مجموعه مقالات پانزدهمین کنفرانس بین‌المللی انجمن رمز ایران
دانشکده مهندسی برق، دانشگاه تربیت دبیر شهید رجایی، تهران، ۶-۷ شهریور ۱۳۹۷
(مقاله مدعو)

شناسایی منشاء انتشار کرم‌واره و بازسازی مسیر آن به شیوه احتمالی

طلا تفضلی[۱]
[۱] عضو هیئت علمی، پژوهشکده امنیت ارتباطات و فناوری اطلاعات، پژوهشگاه ارتباطات و فناوری اطلاعات ، تهران، ایران
tafazoli@itrc.ac.ir

چکیده

تحقیق ارائه شده، گزارشی از رساله دکترای اینجانب می‌باشد که زیر نظر استاد ارجمند آقای دکتر بابک صادقیان در دانشکده کامپیوتر و فناوری اطلاعات دانشگاه صنعتی امیرکبیر انجام گرفته است[۱].

شناسایی منشاء و بازسازی مسیر انتشار کرم‌واره از مسائل مهم در امنیت اطلاعات و فورنسیک دیجیتال می‌باشد. این اطلاعات به مامور پی‌جویی کمک می‌نماید تا مظنونین اولیه را حدس بزند و پی‌جویی‌های بعدی را بر روی کامپیوترهای مورد ظنّ انجام دهد. همچنین مدیران امنیت شبکه و سیستمها با این اطلاعات، نقاط ضعف امنیتی و آسیب‌پذیریهای سیستمها و شبکه‌های خود را شناسایی می‌نمایند.

هدف از این رساله، شناسایی منشاء و بازسازی مسیر انتشار کرم‌واره پویشگر ترجیحی است. ایده ما در حل این مسئله، استفاده از مدلسازی حرکت رو-به-عقب زمانی و همچنین رویکرد تکمیل مرحله به مرحله‌ای می‌باشد تا پس از وقوع حمله کرم‌واره با داشتن اطلاعات جمع‌آوری شده در سطح شبکه، منشاء را شناسایی نموده و مسیر انتشار کرم‌واره را بازسازی نماییم. در این رساله تابع توزیع احتمالی را شکل دادیم که با دریافت ویژگیهایی از شبکه، احتمال آلودگی نودها به کرم‌واره پویشگر ترجیحی را در هر زمان تخمین بزنیم. همچنین بر مبنای این تابع توزیع، الگوریتمی را شکل دادیم که منشاء انتشار کرم‌واره را شناسایی می‌نماید و مسیر انتشار کرم‌واره را بازسازی می‌کند. به منظور نیل به این هدف، رویکرد چهار مرحله‌ای را اجرا نمودیم. دقت الگوریتم برای تشخیص یالها، به طور متوسط به میزان ۹۰٪ می‌باشد.

کلمات کلیدی

مدل انتشار آلودگی، شناسایی منشاء انتشار کرم‌واره، بازسازی مسیر انتشار کرم‌واره، تابع توزیع احتمالی، مدل رو-به-عقب، الگوریتم تخمین تابع توزیع، شبکه بیزین.

۱- مقدمه

یکی از مسائل دشواری که از سالها پیش مطرح بوده، شناسایی منشاء انتشار اطلاعات و همه‌گیری و بازسازی مسیر آن بر مبنای اطلاعات محدود درباره ساختار شبکه و وضعیت متغیر نودها در شبکه می‌باشد. شناسایی منشاء و بازسازی مسیر انتشار در حوزه‌های مختلف علم کاربرد دارد، از قبیل انتشار اطلاعات، انتشار شایعه، انتشار همه‌گیری، انتشار خطا، انتشار ویروس‌واره و کرم‌واره. یکی از کاربردهای مهم شناسایی منشاء و بازسازی مسیر انتشار ویروس‌واره، کرم‌واره و اطلاعات در مبحث فورنسیک دیجیتال می‌باشد.

شناسایی منشاء و بازسازی مسیر انتشار کرم‌واره به مامور پی‌جویی این امکان را می‌دهد که بداند حمله چگونه در شبکه بوقوع پیوسته است و سرنخهایی برای پی‌جوییهای بعدی و پیگرد مظنونین در اختیار وی قرار می‌دهد.

نودهایی که به شیوه احتمالی به عنوان منشاء شناسایی شده‌اند، مورد پی‌جویی بیشتر قرار می‌گیرند. جمع‌آوری شواهد بیشتر از روی منشاء توسط مامور پی‌جویی، کمک به شناسایی مجرمین و عوامل انتشار کرم‌واره در شبکه می‌نماید. همچنین تحلیل‌گران امنیتی می‌توانند آسیب‌پذیری‌های امنیتی

بهره‌برداری شده بر روی کامپیوترها را با شناسایی منشاء و مسیر احتمالی کشف نمایند. شناسایی منشاء و بازسازی مسیر انتشار کرم‌واره از اهمیت بسزایی برخوردار است، مثلاً در حمله استاکس‌نت، شناسایی کامپیوتری که مسئول ورود این کرم‌واره به کشور و سازمانهای حیاتی بوده، به ماموران پی‌جویی در پیگرد مظنونین و شناسایی آسیب‌پذیریها و جلوگیری از انتشار بیشتر کرم‌واره کمک می‌نموده است.

با توجه به اهمیت مبحث شناسایی منشاء و بازسازی مسیر انتشار کرم‌واره و محدودیتهای موجود در روشهای مطرح، در این رساله بر آنیم تا مسئله شناسایی منشاء و بازسازی مسیر انتشار کرم‌واره پویشگر ترجیحی را حل نماییم به گونه‌ای که محدودیتهای روشهای فعلی را کاهش دهیم. برای اینکار تابع توزیع احتمالی را شکل می‌دهیم تا احتمال آلودگی نودها به کرم‌واره پویشگر ترجیحی را در هر زمان تخمین بزنند. همچنین بر مبنای این تابع توزیع، الگوریتمی را شکل می‌دهیم که منشاء انتشار کرم‌واره را شناسایی می‌نماید و مسیر انتشار کرم‌واره را بازسازی می‌کند.

ایده ما در حل این مسئله، استفاده از مدلسازی حرکت رو-به-عقب زمانی و همچنین رویکرد تکمیل مرحله به مرحله‌ای می‌باشد تا پس از وقوع حمله کرم‌واره با داشتن اطلاعات جمع‌آوری شده در سطح شبکه، منشاء را شناسایی نموده و مسیر انتشار کرم‌واره را بازسازی نماییم.

در رویکرد تکمیل مرحله به مرحله‌ای، در ابتدا مدلی اولیه از حرکت رو-به-عقب شکل می‌گیرد و در هر مرحله بعدی تکمیل می‌شود تا بتوانیم بر اساس مدل تکمیل شده منشاء انتشار کرم‌واره را شناسایی کرده و مسیر آن را بازسازی نماییم. هدف اصلی، داشتن مدلی برای تعیین احتمال آلودگی نود به کرم‌واره پویشگر ترجیحی در هر زمان قبلی از انتشار کرم‌واره است. برای نیل به این هدف، مدل در چهار مرحله تکمیل می‌شود. در مرحله اول، مدلی برای تخمین تعداد نودهای آلوده به کرم‌واره پویشگر تصادفی در هر زمان (مدل تصادفی) شکل می‌گیرد. در مرحله دوم، به کمک مدل شکل گرفته در مرحله اول، مدلی برای تخمین احتمال آلودگی نودها به کرم‌واره پویشگر تصادفی در هر زمان شکل می‌گیرد. در مرحله سوم، به کمک الگوریتم EDA، یک مدل تصادفی را برای تخمین رو-به-عقب احتمال آلودگی نودها به کرم واره پویشگر ترجیحی می‌سازیم. این مدل با مدل تصادفی مرحله یک ترکیب می شود. ایجاد مدلهای اولیه مراحل یک و دو، کمک به تکمیل و اصلاح مدل ایجاد شده در مرحله سوم به کمک الگوریتمهای یادگیری می‌نماید. در مرحله چهارم، الگوریتمی برای بازسازی مسیر انتشار کرم‌واره بر مبنای مدل شکل گرفته در مرحله سه ایجاد می‌گردد.

دستاورد اصلی این رساله، ایجاد مدلهای تصادفی رو-به-عقب به منظور تخمین احتمال آلودگی نودها به کرم‌واره‌های پویشگر تصادفی و ترجیحی در هر زمان و ارائه الگوریتمهایی برای شناسایی منشاء و بازسازی مسیر آنها می‌باشد.

این مقاله در ۴ بخش ارائه می‌گردد. در بخش ۲ کارهای مرتبط تشریح می شود. در بخش ۳ به صورت خلاصه رویکرد چهار مرحله‌ای اتخاذ شده در رساله تشریح می‌شود. در بخش چهارم نتیجه‌گیری انجام می‌گیرد.

۲- کارهای مرتبط [*]

در سال‌های اخیر، محققان مجموعه‌ای از روش‌ها را برای شناسایی منشاء انتشار اطلاعات در شبکه‌ها معرفی نموده‌اند. تعداد محدودی از این روش‌ها برای شناسایی منشاء انتشار کرم‌واره مطرح گردیده‌اند و سایر روش‌ها برای شناسایی منشاء انتشار اطلاعات، شایعه، ویروس‌واره‌ها و خطا کاربرد دارند. برخی از این روش‌ها در سطح IP و بسته، منشاء را شناسایی می‌نمایند و بسیاری از روش‌ها منشاء را درسطح کاربرد و ساختار فیزیکی شبکه شناسایی می‌کنند. این روشها به سمت عقب مدلسازی می‌نمایند.

اولین تحقیق انجام شده در زمینه شناسایی منشاء و بازسازی مسیر انتشار کرم‌واره در سال ۲۰۰۵ در دانشگاه کارنگی ملون صورت گرفت[۲]. در این تحقیق، رویکرد رو به عقبی پیشنهاد شد و فرض گردید که کل اتصالات شبکه ذخیره‌سازی می‌شود. از سال ۲۰۱۰ تاکنون، مجموعه‌ای از روشها به منظور شناسایی منشاء اطلاعات و بازسازی مسیر آن پیشنهاد گردیده‌اند[۹-۳، ۲، ۱۴-۱۰]. اهم این روشها مبتنی بر فرضیات ذیل می‌باشند: گراف اتصالات یا گراف آلودگی داده شده است، وضعیت تمام یا بخشی از نودها در یک زمان از انتشار اطلاعات داده شده است. این روشها تلاش می‌نمایند که مسئله شناسایی منشاء را بر روی گراف حل نمایند. اهم این روشها با داشتن توپولوژی [†] شبکه و نودهای آلوده، نود منشاء را بر اساس معیارهای مختلفی بر روی گراف حل می‌نمایند، از قبیل مرکزیت شایعه، مرکزیت گراف و ... . این روشها علاوه بر داشتن گراف اتصالات یا گراف آلودگی بر مبنای فرضیات، به سه دسته ذیل تقسیم می‌گردند:

- تصویر لحظه‌ای از وضعیت کلیه نودها موجود است.
- تصویر لحظه‌ای از وضعیت برخی از نودها موجود است.
- بر روی شبکه سنسورهایی نصب گردیده که وضعیت آلودگی آنها به طور کامل ثبت می‌گردد.

دسته دیگری از روشهای مطرح برای بازسازی مسیر انتشار کرم‌واره از روشهای بازسازی گراف استفاده می‌نمایند. از جمله فرضیات مهم این روشها وجود اطلاعات یک یا چند شار می‌باشد[۱۵].

دسته دیگری از روشها فرض می‌نمایند که مشاهدات دارای نویز می‌باشد و سعی در شناسایی منشاء اطلاعات با این داده‌های دارای نویز می‌نمایند [۱۷-۱۶].

از سال ۲۰۱۵ تا کنون، چندین روش پیشنهاد گردیده‌اند که برخی از آنها رویکرد متفاوتی را برای شناسایی منشاء بکار برده‌اند، از قبیل استخراج ویژگی [۲۰-۱۸] .

نکات ذیل درخصوص روشهای موجود مطرح می‌باشد. اولاً، روشهای فعلی عملیاتی نمی‌باشند، زیرا که اغلب نرخ خطای بالایی دارند. دوماً، اکثر روشهای فعلی برای پیدا کردن منشاء انتشار بسیار زمانبر می‌باشند. همچنین مسائل باز متعددی در این روشها مطرح است، از قبیل دیدگاه روشها از نظر توپولوژی، تعداد منشاء، تعداد شبکه‌ها، پویایی زمانی و پیچیدگی و مقیاس‌پذیری[۲۱].

---

[*] exploit
[†] topology
[‡] cascade
[§] Temporal dynamics

۳- رویکرد چهار مرحله‌ای

ایده ما در حل این مسئله، استفاده از مدلسازی حرکت رو-به-عقب زمانی و همچنین رویکرد تکمیل مرحله به مرحله‌ای می‌باشد تا پس از وقوع حمله کرم‌واره با داشتن اطلاعات جمع‌آوری شده در سطح شبکه، منشاء را شناسایی نموده و مسیر انتشار کرم‌واره را بازسازی نماییم. هدف، کاهش محدودیتهای روشهای موجود از قبیل پیچیدگی محاسباتی و کاهش حجم ذخیره‌سازی می‌باشد.

در رویکرد تکمیل مرحله به مرحله‌ای، در ابتدا مدلی اولیه از حرکت رو-به-عقب شکل می‌گیرد و در هر مرحله بعدی تکمیل می‌شود تا بتوانیم بر اساس مدل تکمیل شده منشاء کرم‌واره را شناسایی کرده و مسیر آن را باز سازی نماییم. هدف اصلی، داشتن مدلی برای تعیین احتمال آلودگی نود به کرم‌واره پویشگر ترجیحی در هر زمان قبلی از انتشار کرم‌واره است. برای نیل به این هدف، مدل در چهار مرحله تکمیل می‌شود. در مرحله اول، مدلی برای تخمین تعداد نودهای آلوده به کرم‌واره پویشگر تصادفی در هر زمان (مدل تصادفی) شکل می‌گیرد. در مرحله دوم، به کمک مدل شکل گرفته در مرحله اول، مدلی برای تخمین احتمال آلودگی نودها به کرم‌واره پویشگر تصادفی در هر زمان شکل می‌گیرد. در مرحله سوم، به کمک الگوریتم EDA، یک مدل تصادفی را برای تخمین رو-به-عقب احتمال آلودگی نودها به کرم واره پویشگر ترجیحی می‌سازیم. این مدل با مدل تصادفی مرحله یک ترکیب می شود. ایجاد مدلهای اولیه مراحل یک و دو، کمک به تکمیل و اصلاح مدل ایجاد شده در مرحله سوم به کمک الگوریتمهای یادگیری می‌نماید. در مرحله چهارم، الگوریتمی برای بازسازی مسیر انتشار کرم‌واره بر مبنای مدل شکل گرفته در مرحله سه ایجاد می‌گردد.

۳-۱- مرحله یک- مدل‌سازی انتشار کرم‌واره رو-به-عقب

در این مرحله مسئله پیشگویی تعداد نودهای منشاء آلودگی و همچنین تعداد نودهای آلوده در هر مرحله از انتشار کرم‌واره پویشگر تصادفی را در نظر می‌گیریم و بر این اساس سه مدل مکمل پیشنهاد می‌نماییم[۲۳-۲۲]:

- مدل قطعی رو-به-عقب
- مدل تصادفی رو-به-عقب و
- مدل تصادفی مارکوفی رو-به-عقب برای تعداد نودهای آلوده

در مدلهای رو-به-عقب، زمان رو به عقب اجرا می‌گردد و مدلسازی در خلاف جهت انتشار کرم‌واره انجام می‌گیرد. در این مرحله ما فرض می‌نماییم که اطلاعات ذیل را در اختیار داریم:

- پارامترهای انتشار آلودگی با مدل SIR
- تعداد نودهای آسیب‌پذیر، آلوده و بازیابی شده در یک زمان از انتشار کرم‌واره

در این مرحله ابتدا مدل قطعی رو-به-عقب را توسعه می‌دهیم. این مدل، مدل جدید SIR است که در آن پارامتر جدیدی به نام نرخ susceptibility[§] را تعریف می‌نماییم. در این مدل تعداد نودهای آسیب‌پذیر، آلوده و بازیابی شده را رو به عقب تعیین می‌نماییم. سپس مدل تصادفی رو-به-عقب را پیشنهاد می‌نماییم. در این مدل فرض می‌کنیم که وضعیت اولیه سیستم (تعداد نودهای آسیب‌پذیر، آلوده و بازیابی شده) در مدل رو-به-عقب داده شده است و ما وضعیت نهایی سیستم در مدل رو-به-عقب را تخمین می‌زنیم (وضعیت اولیه در انتشار کرم‌واره). در این مدل ، پارامتر فشار اتهام را تعریف می‌نماییم تا به صورت احتمالی تعداد نودهای متهم که در ابتدا آلوده بوده‌اند را مشخص کنیم. در نهایت، مدل تصادفی مارکوفی رو-به-عقب بر مبنای زنجیره زمان پیوسته مارکوف تعریف می‌گردد. فرایند رو-به-عقب خاصیت مارکوفی دارد، زیراکه وضعیت سیستم در زمان گذشته هیچ تاثیری بر وضعیت سیستم در زمان آینده ندارد، اگر وضعیت سیستم در زمان حال مشخص باشد. در این مدل فرض می‌نماییم تعداد نودهای آسیب‌پذیر، آلوده و بازیابی شده در یک زمان دلخواه از مدل رو-به-عقب داده شده است و مدل قطعی رو-به-عقب را بر مبنای زنجیره مارکوف زمان پیوسته بیان می‌نماییم و در هر زمان از انتشار کرم‌واره تعداد نودهای آلوده تخمین زده می‌شود.

مهاجمان ممکن است از چندین میزبان آلوده اولیه برای سرعت بخشیدن به انتشار کرم‌واره استفاده نمایند. مدل تصادفی رو-به-عقب به سمت عقب تا زمان تعادل حرکت می‌کند و در این زمان مشخص می‌شود که چند نود منشاء اولیه وجود داشته‌اند. در مدل تصادفی رو-به-عقب، احتمال تعداد نودهای آلوده اولیه تعیین می‌گردد، که مربوط به زمان تعادل در انتشار کرم‌واره می‌باشد. در مدل مارکوفی رو-به-عقب تعداد نودهای آلوده در هر زمان از انتشار کرم‌واره رو به عقب به صورت احتمالی مشخص می‌گردد.

در این مدل، نیاز است که تعداد نودهای آسیب‌پذیر، آلوده و بازیابی شده را در یک زمان از انتشار کرم‌واره داشته باشیم.

تابع توزیع احتمالی تعداد نودهای آلوده در هر زمان از انتشار کرم‌واره پویشگر تصادفی، در مراحل بعدی رساله مورد استفاده قرار می‌گیرد و تکمیل می‌گردد تا منشاء و مسیر انتشار کرم‌واره به صورت احتمالی شکل بگیرد. ما به این سوالات می‌پردازیم: چند نود منشاء در انتشار کرم‌واره وجود دارند؟ چند نود آلوده در هر زمان از انتشار کرم‌واره وجود دارند؟

۳-۲- مرحله دو- توزیع احتمال آلودگی کرم‌واره با مدل رو-به-عقب *

در این مرحله، مسئله تخمین احتمال آلودگی نودها به سمت عقب در هر زمان از انتشار کرم‌واره را با شبکه بیزین بررسی می‌نماییم. در مرحله قبل تعداد نودهای آلوده را به سمت عقب از طریق سه مدل نشان دادیم.

در این مرحله احتمال آلودگی نودها برای انتشار کرم‌واره پویشگر تصادفی با شبکه‌های بیزین نشان داده می‌شود و نشان می‌دهیم که احتمال آلودگی نود در هر زمان به درجه نود و تعداد نودهای آلوده در آن زمان بستگی دارد (شکل ۱). در این مرحله فرض می‌گردد که دانش قبلی درباره پارامترهای آلودگی کرم‌واره موجود است. همچنین تعداد نودهای آسیب‌پذیر، آلوده و بازیابی شده در یک زمان از انتشار کرم‌واره داده شده است. Out_degree[§] نود در هر زمان برای تخمین احتمال آلودگی در آن زمان مورد نیاز می‌باشد.

---

* Deterministic Back-to-Origin model

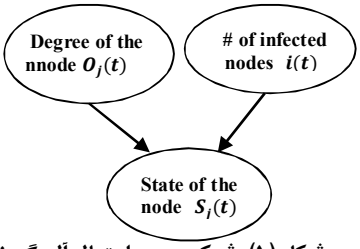
backwards†

Out_degree‡

§ تعداد کل جریانهای خروجی از e بین زمان t و $t + \Delta t$

§

همچنین فرض می‌شـــود که مدل توزیع درجه نودها بر روی زمان برمبنای تاریخچه‌ای*** از مشاهدات درجه نودها در شبکه در زمانیکه انتشار کرم‌واره اتفاق افتاده، یادگرفته شده است. ما از شبیه سازی برای بررسی دقت توزیع احتمال شکل‌گرفته، استفاده نمودیم. نتایج شبیه سازی نشان می‌دهد که تابع توزیع احتمال، احتمال آلودگی نودها را در زمانهای قبل با دقت بالایی تخمین می‌زند. از این روش برای استنتاج مسیر انتشار کرم‌واره می‌تواند استفاده شود.



شکل (۱): شبکه بیزین احتمال آلودگی نود

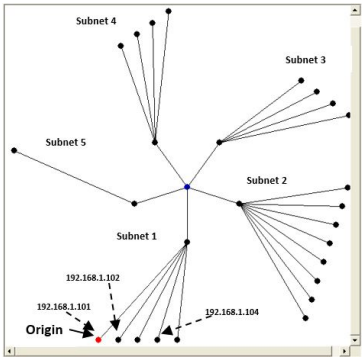
۳-۳- تهیه مجموعه داده و استخراج ویژگیها

به منظور استخراج ویژگیها از ترافیک شبکه، مجموعه داده‌ای مورد نیاز می‌باشـــد که حاوی ترافیک پس‌زمینه نرمال و ترافیک کرم‌واره به صـــورت همزمان باشـــد. هدف ما اســـتفاده از کرم‌واره پویشـــگر ترجیحی برای ایجاد مجموعه داده‌ای اســـت تا در یادگیری مدل بکار گرفته شـــود. همانگونه که در [۱] گفته شـــد، این مجموعه داده متاســـفانه در اختیار عموم نیســـت. بنابراین مجموعه داده‌ای را با ویژگیهای مورد نیاز ایجاد نمودیم.

کرم‌واره پویشگر ترجیحی با یک یا چند نود منشاء آلوده آغاز می‌نماید که در ابتدا کد بدنهاد را دارند. کد بدنهاد از مبدا به همسایگانش در همان زیرشبکه با احتمال بالاتری منتقل می‌گردد و به ســـایر نودها در ســـایر زیرشـــبکه‌ها با احتمال پایین‌تری منتقل می‌شـــود. ما مجموعه داده را به روش ذیل تولید نمودیم:

- مجموعه داده ISCX [۲۰]: این مجموعه داده‌ایدر یک پیاده‌ســـازی فیزیکی با استفاده از دستگاههای واقعی ایجاد شده است. در این رساله، زیرمجموعـه‌ای از دنبالـه نرمـال این مجموعـه داده (روز ۱) کـه در مجموعه‌های داده‌ای آموزش و تست آن آمده، را در نظر گرفته‌ایم (۴۰ ثانیه، ۵۰۰۰۰ میلی ثانیه). بستر تولید ترافیک شامل ۲۲ ایستگاه کاری اســـت که در ۵ زیرشـــبکه قرار گرفته‌اند و در محدوده آدرســـی ۱۹۲,۱۶۸,۱,۰/۲۴ تا ۱۹۲,۱۶۸,۵,۰/۲۴ قرار دارند.
- GTNetS [۲۱]:GTNetS امکان شـــبیه‌ســـازی در مقیاس بزرگ در سطح بسته را فراهم می‌نماید. GTNetS امکان مطالعه رفتار کرم‌واره اینترنتی را در شـــرایط مختلفی فراهم می‌نماید و امکان جمع‌آوری بسته‌های در سطح شبکه را فراهم می‌کند. ما شبیه سازیهایی را برای کرم‌واره TCP پویشـــگر ترجیحی در GTNetS ایجاد نمودیم. در GTNetS نودها فقط دو حالت دارند: آسیب‌پذیر و آلوده (مدل SI). در این رساله، نودها سه حالت دارند: آسیب‌پذیر، آلوده و بازیابی شده (مدل SIR)، بنابراین ما کد GTNetS را تغییر دادیم تا حاوی وضـــعیت R نیز بگردد. (جزئیات و نحوه تغییر کد GTNetS به منظور افزودن وضـــعیت بازیابی شـــده در [۱] آمده اســـت. معماری شـــبیه‌ســـازی در GTNetS در شکل (۲) آمده است.

لیســـت ویژگیهای انتخاب شـــده در رســـاله در جدول (۱) آمده اســـت. این ویژگیها با بررسی و مطالعه تحقیقات مرتبط با تشخیص انتشار کرم‌واره انتخاب گردیده‌اند.



شکل (۲): معماری شبیه‌سازی در GTNetS

۴-۳- مرحله سه- استخراج تابع توزیع احتمال به منظور پی‌جویی زمان آلودگی و بازیابی نودهای آلوده به کرم‌واره پویشگر ترجیحی

هدف از این مرحله، ارائه روشی است که به صورت خودکار، زمانی را که برای نخســـتین بار کرم‌واره شـــروع به آلوده‌ســـازی ماشـــینها در شـــبکه نموده و زمانی که نود بازیابی شـــده را تخمین بزنیم. نودهایی که قبل از ســـایر نودها آلوده شده‌اند، منشاءهای احتمالی کرم‌واره می‌باشند. ما ابتدا مدل احتمالی را بر مبنای مقادیر تاریخی ۱۲ ویژگی شـــبکه (جدول (۱)) یاد می‌گیریم و ســـپس وضـــعیت آلودگی هر نود را با داشـــتن مقادیر ویژگی در هر زمان اســـتنتاج می‌نماییم. دوما، ما مدل یاد گرفته شـــده در این بخش را با مدل مارکوفی تصادفی رو-به-عقب (مرحله ۱) ترکیب می‌نماییم تا تعداد ویژگیهای مورد نیاز را کاهش دهیم. همچنین فرض می‌گردد که دانش قبلی در مورد اینکه کدام نودها آلوده بوده‌اند و چه زمانی آلوده بوده‌اند، نداریم.

در این مرحله، تابع توزیع احتمالی ساده شکل گرفته در مرحله قبل را به کمک ویژگیهای اســـتخراج شـــده در این بخش و با رویکرد الگوریتم EDA تکمیل می‌نماییم. ابتدا تابع توزیع احتمالی را برای تخمین وضـــعیت آلودگی نودها در هر زمان یاد می‌گیریم، در زمانیکه کرم‌واره پویشـــگر ترجیحی منتشـــر می‌گردد و اینکار را با داشتن مقادیر تاریخچه‌ای از ویژگیهای استخراج شده از لایه IP و کاربرد انجام می‌دهیم. سپس به کمک تابع احتمالی، درباره وضعیت آلودگی نودها، با داشتن مقادیر ویژگیها بر حسب زمان استنتاج می‌نماییم.

ما روش چهار مرحله‌ای را برای توسعه تابع احتمالی پیشنهاد می‌نماییم تا زمان آلودگی و بازیابی هر نود را به صـــورت احتمالی تخمین بزنیم: اســـتخراج ویژگیها (بخش ۳-۲)، ایجاد مدل اولیه، یادگیری مدل احتمالی با رویکرد EDA، اســـتنتاج احتمال آلودگی نودها با داشـــتن مقادیر ویژگیها در هر زمان. سپس مدل توسعه یافته در این مرحله (شکل (۳)) را با مدل تصادفی مارکوفی ترکیب می‌نماییم و مدل دومی برای کرم‌واره پویشگر ترجیحی توسعه داده

***historical

جدول (۱): ویژگیهای انتخاب شده در این رساله

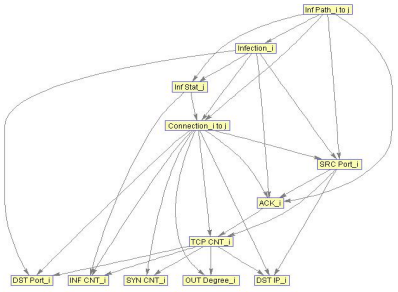
ردیف	ویژگی	شرح
۱	ACK	تعداد بسته‌های ACK دریافت شده توسط میزبان iام در هر میلی‌ثانیه.
۲	$DST\ IP$	تعداد آدرسهای مقصد منحصر به فردی که میزبان iام به آن در هر میلی‌ثانیه متصل می‌گردد.
۳	$DST\ Port$	تعداد پورتهای مقصد منحصر به فردی که میزبان iام در هر میلی‌ثانیه به آن متصل می‌گردند.
۴	$SRC\ Port$	تعداد پورتهای مبدا منحصر به فردی که میزبان iام به آن در هر میلی‌ثانیه متصل می‌گردند.
۵	$SYN\ CNT$	تعداد بسته‌های SYN ارسال شده توسط میزبان iام در هر میلی‌ثانیه
۶	$TCP\ CNT$	تعداد بسته‌های TCP ارسال شده توسط میزبان iام در هر میلی‌ثانیه
۷	$INF\ CNT$	تعداد نودهای آلوده در هر زیرشبکه در هر میلی‌ثانیه
۸	$OUT\ Degree$	تعداد اتصالاتی که میزبان iام در هر میلی‌ثانیه برقرار می‌نماید.
۹	$Infection$	وضعیت میزبان iام در هر میلی‌ثانیه (۱=آلوده، ۰=آسیب‌پذیر یا بازیابی شده)
۱۰	$Inf\ Path\ to\ j$	اگر آلودگی از نود i به نود j منتشر شود یا از نود j به نود i منتشر گردد، مقدار این ویژگی ۱ می‌گردد.
۱۱	$Connection\ to\ j$	اگر جریان شبکه بین نودهای i به j در زمان t وجود داشته باشد، این ویژگی در آن زمان ۱ می‌گردد.
۱۲	$Inf\ Stat\_j$	وضعیت آلودگی سایر نودها در هر زمان

می‌شود[۲۲].

همچنین آزمایشاتی را انجام دادیم تا مدلهای اول و دوم خود را ارزیابی نماییم. برای این منظور مجموعه داده‌ای را برای ساخت مدل و تست آن تولید نمودیم (بخش ۳-۲). آزمایشات نشان می‌دهد که مدلها می‌توانند نود منشاء و زمان آلودگی نودها را با دقت قابل قبولی تخمین بزنند.

رویکرد اتخاذ شده در این رساله رویکرد احتمالاتی می‌باشد. در مقایسه با رو شهای قطعی از قبیل شبکه‌های عصبی، رویکرد احتمالاتی زمانی ا ستفاده می‌شود که عدم قطعیت وجود دارد. بزرگترین مسئله در انتشار آلودگی، عدم قطعیت می‌باشد، بنابراین انتخاب رویکرد احتمالاتی در حل مسئله منشاء انتشار کرم‌واره و بازسازی مسیر با ماهیت انتشار مطابقت دارد.

رویکرد اتخاذ شده در این بخش، رویکرد تکاملی است. در رویکرد تکاملی، در هر مرحله از الگوریتم مجموعه‌ای از راه‌حلها به عنوان نماینده‌ی جمعیت انتخاب می‌گردند و با اجرای مکرر الگوریتم، این جمعیت و مدل تکمیل می‌شود. رویکردهای تکاملی بر مبنای مدلهای احتمالاتی عمل می‌نمایند. الگوریتم انتخاب شده در این مرحله، الگوریتم EDA می‌باشد که جزء دسته الگوریتمهای تکاملی است.



شکل (۳): مدل احتمالی آلودگی کرم‌واره

۵-۳- مرحله ۴- بازسازی مسیر انتشار کرم‌واره پویشگر ترجیحی

هدف از این مرحله، بازسازی مسیر انتشار کرم‌واره پویشگر ترجیحی با استفاده از اطلاعات بدست آمده از مرحله سوم می‌باشد. به منظور نیل به این هدف، الگوریتم توزیع شده‌ای را پیشنهاد می‌نماییم. به کمک آزمایشات تجربی دقت این الگوریتم مورد بررسی قرار گرفته است.

یکی از ویژگیهای قابل استنتاج از مدل ارائه شده در مرحله ۳، احتمال وجود مسیر بین دو نود i و j می‌باشد. بنابراین با استنتاج این ویژگی به کمک سایر ویژگیها، در زمانیکه احتمال وجود یال بین دو نود از مقدار آستانه بگذرد، بین آن دو نود یالی را در نظر می‌گیریم.

در این مرحله [۲۷] اقدام به بازسازی مسیر انتشار کرم‌واره نمودیم و دو الگوریتم را برای این منظور پیشنهاد کردیم. الگوریتم اول از مدل توسعه یافته در گام سه استفاده می‌نماید و دقت خوبی دارد. الگوریتم بازسازی مسیر انتشار کرم‌واره، به صورت توزیع شده و احتمالی عمل می‌نماید و مبتنی بر این فرضیات می‌باشد: جمع‌آوری ۹ ویژگی از شبکه در هر لحظه بر روی هر نود. نیازی به ذخیره‌سازی این ویژگیها در شبکه نیست، این ویژگیها به الگوریتم توزیع شده بازسازی مسیر داده می‌شود و میزان احتمال وجود مسیر بین هر نود با سایر نودهای شبکه در آن لحظه محاسبه می‌گردد. چنانچه این احتمال بالاتر از مقدار آستانه‌ای باشد، به عنوان مسیر احتمالی ثبت می‌گردد. بر اساس محاسبات انجام شده، این الگوریتم دقت خوبی دارد. در مقایسه با سایر روشهای موجود، با کاهش حجم ذخیره‌سازی و سربار محاسباتی توانستیم مسیر انتشار کرم‌واره پویشگر ترجیحی را با دقت خوبی بازسازی نماییم. الگوریتم دوم از ایده درخت پوشای حداکثری در انتشار کرم‌واره استفاده می‌نماید. این الگوریتم بر این فرض استوار است که ۹ ویژگی در هر واحد زمانی در شبکه جمع‌آوری می‌گردد و به الگوریتم داده می‌شود. این الگوریتم نیاز دارد که نود ناظری در شبکه مسیر را بازسازی نماید. بخشی از آن به صورت توزیع شده بر روی هر نود و بخشی از آن به صورت متمرکز بر روی نود ناظر اجرا می‌گردد.

۴- جمع‌بندی و نتیجه‌گیری

در این رساله، منشاء انتشار کرم‌واره پویشگر ترجیحی را به صورت احتمالی شناسایی می‌نماییم و مسیر انتشار آن را به کمک مدلهای احتمالی بازسازی می‌کنیم. شناسایی منشاء و بازسازی مسیر انتشار کرم‌واره به مامور پی‌جویی کمک می‌نماید تا مظنونین احتمالی را حدس بزند و آسیب‌پذیریهای موجود بر روی کامپیوترهای مظنون را شناسایی نماید و نقطه ورود کرم‌واره به شبکه را تشخیص دهد. رویکرد اتخاذ شده در این زمینه، رویکرد تکمیلی است و مدل احتمالی را به صورت تکمیلی توسعه دادیم، تا اهداف پایان‌نامه محقق گردند. برای نیل به این اهداف رویکرد چهار مرحله‌ای را پیشنهاد نمودیم.

به طور کلی در این چهار گام با رویکرد تکمیلی توانستیم مدل احتمالی را توسعه دهیم که منشاء انتشار کرم‌واره پویشگر ترجیحی و پویشگر تصادفی را به شیوه احتمالی معین می‌نماید و مسیر انتشار کرم‌واره پویشگر ترجیحی را باز سازی می‌کند. برای استفاده از این مدل نیاز داریم که پارامترهای این مدل را در شبکه هدف یاد بگیریم. چنانچه بخواهیم مدل یاد گرفته شده را در شبکه دیگری استفاده نماییم، باید از مدل یاد گرفته شده به عنوان یک مدل اولیه

استفاده کنیم و پارامترهای مدل را در شبکه جدید به تدریج بروزرسانی نماییم. برای این منظور، در رویکرد پیشنهادی، نگاشتی را بین مقادیر ویژگیها در زمان انتشار کرم‌واره در دو شبکه (شبکه اولیه یاد گرفته شده و شبکه جدید) انجام دادیم.

مطابق رویکرد پیشنهادی در این رساله، عاملهایی باید بر روی ماشینهای شبکه مورد بررسی نصب گردند. این عاملها امکان مانیتورینگ ماشینهای شبکه مورد بررسی را فراهم می‌نمایند و مقادیر ۱۱ ویژگی را از سرآیند ترافیک شبکه استخراج نموده و به الگوریتم توزیع شده پیشنهادی می‌دهند. این الگوریتم امکان بازسازی مسیر و شناسایی منشاء انتشار کرم‌واره پویشگر ترجیحی را فراهم می‌نماید. روش پیشنهادی انتهایی برای شناسایی منشاء و بازسازی مسیر از لحاظ مقدار محاسبات و تعداد ارتباطات از مرتبه چند جمله‌ای و کارا بوده، و کاملا عملیاتی می‌باشد و قابل بکارگیری است.

سپاسگزاری

از استاد ارجمندم جناب آقای دکتر صادقیان بسیار سپاسگزارم که در راه کسب علم و معرفت مرا یاری نمودند و بدون راهنماییهای ایشان تامین این پایان‌نامه بسیار دشوار می‌نمود.

مراجع

[1] طلا تفضلی ، *شناسایی منشاء انتشار کرم‌واره و بازسازی مسیر آن به شیوه احتمالی*، رساله دکتری، دانشکده مهندسی کامپیوتر و فناوری اطلاعات، دانشگاه صنعتی امیرکبیر، مهر ۱۳۹۶.

[2] Y. Xie, V. Sekar, D. A. Maltz, M. K. Reiter, H. Zhang," Worm Origin Identification Using Random Moonwalks", *IEEE Symposium on Security and Privacy*, p.p. 242--256, 2005, doi: 10.1109/SP.2005.23..

[3] Xiang Y., Li Q., "Online tracing scanning worm with sliding window", *Lecture notes in computer science*, vol 4990, p.p. 482-496, 2007.

[4] Xiang Y, Li Q., Guo D., "Online accumulation: reconstruction of worm propagation path", *Lecture notes in computer science*, vol. 5245, p.p. 162-172, 2008.

[5] Shi W., Li Q., Kang J., Guo D., "Reconstruction of worm propagation path by causality", *Proceedings of IEEE international conference on networking, architecture and storage*, 2009.

[6] Shah D., Zaman T., "Detecting sources of computer viruses in networks: theory and experiment", *Proceedings of 10th international conference on Measurement and modeling of computer systems ACM SIGMETRICS*, p.p. 203--214, 2010, doi: 10.1145/1811039.1811063.

[7] Dong W., Zhang W., Tan C. W. ,"Rooting out the rumor culprit from suspects", *Proceedings of the IEEE International Symposium on Information Theory (ISIT)*, 2013, doi: 10.1109/ISIT.2013.6620711.

[8] Luo W., Peng Tay W., Leng M., "Identifying infection sources and regions in large networks", *IEEE*, 61(11), p.p. 2850--2865, 2013, doi:10.1109/TSP.2013.2256902.

[9] Prakash B. A., Vreeken J., Faloutsos C., "Spotting Culprits in Epidemics: How Many and Which Ones?", *12th IEEE International Conference on Data Mining*, p.p.11-20, 2012, doi:10.1109/ICDM.2012.136.

[10] Nguyen H. T., Ghosh P., Mayo M. L., Dinh T. N., "Multiple information sources identification with provable guarantees", *Proceedings of the 25th ACM international conference on Information and knowledge management*, p.p. 1663-1672, 2016.

[11] Sefer E., Kingsford C., "Diffusion archaeology for diffusion progression history reconstruction", *Journal of knowledge and information system*, 49(2), p.p. 403-427, 2016.

[12] Shah D., Zaman T., "Finding rumor sources on random tree", *Operations research*, 64(3), p.p. 736-755, 2015.

[13] Meriom E. A., Milling C., Caramanis C., Mannor S., Orda A., Shakkottai S., "Localized epidemic detection in network with overwhelming noise", *Proceedings of the 2015 international conference on measurement and modeling of computer systems*, p.p. 441-442, 2015.

[14] Milling C., Caramanis C., Mannor S., Orda A., Shakkottai S., "Detecting epidemics using highly noisy data", *Proceedings of the 14th international symposium on mobile and ad hoc networking and computing*, p.p. 177-186,2013.

[15] Feizi S., Medard M., Quon G., Kellis M., Duffy K., "Network infusion to infer information sources in networks", arXiv: 1606.07383, 2016.

[16] Meriom E. A., Milling C., Caramanis C., Mannor S., Orda A., Shakkottai S., "Localized epidemic detection in network with overwhelming noise", *Proceedings of the 2015 international conference on measurement and modeling of computer systems*, p.p. 441-442, 2015.

[17] Milling C., Caramanis C., Mannor S., Orda A., Shakkottai S., "Detecting epidemics using highly noisy data", *Proceedings of the 14th international symposium on mobile and ad hoc networking and computing*, p.p. 177-186,2013.

[18] Farajtabar M., Rodriguez M.G., Zamani M., Du N., Zha H., Song L., "Back to the Past: Source Identification in Diffusion Networks from Partially Observed Cascades", *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics*, p.p. 232-240, 2015.

[19] Rozenshtein P., Gionis A., Prakash B. A., Vreeken J., "Reconstructing an Epidemic Over Time", *22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, p.p. 1835-1844, 2016, doi:10.1145/2939672.2939865.

[20] Kang C., Park N., Prakash B. A., Serra E., Subrahmanian V. S., "Ensemble Models for Data-driven Prediction of Malware Infections", *9th ACM International Conference on Web Search and Data Mining*, p.p. 583-592, 2010, doi:10.1145/2835776.2835834.

[21] Jiang J., Wen S., Yu S., Xiang Y., Zhou W., "Identifying propagation sources in networks: State-of-the-art and comparative studies", *IEEE Communications Surveys and Tutorials*, 19(1), p.p. 465-481, 2014, doi:10.1109/COMST.2016.2615098.

[22] Tafazzoli T., Sadeghiyan B., "A stochastic model for the size of worm origin", ICGIP 2015, winner of the session.

[23] Tafazzoli T., Sadeghiyan B.,"A stochastic model for the size of worm origin", *Security and Communication Networks*, 9(10), p.p. 1103-1118, 2016. doi:10.1002/sec.1403.

[24] A. Shiravi, H. Shiravi, M. Tavallaee, A. A. Ghorbani, Toward developing a systematic approach to generate benchmark datasets for intrusion detection, *Computers and Security*, 31(3), p.p. 357-374, (2012), doi:http://dx.doi.org/10.1016/j.cose.2011.12.012.

[25] Riley G.F., Using the Georgia Tech Network Simulator, Technical Report, 2006.

[26] Tafazzoli T., Sadeghiyan B., Probability distribution function for investigating node infection and removal times, submitted to elsevier *Performance Evaluation*.

[27] Tafazzoli T., Sadeghiyan B., Probabilistic identification of origin and reconstruction of propagation path for preferential scanning worm, To be submitted.