نشریه علمی "پدافند الکترونیکی و سایبری"
سال نهم، شماره ۲، تابستان ۱۴۰۰، ص ۷-۱

علمی – پژوهشی

بهبود امنیت شبکه‌های سایبری مبتنی بر شناسایی انجمن‌ها با استفاده از الگوریتم خوشه‌بندی طیفی

مرتضی جویبان[۱]، سوده حسینی[۲*]

۱- دانشجوی دکتری ، دانشگاه شهید باهنر کرمان،کرمان، ایران، ۲- دانشیار، دانشگاه شهید باهنر کرمان،کرمان، ایران،

۲- (دریافت: ۱۴۰۲/۱۱/۲۵، بازنگری: ۱۴۰۳/۰۱/۲۳، پذیرش: ۱۴۰۳/۰۲/۱۹، انتشار: ۱۴۰۳/۰۳/۱۳)

۳- DOR: https://dorl.net/dor/00000000000000000000

چکیده

شبکه‌های سایبری باتوجه‌به ساختار و نحوه ارتباطات درون شبکه، نوعی از شبکه‌های پیچیده و بی‌مقیاس در نظر گرفته می‌شوند. تشخیص انجمن‌ها از مهم‌ترین روش‌های تجزیه‌وتحلیل شبکه به‌منظور درک ساختار و روابط میان اعضای شبکه است. با توسعه شبکه‌های سایبری، چالش‌های جدیدی از نظر امنیت اطلاعات برای کاربران ایجاد شده است.

یکی از اهداف شناسایی انجمن‌ها در شبکه‌های سایبری، جلوگیری از انتشار بدافزارها و حملات سایبری است. برای این منظور، جهت جلوگیری و مقابله با حمله و نفوذ به شبکه می‌بایست انجمن‌های موجود در شبکه شناسایی شده تا با ایمن‌سازی و احیاء انجمن‌ها و همچنین پیاده‌سازی سیاست‌های دفاعی متناسب با هر انجمن، آسیب‌ها و حملات صورت‌گرفته توسط مهاجمین به میزان قابل‌توجهی کاهش یابد. در این مقاله روشی برای تشخیص انجمن‌های شبکه‌های سایبری توسط الگوریتم خوشه‌بندی طیفی ارائه شده است. همچنین با استفاده از خاصیت ماتریس لاپلاس نرمال شده در این الگوریتم، می‌توان تعداد انجمن‌های مناسب شبکه سایبری را پیش‌بینی نمود. به‌منظور ارزیابی فرایند تشخیص، از دو معیار ضریب نیم‌رخ و نمایه ژاکارد استفاده می‌شود. نتایج به‌دست‌آمده از معیارهای ارزیابی، کارایی روش پیشنهادی را تأیید می‌نماید.

کلیدواژه‌ها: شبکه‌های سایبری، امنیت اطلاعات، تحلیل شبکه، تشخیص انجمن، خوشه‌بندی طیفی.

۱- مقدمه

اینترنت یک شبکه جهانی است که اشخاص و سازمان‌ها می‌توانند در هر زمان و به هر مکانی از طریق این شبکه متصل شوند. یک شبکه سایبری[۲] می‌تواند متشکل از اینترنت جهانی، دستگاه‌های الکترونیکی هوشمند، سرورها، سوئیچ‌ها و اتصالات شبکه و سایر دستگاه‌ها باشد. کلمه سایبر به هر ارتباطی با فناوری اطلاعات اشاره دارد و به تمام جنبه‌های محاسبات مانند ذخیره، محافظت، دسترسی، پردازش و انتقال داده‌ها مربوط می‌شود. اشکال مختلفی از شبکه‌های سایبری در جهان وجود دارد و هرکدام از روش‌های مختلفی برای انتقال داده‌ها استفاده می‌کنند[1]. سازمان‌های جهانی دارای شبکه‌های سایبری بزرگی هستند که برای ذخیره و انتقال داده‌ها مورد استفاده قرار می‌گیرند. بیشتر این داده‌ها محرمانه است که شامل اطلاعات شخصی مشتریان و کارمندان، اسناد تجاری محرمانه و یا مالکیت معنوی سازمان‌ها می‌باشند. در حالی که اتصال بالای شبکه‌ها و دستگاه‌های سایبری مدرن باعث پیشرفت‌های زیادی در عملکرد و کارایی سیستم‌های شبکه‌ای، فناوری‌های ارتباطی و توسعه سیستم‌های اطلاعاتی شده و تمامی افراد جامعه و سازمان‌ها را درگیر ساخته است، اما از طرف دیگر منجر به افزایش آسیب‌پذیری آن‌ها در برابر حمله یا نفوذ شده است. فضای سایبری در برابر بسیاری از حملات و خطرات آسیب‌پذیر بوده و امنیت اطلاعات اصلی‌ترین چالش در این شبکه‌ها می‌باشد. حملات سایبری[۳] می‌تواند از یک حمله جزئی به یک فرد تا حملات بزرگ به سازمان‌ها بوده که منجر به خسارات بزرگ

* رایانامه نویسنده مسئول: so_hosseini@uk.ac.ir

[2] Cyber Network

[2] Cyber Attacks

مـی‌شـوند. ایـن حمـلات بـرای انتشـار اطلاعـات نادرسـت، از کارانداختن خدمات تـاکتیکی، دسترسـی بـه اطلاعـات حسـاس، جاسوسی، سرقت داده‌ها و خسارات مـالی اسـتفاده مـی‌شـود. در صورتی که هدف حملات دولت یا زیرساخت‌هـای حیـاتی باشـد، می‌تواند به‌عنوان یک موضوع امنیت ملی مطرح گردد[2]. به‌همین دلیل، تشخیص حملات و شناسایی آن‌ها قبل از این‌که آسیبی به شبکه وارد نمایند بسیار مهم است. تشخیص سریع حملات کمک قابل توجهی در کنترل آسیب‌هـای ناشـی از حمـلات بـه شـبکه می‌کند.

اخیراً امنیت شبکه‌هـای سـایبری بـه موضـوع بسـیار مهمـی تبدیل شده است. امنیـت سـایبری بـه مجموعـه‌ای از روش‌هـا و فرآیندهایی اشاره دارد که هدف آن‌ها محافظـت از سیسـتم‌هـای کامپیوتری، شبکه‌ها و داده‌ها از نظر نرم‌افزاری و سخت‌افـزاری در برابر حملات سایبری می‌باشد. در حالی که امنیت شبکه به‌عنوان زیرمجموعه‌ای از حوزه امنیـت سـایبری بـه محافظـت از داده‌هـا هنگام عبور و تبادل در شـبکه مـی‌پـردازد. تحلیـل‌گـران امنیـت سایبری همواره باید تعداد زیادی از بسته‌هایی که در شبکه تبادل می‌شوند را به‌منظـور تجزیـه و تحلیـل شـبکه‌هـای بـزرگ بـرای تشخیص حملات سایبری مورد بررسـی قـرار دهنـد[3]. افـزایش مقیاس شبکه‌های سایبری، حجم داده‌ها و ترافیک سـایبری بـین میلیاردها آدرس پروتکل اینترنتی[1] و همچنین افزایش پیچیـدگی حملات سـایبری، چـالش‌هـای متعـددی را در تشـخیص نفـوذ[2] به‌وجود آورده است. مسـأله اصـلی شناسـایی لحظـه‌ای حمـلات به‌همراه پردازش حجم عظیم ترافیک تولید شده توسط شبکه‌های امروزی است[4]. تکنیک‌های تحلیل گراف به‌طـور گسـترده‌ای در داده‌های ترافیک سایبری استفاده می‌شـوند. بـا ایـن حـال، ایـن تکنیک‌ها به‌دلیل نیاز بـه تجزیـه و تحلیـل روی تعـداد زیـادی از گره‌های شبکه از پیچیدگی محاسباتی بالایی برخـوردار هسـتند. بنابراین تکنیک‌های پارتیشن‌بندی گراف برای تقسیم گراف‌ها بـه زیرگراف‌هایی که به‌عنوان انجمن شناخته شده و مجموعاً نمایانگر شبکه هستند، بیشتر مورد استفاده قرار مـی‌گیرنـد. مطالعـات در حوزه تحلیل و آنالیز شبکه نشان می‌دهد که بسیاری از شـبکه‌هـا دارای ساختار انجمنی بوده و هر ساختار شبکه‌ای به‌طور طبیعـی به چندین انجمن تقسیم می‌شود. شناسایی انجمن‌ها[3] در شبکه و تجزیه و تحلیل انجمن‌های به‌دست‌آمده به‌جای تحلیل کل شبکه، از راه‌های مؤثر در بررسی پیکربندی شبکه‌های سایبری است[5]. تشخیص انجمن‌ها در شبکه بـه متخصصـان و مـدافعان سـایبری کمک می‌کند تا تجزیه و تحلیل شبکه در سطوح مختلف به‌خوبی صورت گرفته و در نتیجه کـارایی و اثربخشـی افـزایش مـی‌یابـد. انجمن در شبکه نشان دهنده‌ی گروهی از اعضای شبکه بوده کـه دارای تعامل بالا و اتصالات متراکم‌تری با یکدیگر نسبت به سـایر بخش‌های شـبکه هسـتند. شناسـایی انجمـن‌هـا در شـبکه‌هـای سایبری تشخیص حمـلات و قربانیـان آسـیب دیـده را آسـان‌تـر سـاخته و سـاختار کارآمـدتری از شـبکه‌هـای بـزرگ را ارائـه مـی‌دهـد[6]. بـرای درک اهمیـت موضـوع مـی‌تـوان یـک کـرم کامپیوتری[4] را در نظر گرفت که به‌عنوان بدافزار[5] شناخته می‌شود و قادر به کپی کردن خود از یک کامپیوتر به کامپیوتر دیگر بدون دخالت هیچ شخصی است. کـرم‌هـای کـامپیوتری اغلـب از یـک شبکه کامپیوتری و با تکیه بر آسیب‌پذیری و ضعف‌هـای سیسـتم قربانی برای انتشار خود استفاده می‌کنند. اگر این بدافزار، یکـی از کامپیوترها که به‌عنوان یک گره از شبکه شناخته می‌شود را آلوده نماید، پتانسیل آسیب رساندن به شبکه میزبان خود را با مصـرف پهنای باند و بارگذاری بیش از حد سرورها با انتشار و تأثیرگذاری بر سایر گره‌های موجود در شبکه را خواهد داشت. اگر انجمن‌های شبکه از قبل شناسایی شده باشند، آسیب‌پذیری گره‌های موجود در انجمن که گره آلوده به آن تعلق دارد نسبت به سایر گره‌هـای شـبکه بیشـتر خواهـد بـود. بنـابراین بـا شناسـایی گـره‌هـای آسیب‌پذیرتر می‌توان تعـداد گـره‌هـایی کـه بـرای بررسـی اولیـه انتخاب می‌شوند را کاهش داد. ساختار انجمنی شـبکه‌هـا بـه مـا کمک می‌کند تا آسیب‌پـذیرترین مجموعـه گـره‌هـا را در شـبکه شناسایی کنیم[7]. شبکه‌هـای پیچیـده[6] بـه‌عنـوان مـدلی بـرای بسیاری از شبکه‌های واقعی مانند اینترنت، شبکه‌هـای اجتمـاعی، شبکه‌های بیولوژیکی و... استفاده می‌شوند. شبکه‌های سـایبری را می‌توان نوعی از شبکه‌های بی مقیـاس[7] ماننـد اینترنـت در نظـر گرفت که دارای ضرایب خوشه‌بنـدی[8] بـالایی هسـتند. بـه‌عنـوان مثال یک انجمن سایبری را می‌توان مجموعـه‌ای از صـفحات وب بسیار متصل در شبکه اینترنت جهانی درنظر گرفت که موضوعات یا علایق مشابهی را به اشتراک مـی‌گذارنـد. بررسـی و شناسـایی انجمن‌ها، در شناخت ماهیت شبکه‌ها، امنیت شبکه و مدل توسعه شبکه نقشی اساسی دارد[8]. با افـزایش مقـادیر و حجـم داده‌هـا به‌دلیل توسـعه تکنولـوژی کـامپیوتر، چگـونگی اسـتفاده مـؤثر و نمـایش داده‌هـا بـا اسـتفاده از روش‌هـای کـاهش ابعـاد داده و شناسایی اجزای اصلی داده‌ها به‌منظـور کشـف مفـاهیم پنهـان از اطلاعات یک موضوع مهم، ضروری و چالش برانگیـز در تجزیـه و تحلیل داده‌های سایبری می‌باشد[9]. الگوریتم‌های مختلفی بـرای

[1] IP Address
[2] Intrusion detection
[3] Community Detection
[4] Computer Worm
[5] Malware
[6] Complex Networks
[7] Scale-Free Networks
[8] Clustering Coefficients

شناسایی انجمن‌ها در شبکه‌های سایبری که نوعی از شبکه‌های پیچیده محسوب می‌شوند، وجود دارد. در این مقاله جهت بهبود برخی از مشکلات ذکر شده درباره امنیت سایبری، بهبود شناسایی حملات و کاهش اثرات حملات سایبری به معرفی الگوریتم خوشه‌بندی طیفی[1] برای کشف انجمن‌های شبکه‌های سایبری پرداخته شده است. مزیت الگوریتم خوشه‌بندی طیفی استفاده از ماتریس لاپلاسی گراف[2] متناظر شبکه برای نمایش مجموعه داده است. زیرا مقادیر ویژه[3] و بردارهای ویژه[4] این ماتریس می‌توانند بینشی از ساختارهای طبیعی انجمن‌های موجود در شبکه‌ها را ارائه داده و اجازه تمایز خودکار بین انجمن‌ها را می‌دهند. خوشه‌بندی طیفی از تجزیه طیفی داده‌های مربوط به شبکه برای کاهش ابعاد و اجرای الگوریتم k-میانگین[5] در فضایی با ابعاد پایین‌تر استفاده می‌کند[10]. بنابراین، خوشه‌بندی طیفی بر پایه‌ی کاهش ابعاد داده‌ها بوده و این کاهش ابعاد به‌ویژه برای تشخیص انجمن‌های مجموعه داده‌های چند بعدی شبکه‌های بزرگ مفید است. نتایج به‌دست‌آمده از آزمایشات عملکرد بالای این روش در فرایند تشخیص انجمن‌های شبکه‌های سایبری شبیه‌سازی شده را نشان می‌دهد. به‌منظور بررسی کیفیت فرایند تشخیص انجمن‌های شبکه از معیار مقدار ضریب نیم‌رخ[6] استفاده می‌شود که به‌صورت ترکیب شباهت درون انجمنی و بین انجمنی بوده و از لحاظ ساختاری کیفیت تشخیص انجمن‌ها را مورد ارزیابی قرار می‌دهد. دومین معیار بررسی کیفیت تشخیص انجمن‌ها نمایه ژاکارد[7] می‌باشد. در ادامه در بخش ۲، مدل پیشنهادی، الگوریتم خوشه‌بندی طیفی و روش‌های ارزیابی معرفی می‌شود. در بخش ۳ به آزمایشات و نتایج مدل پیشنهادی پرداخته می‌شود. در پایان نتیجه‌گیری در بخش ۴ مطرح می‌گردد.

۲- روش پیشنهادی

شبکه‌های سایبری به این دلیل که دارای ساختارهای توپولوژیکی متنوعی هستند و این ساختارها را نمی‌توان به‌دلیل اهمیت آن‌ها در تشکیل شبکه کم ارزش تلقی نمود، به‌عنوان شبکه‌های پیچیده[8] در نظر گرفته می‌شوند. ساختار شبکه‌های سایبری به گونه‌ای است که توزیع ارتباطات یا پیوندهای بین عناصر شبکه همگن و یکسان نمی‌باشد[11]. این ویژگی از شبکه‌ها که در برخی از بخش‌های آن پیوندها چگال‌تر و در بخش‌های دیگر تنک هستند، سبب ایجاد ساختار انجمنی یا خوشه‌ای می‌شود. ساختار انجمنی یکی از ویژگی‌های اساسی در شبکه‌های پیچیده بوده و یک مفهوم بنیادین در مطالعه و پالایش شبکه‌ها می‌باشد[12].

دو عامل اصلی تشکیل‌دهندهٔ شبکه‌ها، عناصر شبکه و روابط بین آن‌ها می‌باشد. شبکه‌های سایبری را می‌توان به‌صورت گراف در نظر گرفت که عناصر شرکت‌کننده در شبکه به‌صورت گره و ارتباط بین آن‌ها با استفاده از یک پیوند نشان داده می‌شود. گره‌ها نمایانگر کامپیوترها، سرورها، ماشین‌های میزبان و دستگاه‌های کامپیوتری خواهد بود. پیوند بین این گره‌ها نیز می‌تواند بر اساس تبادلات درون‌شبکه‌ای، نوع داده‌های ارسالی و... باشد. حملات سایبری با تجزیه‌وتحلیل ترافیک شبکه برای تشخیص رفتار غیرعادی بین گره‌ها قابل‌شناسایی هستند. نگرانی اساسی در این فرایند، زمان و میزان تلاشی است که برای تجزیه‌وتحلیل شبکه باتوجه‌به وجود تعداد زیاد میزبان موردنیاز می‌باشد. افزایش اندازه شبکه‌ها این مشکل را بدتر می‌سازد. با استفاده از شناسایی انجمن‌ها در شبکه‌های سایبری می‌توان انجمن‌ها و همچنین گره‌های دارای اهمیت بالا در هر انجمن را زودتر و آسان‌تر شناسایی نموده و مکانیزم‌های دفاعی در برابر حملاتی مانند انکار سرویس[9]، حملات بدافزاری، روت کیت‌ها[10]، فیشینگ[11] و... را پیاده‌سازی کرد.

شبکه‌های سایبری می‌توانند به‌صورت ساختار یک گراف $G = (V, E)$ در نظر گرفته شوند. در این ساختار V نشان‌دهنده‌ی مجموعه‌ای از اجزای کامپیوتری، افراد، سرورها یا رأس‌های شبکه و E مجموعه‌ای از پیوندها می‌باشد که دو رأس از V را به‌یکدیگر متصل می‌نماید. یک نمایش رایج از ساختار توپولوژیکی شبکه G، ماتریس مجاورت[12] مربوط به آن شبکه می‌باشد. ماتریس مجاورت یا ماتریس همجواری یک شبکه با گراف غیرجهت‌دار، یک ماتریس متقارن با عناصر نامنفی است. برای این منظور از ماتریس دودویی متقارن غیرمنفی $A = [a_{ij}] \in R_{+}^{N\times N}$ استفاده می‌گردد. فرض کنید شبکه سایبری مورد مطالعه به‌صورت یک گراف غیرجهت‌دار متشکل از N گره نمایش داده شود. ارتباط هر زوج گره (i, j) به‌صورت عدد نامنفی a_{ij} نمایش داده می‌شود. به‌طوریکه $a_{ij} = 1$ به‌معنای وجود ارتباط بین گره i و j بوده و $a_{ij} = 0$ به‌معنای عدم ارتباط است. ضمناً فرض می‌شود این گراف متناظر با شبکه سایبری شامل k

[1] Spectral Clustering
[2] Graph Laplacian Matrix
[3] Eigenvalues
[4] Eigenvectors
[5] K-means Algorithm
[6] Silhouette Value
[7] Jaccard Index
[8] Complex Networks
[9] Denial-of-Service
[10] Rootkits
[11] Phishing
[12] Adjacency Matrix

زیرگراف همبند یا انجمن می‌باشد که انجمن‌ها به‌صورت $C_1, C_2, \dots, C_k$ نمایش داده می‌شوند. مسأله تشخیص انجمن‌ها مربوط به اختصاص گره‌های موجود در شبکه سایبری بر اساس اطلاعات توپولوژیکی به k گروه مختلف می‌باشد. شکل (۱)، شمای کلی الگوریتم پیشنهادی بر روی یک گراف ساده غیرجهت‌دار که نمایانگر یک شبکه سایبری متشکل از ۴ انجمن را نشان می‌دهد. گره‌ها با رنگ‌های متفاوت متعلق به انجمن‌های مختلف بوده و می‌توانند با به اشتراک گذاشتن اطلاعات مشترک، با یکدیگر ارتباط داشته باشند. در ادامه به معرفی الگوریتم خوشه‌بندی طیفی و استفاده از آن برای تشخیص انجمن‌های شبکه پرداخته می‌شود.

۲-۱- الگوریتم خوشه‌بندی طیفی

در سال‌های اخیر الگوریتم‌های مختلفی جهت شناسایی انجمن‌های شبکه مورد استفاده قرار گرفته است. الگوریتم خوشه‌بندی طیفی یکی از پرکاربردترین روش‌ها برای تجزیه و تحلیل داده‌های اکتشافی بوده و رویکردی متداول مبتنی بر گراف برای خوشه‌بندی بدون نظارت بر داده‌ها دارد. یکی از چالش‌های روش‌های کلاسیک مانند الگوریتم k-میانگین این است که در برخورد با داده‌های با ابعاد بالا به‌علت بزرگ شدن فضای مورد بررسی، قادر به تضمین دست‌یابی به مقدار بهینه نخواهند بود. برای حل این موضوع می‌توان از روش خوشه‌بندی طیفی استفاده نمود که داده‌ها ابتدا به ابعاد کوچک‌تری تبدیل شده و سپس در فضای جدید، عملیات تشخیص انجمن انجام می‌شود. این روش در مقایسه با روش‌های پیشین مانند خوشه‌بندی سلسله مراتبی[1]، k-میانگین و ... به سادگی قابل پیاده‌سازی بوده و سریع‌تر می‌باشد و قابل ترکیب با سایر الگوریتم‌های تشخیص انجمن نیز است. همچنین بسیاری از روش‌های تشخیص انجمن مانند روش k-میانگین، ساختار انجمن‌های شبکه‌ها را به‌صورت یکنواخت و دایره‌ای فرض می‌کنند که در دنیای واقعی مخصوصا شبکه‌های سایبری اینگونه نمی‌باشد. در این گونه از شبکه‌ها، خوشه‌بندی طیفی بهتر عمل می‌کند، زیرا برای اشکال و اندازه‌های مختلف از داده کارآمد می‌باشد. مزیت مهم دیگر خوشه‌بندی طیفی توانایی تعیین تعداد انجمن‌های شبکه است که در مورد شبکه‌های سایبری دنیای واقعی کارآمد خواهد بود[13]. خوشه‌بندی طیفی به‌دلیل استفاده از مفاهیم متعدد ریاضی مانند ماتریس‌های وزنی[2]، ماتریس‌های درجه[3]، ماتریس‌های شباهت[4]، ماتریس لاپلاس، مقادیر و بردارهای ویژه به‌منظور گروه‌بندی کردن نقاط مشابه در یک گروه، یک الگوریتم پیشرفته‌تر در مقایسه با الگوریتم k-میانگین محسوب می‌شود[14].

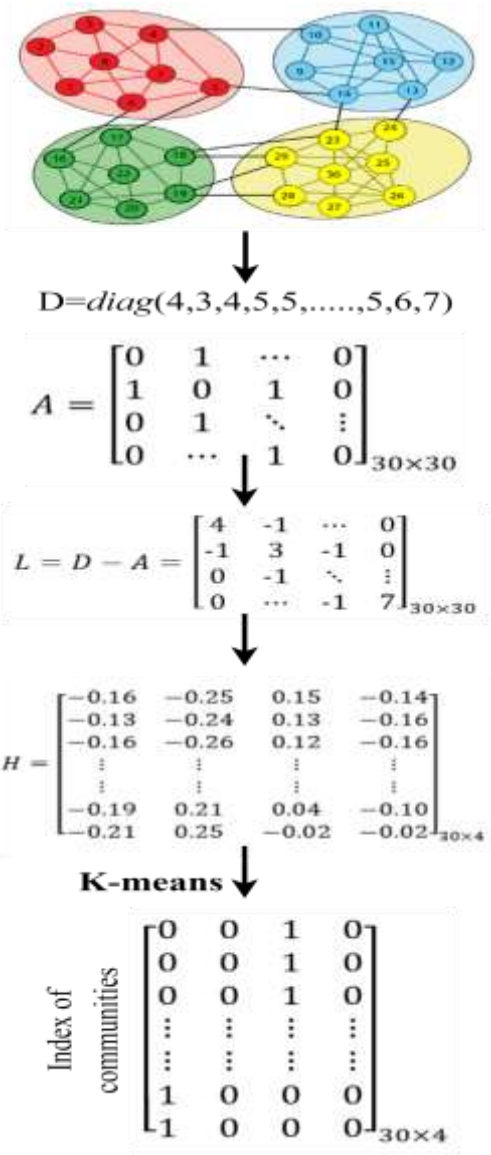
هدف از استفاده این الگوریتم شناسایی بخش‌هایی از گراف شبکه است که برش یا تعداد پیوندهای بین دو مجموعه غیرمتصل از شبکه را کمینه می‌سازد. اگر دو مجموعه A و B را افرازهایی از مجموعه گره‌های V گراف معادل شبکه سایبری در نظر بگیریم که $A \cup B = V$ و $A \cap B = \emptyset$ باشد. افرازهای A و B باید به‌صورتی انتخاب شوند که برش بین دو مجموعه را کمینه سازند. به‌عبارت دیگر هدف الگوریتم خوشه‌بندی طیفی، افراز k تایی گراف با حذف یال‌هایی از گراف است که باعث ایجاد k زیرگراف مجزا می‌شود[15]. انتخاب تابع برش تأثیر زیادی بر روی اندازه انجمن‌های شبکه دارد. فرض کنید شبکه سایبری به k انجمن غیرهمپوشان به‌صورت $C_1, C_2, \dots, C_k$ تقسیم شده باشد. دو متغیر برش نسبی[5] و برش نرمالیزه[6] به‌صورت روابط (۱) و (۲) تعریف می‌شوند.

$$Ratio\ cut(c_1, \dots, c_k) = \sum_{i=1}^{k} \frac{cut(c_i, \bar{c}_i)}{|c_i|} \qquad (۱)$$

$$Normalized\ cut(c_1, \dots c_k) = \sum_{i=1}^{k} \frac{cut(c_i, \bar{c}_i)}{vol(c_i)} \qquad (۲)$$

در این روابط $\bar{c}_i$ متمم c_i بوده و $vol(c_i)$ مجموع درجات رئوس در انجمن c_i می‌باشد. در برش نسبی، اندازه زیرگراف یا انجمن c_i برابر با تعداد رئوس در آن انجمن می‌باشد.

[1] Hierarchical Clustering
[2] Weight Matrix
[3] Degree Matrix
[4] Similarity Matrix
[5] Ratio Cut
[6] Normalized Cut



شکل (۱): ساختار الگوریتم خوشه‌بندی طیفی برای تشخیص انجمن‌های شبکه سایبری شبیه‌سازی شده

برش نرمالیزه، یکی از شاخص‌های مهمی است که می‌تواند تفسیر مناسبی از برش گراف را نشان دهد. در روابط فوق، مقادیر را می‌توان به‌صورت رابطه (۳) محاسبه نمود. هدف این توابع کم کردن پیوندهای بین انجمن‌های موجود در شبکه می‌باشد تا فرایند افراز و جداسازی به‌درستی انجام شود. به‌طور کلی، در الگوریتم خوشه‌بندی طیفی ماتریس‌های مجاورت، درجه و لاپلاس مطرح می‌شود. ماتریس مجاورت $(A = (a_{ij})_{n\times n})$ نشان دهنده وجود ارتباط یا پیوند میان اعضای متناظر گراف بدون جهت و بدون وزن بوده و به‌صورت متقارن می‌باشد.

$$Cut(c_i, \bar{c}_i) = \frac{1}{2}\sum_{i=1}^{k}(c_i, \bar{c}_i) \qquad (۳)$$

$$vol(c_i) = \sum_{i\in A} d_i$$

$$W(c_i, \bar{c}_i) = \sum_{i\in A, j\in \bar{A}} a_{ij}$$

ازآنجایی‌که اکثر شبکه‌های سایبری دارای روابط دوسویه و گراف‌های متناظر آن‌ها بدون جهت هستند، بنابراین $a_{ij} = a_{ji}$ است. ماتریس درجه(D)، یک ماتریس قطری[1] است که حاوی اطلاعاتی در مورد درجه هر رأس است. به‌عبارت دیگر تعداد یال‌های متصل به هر رأس بر روی قطر اصلی این ماتریس مانند رابطه (۴) نشان داده می‌شود. همچنین می‌توان درجه گره‌ها را با در نظر گرفتن مجموع هر سطر در ماتریس مجاورت محاسبه نمود.

$$D = \begin{bmatrix} \sum_{j=1}^{n} a_{1j} & 0 & \cdots & \cdots \\ 0 & \sum_{j=1}^{n} a_{2j} & \cdots & \cdots \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & \cdots & \sum_{j=1}^{n} a_{nj} \end{bmatrix} \qquad (۴)$$

$$and \quad v_i \in V\,, d_i = \sum_{j=1}^{n} a_{ij}$$

ماتریس لاپلاس (L)، برای یافتن بسیاری از خواص مفید گراف‌ها یا نقاط داده مورداستفاده قرار می‌گیرد و همچنین از ویژگی‌هایی است که در خوشه‌بندی طیفی مورداستفاده قرار می‌گیرد. باتوجه‌به گراف ساده G با n رأس و استفاده از روش برش نسبی، ماتریس لاپلاسی آن به‌صورت رابطه (۵) تعریف می‌گردد. خوشه‌بندی طیفی غیرنرمال از ماتریس لاپلاس ساده و غیرنرمال استفاده می‌کند.

$$L_{n\times n} = D - A \qquad (۵)$$

در رابطه (۵) که ماتریس لاپلاس غیرنرمال می‌باشد، D ماتریس قطری حاوی مجموع درجات رئوس و A ماتریس مجاورت مربوط به شبکه می‌باشد. برای نرمال کردن ماتریس لاپلاس با در نظر گرفتن برش نرمالیزه، ماتریس لاپلاسی نرمال متقارن به‌صورت رابطه (۶) تعریف می‌شود. ماتریس حاصل یک ماتریس نیمه معین مثبت[2] با n مقدار ویژه با ارزش واقعی غیرمنفی است که به‌ترتیب صعودی مرتب شده است $0 = \lambda_1 \le \lambda_2 \le \cdots \le \lambda_n$. خوشه‌بندی طیفی نرمال شده به حالتی از

[1] Diagonal Matrix
[2] Positive Semi-Definite Matrix

خوشه‌بندی طیفی گفته می‌شود که از ماتریس لاپلاس نرمال شده استفاده می‌کند. ماتریس متقارن محاسبه شده، ویژگی‌هایی از توپولوژی شبکه را توصیف می‌کند[16].

$$L_{sym} = D^{\frac{-1}{2}} L D^{\frac{-1}{2}} = I_{n\times n} - D^{\frac{-1}{2}} A D^{\frac{-1}{2}} \qquad (۶)$$

به صورتی که $I_{n\times n}$ ماتریس همانی[1] می‌باشد. با بکار بردن تئوری طیفی گراف، یک جواب تقریبی با استفاده از بردار ویژه متناظر با کوچک‌ترین مقدار ویژه‌ی ماتریس لاپلاس نرمال L محاسبه می‌گردد. با تعریف بردارهای شاخص $h_j = (h_{1j}, \dots, h_{nj})$ به‌صورت رابطه (۷)، ماتریس H که ستون‌های آن حاوی این بردارها است، تعیین می‌شود.

$$h_{ij} = \begin{cases} 1/\sqrt{vol(A_j)} & , \quad if\ V_i \in A_j \\ 0 & , otherwise \end{cases} \quad , (i = 1, \dots, n; j = 1, \dots, k). \qquad (۷)$$

می‌توان مسئله کمینه‌سازی برش نرمال شده را همانند رابطه (۸) فرموله‌بندی کرد[17].

$$\min_{C_1,\dots,C_k} Tr(H'LH) \quad subject\ to\ H'DH = I \qquad (۸)$$

تشخیص انجمن‌ها بر اساس خوشه‌بندی طیفی بر محاسبه k کوچک‌ترین مقادیر ویژه و یافتن بردارهای ویژه مربوطه تمرکز دارد. ماتریس H را می‌توان توسط k بردارهای ویژه به‌عنوان ستون ساخت. هر سطر از این ماتریس مربوط به یک گره متناظر با هر موجودیت در شبکه و متعلق به یکی از k انجمن‌های شبکه می‌باشد. در مرحله آخر می‌توان از الگوریتم‌های دیگر مانند k-میانگین بر روی ماتریس حاصل از مراحل فوق که حاوی بردار ویژه ماتریس می‌باشد، انجمن‌های شبکه را تشخیص داد. به‌طورکلی، الگوریتم خوشه‌بندی طیفی با محاسبه ماتریس‌های لاپلاس از گراف متناظر با شبکه و استخراج مقادیر و بردارهای ویژه و با استفاده از الگوریتم‌های طبقه‌بندی، داده‌های ورودی را گروه‌بندی می‌کند. الگوریتم ۱ مراحل تشخیص انجمن‌ها با استفاده از الگوریتم خوشه‌بندی طیفی را نشان می‌دهد

الگوریتم ۱: الگوریتم خوشه‌بندی طیفی برای تشخیص انجمن‌ها

ورودی: ماتریس مجاورت (A) شبکه G و تعداد انجمن‌ها k

خروجی: انجمن‌های های $C_1, C_2, \dots, C_k$

۱- نرمال کردن ماتریس مجاورت ورودی الگوریتم

۲- محاسبه ماتریس لاپلاسین L با استفاده از ماتریس مجاورت متناظر با گراف شبکه

۳- محاسبه ماتریس لاپلاس نرمال شده L_{sym}

۴- محاسبه مقادیر ویژه و مرتب‌کردن آن‌ها ($0 = \lambda_1 \le \cdots \le \lambda_k$)

۵- محاسبه بردارهای ویژه ($u_1, \cdots, u_k$) متناظر با k کوچک‌ترین مقدار ویژه ($\lambda_1, \cdots, \lambda_k$) از ماتریس L

۶- محاسبه ماتریس $H_{n\times k}$ حاوی ستون‌هایی از k برداروِیژه

۷- استفاده از الگوریتم k-میانگین بر روی H برای تناظر با انجمن‌ها

۲-۲- روش‌های ارزیابی

پس از تجزیه و تشخیص انجمن‌های شبکه، می‌بایست کیفیت فرایند تشخیص و محاسبات مورد بررسی قرار گیرد. برای این منظور از دو معیار مقدار ضریب نیم‌رخ و نمایه ژاکارد استفاده می‌شود. با استفاده از معیار ضریب نیم‌رخ می‌توان انجمن‌های شناسایی شده را از نظر ساختاری و میزان تراکم انجمن‌های موجود در شبکه سایبری مورد ارزیابی قرارداد. به‌عبارت‌دیگر این معیار برای اندازه‌گیری میزان نزدیکی و فاصله اعضای حاضر در شبکه می‌باشد. بعضی از اعضای حاضر در شبکه ممکن است در همسایگی با چند انجمن قرار داشته باشند که به‌عنوان اعضای مرزی شناخته می‌شوند. با استفاده از این روش می‌توان تشخیص داد که کدام یک از عناصر موجود در شبکه به‌صورت اعضای مرزی هستند. ازآنجایی‌که شناسایی اعضای دارای اهمیت بالا در شبکه‌های سایبری که دارای پیوندهای زیاد و درجه بالاتری هستند برای افزایش امنیت و مکانیزم‌های دفاعی ضروری می‌باشد، بنابراین با استفاده از این ویژگی معیار ضریب نیم‌رخ می‌توان انجمن‌ها و اعضای شاخص حاضر در شبکه را شناسایی نمود. با استفاده از رابطه (۹)، فاصله هر عضو موجود در انجمن را با هر کدام از اعضای موجود در همان انجمن محاسبه می‌شود. به عبارتی میانگین فاصله عضو o_i را با دیگر اعضای موجود در همان انجمن را محاسبه نموده و در مقدار a قرار می‌گیرد. این‌کار با استفاده از ارزیابی ماتریس ورودی و ماتریس خروجی حاصل از فرایند تشخیص انجمن‌های شبکه قابل انجام می‌باشد.

$$a(O_i) = \frac{1}{|C_A| - 1} \sum_{O_j \in C_A, O_j \neq O_i} d(O_i, O_j) \qquad (۹)$$

همچنین در رابطه (۱۰)، فاصله عضو o_i را با کل اعضای موجود در انجمن‌های دیگر به‌غیر از همان انجمنی که در آن قرار دارد محاسبه نموده و در مقدار b قرار می‌گیرد.

$$b(O_i) = \min_{C_B \neq C_A} \frac{1}{|C_B|} \sum_{O_j \in C_B} d(O_i, O_j) \qquad (۱۰)$$

با استفاده از نتایج روابط (۹) و (۱۰) مقادیر ضریب نیم‌رخ در رابطه (۱۱) قابل محاسبه خواهد بود.

$$sil(O_i) = \frac{b(O_i) - a(O_i)}{\max\{a(O_i), b(O_i)\}} \qquad (۱۱)$$

مقادیری که در محاسبه ضریب نیم‌رخ حاصل می‌شود در بازه $[-1 \quad 1]$ قرار دارند. حال هرچه این مقادیر به‌دست‌آمده بزرگتر باشند، میانگین فاصله آن عضو در انجمن کمتر از میانگین فاصله

[1] Identity Matrix

با انجمن‌های همسایه است، بنابراین فرایند تشخیص انجمن به‌خوبی صورت گرفته است[18]. هرچه مقدار ضریب نیم‌رخ کمتر باشد، میانگین فاصله داخل انجمن بزرگ‌تر از میانگین فاصله با انجمن‌های همسایه می‌باشد، بنابراین دقت تشخیص پایین خواهد بود. همچنین رابطه (۱۲) میانگین مقدار ضریب نیم‌رخ را برای تمام اعضای انجمن محاسبه می‌نماید. با استفاده از این رابطه می‌توان میزان دقت ساختاری اعضای حاضر در هر انجمن را مورد ارزیابی قرار داد.

$$sil(C_i) = \frac{1}{|C_i|}\sum_{a_j \in C_i} sil(O_j) \qquad (۱۲)$$

معیار ژاکارد نیز به‌عنوان یک معیار دیگر برای درستی آزمایش‌های مورداستفاده قرار می‌گیرد. نمایه ژاکارد یا ضریب شباهت ژاکارد، معیاری برای مقایسه یا تفاوت مجموعه نمونه‌های آماری است. به‌منظور تعیین درستی و دقت عملکرد عملیات تشخیص انجمن‌های شبکه، از مجموعه‌داده‌هایی استفاده می‌شود که تعداد انجمن‌ها و همچنین اعضای حاضر در هر انجمن به‌عنوان دانش پیشین از قبل مشخص باشد. میزان شباهت دو مجموعه نمونه که در اینجا ماتریس‌های خروجی و حقیقی ما هستند، با نمایه ژاکارد طبق رابطه (۱۳) که از تقسیم تعداد اشتراک دو مجموعه بر تعداد اجتماع آن‌ها محاسبه می‌شود[19].

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \qquad (۱۳)$$

در رابطه (۱۳)، می‌توان A را به‌عنوان ماتریس مجاورت انجمن‌های حاضر و قابل انتظار در شبکه و همچنین ماتریس نشان دهنده‌ی انجمن‌ها که از فرایند تشخیص انجمن با استفاده از الگوریتم خوشه‌بندی طیفی به‌دست می‌آید را به‌عنوان ماتریس B در نظر گرفت. اگر مقدار نمایه ژاکارد برابر با یک باشد، می‌توان گفت فرایند تشخیص اعضا در شبکه برای هر انجمن دقیقاً برابر با انجمن‌های از قبل تعیین شده و قابل انتظار است. هر چه میزان مقدار به‌دست‌آمده به مقدار یک نزدیک‌تر باشد، صحت و دقت فرایند تشخیص انجمن‌ها بالاتر خواهد بود.

۳- آزمایشات و نتایج

در این بخش، عملکرد الگوریتم خوشه‌بندی طیفی در تشخیص انجمن‌ها مورد بررسی قرار می‌گیرد. ازآنجاکه مناسب‌ترین روش نمایش شبکه‌ها مانند شبکه‌های سایبری استفاده از گراف‌ها و ماتریس‌های مجاورت می‌باشد، الگوریتم خوشه‌بندی طیفی را می‌توان برای تمامی مجموعه‌داده‌هایی که به‌صورت ماتریسی بوده و یا ساختار شبکه‌ای دارند جهت آنالیز شبکه و تشخیص انجمن‌های شبکه مورداستفاده قرارداد. این الگوریتم به دلیل کار با ساختمان داده‌های ماتریسی، کاهش ابعاد داده‌های ورودی، استخراج ویژگی‌های گراف از ماتریس‌های لاپلاس، محاسبه ویژگی با ابعاد بالا به‌صورت مسائل غیرخطی و تأثیر وزن اتصالات درون‌شبکه‌ای از عملکرد بهتری نسبت به سایر الگوریتم‌ها مانند k-میانگین، سلسله‌مراتبی و الگوریتم بر پایه چگالی نقاط برخوردار می‌باشد. همان‌طور که در ادامه نشان داده خواهد شد، این روش بادقت بالایی انجمن‌های موجود در شبکه را تشخیص داده و می‌توان با کمک این روش یک شبکه سایبری واقعی را به تعداد انجمن‌های مناسب تقسیم نموده و سیاست‌های امنیتی متناسب با هر انجمن را پیاده‌سازی نمود. با استفاده از مقادیر ضریب نیم‌رخ که کیفیت انجمن‌ها را از نظر ساختاری بررسی می‌کند، انجمن‌ها و اعضایی از شبکه را که دارای میانگین درجه بالاتری هستند را شناسایی نموده و باتوجه‌به ارتباطات بالای این اعضا، سرویس‌های دفاعی را بر روی آن‌ها اجرا نمود تا باعث جلوگیری از حملات احتمالی در شبکه شوند. در ادامه، باتوجه‌به خاصیت ماتریس لاپلاس نرمال شده در این الگوریتم، تعداد انجمن‌های مناسب شبکه مورد بررسی قرار می‌گیرد.

۳-۱- شبیه‌سازی شبکه سایبری

ابتدا از شبکه شبیه‌سازی شده در شکل (۱) برای نشان دادن کاربرد روش خوشه‌بندی طیفی در تشخیص انجمن‌های شبکه‌های سایبری استفاده می‌شود. سپس روش پیشنهادی را بر روی داده واقعی نشان داده و با استفاده از معیارهای ارزیابی که در بخش (۲) بحث شد، درستی نتایج به‌دست‌آمده مورد بررسی قرار می‌گیرد. شکل (۲)، توپولوژی شبکه سایبری شبیه‌سازی شده را نشان می‌دهد. همچنین ساختار کلی و انجمن‌های شبکه بعد از اجرای الگوریتم به‌منظور شناسایی انجمن‌ها در شکل (۳) نشان داده شده است. نتایج حاصل از فرایند شناسایی انجمن‌های شبکه شبیه‌سازی شده برای مقادیر مختلف از تعداد انجمن‌ها و اعضای حاضر در هر انجمن در جدول (۱) ارائه شده است. از

جدول (۱): لیست اعضای انجمن‌ها

	k=۲	k=۳	k=۴
انجمن ۱	۱, ۲, ۳, ۴, ۵, ۶, ۷, ۸, ۹, ۱۰, ۱۱, ۱۲, ۱۳,۱۴ ,۱۵	۱, ۲, ۳, ۴, ۵, ۶, ۷, ۸	۱, ۲, ۳,۴ ۵, ۶, ۷, ۸
انجمن ۲	۱۶, ۱۷, ۱۸, ۱۹, ۲۰, ۲۱, ۲۲, ۲۳, ۲۴, ۲۵, ۲۶, ۲۷, ۲۸, ۲۹, ۳۰	۹, ۱۰, ۱۱, ۱۲, ۱۳, ۱۴, ۱۵	۹, ۱۰, ۱۱, ۱۲, ۱۳, ۱۴,۱۵
انجمن ۳		۱۶, ۱۷, ۱۸, ۱۹, ۲۰, ۲۱, ۲۲, ۲۳, ۲۴, ۲۵, ۲۶, ۲۷, ۲۸, ۲۹, ۳۰	۱۶,۱۷,۱۸, ۱۹, ۲۰,۲۱ , ۲۲
انجمن۴			۲۳,۲۴, ۲۵, ۲۶,

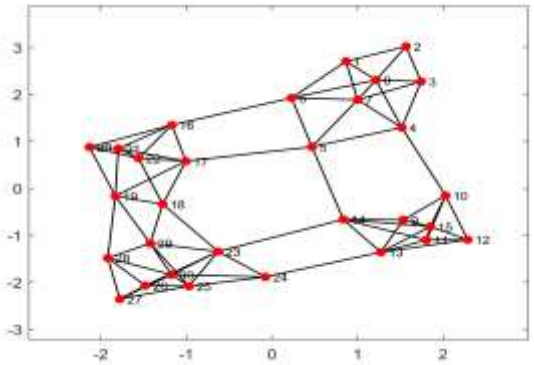
۲۷,۲۸, ۲۹,۳۰

آنجایی که شبکه فرضی با $k = 4$ انجمن در نظر گرفته شده است، معیارهای ارزیابی نیز دارای دقت و صحت بالایی با در نظر گرفتن چهار انجمن هستند. به‌منظور ارزیابی و بررسی دقت تشخیص انجمن‌ها و سنجش عملکرد الگورتم، ابتدا از نمایه ژاکارد تعریف شده در رابطه (۱۳) استفاده می‌شود.

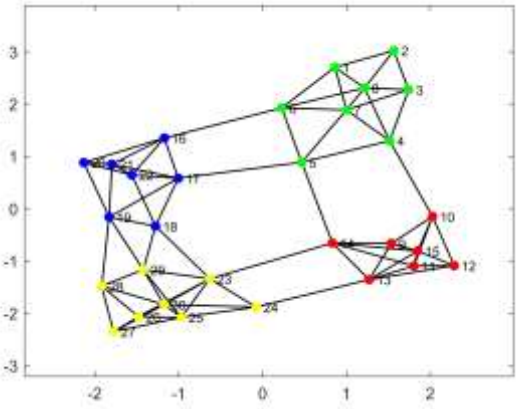
با مقایسه نتایج به‌دست‌آمده از اعضای حاضر در هر انجمن با انجمن‌های از قبل تعریف شده در شکل (۱)، مقدار ۱ حاصل می‌شود که نشان دهنده‌ی دقت بالای الگوریتم در تشخیص انجمن‌های شبکه سایبری فرضی است.

نکته قابل‌توجه در کاربرد الگوریتم خوشه‌بندی برای تشخیص انجمن‌ها این است که این روش تأکیدی بر روی منحصربه‌فرد بودن انجمن‌های به‌دست‌آمده حاصل از این روش ندارد؛ بنابراین اعضای تشخیص‌داده‌شده انجمن‌ها در شبکه ممکن است با اعضای از پیش تعیین شده مربوط به آن انجمن یکسان نباشد؛ اما اعضای هم انجمن در یک انجمن یکسان قرار خواهند گرفت. با استفاده از معیار ضریب نیم‌رخ نیز می‌توان برای هر عضو و هر انجمن، عملکرد تشخیص انجمن‌ها را از نظر ساختاری مورد بررسی قرارداد و اعضای مرزی و بااهمیت در هر انجمن را تعیین نمود. نتایج به‌دست‌آمده با استفاده از روابط (۱۱) و (۱۲) نشان می‌دهند که فرایند تشخیص انجمن‌های شبکه سایبری شبیه‌سازی شده از عملکرد بسیار خوبی برخوردار بوده است.

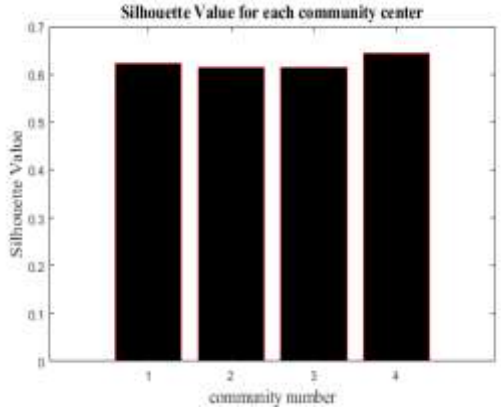
همان‌طور که در شکل (۴) نشان داده شده است، برای تمامی چهار انجمن در نظر گرفته شده مقادیر ضریب نیم‌رخ برابر ۰/۶۲۳۹ و نزدیک به مقدار مقدار یک بوده و این امر میزان دقت تشخیص از نظر ساختاری و چگال بودن اعضای انجمن‌ها را نشان می‌دهد.



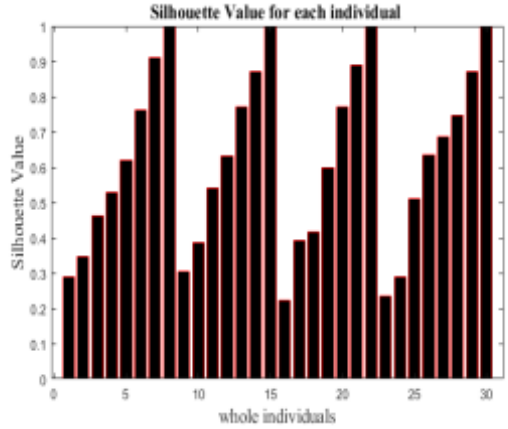
شکل (۲): توپولوژی شبکه سایبری شبیه‌سازی شده



شکل (۳): انجمن‌های شناسایی شده توسط الگوریتم



شکل (۴): مقادیر ضریب نیم‌رخ برای هر انجمن در شبکه



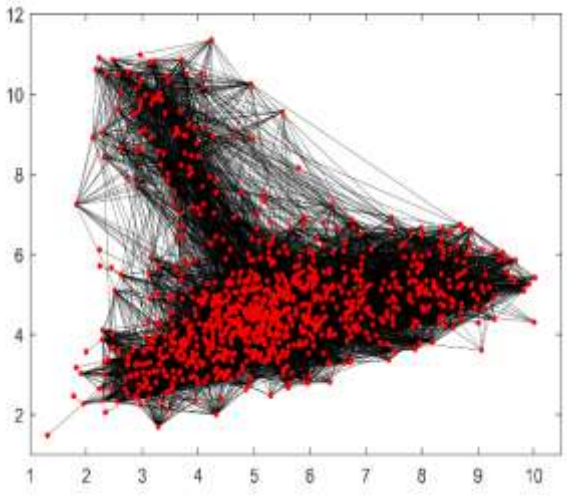
شکل (۵): مقادیر ضریب نیم‌رخ برای هر عضو در شبکه

همچنین در شکل (۵) می‌توان اعضای مرکزی و مرزی در انجمن‌ها را مشاهده نمود. که این مقدار در گره‌های با درجه بالا و با اهمیت برابر با مقدار ضریب نیم‌رخ ۱ می‌باشد. بنابراین این اعضا در شبکه دارای ارتباطات بالاتری بوده و از نظر امنیتی دارای نقش بیشتری در انتشار و گسترش بدافزارها در شبکه را دارا می‌باشند.

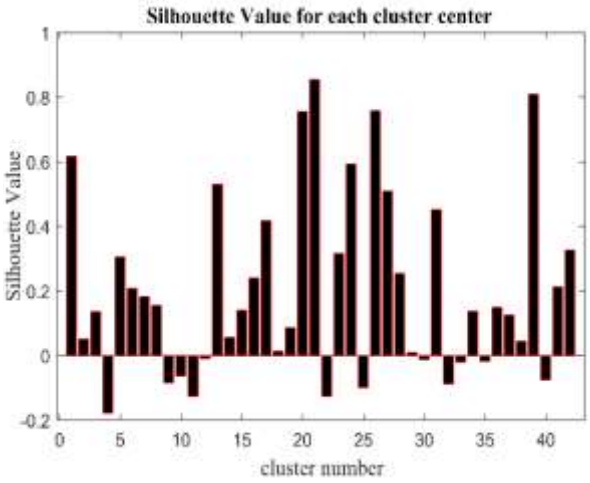
۳-۲- آزمایش مدل بر روی یک شبکه دنیای واقعی

پس از بررسی روش بر روی شبکه شبیه‌سازی شده، فرایند

تشخیص انجمن را بر روی داده‌های واقعی که نمونه‌ای از یک شبکه سایبری است، مورد سنجش قرار می‌گیرد. مجموعه‌داده موردمطالعه، مجموعه داده‌ای است که دارای جوامع حقیقی بوده و تعداد انجمن‌های حاضر در شبکه جهت سنجش فرایند تشخیص از قبل تعیین شده می‌باشد. برای این منظور از مجموعه‌داده ایمیل یک مؤسسه بزرگ تحقیقاتی در اروپا استفاده می‌شود. در این مجموعه‌داده، اطلاعات ناشناس تمامی ایمیل‌های دریافتی و خروجی بین اعضای این مؤسسه تحقیقاتی موجود است. درصورتی‌که فردی به شخص دیگر ایمیلی را ارسال نماید، یک یال (u,v) در شبکه از شخص u به شخص v وجود خواهد داشت. ایمیل‌ها فقط ارتباط بین اعضای موسسه را نشان می‌دهند و مجموعه داده شامل پیام‌های ورودی یا پیام‌های خروجی به سایر نقاط جهان نیستند. موسسه مورد مطالعه شامل ۱۰۰۵ کارمند بوده که مجموع ایمیل‌های تبادل شده بین آن‌ها ۲۵۵۷۱ ایمیل می‌باشد. بنابراین می‌توان موسسه مورد نظر را به‌صورت شبکه‌ای شامل ۱۰۰۵ گره و ۲۵۵۷۱ پیوند بین اعضای شبکه در نظر گرفت. شکل (۶)، شمای کلی شبکه ایمیلی موسسه تحقیقاتی را نشان می‌دهد.



شکل (۶): توپولوژی شبکه ایمیلی موسسه تحقیقاتی



شکل (۷): مقادیر ضریب نیم‌رخ برای هر انجمن در شبکه

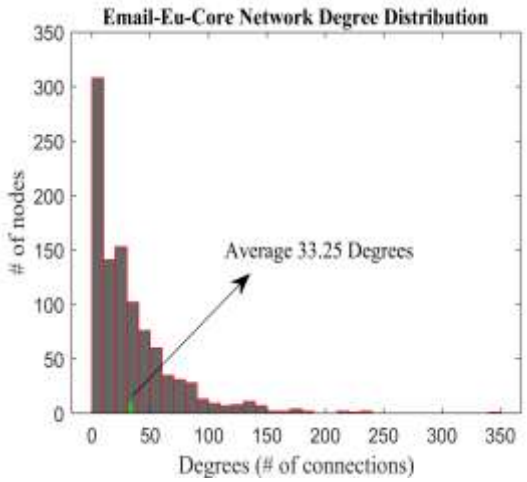
طبق اطلاعات داده شده تعداد انجمن‌های این شبکه برابر با ۴۲ انجمن می‌باشد. ورودی الگوریتم ارائه شده یک ماتریس مجاورت متقارن ۱۰۰۵× ۱۰۰۵ می‌باشد. ارزیابی کیفیت فرایند تشخیص انجمن‌ها با استفاده از مقادیر نمایه ژاکارد مقدار ۰/۹۲۶را نشان می‌دهد که نشان‌دهنده‌ی دقت بالا در فرایند تشخیص انجمن‌های شبکه مورد نظر می‌باشد. همچنین جهت بررسی کیفیت ساختاری انجمن‌های شبکه از معیار ضریب نیم‌رخ استفاده می‌شود که نتایج به‌دست‌آمده در شکل (۷) نشان داده شده است. با توجه به مقادیر حاصل شده توسط مقدار ضریب نیم‌رخ برای انجمن‌ها می‌توان نتیجه گرفت که انجمن‌هایی که دارای مقادیری نزدیک به ۱ هستند از تراکم و همبستگی بالایی برخوردار بوده و درنتیجه میانگین درجات در این انجمن‌ها بالاتر خواهد بود. بنابراین می‌توان اعضایی که تأثیر بالاتری با توجه به ارتباطات بالایی که در شبکه داشته را شناسایی نموده و سرویس‌ها و سیاست‌های امنیتی را در این انجمن‌ها اعمال نمود. همچنین نتایج حاصل از این ارزیابی نشان دهنده کیفیت مطلوب تشخیص انجمن‌ها در شبکه مربوط به داده‌های ایمیلی موسسه تحقیقاتی برای هر انجمن می‌باشد.

۳-۳- تعیین تعداد انجمن‌های شبکه سایبری

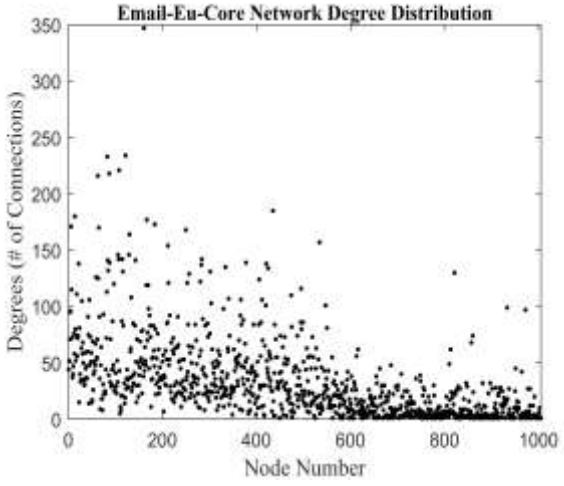
یکی از چالش‌های موجود برای شناسایی انجمن‌های شبکه‌های سایبری مانند شبکه اینترنت جهانی، یافتن تعداد زیرشبکه‌ها یا انجمن‌های شبکه سایبری است. به‌عنوان‌مثال در فضای وب جهانی[1] بیشتر صفحات به‌شدت به یکدیگر مرتبط هستند. تعیین تعداد انجمن‌های حاضر در شبکه که همان زیرگراف‌های مربوط به گراف متناظر با شبکه سایبری است، نقش بسیار زیادی در فرایند پیاده‌سازی سیاست‌های امنیتی خواهد داشت.

[1] World Wide Web

همان‌طور که در بخش (۱) اشاره شد، شبکه‌های سایبری نوعی از شبکه‌های بدون مقیاس هستند که توزیع درجه در این شبکه‌ها یکنواخت نبوده و دارای ضرایب خوشه‌بندی بالایی در بعضی از مناطق خاص (مانند هاب‌ها[1] در اینترنت) هستند. به‌عبارت دیگر در شبکه‌های سایبری تعداد کمی از اعضای شبکه دارای روابط زیادی بوده و نقش کلیدی در شبکه را ایفا می‌کنند و باقیمانده اعضای حاضر در شبکه دارای درجه و ارتباطات یکسانی هستند. شکل (۸) توزیع درجات مربوط به شبکه ایمیلی که نمونه‌ای از شبکه‌های دنیای واقعی و سایبری می‌باشد را نشان می‌دهد. توزیع درجات مربوط به هر یک از اعضای حاضر در شبکه نیز در شکل (۹) نشان داده شده است. طبق مشاهدات تعداد زیادی از اعضای ۱۰۰۵ عضوی شبکه دارای ارتباطات و درنتیجه درجه کم‌تری هستند به‌صورتی که میانگین درجات این شبکه ۳۳/۲۵ می‌باشد. این خاصیت از شبکه‌های سایبری باعث تشکیل ساختارهای گروهی و انجمنی می‌شود.



شکل (۸): توزیع درجه شبکه ایمیلی موسسه تحقیقاتی

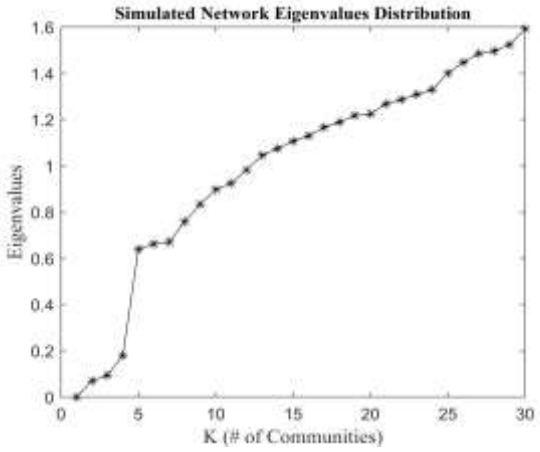


شکل (۹): توزیع درجات هر عضو شبکه ایمیلی موسسه تحقیقاتی

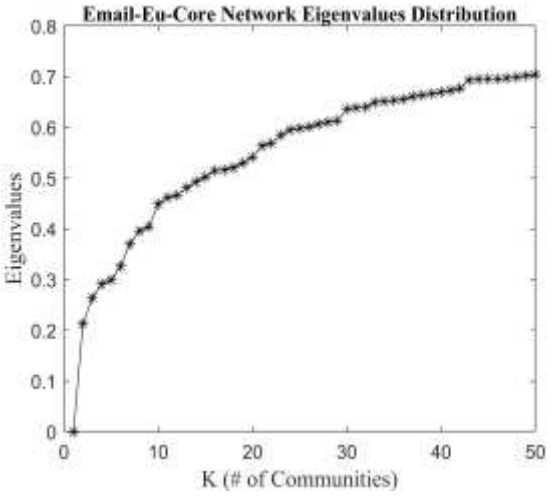
[1] Hubs

سیاست‌های امنیتی مربوط به شبکه‌های سایبری بسته به تعداد اعضای حاضر در انجمن، تعداد افرادی که دارای ارتباط و درجه بالایی هستند، تعداد اعضای مرزی حاضر در هر انجمن و همچنین تعداد انجمن‌های کل شبکه متفاوت خواهد بود. در مسئله تشخیص انجمن‌ها، تعیین تعداد انجمن‌ها مخصوصاً در شبکه‌های با مقیاس بزرگ بسیار مهم می‌باشد. زیرا بسیاری از الگوریتم‌های تشخیص انجمن، تعداد انجمن‌های شبکه را به‌عنوان ورودی نیاز داشته و سپس فرایند تشخیص انجمن‌ها آغاز می‌شود.

باتوجه‌به خاصیت ماتریس لاپلاس نرمال شده در الگوریتم خوشه‌بندی و همچنین روش اکتشافی مبتنی بر فاصله[2] می‌توان تعداد انجمن‌های مناسب شبکه را پیش‌بینی نمود. اگر m مقدار ویژه اول نسبتاً کوچک $(0 = \lambda_1 \leq \cdots \leq \lambda_m \leq \cdots \leq \lambda_n)$ مربوط به ماتریس لاپلاس نرمال که در بخش (۲) توضیح داده شد را در نظر گرفته، و سپس مقدار ویژه بعدی ناگهان بزرگ‌تر شود،

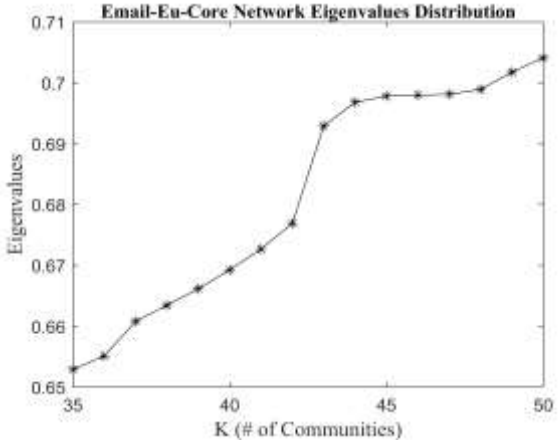


شکل (۱۰): توزیع مقادیر ویژه شبکه سایبری شبیه‌سازی شده



شکل (۱۱): توزیع ۵۰ مقدار اول مقادیر ویژه شبکه ایمیلی

[2] Distance-based Heuristic Method



شکل (۱۲): توزیع ۵۰-۳۵ مقدار اول مقادیر ویژه شبکه ایمیلی

در این صورت تعداد انجمن‌های مناسب شبکه برابر با m خواهد بود ($k = m$). هر شبکه سایبری n عضوی را می‌توان به صورت یک انجمن n عضوی و یا n انجمن تک عضوی در نظر گرفت. شکل (۱۰) توزیع مقادیر ویژه برای شبکه سایبری شبیه‌سازی شده را نشان می‌دهد. از آنجایی که شبکه مفروض از ۴ انجمن تشکیل شده است، مقادیر ویژه زمانی که $k = 1,2,3,4$ می‌باشد، به‌ترتیب کوچک می‌باشد. اما زمانی که $k = 5$ است، مقدار ویژه به‌وضوح بزرگ‌تر می‌شود. بنابراین تعداد انجمن‌های این شبکه شبیه‌سازی شده برابر با $k = 4$ خواهد بود. همچنین در شبکه با مقیاس بزرگ‌تر که ساختاری پیچیده‌تر و مشابه به یک شبکه سایبری دارد نیز برای تعداد انجمن‌های ۱ تا ۵۰ در شکل (۱۱) نشان داده شده است. تعداد انجمن‌های مناسب این شبکه در بازه [50 1] با توجه به مقادیر ویژه و اختلاف چشم چشمگیر بین این مقادیر قابل بررسی هستند. از آنجا که تعداد انجمن‌های این شبکه از قبل۴۲ انجمن درنظر گرفته شده است، جهت صحت عملکرد این روش برای تعیین تعداد انجمن‌های شبکه به‌صورت دقیق‌تر، مقادیر برای تعداد انجمن‌های [50 35] مورد بررسی قرار می‌گیرد. طبق شکل (۱۲)، مقادیر ویژه برای تعداد انجمن‌های ۴۲ و ۴۳ تفاوت نسبتاً بالایی دارند، بنابراین تعداد انجمن‌ها را می‌توان $k = 42$ در نظر گرفت. از نتایج مشاهده شده می‌توان کارایی خاصیت ماتریس لاپلاس نرمال شده در الگوریتم خوشه‌بندی طیفی برای تعیین تعداد انجمن‌های شبکه را نتیجه گرفت.

۴- نتیجه‌گیری

شناسایی حملات سایبری یکی از چالش‌های امنیتی در حوزه اینترنت است. ساختار انجمنی شبکه‌های سایبری باتوجه‌به نوع، اهمیت و میزان ارتباطات، تأثیر زیادی در نوع سیاست‌های مناسب با این حملات دارد. فعالیت‌های مخرب به‌شدت در شبکه‌ها گسترش پیدا کرده است. این نوع از فعالیت‌ها باعث ایجاد انجمن‌هایی همراه با آلودگی‌های متنوع باتوجه‌به نوع ارتباطات و تبادل‌های اطلاعاتی شده است؛ بنابراین اگر یکی از انجمن‌های شبکه دارای یک گره آلوده و مخرب باشد، به‌احتمال بسیار بالایی گره‌های دیگر آن انجمن نیز آلوده خواهند شد. برای این منظور می‌توان از الگوریتم‌های تشخیص انجمن استفاده کرد. استفاده از انجمن‌ها برای تجزیه‌وتحلیل الگوهای ارتباطی و تشخیص میزان گسترش فعالیت‌های مخرب آسان‌تر خواهد بود. همچنین می‌توان انجمنی را که برای برنامه‌های امنیت سایبری بهتر عمل می‌کند را شناسایی نمود.

در این مقاله با استفاده از الگوریتم خوشه‌بندی طیفی، فرایند تشخیص انجمن‌های موجود در دو شبکه شبیه‌سازی شده و واقعی ارائه شد که می‌توانند ساختاری مانند شبکه‌های سایبری داشته باشند. با استفاده از روش ارائه شده می‌توان انجمن‌ها و اعضایی که دارای تبادلات اطلاعاتی بالایی بوده و ارتباطات بیشتری با دیگر از اعضای حاضر در شبکه می‌باشند را شناسایی نموده و جهت پیشگیری و تشخیص حملات سایبری و اعمال سیاست‌های لازم برای بهبود امنیت شبکه اقدام نمود. نتایج مشاهده شده و ارزیابی حاصل از فرایند تشخیص انجمن‌های شبکه سایبری نشان می‌دهد که این روش از عملکرد مطلوبی برخوردار است؛ بنابراین می‌توان از این روش در پژوهش‌های مربوط به پیکربندی و تحلیل شبکه، امنیت شبکه، شناسایی حملات احتمالی در شبکه، تشخیص انجمن‌ها و زمینه‌های یادگیری ماشین استفاده نمود.

۵- مراجع

[1] B. Falahati and Y. Fu, “A study on interdependencies of cyber-power networks in smart grid applications,” *2012 IEEE PES Innov. Smart Grid Technol. ISGT 2012*, pp. 1–8, 2012, doi: 10.1109/ISGT.2012.6175593.

[2] L. X. Yang, P. Li, X. Yang, and Y. Y. Tang, “Security Evaluation of the Cyber Networks under Advanced Persistent Threats,” *IEEE Access*, vol. 5, pp. 20111–20123, 2017, doi: 10.1109/ACCESS.2017.2757944.

[3] M. D. Mohsen Hesabi, “An Improved Method for Malware Attack Detection in Cloud Computing Using Collective Learning,” *Sci. J. Electron. Cyber Def.*, vol. 10, no. 4, pp. 33–39, 2023, [Online]. Available: https://ecdj.ihu.ac.ir/article_207864.html. (in persian).

[4] Y. Javed, M. A. Khayat, A. A. Elghariani, and A. Ghafoor, “PRISM: A Hierarchical Intrusion Detection Architecture for Large-Scale Cyber Networks,” *IEEE Trans. Dependable Secur. Comput.*, pp. 1–17, 2023, doi: 10.1109/TDSC.2023.3240315.

[5] H. Deshmukh and J. Springer, “Identifying Bipartite Subgraphs for Community Detection in Very Large Scale Cyber Networks,” *Proc. - 2018 IEEE Int. Conf. Big Data, Big Data 2018*, pp. 4789–4797, 2019, doi: 10.1109/BigData.2018.8622614.

[6] A. Croitoru, N. Wayant, A. Crooks, J. Radzikowski, and A. Stefanidis, “Linking cyber and physical spaces through community detection and clustering in social media feeds,” *Comput. Environ. Urban Syst.*, vol. 53, pp. 47–64, 2015, doi:

869–888, 2022, doi: 10.1007/s10586-021-03430-0.

[16] Z. A. El Mouden, A. Jakimi, and M. Hajar, "An application of spectral clustering approach to detect communities in data modeled by graphs," in *ACM International Conference Proceeding Series*, 2019, vol. Part F1481, doi: 10.1145/3320326.3320330.

[17] J. Qiu, J. Peng, and Y. Zhai, "Network community detection based on spectral clustering," in *Proceedings - International Conference on Machine Learning and Cybernetics*, 2014, vol. 2, pp. 648–652, doi: 10.1109/ICMLC.2014.7009685.

[18] M. Shutaywi and N. N. Kachouie, "Silhouette analysis for performance evaluation in machine learning with applications to clustering," *Entropy*, vol. 23, no. 6, pp. 1–17, 2021, doi: 10.3390/e23060759.

[19] L. da F. Costa, "Further Generalizations of the Jaccard Index," 2021, [Online]. Available: http://arxiv.org/abs/2110.09619.

10.1016/j.compenvurbsys.2014.11.002.

[7] P. Mane, S. Shanbhag, T. Kamath, P. Mackey, and J. Springer, "Analysis of Community Detection Algorithms for Large Scale Cyber Networks," *Proc. 2016 Inf. Secur. Res. Educ. Conf.*, no. 2016, 2016.

[8] X. Hu, J. Han, and N. Cercone, "Discovering Cyber Communities from the WWW," *Proc. - IEEE Comput. Soc. Int. Comput. Softw. Appl. Conf.*, pp. 590–594, 2003, doi: 10.1109/cmpsac.2003.1245400.

[9] L. Hu, X. Pan, Z. Tang, and X. Luo, "A Fast Fuzzy Clustering Algorithm for Complex Networks via a Generalized Momentum Method," *IEEE Trans. Fuzzy Syst.*, vol. 30, no. 9, pp. 3473–3485, 2022, doi: 10.1109/TFUZZ.2021.3117442.

[10] T. R. Smith and N. Bosanac, "Constructing a set of motion primitives in the circular restricted three-body problem via clustering," *Adv. Astronaut. Sci.*, vol. 171, pp. 1283–1302, 2020.

[11] N. Nejjari, S. Lahlou, O. Fadi, K. Zkik, M. Oudani, and H. Benbrahim, "Conflict spectrum: An empirical study of geopolitical cyber threats from a social network perspective," *2021 8th Int. Conf. Soc. Netw. Anal. Manag. Secur. SNAMS 2021*, 2021, doi: 10.1109/SNAMS53716.2021.9732155.

[12] F. Gasparetti, G. Sansonetti, and A. Micarelli, "Community detection in social recommender systems: a survey," *Appl. Intell.*, vol. 51, no. 6, pp. 3975–3995, 2021, doi: 10.1007/s10489-020-01962-3.

[13] Amna, N. M. Nawi, M. Aamir, and M. F. Mushtaq, *The Comparative Performance Analysis of Clustering Algorithms*, vol. 457 LNNS. Springer International Publishing, 2022.

[14] U. Von Luxburg, "A tutorial on spectral clustering," *Stat. Comput.*, vol. 17, no. 4, pp. 395–416, 2007, doi: 10.1007/s11222-007-9033-z.

[15] K. Berahmand, M. Mohammadi, A. Faroughi, and R. P. Mohammadiani, "A novel method of spectral clustering in attributed networks by constructing parameter-free affinity matrix," *Cluster Comput.*, vol. 25, no. 2, pp.

Improving the Security of Cyber Networks Based on Community Detection Using Spectral Clustering Algorithm

(Received: ……... ; Accepted: ……)

Abstract

Cyber networks are considered to be a type of complex and free-scale networks due to the structure and mode of communication within the network. Identifying communities is one of the most important methods of network analysis in order to understand the structure and relationships between network members. With the development of cyber networks, new challenges have been created for users in terms of information security

One of the goals of identifying the communities in cyber networks is to prevent the spread of malware and cyber attacks. For this purpose, in order to prevent and deal with network attacks and intrusions, the communities in the network should be identified in order to significantly reduce the damage and attacks by attackers by securing and reviving the communities as well as implementing defense policies appropriate to each community. In this article, a method for detecting cyber communities by spectral clustering algorithm is presented. Also, by using the property of the normalized Laplace matrix in this algorithm, it is possible to predict the number of suitable cyber communities. In order to evaluate the detection process, two criteria Silhouette Value and Jaccard index are used. The results obtained from the evaluation criteria confirm the effectiveness of the proposed method.

Keywords: Cyber networks, Information security, Network analysis, Community detection, Spectral Clustering.

نشریه علمی "پدافند الکترونیکی و سایبری"
سال نهم، شماره ۲، تابستان ۱۴۰۰، ص ۷-۱