در سطح بالای سلسله‌مراتب زمان، والد فیلترهای بلوم در سطح پایین خودش است به شرطی که بازه زمانی فیلتر بلوم در سطح پایین، زیرمجموعه‌ای از بازه زمانی فیلتر بلوم سطح بالا باشد. به‌طور مثال فیلتر بلوم سطح بالا برای درج داده‌های ساعت ۱۲-۲۴ در نظر گرفته شده است. فیلتر بلوم در سطح پایین که بسته‌های ساعت ۲۴-۲۲ در آن درج می‌شود، فرزند فیلتر بلوم سطح بالا خواهد بود.

برای انجام پرس‌وجو ابتدا رشته پرس‌وجو به قطعه‌هایی تقسیم می‌شود. سپس باید برای تک‌تک این قطعه‌ها در ساختار سلسله‌مراتب زمان پرس‌وجو انجام شود. روند پرس‌وجو از بالا به پایین است به‌طوری که ابتدا هر قطعه در فیلتر بلوم بالاترین سطح پرس‌وجو می‌شود. اگر پاسخ پرس‌وجو برای قطعه‌ی اول مثبت بود، بقیه‌ی قطعه‌ها نیز در این فیلتر بلوم پرس‌وجو خواهند شد. اگر تمامی قطعه‌های رشته پرس‌وجو در فیلتر بلوم بالاترین سطح وجود داشته باشند، پرس‌وجو در فرزندهای این فیلتر بلوم ادامه خواهد یافت. درصورتی‌که پاسخ فیلتر بلوم والد به پرس‌وجویی منفی باشد، پرس‌وجو در همان فیلتر بلوم متوقف شده و فرزندهای آن فیلتر بلوم پرس‌وجو نخواهند شد.

در حالت کلی پاسخ به پرس‌وجو در سلسله‌مراتب زمان، دارای ۲ مرحله است. در مرحله اول فیلترهای بلوم سلسله‌مراتب زمان که قطعه‌ها به تنهایی در آن‌ها درج شده، طبق روال ذکر شده، صرفاً برای گرفتن جواب بله یا خیر (وجود داشتن یا نداشتن رشته پرس‌وجو در سلسله‌مراتب زمان) پرس‌وجو می‌شوند. اگر حداقل یک شاخه از سلسله‌مراتب از ریشه تا برگ وجود داشته باشد که تمامی فیلترهای بلوم بر روی آن شاخه، پاسخ مثبت به پرس‌وجو داده باشند، آن‌گاه پاسخ سلسله‌مراتب زمان به پرس‌وجو «مثبت» است و رشته مورد نظر احتمالاً در سلسله‌مراتب درج شده است. این روند در شکل (۵) نشان داده شده است.

شکل (۵): پاسخ مثبت یا منفی به حضور رشته پرس‌وجو در سلسله‌مراتب زمان

در مرحله بعد، پس از این‌که برای حضور رشته پرس‌وجو در سلسله‌مراتب زمان، پاسخ «مثبت» برگردانده شد، باید در

فیلترهای بلوم متناظر با برگ‌های سلسله‌مراتب که قطعه‌های الحاق‌شده با شناسه جریان در آن‌ها درج شده، پرس‌وجو انجام شود. پرس‌وجوی مرحله دوم در فیلترهای بلوم پایین‌ترین سطح که فیلترهای بلوم برگ متناظر با آن‌ها، پاسخ مثبت به پرس‌وجوی مرحله اول داده‌اند، انجام خواهد شد. به‌عنوان مثال در شکل (۶) مشاهده می‌شود که فیلتر بلوم برگ ۲ و ۳ پاسخ مثبت به پرس‌وجوی مرحله اول داده‌اند. پرس‌وجوی مرحله دوم در فیلترهای بلوم ۲ و ۳ پایین‌ترین سطح که در آن قطعه‌ها با شناسه جریان الحاق و درج شده‌اند، انجام خواهد شد.

پس از تعیین فیلترهای بلوم سطح پایین برای پرس‌وجوی مرحله دوم، قطعه‌های رشته پرس‌وجو با تک‌تک شناسه‌های جریان در فهرست هر یک از این فیلترهای بلوم الحاق شده و در فیلتر بلوم متناظر، پرس‌وجو صورت می‌گیرد. اگر پاسخ پرس‌وجو برای الحاق با یک شناسه جریان مثبت باشد، این شناسه حاوی مبدأ و مقصد احتمالی رشته پرس‌وجو است.

شکل (۶) تعیین فیلترهای بلوم سطح پایین برای پرس‌وجوی مرحله دوم

همچنین در روند پرس‌وجو باید به این نکته توجه داشت که برای چه بازه زمانی بررسی و انتساب انجام می‌شود. اگر در طول مدت زمان مورد نظر، به خاطر پرشدن فیلتر بلوم تعویض از یک سلسله‌مراتب زمان به سلسله‌مراتب دیگر رخ داده باشد، در روند پرس‌وجو باید در تمامی سلسله‌مراتب‌های موجود پرس‌وجو انجام شود. همان‌طور که ملاحظه می‌شود با کوچک‌تر و تکه‌تکه شدن فیلترها در سطوح پایین‌تر، ریزدانگی زمانی و کاهش تعداد شناسه‌های جریان برای پرس‌وجو حاصل می‌شود.

به‌عبارت دیگر می‌توان با ریزدانگی بیشتری زمان مشاهده نمونه ترافیک مخرب را اعلام کرد. همچنین با کاهش تعداد شناسه‌های جریان که اشتباه گزارش شده‌اند، خطای کلی سامانه کمتر می‌شود. با کمتر شدن خطای کلی سامانه، روش ارائه‌شده نسبت به روش پیشین، از بازه زمانی طولانی‌تری برای انتساب پشتیبانی خواهد کرد.

۵- ارزیابی راه‌کار پیشنهادی

مسئله کلی که در این پژوهش مطرح می‌شود، تغییر در ماهیت روش‌های انتساب بار داده نیست. به عبارت دیگر تلاش نمی‌شود تا با ارائه روش جدید، خطای فیلتر بلوم کاهش یافته و یا نسبت کاهش حجم داده بهبود یابد. بلکه در راه‌کار ارائه‌شده، نحوه به‌کارگیری روش‌های انتساب بار داده موجود، مورد بررسی قرار می‌گیرد. در بخش ارزیابی، حالت معمول استفاده از روش‌های انتساب بار داده با حالتی که از سلسله‌مراتب زمان استفاده شده، بررسی می‌شود.

در ارزیابی که بر روی روش‌های انتساب بار داده انجام شده است، روش هم‌پوشانی قطعه‌ی اندازه متغیر با غربالگری[1] (WBS) [۱۶] نسبت به روش‌های دیگر کارایی بهتری داشته است. در این پژوهش نیز برای مقایسه روش‌های قبلی و سلسله‌مراتب زمان، بر روی فیلترهای بلوم از روش WBS برای درج و پرس‌وجو استفاده شده است. به‌عبارت دیگر در ارزیابی‌های صورت گرفته، روش انتساب بار داده WBS که بهترین روش انتساب بار داده است، انتخاب شده و یک بار در حالت عادی و معمولی و بار دیگر در روش سلسله‌مراتب زمان مورد استفاده قرار می‌گیرد. در ادامه این ۲ وضعیت با یکدیگر مورد ارزیابی قرار می‌گیرند.

در مرحله اول برای داشتن سامانه انتساب بار داده، روش هم‌پوشانی با قطعه متغیر با غربالگری (WBS) با زبان پایتون پیاده‌سازی گردید. اندازه پنجره الگوریتم غربالگری مطابق با پژوهش [۱۶] ۳۲ بایت و مقدار هم‌پوشانی ۴ بایت در نظر گرفته شده است. تابع درهم‌ساز در این الگوریتم نیز، تابع درهم‌سازی MD5 است.

روند مشخص شدن قطعه‌ها در روش WBS بدین شکل است که ابتدا پنجره‌ای به طول k بر روی بار داده حرکت داده شده و در هر موقعیت با استفاده از تابع MD5 مقدار درهم‌سازی آن محاسبه شده و در آرایه‌ای ذخیره می‌شود. سپس پنجره‌ای با اندازه w بر روی آرایه مقادیر درهم‌سازی‌شده حرکت داده شده و در هر موقعیت، بیشینه مقدار انتخاب می‌شود که این بیشینه مقدار مرز بین قطعه‌ها خواهد بود.

پس از پیاده‌سازی روش WBS، ساختار داده سلسله‌مراتبی تعریف شده که به‌عنوان ورودی، تعداد سطوح سلسله‌مراتب، تعداد فرزندهای هر فیلتر بلوم سطح بالا، تعداد درج و مدت زمان نگهداری داده در فیلتر بلوم بالاترین سطح را دریافت می‌کند. پس از دریافت این پارامترها در این سلسله‌مراتب مجموعه‌ای از فیلترهای بلوم با ویژگی‌های توضیح داده شده در فصل قبل،

[1] Winnowing Block Shingling (WBS)

ایجاد شده است. بر روی هریک از این فیلترهای بلوم در سلسله‌مراتب زمان، درج بسته‌ها با روش WBS پیاده‌سازی و اجرا می‌گردد.

در ساختار سلسله‌مراتب زمان، بار داده با روش WBS قطعه‌بندی شده و قطعه‌های قابل درج در فیلترهای بلوم تولید می‌شود. قطعه‌ها سپس در هر سطح، در فیلتر بلوم متناسب با زمان رسیدن بسته، درج می‌شوند. درج در تمامی سطوح انجام شده و در فیلترهای بلوم موجود در پایین‌ترین سطح نیز، هر کدام از قطعه‌ها با شناسه جریان بسته، الحاق شده و درج صورت می‌گیرد. در انتها در فهرست هر یک از فیلترهای بلوم پایین‌ترین سطح، شناسه جریان اضافه می‌شود.

پس از پیاده‌سازی سلسله‌مراتب زمان با ویژگی‌های مورد نظر و روش WBS در وضعیت معمولی، درج داده‌ها شروع می‌شود. مجموعه داده ورودی به هریک از روش‌ها یک پرونده در قالب PCAP است. مجموعه داده مورد استفاده، ۴ گیگابایت از ترافیک شبکه مربوط به سوئیچ مرکزی دانشکده کامپیوتر دانشگاه صنعتی شریف حاوی بسته‌های تی‌سی‌پی[1] و یودی‌پی[2] است. این ترافیک حاوی ۷۵۱۹۶۷۰ بسته بوده که در بازه ۷۸۴ ثانیه جمع‌آوری شده است. در این ترافیک ۵۰۰۴۰۲ شناسه جریان منحصربفرد مشاهده شده است. شناسه جریان در این ارزیابی، ترکیبی از آدرس اینترنت مبدأ و مقصد تعریف شده است.

همچنین باید اشاره کرد که این پیاده‌سازی و ارزیابی‌ها بر روی سامانه عامل لینوکس توزیع دبیان با حافظه دسترسی تصادفی[3] ۸ گیگابایت اجرا شده است. مبنای تعویض از یک سلسله‌مراتب به سلسله‌مراتب جدید نیز براساس زمان است. اگر زمان رسیدن بسته جاری، از زمان پایان فیلتر بلوم در بالاترین سطح بیشتر باشد، این سلسله‌مراتب در حافظه دیسک ذخیره شده و درج بسته‌ها در سلسله‌مراتب جدید انجام می‌شود.

در ادامه، پس از درج بسته‌ها در روش سلسله‌مراتب زمان و وضعیت معمولی، برای ارزیابی و مقایسه نتایج، پرس‌وجوهایی در هر دو روش انجام می‌شود تا مقدار خطای مثبت- نادرست و تعداد جریان‌های اشتباه گزارش‌شده بررسی شود. برای انجام پرس‌وجو، رشته‌های تصادفی با طول‌ها ۱۰۰ بایت تولید شده و با استفاده از این رشته‌ها پرس‌وجو انجام می‌شود.

دلیل استفاده از رشته‌های پرس‌وجو به طول ۱۰۰ بایت این است که هرچه طول رشته پرس‌وجو بیشتر باشد، احتمال وقوع خطا در فیلترهای بلوم کاهش می‌یابد. همچنین با رشته‌های پرس‌وجو با طول کمتر نیز احتمال وقوع خطا بیشتر بوده و شناسه‌های جریان زیادی در روند بررسی باید پرس‌وجو شود که زمان ارزیابی را تحت تاثیر قرار خواهد داد. از این رو رشته با طول ۱۰۰ بایت، نمونه مطلوبی برای ارزیابی خواهد بود.

در این بخش همچون پژوهش [۱۷] فرض می‌شود، رشته تولیدشده تصادفی بوده و با احتمال بسیار بالایی در فیلترهای بلوم هر دو وضعیت وجود ندارد. در این صورت هر پاسخ مثبتی که به پرس‌وجو برگردانده شود، یک خطای مثبت- نادرست بوده و جریانی که به‌عنوان حامل احتمالی رشته پرس‌وجو معرفی می‌شود، جریان اشتباه خواهد بود.

۵-۱- بررسی سناریوی مقایسه و تحلیل نتایج

در این بخش از ارزیابی، مصرف حافظه در هر دو وضعیت سلسله‌مراتب زمان و WBS معمولی یکسان در نظر گرفته شده و نسبت کاهش حجم داده برابر با ۱۵۰ به ۱ است.

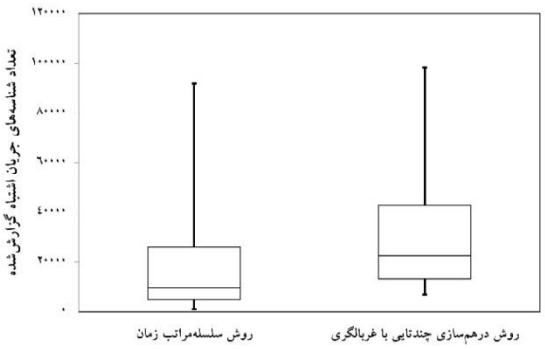
در بزرگ‌ترین فیلترهای بلوم در هر دو روش ترافیک بازه‌های زمانی ۸۰ ثانیه‌ای ذخیره و نگهداری می‌شود. در این ارزیابی در روش WBS معمولی، مصرف حافظه برای درج چندین باره بسته‌ها در فیلترهای بلوم استفاده شده است. به‌عبارت دیگر مقایسه بین روش سلسله‌مراتب زمان و نمونه چندتایی از روش WBS انجام خواهد شد. در نتیجه در هر دو روش، ۳ سطح فیلتر بلوم برای درج بسته‌ها استفاده شده و ۱۰ بار تعویض از سلسله‌مراتب زمان به سلسله‌مراتب دیگر و از یک فیلتر بلوم به فیلتر دیگر در روش WBS چندتایی انجام می‌شود. این حالت از مقایسه در شکل (۷) نمایش داده شده است.

شکل (۷): مقایسه در حالت مصرف حافظه یکسان، نسبت کاهش حجم داده یکسان

پس از درج ترافیک در روش سلسله‌مراتب زمان و روش WBS چندتایی، تعداد ۱۰۰۰ رشته پرس‌وجو با طول ۱۰۰ بایت در هر یک از روش‌ها پرس‌وجو شده و نتایج آن در این بخش ارائه می‌گردد. در ادامه نمودار توزیع تعداد شناسه‌های جریان اشتباه

[1] TCP
[2] UDP
[3] RAM

گزارش‌شده در روش سلسله‌مراتب زمان و روش WBS چندتایی برای رشته پرس‌وجو با طول ۱۰۰ بایت در شکل (۸) نشان داده شده است.



شکل (۸): توزیع تعداد شناسه‌های جریان اشتباه گزارش‌شده برای رشته ۱۰۰ بایتی

همان‌طور که ملاحظه می‌شود در روش سلسله‌مراتب زمان به دلیل این‌که تعداد شناسه‌های جریان که در پایین‌ترین سطح سلسله‌مراتب پرس‌وجو می‌شوند، نسبت به WBS چندتایی کمتر بوده، در مجموع تعداد شناسه‌های اشتباه کمتری برای ۱۰۰۰ پرس‌وجو گزارش شده است. همچنین میانگین تعداد شناسه‌های جریان گزارش‌شده در روش سلسله‌مراتب زمان ۱۹۹۲۶ شناسه و در روش WBS چندتایی این میانگین برابر با ۲۸۳۲۷ شناسه بوده است.

در نمودار توزیع تعداد شناسه‌های اشتباه در شکل (۸)، بیشینه مؤثر در سلسله‌مراتب زمان و WBS چندتایی به ترتیب برابر با ۲۵۹۴۱ و ۴۲۸۹۸ شناسه است. به‌عبارت دیگر در ۲۵ درصد از مواردی که پاسخ مثبت به پرس‌وجوی مرحله اول داده شده، در روش سلسله‌مراتب زمان، تعداد شناسه‌هایی بیش از ۲۵۹۴۱ و در نمونه WBS چندتایی بیش از ۴۲۸۹۸ اشتباه گزارش شده است.

نتیجه‌گیری نهایی این است که در روش سلسله‌مراتب زمان به‌دلیل تقسیم فیلترهای بلوم در سطوح پایین‌تر به فیلترهای بلوم کوچک‌تر و با توجه به اینکه احتمال رخ ندادن خطا در برخی از شاخه‌های سلسله‌مراتب وجود دارد، فضای پرس‌وجو بر روی شناسه‌های جریان محدود خواهد شد. با محدود شدن تعداد شناسه‌های جریان برای پرس‌وجو، در تعداد شناسه‌های جریان اشتباه گزارش‌شده بهبود حاصل می‌شود.

همچنین در روش سلسله‌مراتب زمان ریزدانگی زمانی بیشتری نسبت به روش WBS چندتایی وجود دارد. به‌طوری‌که در سلسله‌مراتب زمان، فیلترهای بلوم پایین‌ترین سطح، شناسه‌های جریان برای بازه زمانی ۲۰ ثانیه را نگهداری می‌کنند در حالی‌که شناسه‌های موجود در فیلترهای بلوم روش WBS چندتایی برای بازه‌های ۸۰ ثانیه‌ای هستند.

همان‌طور که ملاحظه می‌شود، روش سلسله‌مراتب زمان در حالت کلی مقدار خطای کمتری نسبت به روش WBS معمولی داشته و ۸۴۰۰ مورد شناسه جریان کمتری را نسبت به روش قبلی به اشتباه گزارش می‌کند. در نهایت خطای سامانه در این سناریو از ۵/۶۶ به ۳/۹۸ درصد کاهش می‌یابد. با توجه به این‌که خروجی نهایی سامانه انتساب بار داده شناسه‌های جریان حامل رشته پرس‌وجو است، روشی که در مجموع شناسه‌های اشتباه کمتری گزارش دهد، مطلوبیت بیشتری خواهد داشت. همچنین بدیهی است هرچه تعداد شناسه‌های اشتباه تعداد کمتری داشته باشند، خطای نهایی سامانه انتساب بار داده کمتر بوده و از بررسی در بازه‌های زمانی طولانی‌تر پشتیبانی می‌کند.

۶- نتیجه‌گیری

همان‌طور که در این پژوهش اشاره شد، تاکنون در راستای عملیاتی کردن روش‌های انتساب بار داده در شبکه‌های واقعی و برای طولانی‌مدت پژوهش قابل قبولی ارائه نشده است. خروجی این روش‌ها باید شناسه‌های جریانی باشد که مشکوک به ارسال نمونه ترافیک مخرب هستند. در این پژوهش اساسی‌ترین مشکل عملیاتی کردن روش‌های انتساب بار داده، زیاد بودن پرس‌وجوها در بازه زمانی طولانی مدت مطرح شده است.

پرس‌وجوی زیاد به دلیل زیاد بودن تعداد شناسه‌های جریان برای پرس‌وجو و همچنین تعویض بین فیلترهای بلوم و زیاد شدن تعداد فیلترهای بلوم ناشی شده است. راه‌کاری که در این پژوهش ارائه شد، راه‌کاری مبتنی بر سلسله مراتب زمان بود که در آن احتمال رخ ندادن خطا در برخی از شاخه‌های آن وجود داشت که باعث می‌شد شناسه‌های جریان موجود در آن شاخه پرس‌وجو نشوند و در نهایت مقدار شناسه‌هایی که اشتباه گزارش می‌شوند کاهش یابد. همچنین برای جلوگیری از زیاد شدن تعویض‌ها بین فیلترهای بلوم، بزرگ‌ترین اندازه حافظه یکپارچه در بالاترین سطح سلسله‌مراتب انتخاب شده تا تعداد فیلترهای بلوم برای پرس‌وجو کاهش یابد.

در بخش ارزیابی دو معیار تاثیرگذار حافظه و نسبت کاهش حجم داده در نظر گرفته شده و مقایسه بین روش سلسله‌مراتب زمان و روش WBS چندتایی انجام شد. در این ارزیابی مشاهده شد که روش سلسله‌مراتب زمان ۸۴۰۰ شناسه جریان کمتری را به اشتباه گزارش می‌کند و خطای سامانه از ۵/۶۶ به ۳/۹۸ درصد کاهش می یابد. همچنین هرچه تعداد اشتباهات سامانه کمتر

باشد، سامانه انتساب بـار داده، در بـازه‌هـای زمـانی طـولانی‌تـری مقبولیت خواهد داشت.

۷- منابع

[1] E. S. Pilli, R. C. Joshi, and R. Niyogi, "Network forensic frameworks: Survey and research challenges", *digital investigation*, vol. 7, pp. 14-27, 2010.

[2] K. Shanmugasundaram, H. Brönnimann, and N. Memon, "Payload attribution via hierarchical bloom filters", in *Proceedings of the 11th ACM Conference on Computer and Communications Security*, Washington, DC, USA, pp. 31-41, 2004.

[3] D. D. Clark and S. Landau, "Untangling attribution", *Harv. Nat'l Sec. J.*, vol. 2, pp. 323-353, 2011.

[4] B. H. Bloom, "Space/time trade-offs in hash coding with allowable errors", *Communications of the ACM*, vol. 13, pp. 422-426, 1970.

[5] A. Almulhem and I. Traore, "A survey of connection-chains detection techniques", in 2007 IEEE Pacific Rim Conference on Communications, Computers and Signal Processing, Victoria, B.C., Canada, pp. 219-222, 2007.

[6] S. C. Lee and C. Shields, "Challenges to automated attack traceback", IT professional, vol. 4, pp. 12-18, 2002.

[7] A. Mairh, D. Barik, K. Verma, and D. Jena, "Honeypot in network security: a survey", in International conference on communication, computing & security, ODISHA, India, pp. 600-605, 2011.

[8] A. C. Snoeren, C. Partridge, L. A. Sanchez, C. E. Jones, F. Tchakountio, S. T. Kent, *et al.*, "Hash-based IP traceback," *ACM SIGCOMM Computer Communication Review*, vol. 31, pp. 3-14, 2001.

[9] C. Gong and K. Sarac, "IP traceback based on packet marking and logging", in *IEEE International Conference on Communications*, Seoul, Korea, pp. 1043-1047, 2005.

[10] S. M. Bellovin, M. Leech, and T. Taylor, "ICMP traceback messages", Internet draft: draftietfitrace 03. txt, 2003.

[11] C. Gong and K. Sarac, "IP traceback based on packet marking and logging", in IEEE International Conference on Communications, Seoul, Korea, pp. 1043-1047, 2005.

[12] L. Fan, P. Cao, J. Almeida, and A. Z. Broder, "Summary cache: a scalable wide-area web cache sharing protocol", IEEE/ACM Transactions on Networking (TON), vol. 8, pp. 281-293, 2000.

[13] F. M. Cuenca-Acuna, C. Peery, R. P. Martin, and T. D. Nguyen, "Planetp: Using gossiping to build content addressable peer-to-peer information sharing communities", in High Performance Distributed Computing, 2003. Proceedings. 12th IEEE International Symposium on, pp. 236-246, 2003.

[14] A. Broder and M. Mitzenmacher, "Network applications of bloom filters: A survey", Internet mathematics, vol. 1, pp. 485-509, 2004.

[15] K. Shanmugasundaram, N. Memon, A. Savant, and H. Bronnimann, "ForNet: A distributed forensics network", in International Workshop on Mathematical Methods, Models, and Architectures for Computer Network Security, Petersburg, Russia, pp. 1-16, 2003.

[16] M. Ponec, P. Giura, J. Wein, and H. Brönnimann, "New payload attribution methods for network forensic investigations", ACM Transactions on Information and System Security (TISSEC), vol. 13, pp. 15-47, 2010.

[17] Chen, Yan, et al. "CAS: Content Attribution System for Network Forensics." International Conference on Trustworthy Computing and Services. Springer, Berlin, Heidelberg, 2014.

[18] Wei, Yichen, et al. "Winnowing multihashing structure with wildcard query." Asia-Pacific Web Conference. Springer, Cham, 2014.

[19] Hosseini, S. Mohammad, and Amir Hossein Jahangir. "An Effective Payload Attribution Scheme for Cybercriminal Detection Using Compressed Bitmap Index Tables and Traffic Downsampling." IEEE Transactions on Information Forensics and Security 13.4 (2018): 850-860.

[20] M. H. Haghighat, M. Tavakoli, and M. Kharrazi, "Payload attribution via character dependent multi-bloom filters", *IEEE Transactions on Information Forensics and Security*, vol. 8, pp. 705-716, 2013.

[21] L. Zhang and Y. Guan, "TOPO: A topology-aware single packet attack traceback scheme", in Securecomm and Workshops, Securecomm and Workshops, 2006, pp. 1-10, 2006.

Journal of Electronical & Cyber Defence
Vol. 7, No. 3, 2019, Serial No. 27

Improving Payload Attribution Techniques in Computer Network Criminology with Time based Hierarchical Bloom Filter

Z. Sasan, M. Kharazi*

*Sharif University of Technology

(Received: 10/09/2018, Accepted: 13/10/2018)

ABSTRACT

In the light of increased network attacks, payload attribution is an essential part of any forensics analysis of the attack. Usually attribution has to be done based on the payload of the packets. In such techniques network traffic should be stored in its entirety while user privacy is preserved. Bloom filters have been an ideal tool for such requirements. Previous works in this area have tried to minimize the false positive error rate associated with the bloom filter while improving on the data reduction ratio but there has not been any notable research on practical implementations in computer networks. A payload attribution technique should provide a list of connections which are suspects of carrying a specific payload (i.e. malware signature). The problem arises with the fact that there are too many queries required, given the large number of connections and the number of bloom filters involved over long time periods, which results in a large aggregate error rate. In this work, we propose a technique with which a time-based hierarchical bloom filter configuration is proposed to tackle the noted problem. Our evaluation shows that with this proposed technique we are able to limit the false positive error rate of the system as compared to the previously proposed techniques. This leads to an overall error reduction in the payload attribution system. More specifically, the error rate compared to previous work drops from 5.66% to 3.98% which results in reducing the number of incorrectly identified flows by 8400.

Keywords: Network Forensics, Payload Attribution

* Corresponding Author Email: kharrazi@sharif.edu