نشریه علمی "پدافند الکترونیکی و سایبری"
سال هفتم، شماره ۲، تابستان ۱۳۹۸، ص ۳۲-۲۵

روش بهبودیافته برای بی‌نام‌سازی داده‌های گراف با تأکید بر حفظ متوسط فاصله بین گره‌ها

سیده حمیده عرفانی[۱]، رضا مرتضوی[۲*]

۱- مربی، ۲- استادیار، دانشکده فنی و مهندسی، دانشگاه دامغان

(دریافت: ۹۷/۰۱/۱۹، پذیرش: ۹۷/۰۷/۲۱)

چکیده

در دهه اخیر هم‌زمان با افزایش نیاز به نمایش انواع داده به‌صورت گراف، مانند داده‌های شبکه‌های اجتماعی یا شبکه‌های ارتباطی، بسیاری از محققین به ارائه روش‌های حفظ حریم خصوصی در این نوع از اطلاعات و داده‌ها پرداخته‌اند. اگرچه که تاکنون روش‌های زیادی برای بی‌نام‌سازی داده‌های جدولی ارائه شده است، اما موضوع بی‌نام‌سازی گراف را می‌توان جزء موضوعات جدید تحقیقاتی به‌حساب آورد. یکی از مدل‌های مهم برای جلوگیری از خطر افشای شناسه کاربران در داده‌های گراف به تعداد اتصالات همسایه‌های یک گره در گراف توجه دارد. در این مقاله یک روش بی‌نام‌سازی جهت حفظ حریم خصوصی بر اساس این مدل پیشنهاد می‌شود که در آن بر اساس افزایش و کاهش یال‌های گراف خطر نقض حریم خصوصی کاهش می‌یابد. اعمال روش فوق بر روی تعدادی گراف‌های واقعی با اندازه‌های متفاوت نشان می‌دهد که روش فوق می‌تواند سودمندی بهتری را برحسب یکی از معیارهای مهم در گراف‌ها یعنی متوسط فاصله گره‌ها از یکدیگر به‌دست آورد و ساختار گراف تغییرات کمتری خواهد داشت.

کلیدواژه‌ها: حریم خصوصی، داده‌های شخصی، شبکه‌های اجتماعی، بی‌نام‌سازی گراف

۱- مقدمه

پدیده‌های زیادی وجود دارند که می‌توان آن‌ها را به‌صورت گراف نمایش داد. شبکه‌های اجتماعی، شبکه‌های ارتباطی و یا حتی اینترنت مثال‌هایی هستند که می‌توانند اهمیت مطالعه بر روی داده‌های گراف را به‌روشنی نشان دهند [۱]. به‌عنوان مثال با مدل کردن شبکه‌های اجتماعی مانند فیس‌بوک، توییتر و ... به‌صورت ساختمان داده گراف که در آن رأس‌های گراف معرف کاربران و یال‌های گراف بیان‌کننده روابط اجتماعی میان کاربران شبکه است، می‌توان تحلیل‌های بسیار مهمی را در زمینه پدافند غیرعامل انجام داد. به‌علاوه، تجزیه و تحلیل این نوع از اطلاعات برای بسیاری از اهداف و کاربردها مانند داده‌کاوی، تبلیغات، پشتیبانی از تصمیمات تجاری و دیگر کاربردهای تحقیقاتی مهم است [۲]. از طرف دیگر شرکت‌های تجاری صاحب چنین داده‌های باارزشی از انتشار این اطلاعات بدون تغییر امتناع می‌ورزند، زیرا بسیاری از این پایگاه داده‌های گراف-شکل را حاوی اطلاعات محرمانه‌ای راجع به افراد می‌دانند که نباید انتشار یابند. به همین دلیل بر روی داده‌های گراف-شکل معمولاً قبل از انتشار، عملیات بی‌نام‌سازی انجام می‌گیرد. در دهه اخیر بسیاری از محققین به انجام تحقیقات در زمینه حفظ حریم خصوصی کاربران شبکه‌های اجتماعی پرداخته‌اند [۴-۳].

به‌منظور جلوگیری از انتشار اطلاعات حریم خصوصی داده‌های گراف-شکل، روش‌های بی‌نام‌سازی تغییراتی در گراف اصلی ایجاد می‌کنند. این تغییرات می‌تواند شامل اضافه کردن و یا کاهش تعداد رأس‌ها و یا یال‌های گراف باشد. درحالی‌که این تغییرات در بی‌نام‌سازی باید به‌گونه‌ای انجام شود که تحلیل درست اطلاعات گراف را مخصوصاً در سطح معیارهای کلان[۱] (ماکروسکوپیک) به دنبال داشته باشد. به‌عبارت‌دیگر گراف بی‌نام شده باید تا حد امکان به گراف اصلی نزدیک باشد، بدون آن‌که بر اساس اطلاعات منتشرشده امکان انتساب برچسب به شخص یا موجودیتی در دنیای واقعی و با احتمال صحت بالا وجود داشته باشد. بنابراین، در مسئله بی‌نام‌سازی گراف، باید تعادلی میان حریم خصوصی صاحبان داده‌ها و نیز کیفیت داده‌ها ایجاد گردد [۵]. با توجه به ذات انتزاعی مفهوم حریم خصوصی، مدل‌های متعددی برای بی‌نامی گراف تعریف شده است. همچنین برای عملیاتی کردن این مدل‌ها و برآوردن نیازمندی‌های آن‌ها، روش‌های متعددی پیشنهاد شده است. این روش‌ها مبتنی بر عملگرهای تغییردهنده مختلفی در گراف هستند که بر روی ویژگی‌های مهم گراف تأثیر می‌گذارند. ازآنجاکه تغییر در تعداد رأس‌های گراف به معنی اضافه کردن و یا کاهش موجودیت(هایی) در گراف اصلی است که در دنیای واقعی وجود خارجی ندارند، در این مقاله تنها بر روی روش‌های افزایش و کاهش یال تمرکز می‌کنیم. به‌بیان‌دیگر، اگر گراف موردنظر

* رایانامه نویسنده مسئول: r_mortazavi@du.ac.ir

1- Macroscopic

نشان‌دهنده وجود ارتباط بین تعدادی موجودیت باشد (مانند گراف شبکه‌های اجتماعی)، تعدادی ارتباطات غیرواقعی به‌صورت یال‌های جدید به گراف اولیه اضافه می‌شود یا تعدادی از ارتباط‌های واقعی در نمایش گرافی حذف می‌شوند.

در این مقاله بر اساس مدل حریم خصوصی (K, L)–بی‌نامی که توسط Feder و همکاران [۶] و به‌عنوان نسخه ساده‌شده‌ای از مدل K–بی‌نامی ارائه شده است، روی بی‌نام‌سازی گراف‌ها برای جلوگیری از شناسایی هویت کاربران بر اساس اطلاع مهاجم از شناسه یک همسایه از گره قربانی در ضمن حفظ سودمندی گراف متمرکز می‌شویم. به‌طور خاص در این مقاله مدل (K, L)–بی‌نامی به ازای $L = 1$ به دو روش بر روی تعدادی پایگاه داده اعمال شده است: روش اول (مشابه روش Feder و همکارانش [۶] ولی مبتنی بر برنامه‌ریزی ریاضی) و تنها با اضافه کردن یال به گراف، آن را بی‌نام می‌کند ولی روش دوم که به‌عنوان نوآوری این مقاله ارائه شده است، به‌منظور حفظ ویژگی‌های اصلی گراف اولیه، از هر دو عملگر افزودن و سپس حذف یال از گراف استفاده می‌کند تا بتواند ویژگی مهم متوسط فاصله بین گره‌ها را حفظ کند.

ادامه مقاله به‌صورت زیر سازمان‌دهی شده است. در بخش ۲، مروری بر کارهای گذشته در زمینه بی‌نام‌سازی گراف آورده شده است. بخش ۳، برخی مفاهیم پایه‌ای را یادآوری می‌کند. در بخش ۴، روش پیشنهادی شرح داده می‌شود. در بخش ۵، نتایج عددی به‌دست‌آمده از اعمال روش پیشنهادی و مقایسه آن با روش مشابه گزارش می‌شود. بخش ۶ نیز به جمع‌بندی مقاله می‌پردازد.

۲- کارهای مرتبط

رایج‌ترین روش‌های ارائه‌شده به‌منظور حفظ حریم خصوصی اطلاعات موجود در گراف، روش‌های بی‌نام‌سازی هستند که فرآیند اجرای آن‌ها را می‌تواند به‌صورت کاملاً تصادفی و یا شامل تابع هدف دارای محدودیت در نظر گرفت [۵]. روش‌های تصادفی بر اساس ایجاد نویز تصادفی در داده‌های اصلی عمل می‌کنند. ایجاد نویز داده‌های گراف-شکل دارای دو رویکرد کلی است: (۱) اضافه یا حذف یال در گراف اصلی: این روش تعداد گره‌های گراف را حفظ می‌کند، (۲) جابه‌جایی تصادفی. در این رویکرد یال‌ها میان زوج نودها جابه‌جا می‌شوند. این راهبرد درجه رأس‌ها را حفظ کرده و تعداد یال‌ها را افزایش می‌دهد. طبیعتاً فن تغییر با یال‌های تصادفی را می‌توان به‌عنوان یک روش درهم‌ریختگی مبتنی بر نویز افزایشی در نظر گرفت [۵]. Hay و همکارانش [۷] روشی را به نام درهم‌آمیختگی تصادفی به‌منظور بی‌نام‌سازی گراف بدون برچسب ارائه کردند که بر مبنای حذف P یال از گراف و سپس اضافه کردن P یال کاذب به گراف، به‌صورت تصادفی می‌باشد. Ying و Wu [۸] روشی را پیشنهاد کرده‌اند که خصوصیات طیفی گراف اصلی را حفظ می‌کند. همچنین Ying و همکارانش [۹] روش افزایش و حذف یال به‌صورت تصادفی ارائه کردند که در آن گراف بر اساس دنباله درجه‌ها به بلوک‌هایی تقسیم شده و تغییرات به‌صورت تصادفی بر روی هر بلوک اجرا می‌شود (نه بر روی مجموعه کلیه رأس‌های گراف).

یکی دیگر از راه‌کارهای به‌کار گرفته‌شده در جهت بی‌نام‌سازی گراف اضافه و حذف یال به‌منظور رفع محدودیت‌های مشخصی است، که معمولاً این محدودیت‌ها شامل رسیدن حریم خصوصی گراف به یک سطح مشخص و بر اساس نیازمندی‌های مدل حریم خصوصی است. به‌عنوان مثال تعریف K-بی‌نام‌سازی که به‌منظور حفظ حریم خصوصی توسط Sweeney [۱۰] ارائه شده است را در نظر بگیرید. با توجه به این تعریف یک مهاجم نمی‌تواند با احتمال بیشتر از $1/K$ یک فرد را در پایگاه داده جدولی شناسایی کند. ایده اصلی پیاده‌سازی این مدل بر روی داده‌های گراف این است که درجه رأس‌ها را به‌عنوان شبه شناسه در نظر می‌گیرد و معادل با آن مدل بی‌نامی K–درجه را تعریف می‌کند. این روش مبتنی بر اضافه کردن و حذف یال از گراف است به‌طوری‌که برای کلیه رأس‌های گراف حداقل $K-1$ رأس دیگر با همان درجه وجود داشته باشد [۱۰]. Liu و Terzi روشی را پیشنهاد دادند که گرافی مانند G و عدد صحیح K را دریافت کرده و گراف K–بی‌نام را به‌گونه‌ای می‌سازد که تعداد تغییرات یال‌ها به حداقل برسد. در مرجع [۱۱]، به ازای هر رأس گراف، زیرگراف ۱-همسایگی در نظر گرفته می‌شود که بی‌نام‌سازی K–همسایگی نامیده می‌شود. Feder و همکارانش [۶] سطح جدیدی از بی‌نام‌سازی را با عنوان (K, L)–بی‌نام‌سازی معرفی کردند که بر اساس آن گراف G، (K, L)–بی‌نام نامیده می‌شود اگر به‌ازای هر رأس در گراف حداقل K رأس دیگر وجود داشته باشد که حداقل L رأس همسایه را به اشتراک گذاشته باشند. ایده اصلی این تعریف از بی‌نام‌سازی این است که کلیه رأس‌های گراف تعدادی از همسایگان خود را با K رأس دیگر به اشتراک بگذارند. بدیهی است بر اساس مدل‌های متفاوتی از مهاجمین و سطح اطلاعات مفروض ایشان، مدل‌های فوق کارایی خواهند داشت. همچنین در یکی از جدیدترین کارهای انجام شده در این حوزه رحمتی و همکاران به خوشه‌بندی محتوایی- ساختاری گراف پرداخته‌اند [۱۲].

۳- مفاهیم پایه‌ای

در این بخش برخی مفاهیم مقدماتی مورداستفاده در این مقاله یادآوری می‌شود.

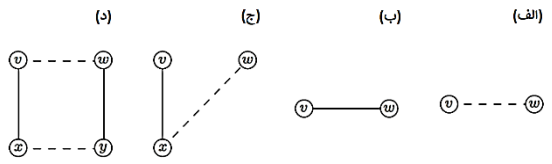
۳-۱- تعریف گراف

ساختمان داده $G = (V, E)$ را گراف ساده و بدون جهت می‌نامیم که در آن، V نشان‌دهنده مجموعه رأس‌های (گره‌های) گراف و E

مجموعه یال‌های گراف است. n را تعداد رأس‌های گراف و m را تعداد یال‌های گراف در نظر می‌گیریم. تعداد گره‌های همسایه یک گره را درجه آن گره در نظر می‌گیریم. کوتاه‌ترین فاصله بین دو رأس در G عبارت است از حداقل تعداد یال‌هایی که باید در یک مسیر با شروع از یکی از آن دو رأس طی شود تا به رأس دیگر رسید. می‌توان بین تمام رئوس گراف این کمترین فاصله را در مرتبه زمانی $O(n^3)$ با الگوریتم فلوید-وارشال به‌دست آورد. متوسط این کوتاه‌ترین فاصله‌ها APL[1] خوانده می‌شود که به‌عنوان یکی از ویژگی‌های مهم گراف شناخته می‌شود که حفظ آن در ضمن اعمال حریم خصوصی می‌تواند برای تحلیل‌های آتی حیاتی باشد. متوسط درجه رئوس را با AVD[2] نشان می‌دهیم. همچنین نسبت m یال از گراف G به حداکثر یال‌های گراف کامل با n رأس را با عنوان Density آن گراف می‌شناسیم. یکی از معیارهای مهم در گراف، مقدار betweenness مربوط به یال‌هاست. این معیار بیان می‌کند که چه تعداد از کوتاه‌ترین مسیرها بین گره‌های گراف از یک یال مشخص می‌گذرند. به‌طور ضمنی هرچقدر این مقدار برای یالی بیشتر باشد، انتظار می‌رود حذف آن یال تأثیر بیشتری روی APL داشته باشد، زیرا تعداد بیشتری از مسیرها را تحت تأثیر قرار می‌دهد.

۳-۲- تغییر در مجموعه یال‌های گراف

بسیاری از روش‌های حفاظت از حریم خصوصی ازجمله روش‌های بی‌نام‌سازی، مبتنی بر اصلاح یال[3] (اضافه کردن یال و یا حذف یال) هستند. چهار عملگر اصلاح یال وجود دارد که در شکل (۱) نشان داده شده است [۵]. معمولاً در روش‌های بی‌نام‌سازی از دو رویکرد افزودن/حذف یال استفاده می‌شود و چرخش و تعویض یال را می‌توان به‌عنوان نمونه‌های خاصی از روش افزودن/حذف یال در نظر گرفت. همان‌طور که در شکل (۱) نشان داده شده است تعویض یال نه‌تنها تعداد یال‌های گراف را تغییر نمی‌دهد بلکه درجه رئوس تحت تأثیر اصلاح را نیز بدون تغییر نگه می‌دارد که این امر در حفظ حریم خصوصی مؤثر خواهد بود.



شکل (۱): عملگرهای پایه‌ای اصلاح یال، خطوط ممتد بیانگر یال‌های موجود و خطوط خط‌چین نمایانگر یال‌های جدید جهت اضافه شدن به گراف هستند. (الف) اضافه کردن یال، (ب) حذف یال، (ج) چرخش یال، (د) تعویض یال [۵].

۳-۳- مدل مهاجم و مدل (K,L)-بی‌نامی

یکی از نگرانی‌های صاحبان داده‌ها در شبکه‌های اجتماعی این است که اگر شناسه یکی از دوستان آن‌ها توسط مهاجم در داده‌های منتشرشده افشا شود، مهاجم خواهد توانست با احتمال بالایی خود آن گره را نیز شناسایی کند. این حالت زمانی اتفاق می‌افتد که همسایه قربانی دارای تعداد همسایه کمی باشد و وجود ارتباط با وی برای مهاجم مسلم باشد. بدیهی است هرچقدر تعداد گره‌های بیشتری توسط مهاجم شناسایی شوند، احتمال شناسایی یک گره به‌عنوان همسایه مشترک این مجموعه بالاتر می‌رود. Feder و همکارانش در [۶] مدل جدیدی از بی‌نام‌سازی در گراف را با عنوان (K,L)-بی‌نامی معرفی کردند که درواقع برای مقابله با مهاجمی با توانایی فوق طراحی شده است. طبق تعریف یک گراف(K,L)-بی‌نام نامیده می‌شود اگر به‌ازای هر رأس v در گراف، حداقل K رأس دیگر وجود داشته باشد که حداقل L رأس همسایه را با v به اشتراک گذاشته باشند. ایده اصلی این تعریف از بی‌نام‌سازی این است که کلیه رأس‌های گراف تعدادی از همسایگان خود را با K رأس دیگر به اشتراک بگذارند. در این مقاله به‌ازای $L = 1$ و مقادیر مختلفی از K، روشی در جهت بی‌نام‌سازی گراف با توجه به تعریف فوق ارائه شده است. به‌طور خاص $(K,1)$-بی‌نامی به این معنی است که درجه کلیه گره‌های گراف حداقل K باشد.

۴- روش پیشنهادی

در این بخش از مقاله به شرح روش پیشنهادی در دو مرحله اضافه کردن یال و حذف یال می‌پردازیم که هم‌زمان با حفظ حریم خصوصی در گراف، کیفیت گراف را نیز در بالاترین سطح ممکن نگه می‌دارد. در ادامه نیز ویژگی‌های پایگاه داده‌های مورد استفاده در این مقاله شرح داده شده است. همان‌طور که گفته شد در مسئله بی‌نام‌سازی به دنبال ایجاد تعادلی میان حفظ حریم خصوصی و کیفیت اطلاعات هستیم. به‌عبارت دیگر به دنبال ایجاد کمترین تغییرات در گراف اصلی هستیم به نحوی که سطحی از حریم خصوصی برآورده شود. به‌همین دلیل ایده اصلی روش پیشنهادی بر مبنای یک مدل ریاضی شکل گرفته است که به دنبال کمترین تغییرات (حفظ کیفیت اطلاعات) و هم‌زمان برآورده کردن قیود مشخصی (حفظ حریم خصوصی) می‌باشد. در ادامه شرح جزئیات روش پیشنهادی در دو مرحله اضافه کردن یال به گراف و حذف یال از گراف آورده شده است.

1- Average Path Length
2- Average Vertices Degree
3- Edge Modifications

۴-۱- مرحله اضافه کردن یال

در این مرحله با توجه به تعریف بی‌نام‌سازی مطرح شده توسط Feder و همکارانش [۶]، پیاده‌سازی مفهوم $(K,1)$–بی‌نام‌سازی با استفاده از یک مدل ریاضی انجام گرفته است که هدف از آن اضافه کردن حداقل تعداد یال به گراف اصلی است به‌طوری‌که گراف $(K,1)$–بی‌نام به‌دست آید. از آنجایی که مدل ریاضی یک مسئله بهینه‌سازی را ارائه می‌دهد که با استفاده از نرم‌افزار گمز[1] راه‌حل بهینه به‌دست خواهد آمد. در قسمت قیود مسئله نیز از تعریف (K,L)–بی‌نام‌سازی مرجع [۶] به ازای $L=1$ استفاده شده است. مدل ریاضی روش پیشنهادی در رابطه (۱) قابل مشاهده می‌باشد.

$$\min_{x_{ij}} \sum_{\substack{1\le i,j\le n \\ i<j}} x_{ij}$$
$$s.t$$
$$\sum_{i<j} x_{ij} + \sum_{i>j} x_{ji} \ge K, \qquad 1 \le i,j \le n\,, \deg(v_i) \le K \qquad (۱)$$
$$x_{ij} = 1\,, \qquad \forall e_{ij} \in E$$
$$x_{ij} \in \{0,1\}$$

بدیهی است که مسئله رابطه (۱) حل‌پذیر (feasible) است، زیرا با فرض $x_{ij}=1$ تمام قیود سخت برآورده می‌شوند، لیکن گراف اولیه به یک گراف کامل تبدیل می‌شود که احتمالا به لحاظ سودمندی نامطلوب است زیرا تغییرات زیادی نسبت به گراف اولیه در آن انجام شده است. در حالت کلی، مسئله فوق که جزو مسائل برنامه‌ریزی صفر و یک محسوب می‌شود، جزو مسائل NP-Complete محسوب می‌شود [۱۳] و نمی‌توان آن را در حالت کلی به صورت کارآمد حل نمود، لیکن برای کاربردهای عملی راه‌کار شاخه و کران (Branch and Bound) وجود دارد که این مقاله نیز از آن بهره می‌برد. به‌طور خلاصه در این روش محدوده فضای مسئله بر اساس مقدار یک متغیر به دو زیرفضا تقسیم می‌شود و جواب‌ها در این زیرفضاها جستجو می‌شوند. برای بهبود کارایی و جلوگیری از بررسی همه حالات، چنانچه الگوریتم در یک زیرفضا به کران بالا یا پایین محدودتری نسبت به بهترین کران‌های قبلی رسید از آن‌ها برای محدود کردن فضای جستجو در سایر زیرفضاهای باقیمانده کمک می‌گیرد تا همه حالات بررسی نشوند.

در رابطه (۱) x_{ij} متغیر تصمیم دودویی مسئله است که بیانگر اتصال $(x_{ij}=1)$ یا عدم اتصال $(x_{ij}=0)$ دو راس v_i و v_j می‌باشد. جهت حفظ یال‌های اولیه گراف و نیز حل سریع‌تر مسئله بهینه‌سازی، در ابتدا به متغیرهای تصمیم متناظر با یال‌های گراف اصلی مقدار اولیه ۱ داده شده است. همچنین از آنجایی که گراف را ساده و بدون جهت فرض کردیم، جهت کاهش حجم محاسبات تنها به‌ازای $i<j$ مسئله را حل خواهیم کرد زیرا داریم $(x_{ij}=x_{ji})$. در قسمت محدودیت‌های مسئله رسیدن به سطحی از حریم خصوصی مدنظر می‌باشد که پارامتر آن یعنی K در تعریف مدل مرجع داده شده است. یادآوری می‌شود در این مدل کلیه رئوس گراف با حداقل K رأس دیگر یک همسایه مشترک خواهند داشت که به موجب این ویژگی در گراف، مهاجم به ازای دانستن شناسه مربوط به یک رأس v به حداقل K رأس همسایه v خواهد رسید. به این ترتیب از شناسایی رأس‌های همسایه به‌صورت منحصر به فرد جلوگیری خواهد شد. این محدودیت موجب می‌شود کمترین درجه هر رأس گراف حداقل K باشد. در حل رابطه (۱) عملیات اصلاح یال در گراف بی‌نام شده به‌ازای اضافه کردن یال در گراف اصلی انجام می‌گیرد (زیرا محدودیت $x_{ij}=1\,,\ \forall e_{ij}\in E$ تضمین‌کننده عدم حذف یال‌ها در گراف اولیه است). حل مسئله مذکور با استفاده از حل‌کننده صنعتی $CPLEX$ به راه‌حل بهینه خواهد رسید. این بدان معنی است که هیچ یالی بیهوده به گراف اضافه نخواهد شد. نتایج عددی تا این مرحله از روش پیشنهادی مشابه روش آقای Feder و همکارانش [۶] است.

۴-۲- مرحله حذف یال

از آنجایی که اضافه کردن یال به گراف ویژگی‌هایی از گراف را مانند تعداد یال‌ها، چگالی یال‌ها، متوسط درجه رئوس، متوسط کوتاه‌ترین مسیر میان گره‌های گراف و ... را دچار تغییر می‌کند، به‌منظور کاهش تفاوت میان گراف اصلی و گراف بی‌نام شده به دست آمده از مرحله اضافه کردن یال، مرحله‌ای برای حذف یال نیز در روش پیشنهادی در نظر گرفته شده است که در آن این امکان وجود دارد که از میان یال‌های گراف اصلی برخی از آن‌ها با حفظ ویژگی (K,L)–بی‌نامی از گراف حذف شوند. به عبارت دیگر پس از افزودن یال‌های جدید معیار متوسط درجه رئوس و متوسط کوتاه‌ترین فاصله نسبت به گراف اصلی به ترتیب افزایش و کاهش می‌یابند که برای کاهش این تغییرات امکان حذف یال‌های اولیه گراف اصلی بررسی می‌گردد. البته به‌منظور حذف هر یال و در صورت عدم ایجاد مشکل در بی‌نامی گراف، حداکثر به تعداد یال‌های اضافه‌شده، یال حذف خواهد شد. در Algorithm 1 بررسی حذف یال (v,w) انجام می‌گیرد و به‌ازای کلیه یال‌های قابل حذف گراف، بررسی فوق در Algorithm 2 تکرار شده تا در نهایت گراف بی‌نام شده جدید حاصل گردد.

1- GAMS Software

Algorithm 1: Check edge deletion
Function EdgeDeletionCheck(v, w, M, K)
Input: An edge e = (v, w) ∈ E, Graph adjacency matrix M, (K,1)-anonymity, model parameter K.
Output: A boolean indicating whether the input edge has been deleted.
1. flag = **true**;
2. M(v,w) = M(w,v) = 0; %delete edge (w,v) %
3. **if** (sum (M(v,:)) < K or sum (M(w,:)) < K)
4. flag = **false**;
5. **end**
6. **return** flag;
end

در خطوط ۳ تا ۵ بررسی می‌شود که حذف یال (v,w) برای رئوس v و w از نظر مدل حریم خصوصی مشکلی ایجاد می‌کند یا خیر. در صورتی که درجه هر یک از دو رأس v و w کمتر از K (پارامتر مدل) شود، مقدار flag برابر false خواهد شد وگرنه مقدار true بر می‌گردد.

Algorithm 2: Edge Deletion
Function EdgeDeletion(M, K, NumAddedEdges, DelAbleEdges)
Input: Graph adjacency matrix M, (K,1)-anonymity, Number of added edges, List of deletable edges ∈ E.
Output: New (K,1)-anonymity graph
1. Num = length (DelAbleEdges);
2. counterDel = 1;
3. sort DelAbleEdges based on their betweenness values
4. **for** i = 1: Num
5. **if** (counterDel <= NumAddedEdges)
6. DelFlag = EdgeDeletionCheck(DelAbleEdges(i,1), DelAbleEdges(i,2), M, K);
7. **if** (DelFlag)
8. M(DelAbleEdges(i,1), DelAbleEdges(i,2)) = 0;
9. M(DelAbleEdges(i,2), DelAbleEdges(i,1)) = 0;
10. counterDel = counterDel +1;
11. **end**
12. **end**
13. **end**
14. **return** graph(M);
End

در Algorithm 2 کلیه یال‌های قابل حذف از گراف (با توجه به مدل حریم خصوصی) با نام DelAbleEdges به همراه تعداد یال‌های اضافه شده به گراف (در مرحله اضافه کردن یال) با عنوان NumAddedEdges به الگوریتم ارسال می‌گردد. ابتدا این یال‌ها به صورت صعودی بر حسب مقدار betweenness مرتب می‌شود تا در حلقه بعدی امکان حذف آن‌ها از گراف بررسی شود. یادآوری می‌شود انتظار می‌رود حذف یال با کمترین مقدار betweenness تأثیر منفی کمی در ویژگی‌های فاصله‌ای در گراف داشته باشد و بنابراین، در یک رویکرد حریصانه امکان حذف آن‌ها بررسی خواهد شد. در حلقه تکرار خطوط ۴ تا ۱۳ به ازای کلیه یال‌های قابل حذف، اگر تعداد یال‌های حذف شده (counterDel) کمتر از تعداد یال‌های اضافه شده (NumAddedEdges) باشد، Algorithm 1 به بررسی امکان حذف یال‌های کاندید حذف می‌پردازد (خط ۶). در صورتی که خروجی Algorithm 1 برابر false باشد یال بعدی بررسی می‌شود و در صورت true بودن خروجی، آن یال از گراف حذف خواهد شد (خط ۸ و ۹).

۵- نتایج تجربی

در این بخش نتایج به‌دست آمده از اعمال روش پیشنهادی و روش مرجع گزارش می‌شود.

۵-۱- پایگاه داده‌های مورد استفاده

در این مقاله از چهار پایگاه داده معتبر، پرکاربرد و حقیقی به منظور آزمایش روش پیشنهادی استفاده شده است.

- Polbooks: شامل شبکه‌ای از کتاب‌های مربوط به سیاست‌های آمریکا که توسط سایت Amazon.com فروخته می‌شود [۱۴].
- Karate: شبکه‌ای که روابط میان ۳۴ عضو باشگاه کاراته Zachary را نشان می‌دهد [۱۵].
- Football: اطلاعات مربوط به شبکه بازی‌های فوتبال آمریکایی بین کالج‌های Division IA که به‌طور منظم در طول فصل پاییز سال ۲۰۰۰ برگزار شده است [۱۶].
- Can_292: ماتریس مجاورت این گراف بر اساس الگوهای منظم در جشنواره کن توسط Lucien Marro جمع‌آوری شده است.

ویژگی‌های ساختاری این پایگاه داده‌ها از قبیل تعداد رأس، تعداد یال، چگالی، APL و نیز AVD گراف در جدول (۱) قابل مشاهده می‌باشد.

جدول (۱): مشخصات ساختاری پایگاه داده‌های گراف مورد آزمایش

G	n	m	density	APL	AVD
Karate	۳۴	۷۸	۰/۱۳۹۰	۲/۴۰۸۲	۴/۵۸۸۲
Polbooks	۱۰۵	۴۴۱	۰/۰۸۰۸	۳/۰۷۸۸	۸/۴۰۰۰
Football	۱۱۵	۶۱۶	۰/۰۹۴۰	۲/۵۰۸۲	۱۰//
Can_292	۲۹۲	۱۱۲۴	۰/۰۲۶۵	۴/۶۰۵۸	۷/۶۹۸۶

۵-۲- معیارهای ارزیابی

بسیاری از روش‌های ارائه‌شده در حوزه بی‌نام‌سازی گراف معیار حفظ کیفیت اطلاعات را میزان تغییرات در تعداد یال‌ها (میزان کاهش و یا افزایش یال‌ها) در نظر گرفته‌اند بدون آن‌که به ساختار گراف توجهی داشته باشند. در حالی که در روش ارائه شده در این مقاله علاوه بر تغییراتی که در تعداد یال‌ها و درنتیجه در چگالی گراف رخ می‌دهد، از معیارهایی جهت ارزیابی تغییرات

ساختاری گراف بی‌نام سازی شده نیز استفاده می‌گردد. در ادامه معیارهای ارزیابی روش پیشنهادی در مقابل روش Feder و همکارانش [۶] آورده شده است. در کلیه موارد مقدار کمتر نشان از برتری روش مربوطه است. علت انتخاب روش بی‌نام‌سازی Feder و همکارانش [۶] جهت مقایسه با روش پیشنهادی و ارزیابی آن، قطعی بودن و نیز نزدیک بودن رویکرد تابع هدف مطرح شده توسط ایشان به روش پیشنهادی این مقاله می‌باشد.

- تغییر در تعداد یال‌ها (Δm)،
- درصد تغییرات نسبی متوسط طول کوتاه‌ترین مسیر در گراف (APL) بی‌نام شده نسبت به گراف اصلی که در رابطه (۲) شرح داده شده است.

$$\Delta APL = \frac{|APL_{new} - APL_{orig}|}{APL_{orig}} \times 100\% \qquad (۲)$$

- درصد تغییرات نسبی متوسط درجه رئوس در گراف بی‌نام شده نسبت به گراف اصلی که از رابطه (۳) به دست خواهد آمد.

$$\Delta AVD = \frac{|AVD_{new} - AVD_{orig}|}{AVD_{orig}} \times 100\% \qquad (۳)$$

در این روابط اندیس $orig$ مربوط به گراف اصلی و اندیس new مربوط به گراف جدید (بعد از حذف یا افزودن یال) است.

۵-۳- تأثیر بی‌نام‌سازی روی متوسط طول مسیر

در این بخش نتایج به‌دست‌آمده از اعمال روش مرجع [۶] (به اختصار $M+$) و روش پیشنهادی (به اختصار $M\pm$) بر روی چهار پایگاه داده آورده شده در جدول (۵-۲) نمایش داده شده است. همان‌طور که مشاهده می‌شود برای هر پایگاه داده و به‌ازای مقادیر مختلف K، تغییر در تعداد یال‌ها، تغییرات متوسط درجه رئوس و نیز تغییرات نسبی متوسط طول کوتاه‌ترین مسیر در گراف بی‌نام شده به دو روش Feder و همکارانش [۶] و روش پیشنهادی قابل مشاهده می‌باشد. مقایسه دو روش $M+$ و $M\pm$ نشان می‌دهد روش پیشنهادی دارای خطای کمتری در متوسط درجه رئوس(ΔAVD) و نیز متوسط طول کوتاهترین مسیر(ΔAPL) می‌باشد. مقادیر پررنگ در جدول‌ها به معنای نتایج با خطای کمتر می‌باشد.

جدول (۲): نتایج $M+$ و $M\pm$ روی گراف Karate

k	M+			M±		
	Δm	ΔAPL	ΔAVD	Δm	ΔAPL	ΔAVD
۲	۱	۰/۱۴۸۰	۱/۲۸۲۱	۰	**۰/۰۷۴۰**	**۰/۰۰۰۰**
۳	۷	۵/۱۰۷۳	۸/۹۷۴۴	۰	**۴/۲۹۳۱**	**۰/۰۰۰۰**
۴	۱۶	۹/۱۰۴۴	۲۰/۵۱۲۸	۴	**۴/۶۶۳۲**	**۵/۱۲۸۲**
۵	۲۸	۱۴/۹۵۱۹	۳۵/۸۹۷۴	۲۰	**۱۲/۲۸۷۲**	**۲۵/۶۴۱۰**
۶	۴۱	۱۸/۶۵۲۸	۵۲/۵۶۴۱	۳۶	**۱۷/۲۴۶۵**	**۴۶/۱۵۳۸**
۷	۵۶	۲۲/۵۰۱۹	۷۱/۷۹۴۹	۵۱	**۲۱/۴۶۵۶**	**۶۵/۳۸۴۶**
۸	۷۰	۲۴/۳۵۲۳	۸۹/۷۴۳۶	۶۶	**۲۳/۷۶۰۲**	**۸۴/۶۱۵۴**
۹	۸۵	۲۷/۶۰۹۲	۱۰۸/۹۷۴۴	۸۳	**۲۷/۳۸۷۱**	**۱۰۶/۴۱۰۳**
۱۰	۱۰۰	۲۹/۲۳۷۶	۱۲۸/۲۰۵۱	۹۸	**۲۹/۰۸۹۶**	**۱۲۵/۶۴۱۰**

جدول (۳): نتایج $M+$ و $M\pm$ روی گراف Polbooks

k	M+			M±		
	Δm	ΔAPL	ΔAVD	Δm	ΔAPL	ΔAVD
۲	۰	۰/۰۰۰۰	۰/۰۰۰۰	۰	۰/۰۰۰۰	۰/۰۰۰۰
۳	۱	۰/۷۰۲۰	۰/۲۲۶۸	۰	**۰/۶۹۶۰**	**۰/۰۰۰۰**
۴	۴	۱/۲۰۷۶	۰/۹۰۷۰	۰	**۱/۱۸۳۸**	**۰/۰۰۰۰**
۵	۱۵	۵/۱۵۱۷	۳/۴۰۱۴	۰	**۵/۰۵۶۵**	**۰/۰۰۰۰**
۶	۳۶	۹/۸۳۳۴	۸/۱۶۳۳	۰	**۹/۵۲۴۱**	**۰/۰۰۰۰**
۷	۶۳	۱۴/۹۶۱۳	۱۴/۲۸۵۷	۰	**۱۳/۸۱۹۲**	**۰/۰۰۰۰**
۸	۹۵	۱۸/۱۸۵۶	۲۱/۵۴۲۰	۱۵	**۱۵/۶۵۱۴**	**۳/۴۰۱۴**
۹	۱۳۰	۲۱/۷۱۳۳	۲۹/۴۷۸۵	۶۴	**۱۹/۹۶۴۳**	**۱۴/۵۱۲۵**
۱۰	۱۷۰	۲۴/۴۴۳۸	۳۸/۵۴۸۸	۱۱۴	**۲۳/۴۰۲۷**	**۲۵/۸۵۰۳**

جدول (۴): نتایج $M+$ و $M\pm$ روی گراف Football

k	M+			M±		
	Δm	ΔAPL	ΔAVD	Δm	ΔAPL	ΔAVD
۲	۰	۰/۰۰۰۰	۰/۰۰۰۰	۰	۰/۰۰۰۰	۰/۰۰۰۰
۳	۰	۰/۰۰۰۰	۰/۰۰۰۰	۰	۰/۰۰۰۰	۰/۰۰۰۰
۴	۰	۰/۰۰۰۰	۰/۰۰۰۰	۰	۰/۰۰۰۰	۰/۰۰۰۰
۵	۰	۰/۰۰۰۰	۰/۰۰۰۰	۰	۰/۰۰۰۰	۰/۰۰۰۰
۶	۰	۰/۰۰۰۰	۰/۰۰۰۰	۰	۰/۰۰۰۰	۰/۰۰۰۰
۷	۰	۰/۰۰۰۰	۰/۰۰۰۰	۰	۰/۰۰۰۰	۰/۰۰۰۰
۸	۱	۰/۲۰۶۸	۰/۱۶۳۱	۰	**۰/۱۹۴۶**	**۰/۰۰۰۰**
۹	۳	۰/۲۴۹۴	۰/۴۸۹۴	۰	**۰/۲۲۵۰**	**۰/۰۰۰۰**
۱۰	۷	۰/۶۶۳۰	۱/۱۴۱۹	۰	**۰/۵۹۰۰**	**۰/۰۰۰۰**

جدول (۵): نتایج M+ و M± روی گراف Can_292

k	M+			M±		
	Δm	ΔAPL	ΔAVD	Δm	ΔAPL	ΔAVD
۲	۰	۰/۰۰۰۰	۰/۰۰۰۰	۰	۰/۰۰۰۰	۰/۰۰۰۰
۳	۰	۰/۰۰۰۰	۰/۰۰۰۰	۰	۰/۰۰۰۰	۰/۰۰۰۰
۴	۱۷	۴/۹۰۵۴	۱/۵۱۲۵	۰	**۴/۸۹۶۲**	**۰/۰۰۰۰**
۵	۴۰	۱۳/۲۸۵۱	۳/۵۵۸۷	۰	**۱۳/۲۶۴۷**	**۰/۰۰۰۰**
۶	۸۳	۱۹/۲۴۵۱	۷/۳۸۴۳	۰	**۱۹/۱۲۵۵**	**۰/۰۰۰۰**
۷	۱۵۲	۲۳/۲۱۰۹	۱۳/۵۲۳۱	۲۷	**۲۲/۷۳۷۸**	**۲/۴۰۲۱**
۸	۲۲۸	۲۸/۲۹۱۵	۲۰/۲۸۴۷	۱۶۲	**۲۷/۸۲۷۰**	**۱۴/۴۱۲۸**
۹	۳۲۶	۳۲/۶۱۸۳	۲۹/۰۰۳۶	۳۱۳	**۳۲/۴۹۵۱**	**۲۷/۸۴۷۰**
۱۰	۴۵۸	۳۷/۰۴۲۴	۴۰/۷۴۷۳	۴۵۴	**۳۷/۰۴۰۴**	**۴۰/۳۹۱۵**

همان‌طور که انتظار می‌رود، هم‌زمان با افزایش سطح حریم خصوصی بر اساس معیار K، تعداد کمتری یال را می‌توان حذف کرد، زیرا وجود یال‌ها برای کاهش احتمال موفقیت مهاجم در بازشناسی مهم است. البته با توجه به اینکه در کاربردهای عملی معمولاً K مقدار خیلی بزرگی نیست، این مورد مشکلی در استفاده عملی از روش پیشنهادی ایجاد نمی‌کند و همچنان این روش می‌تواند تأثیرگذار باشد. به‌علاوه، نتایج نشان می‌دهند، روش پیشنهادی همواره به درصد تغییرات نسبی بهتر یا مساوی در متوسط طول کوتاه‌ترین مسیرها بین گره‌های گراف دست یافته است. برای مثال در مجموعه داده Karate به‌ازای $K = 4$ در روش مربوط به Feder و همکارانش [۶] معیار APL نسبت به گراف اصلی به مقدار ۱۰۴۴/۹ درصد کاهش داشته است، اما در روش پیشنهادی و بعد از حذف تعدادی از یال‌ها، بدون این‌که نیازمندی‌های مدل حریم خصوصی نقض شود، این معیار تنها ۶۶۳۲/۴ درصد کاهش یافته است. این بهبود قابل‌توجه به این دلیل است که در روش Feder و همکارانش [۶] به تعداد ۱۶ یال جدید به گراف اضافه شده است که باعث کاهش قابل‌توجه APL می‌شود، اما با حذف ۱۲ یال از گراف اصلی، مجدداً APL افزایش می‌یابد و خطای کمتری را نشان می‌دهد. تحلیل مشابهی برای معیار AVD نیز برقرار است. برای مثال در مجموعه داده Can_292 بر اساس روش مرجع [۶]، AVD به طور نسبی بیش از ۱۳ درصد رشد داشته است، در حالی که در روش پیشنهادی این رشد تنها کمی بیشتر از ۲ درصد بوده است. به طور کلی از ۳۶ آزمایش انجام شده (بر روی چهار پایگاه داده مختلف و به ازای مقادیر صحیح $2 \le K \le 10$) در ۷۵ درصد موارد (۵۴ مورد از ۷۲ آزمایش)، معیارهای ΔAPL و ΔAVD بهبودیافته است (موارد پررنگ شده در جداول ۳ تا ۵) و در بقیه موارد نتیجه یکسانی در مقایسه با روش Feder و همکارانش [۶] به‌دست آمده است (موارد کم‌رنگ در جداول۳ تا ۵). به عبارت دیگر تغییرات APL و AVD در روش پیشنهادی در کلیه آزمایش‌ها نسبت به روش Feder و همکارانش [۶] تغییرات کمتر و یا مساوی آن داشته است.

۶- نتیجه‌گیری

در این مقاله روشی برای پیاده‌سازی نسخه‌ای از مدل حریم خصوصی (K, L)–بی‌نامی برای گراف‌های ساده ارائه شد. در روش پیشنهادی تأکید بر این بوده است که ضمن حفظ نیازمندی‌های مدل، ویژگی‌های مهم گراف یعنی تعداد یال‌ها، متوسط درجه رئوس و متوسط طول کوتاه‌ترین مسیرها بین گره‌ها حفظ شود. نتایج تجربی موفقیت روش پیشنهادی در ۷۵ درصد از موارد را (۵۴ مورد از ۷۲ آزمایش) در مقایسه با روش مشابه تائید می‌کند. روش ارائه شده را می‌توان در آینده در چندین جهت توسعه داد. چنانچه نسخه تغییریافته از مدل حریم خصوصی (K, L)–بی‌نامی در گراف‌های وزن‌دار تعریف شود، می‌توان روش فوق را برای حفظ نیازمندی‌های آن مدل جدید توسعه داد. استفاده از روش‌های ابتکاری برای اجرای سریع‌تر روش پیشنهادی بر روی گراف‌های خیلی بزرگ می‌تواند ازجمله کارهای آتی محسوب شود. پیاده‌سازی روش‌های مربوط به یادگیری ماشین در گراف‌ها با در نظر گرفتن مدل‌های حریم خصوصی نیز از کارهای قابل توجه برای بررسی در آینده به شمار می‌آید.

۷- منابع

[1] A. Padrola and V. Muntés-Mulero, "Graph anonymization via metric embeddings: Using classical anonymization for graphs," Intelligent Data Analysis, pp. 365–388, 2014.

[2] Sh. Ji, T. Wang, J. Chen, W. Li, P. Mittal, and R. Beyah, "De-SAG: On the De-anonymization of Structure-Attribute Graph Data," IEEE Transactions on Dependable and Secure Computing, pp. 1-14, 2016.

[3] A. Shaharkhiz and H. Shahriari, "A method for preserving privacy in published multi-relational social networks," International Conference on Knowledge Management and Information Sharing - KMIS 2011, France, 2011.

[4] E. Sargolzaee and M. Abdollahi Azgomi, "Preserving the Privacy of Published Social Networks Data Based on Graph Modification," Proc. of the 10th International ISC Conference on Information Security and Cryptology (ISCISC'13), pp. 30-37, 2013. (in Persian) .

[5] F. Rousseau, J. Casas-Roma, and M. Vazirgiannis, "Community-preserving anonymization of graphs," Knowledge and Information Systems, pp. 315-343, 2017.

[6] T. Feder, S. U. Nabar and E. Terzi, "Anonymizing graphs," CoRRabs. 0810/5578, 2008.

[7] M. Hay, G. Miklau, D. Jensen, P. Weis, and S. Srivastava, "Anonymizing social networks," Technical Report No. 07-19. Computer Science Department, University of Massachusetts Amherst, Amherst, 2007.

[8] X. Ying, K. Pan, X. Wu, and L. Guo, "Comparisons of randomization and k-degree anonymization schemes for privacy preserving social network publishing," In: Workshop on social network mining and analysis, pp. 1–10, 2009

[9] X. Ying and X. Wu, "Randomizing social networks: a spectrum preserving approach," In: Proceedings of the SIAM international conference on data mining, 2008.

[10] L. Sweeney, "k-anonymity: a model for protecting privacy". Int J Uncertain Fuzziness Knowl Based Syst, vol. 10, pp. 557–570, 2002.

[11] B. Zhou and J. Pei, "Preserving privacy in social networks against neighborhood attacks," In: Proceedings of the IEEE 24th international conference on data engineering, 2008.

[12] K. Rahmati, H. Naderi, and S. Keshvari, "Content-Structural Graph Clustering and a New Measure for Its Evaluation," Journal of Advanced Defence Science and Technology, vol. 9, pp. 201-210, 2018.

[13] R. M. Karp, "Reducibility among combinatorial problems," Complexity of computer computations, Springer, Boston, MA, pp. 85-103, 1972.

[14] www.orgnet.com

[15] W. W. Zachary, "An information flow model for conflict and fission in small groups," Journal of Anthropological Research, vol. 33, pp. 452–473, 1977.

[16] M. Girvan and M. E. Newman, "Community structure in social and biological networks," Proceedings of the national academy of sciences, vol. 99, pp. 7821–7826, 2002.

Journal of Electronical & Cyber Defence
Vol. 7, No. 2, 2019, Serial No. 26

An Improved Method for Graph Data Anonymization with Emphasis on Preserving the Average Path Length

S. H. Erfani, R. Mortazavi*

*School of Engineering, Damghan University, Damghan, Iran

(Received: 08/04/2018, Accepted: 13/10/2018)

ABSTRACT

In recent decades, in view of the widespread use of graph data in different applications, for instance in social networks, communications networks, etc. many researchers have investigated different anonymization approaches for such data. Although relational data anonymization is mature enough, graph data anonymization is a challenging and relatively new field of research. One of the most important anonymization models against identity disclosure risk in graph data addresses the number of links a node's neighbors have, in the graph. In this paper, an improved method is proposed that realizes this model using both edge addition and deletion to the original graph. The application of the method to a number of different real-world graphs confirms that the method can produce more useful graphs in terms of one of the most important characteristics in such data, i.e., the Average Path Length in the graph and graph structure will undergo less change.

Keywords: Privacy, Personal Data, Social Networks, Graph Anonymization

* Corresponding Author Email: r_mortazavi@du.ac.ir