یک روش گمنام‌سازی برای حفظ حریم خصوصی در انتشار داده‌های جریانی مسیر

علیرضا مودی[1]، مهری رجائی[2]

[1] دانشجوی کارشناسی ارشد، گروه فناوری اطلاعات، دانشگاه سیستان و بلوچستان ، زاهدان،
alireza_moodi@pgs.usb.ac.ir

[2] استادیار، گروه فناوری اطلاعات، دانشگاه سیستان و بلوچستان ، زاهدان،
rajayi@ece.usb.ac.ir

چکیده

امروزه حجم انبوهی از داده‌های مسیر حرکت افراد و اجسام از طریق اتصال به اینترنت به طور پیوسته و جریانی تولید و ذخیره می‌شود. تجزیه و تحلیل و استخراج دانش از داده‌های مسیر به‌روز می‌تواند برای پشتیبانی برنامه‌های مختلف اشیاء درحال حرکت و تحلیلگران مفید باشد. از سوی دیگر به دلیل اینکه داده‌های مسیر دارای ابعاد بالایی است و در بسیاری موارد پس از گذر زمان ممکن است جذابیت و اعتبار خود را از دست بدهد بنابراین انتشار داده‌های جریانی مسیر مورد توجه است اما، انتشار داده‌های مسیر حرکت اجسام مخاطره‌آمیز است. در این مقاله برای حل این مشکل، راه‌کاری برای گمنام‌سازی داده‌های جریانی مسیر بر اساس مدل حریم خصوصی k-گمنامی و l-گوناگونی ارائه شده است.

داده‌های مسیر درون پنجره‌های لغزان بصورت پویا بروزرسانی می‌شوند. برای مقابله با داده‌های با ابعاد بالا ترکیب زمان و مکان، و افزایش کارایی در روش پیشنهادی هر مسیر درون یک پنجره به یک رشته بیتی نگاشت می‌شود. سپس برای گروه‌بندی مسیرها یک الگوریتم حریصانه ارائه شده است که براساس معیار و شاخص تفاوت با سایر مسیرها که با استفاده از عملیات بیتی بر روی رشته‌های متناظر بدست می‌آید، مسیرها را در گروه‌هایی که نیازمندی‌های k-گمنامی و l-گوناگونی برآورده شده باشد، تقسیم می‌کند. و برای هر گروه یک مسیر برگزیده منتشر می‌شود. نتایج شبیه‌سازی نشان می‌دهد که روش پیشنهادی با برآورده کردن نیازمندی‌های حریم خصوصی به‌طور قابل توجهی زمان اجرا و اتلاف اطلاعات ناشی از گمنام‌سازی را در مقایسه با روش‌های موجود کاهش داده است.

کلمات کلیدی

داده‌های جریانی مسیر ، حفظ حریم خصوصی ، حملات حریم خصوصی ،جریان داده، از دست رفتن اطلاعات ، انتشار داده

1- مقدمه

در اینترنت اشیا ، موقعیت جغرافیایی دقیق یک شی و همچنین ابعاد دقیق جغرافیایی یک شی حیاتی خواهد بود[1]. بنابراین ، پیگیری داده‌های مرتبط با یک شی ، از جمله موقعیت مکانی آن در زمان و مکان ، از اهمیت بالایی برخوردار بوده است [2]زیرا فرد پردازش کننده اطلاعات می تواند تصمیم بگیرد که آیا این اطلاعات برای فرایند تصمیم گیری و اقدام ، مهم بوده است یا خیر[3].

به مرور زمان موقعیت زمانی-مکانی کاربران مختلف به صورت داوطلبانه یا غیرداوطلبانه در قالب مسیرهای حرکت اشیا متحرک به صورت خودکار جمع-آوری و ذخیره شده است. از اینرو، ردیابی آنها امکان‌پذیر خواهد شد. به عبارت دیگر، مسیرهای حرکت دنباله‌های زمانی-مکانی از اشیا متحرک هستند که در

طول ساعت‌ها، روزها، ماه‌ها و یا سال‌ها جمع‌آوری شده و شامل اطلاعات ارزشمندی برای کاربردهای مختلف داده‌کاوی می‌باشند[4]. منظور از اشیا متحرک، غالبا اشیا عینی (برای مثال، انسان، وسیله نقلیه، حیوان و محموله)، یا مفاهیم ذهنی (برای مثال بیماری‌های در حال شیوع و یا مکان صفحاتی از وب که مخاطب در حال وب‌گردی است) می‌باشد. از جمله کاربردهای داده-کاوی مسیرهای حرکت اشیاء متحرک، می‌توان به طرح‌ریزی ترافیک شهری، حمل و نقل هوشمند، شناسایی نواحی پر ازدحام یا کم ازدحام شهری یا جهت استفاده شرکت‌های تبلیغاتی و غیره اشاره کرد. در بسیاری از کاربردها یا نهاد مالک داده‌ها (ارائه‌دهنده سرویس) مجهز به ابزار داده‌کاوی نیست و یا نهاد علاقه‌مند به تحلیل داده‌ها جدا از نهاد مالک داده‌ها است. بنابراین در این موارد، ارائه‌دهنده سرویس یا مالک داده‌ها بایستی داده‌ها را به منظور تحلیل و داده‌کاوی منتشر کند. اما انتشار داده‌های مسیر افراد که پیوسته در حال تغییر است، حریم خصوصی افراد صاحب داده‌ها را به خطر می‌اندازد[4].

حفظ حریم خصوصی از موضوعات بسیار با اهمیت برای افراد و سازمان-ها است، به خصوص زمانی که سازمان‌ها با امر داده‌کاوی مواجه هستند و باید اطلاعات کاربران خود را برای مراکز داده‌کاوی ارسال کنند، تا دانش مورد نیاز آنها برای کاربردهای آینده از این داده‌ها استخراج شود. اما باید انتشار داده به گونه‌ای باشد که اطلاعات حساس کاربران از دسترس‌های غیر مجاز در امان بماند. زیرا ممکن است فرد متخاصم اطلاعاتی در مورد فرد هدف و قربانی خود از قبل داشته باشد، و با دستیابی به اطلاعات منتشر شده، فرد هدف را شناسایی و یا به اطلاعات حساس وی با احتمال قابل توجهی دست یابد. بنابراین نیاز است تا تحریفاتی در داده منتشر شده صورت گیرد که حریم خصوصی افراد حفظ شود، به اعمال این تحریفات گمنام‌سازی می‌گویند. اما گمنام‌سازی با وجود حفظ حریم خصوصی بایستی به گونه‌ای باشد که نتایج تحلیل‌های انجام شده بر روی داده گمنام شده، نزدیک به نتایج تحلیل‌های انجام شده بر روی داده اصلی باشد، یا به عبارت دیگر سودمندی داده حفظ شود[5].

امروزه، با توجه به گسترش اینترنت اشیاء، داده‌های مسیر به صورت پیوسته و مداوم با حجم بسیار بالا در حال تولید می‌باشد. بنابراین انبارکردن و ذخیره داده‌ها و بعد گمنام‌سازی آنها غیرعملی است، زیرا حجم اطلاعات مسیریابی در مقیاس‌های زمانی کم مثل ساعت، روز و یا هفته خیلی زیاد می-باشد، ویژگی ابعاد بالا بدین معناست که ترکیب مکان و زمان با یکدیگر، حالات بسیار زیادی را تولید می‌کند. یعنی اگر همه‌ی داده‌ها جمع‌آوری و سپس منتشر شود با چالش‌های زیر روبرو هستیم: اولا حجم داده‌های جمع-آوری شده خیلی بالاست; دوم، گمنام کردن این حجم از داده نیاز به زمان اجرای بالایی دارد. بدین ترتیب، فرآیند گمنام‌سازی سخت و دشوار خواهد شد و از همه مهمتر، ممکن است داده‌هایی که بعد از این مدت زمان منتشر می-شود «برای مثال داده‌های مسیرهای قبلی» صحت و اعتبار داده‌ای خود را از دست داده باشند و مورد توجه تحلیلگر نباشد[6]. زیرا در برخی از کاربردها تحلیلگر نیاز به داده‌های به‌روز و آنی برای تحلیل دارد. داده‌های جریانی دارای حجم زیادی داده می‌باشند، بنابراین داده‌ها را با توجه به زمان آن تقسیم می-کنند و در هر بازه زمانی مسیرهای مربوط به اشیایی که در آن بازه بوده‌اند را منتشر می‌کنند. برای انجام این کار داده‌ها جریانی وارد می‌شوند و تا حد یک پنجره ذخیره شده و گمنام شده و منتشر می‌شوند و بعد گام بعدی زمانی در نظر گرفته می‌شود. بنابراین مسئله حفظ حریم خصوصی در انتشار داده‌های جریانی مسیر مطرح می‌شود.

مدلهای حریم خصوصی نظیر k–گمنامی [7] وL–گوناگونی [7] ابتدا برای گمنام‌سازی داده‌های رابطه‌ای مطرح شد. این داده‌ها به صورت جداولی به صورت صفات شبه شناسه و صفت حساس تقسیم شده بودند. بونچی [7] روش‌های موجود برای گمنام‌سازی مسیرهای حرکت اشیاء متحرک را به دو دسته‌ی کلی تقسیم کرده است: روش‌های مبتنی بر خوشه‌بندی و اغتشاش و روش های مبتنی بر شبه‌شناسه. در دسته اول ، مساله‌ی گمنام‌سازی مسیرهای حرکت را بدون در نظر گرفتن مفهوم شبه‌شناسه‌ها مورد بررسی قرار داده و از مفهوم k–گمنامی برای این منظور استفاده کرده‌اند. روش‌های دسته دوم برخلاف دسته‌ی اول مفهوم شبه‌شناسه را در فرآیند گمنام‌سازی مسیرهای حرکت اشیاء متحرک مورد استفاده قرار داده‌اند[8, 9].

در زمینه انتشار داده‌های پیوسته، نسخه‌های بروز شده یک جدول داده به طور منظم به عنوان مثال به صورت هفتگی منتشر می‌شوند. ونگ و همکاران [10] روشی را برای داده‌های زمانی در قالب رابطه‌ای پیشنهاد کردند. روش آنها براساس جابه‌جایی رکورد زمانی بین نسخه‌های متعدد از یک پنجره منتشر شده بود. ژائو و تائو [11] این مشکل را در انتشار مجدد جدول داده‌های رابطه‌ای بروز شده با روش m– انحراف حل کردند . به اینصورت که درج و حذف رکورد در جدول داده‌های بروز شده اینگونه بود که اگر یک رکورد الزامات حریم‌خصوصی تحمیلی را برآورده نکند، رکوردهای جعلی به منظور دستیابی به m–انحراف در یک جدول بروز شده ایجاد می‌شوند . به کارگیری m–انحراف در جریان مسیر به این دلایل نامناسب هستند. اول، گمنام‌سازی m–انحراف در داده‌های رابطه‌ای است و مسیرها طبیعتا دارای ابعاد بالا هستند از این رو استفاده از روش‌های مبتنی بر مشخصه شبه‌شناسه کاهش اطلاعات قابل توجهی را متحمل می‌شود. دوم m–انحراف وجود داده جریانی را در نظر نمی‌گیرد، که این موضوع باعث می‌شود تا شرایط محدودیت زمانی کوتاه و سریع برای انتشار جدول داده‌های بروز شده را نداشته باشد، درحالیکه باید داده‌های جریانی را از طریق گمنام‌سازی و منتشر کردن داده-هایی که تازه رسیده‌اند به طور همزمان و با سرعت بالا انجام دهد. سوم m–انحراف با افزودن رکورد تقلبی به جدول داده حاصل می‌شود. از سوی دیگر، سودمندی داده در این روش پایین است. زیرا همه‌ی رکوردهای منتشر شده متعلق به همه افراد در حال حرکت واقعی نیستند. این ویژگی به هنگام تجزیه تحلیل داده‌های گمنام شده نتایج معتبری را به دست نمی‌آورد. این الگوریتم برای گمنام‌سازی داده‌های پیوسته برای جریان‌های با ظرفیت نامحدود، داده-های گذرا و زمان واقعی مناسب نیستند. زیرا این ویژگی‌ها نیازمند پردازش پویا و مقیاس پذیر با تاخیر زمانی کوچک هستند.

لی و همکاران [12] حفظ حریم خصوصی افراد در جریان داده‌های عددی را پیشنهاد کرده‌اند. دی ورک و همکاران [13] مجموعه‌ای از الگوریتم ها بر اساس حریم خصوصی تفاضلی پیشنهاد کردند تا به برخی از وظایف شمارشی خاص بپردازد. هر دو روش شامل افزودن نویز هستند و این باعث می شود اعتبار داده‌ها کم شود و انجام داده‌کاوی با دقت انجام نشود.

ژائو [14] یک چهارچوب خوشه‌ای برای k–گمنامی یک جریان داده های رابطه‌ای پیشنهاد کرد. خوشه‌ها با ورود عناصر داده‌های ورودی ساخته می‌شوند. زمانیکه خوشه حاوی داده‌های متعلق به حداقل k فرد جابه‌جا شده باشد، داده‌ها در همان سطح تعمیم منتشر میشوند. برای محدود کردن از دست

دادن اطلاعات به علت تعمیم یک مکانیزم پیش بینی عناصر داده‌های آینده را وارد می‌کنند.

حسینی [14] الگوریتمی برای حفظ حریم خصوصی افراد بر روی داده-های جریانی پیشنهاد کرده است با نام گمنام‌سازی افزایشی جریان مسیر(ITSA) که به صورت افزایشی دنباله‌ای از پنجره‌های کشویی در جریان مسیر را گمنام‌سازی می‌کندو به عنوان اولین کار گمنام‌سازی جریان مسیر در ابعاد بزرگ مطرح است.در این روش در برابر حمله‌های شباهت و انحراف برجسته تا حدی مقاوم است اما برای حفاظت از ویژگی حساس هیچ تدبیری اتخاذ نشده است.همچنین در برابر حملات افشا عضویت ، افشا شناسه و حمله حساسیت در برابر حریم خصوصی افراد محافظت نمی‌کند.

همانطور که عنوان شد اغلب روش‌های گمنام‌سازی موجود [14] برای داده‌های رابطه‌ای و داده‌های با حجم بالاو انباشته ارائه شده است که بدون در نظر گرفتن ویژگی خاص مسیرها برای داده‌های مسیر نیز استفاده شده است و برای گمنام‌سازی داده‌های مسیر مخصوصا داده‌های جریانی مسیر تحقیقات کمی صورت گرفته است [14]. بعلاوه در مساله حفظ حریم خصوصی داده-های جریانی مسیر با دو چالش مهم روبرو هستیم: اول اینکه به دلیل ابعاد بالای داده‌های مسیر ایجاد ترکیبهای متنوع زمان و مکان، پیدا کردن یک معیار شباهت مناسب برای گروه‌بندی و تعمیم مسیرهای مشابه دشوار است. دوم اینکه به دلیل جریانی بودن داده‌ها بایستی زمان اجرای لازم برای گمنام-سازی داده‌ها کم باشد تا داده‌های منتشر شده قبل از انتشار اعتبار و صحت خود را از دست ندهد و برای تحلیلگر مورد استفاده باشد.

در این مقاله یک روش گمنام‌سازی برای داده‌های جریانی مسیر براساس مدل حریم خصوصی k–گمنامی و l–گوناگونی ارائه شده است که هم از کشف مسیر و هم از کشف صفت حساس جلوگیری می‌کند. در این روش جهت کاهش زمان انجام گمنام‌سازی داده‌ها، داده‌های مسیر را در یک پیش پردازش به رشته‌های بیتی تبدیل کرده سپس معیاری برای میزان شباهت مسیرها جهت گروه‌بندی براساس فاصله همینگ ارائه شده است. از انجائیکه انجام عملیات بر روی رشته‌های بیتی بسیار کمتر است، نتایج شبیه-سازی نشان می‌دهد که زمان پردازش روش پیشنهادی بسیار پایین است، بعلاوه میزان کاهش اطلاعات و انحرافات داده‌های منتشر شده نسبیت به روش ITSA [14]کاهش قابل توجهی داشته است.

2- مفاهیم اولیه

در این بخش با چند تعریف برای حفظ حریم خصوصی در انتشار داده‌های جریانی مسیر آشنا می‌شوید:

کاربر: فرد یا دستگاهی است که اطلاعات مکانی و به همراه آن اطلاعات حساس و غیرحساس آن در ارائه دهنده سرویس مبتنی بر مکان وجود دارد، که تحلیلگران و یا سایر ارگان‌ها و سازمانها نیازمند تحلیل این اطلاعات برای تصمیم‌گیری‌های خود می‌باشند.

دشمن، متخاصم، مهاجم و یا حمله کننده: فرد یا سازمان و یا برنامه‌ای که قصد دارد از اطلاعات منتشر شده جهت کشف اطلاعات بیشتر در مورد یک فرد خاص، سوءاستفاده کند.

قربانی یا هدف: فرد یا دستگاهی است که دشمن می‌خواهد حریم خصوصی این عضو را به خطر بیاندازد و به اطلاعات حساس این فرد دست پیدا کند و یا هویت این فرد را فاش کند.

دانش پیش‌زمینه : به اطلاعاتی گفته می‌شود که دشمن از قبل این اطلاعات را در مورد قربانی در اختیار دارد. بدین ترتیب با در دسترسی به داده-های منتشر شده و براساس دانش پیش زمینه خود در مورد قربانی، قربانی را شناسایی کرده و به این ترتیب به بخشی از اطلاعات حساس وی که از قبل در اختیار نداشته است، دست پیدا می‌کند.

صفت شناسه : صفتی که مشخص کننده یک موجودیت خاص است و یکتا و منحصربه‌فرد است . مثل:کدملی

صفات شبه‌شناسه : تعدادی صفت که به تنهایی موجب شناسایی یک موجودیت نمی‌شود اما در کنار یکدیگر یک یکتایی خاص را در موجودیت به وجود می‌اورند که آن را متمایز می‌کند و منجر به کشف هویت موجودیت می-شود.

اطلاعات حساس : اطلاعاتی که برای کاربر مهم هستند و افشای این اطلاعات منجر به نقض حریم خصوصی کاربر می‌شود.

جمع‌آوری‌کننده داده: صاحب داده یا منتشر کننده داده مورد اعتماد است و حفظ حریم خصوصی کاربران را تضمین می‌کند. ممکن است به ابزار داده-کاوی مجهز نباشد و هدف آن از انتشار داده این است که هر کس بنا به نیاز خودش به نتایج تقریبی تحلیلی و آماری در مورد داده‌های جمع‌آوری شده دست یابد.

تحلیل‌گر داده: به ابزار داده‌کاوی مجهز هستند و خبره این کار هستند. هدف آنها دستیابی به نتایج تحلیلی مورد نیاز خود، بر روی داده‌های منتشر شده است.

حریم خصوصی: آن دسته از داده‌هایی که کاربر به عنوان مالک داده نمی‌خواهد در اختیار متخاصمین قرار بگیرد.

حفظ حریم خصوصی در انتشار داده‌های رابطه‌ای: مبنای کار حفظ حریم خصوصی روی داده‌های شبکه‌ای است و حفظ حریم خصوصی روی داده‌های رابطه‌ای فقط شامل شناسه، شبه‌شناسه و صفت حساس است. دانش پیش-زمینه متخاصم فقط می‌تواند شامل صفات شبه شناسه باشد و از حملات بر روی این داده‌ها با روشهای متفاوتی می‌توان جلوگیری کرد.

گمنام‌سازی: تغییر با تحریف مجموعه داده اصلی به منظور حفظ حریم خصوصی.

محو : جلوگیری از انتشار داده‌هایی که موجب نقض حریم خصوصی می‌شوند.

k–گمنامی: در مسیرهای داخل گروه ، هر مسیر حداقل با k-1 مسیر دیگر یکسان و غیر قابل تمایز باشد.

l–گوناگونی: در این روش می‌بایست گروه‌ها حداقل شامل l مقدار حساس متفاوت باشند.

داده‌های جریانی: حجم زیادی از داده‌ها که می‌توانند از منابع مختلفی به طور پیوسته در طی زمان تولید شود و پردازش این حجم زیاد از داده‌ها فرآیندی دشوار و زمان‌بر است.

داده‌های جریانی مسیر: به اطلاعاتی که حاوی موقعیت مکانی و زمانی یک فرد یا یک دستگاه که ممکن است از منابع مختلفی به طور پیوسته در

طی زمان جمع شده باشد و این داده‌ها می‌توانند دارای فیلدهای دیگری نیز باشند که این اطلاعات می‌تواند حساس و غیر حساس باشد.

پنجره لغزان: داده‌های جریانی در یک مدت زمان معلوم (پنجره) ذخیره می‌شوند و در بازه‌های زمانی مشخص این داده‌ها، گمنام شده و منتشر می‌شود.

2-1- تعریف مسئله

یک مجموعه داده مسیر به صورت N=(E,L,S) است که :

- E یک مجموعه اعدادی است به صورت $E = \{e_1, e_2, \dots, e_n\}$ که n=|E| است و e_i شناسه موجودیت i ام به حساب می آید.
- مجموعه $L = \{L_{e1}, \dots, L_{en}\}$ مجموعه مسیرهای موجودیت‌های E را نشان می دهد که $L_i = \{l_{i,1}, \dots, l_{i,t}\}$ و L_i مسیر موجودیت i در زمان 1 تا t را نشان می دهد به طوری‌که $l_{i,t}$ نشان دهنده مکان موجودیت i ام در زمان t می‌باشد. مجموعه مکان‌هایی هستند که باعث می‌شوند یک مسیر ساخته شود و ذات هر مسیر با مسیر دیگر متفاوت است که با توجه با باطن یک مسیر می تواند یک مسیر طولانی یا بلند باشد همانطور که در دنیای واقعی برای رسید
- ن از یک مکان به مکان دیگر ممکن است این مکان ها از همدیگر دور یا نزدیک باشند به همین نسبت می‌تواند مکان‌ها کم یا زیاد باشند بنابراین برای اینکه بگوییم یک مسیر داریم باید از یک مکان به مکان دیگر رفته باشیم که این جابه‌جایی نیازمند زمان است پس حداقل بیشتر از دو مکان و زمان خواهیم داشت (|L|>1) .
- S یک مجموعه از ویژگی‌های حساس است که به هر مسیر یک ویژگی حساس نسبت داده می‌شود که می‌تواند بعضی از مسیرهای متفاوت دارای ویژگی حساس مشابه باشند همانطور که در دنیای واقعی ممکن است برای رسیدن از منزل به محل کار بتوانیم از خیابان‌های متفاوتی عبور کنیم. اما هر ویژگی حساس دارای مجموعه برچسب‌های S با دامنه Ds است که می‌تواند بستگی به تعداد ویژگی حساس در این مجموعه داده متفاوت باشد. صفات حساس مانند اسم مسیر، میزان هزینه‌ها، بیماری جز حریم خصوصی داده‌های جریانی مسیر هستند و باید در برابر حملاتی که باعث نقض حریم خصوصی می‌شنود از آنها محافظت کرد.

2-2- فرضیات

در این پژوهش فرض شده‌است دستگاه‌های متحرک تمایل به ارسال اطلاعات مکانی خود دارند که این اطلاعات مکانی می‌تواند از طریق شبکه یا جی‌پی‌اس و به روش‌های دیگر جمع‌آوری شده و برای یک شخص ثالث قابل اعتماد ارسال شود. به این صورت است که در تمام زمان‌ها موقعیت دستگاه-های متحرک موجود است و ماهیت داده‌ها به گونه‌ای است که همراه با این داده‌های مکانی، داده‌های دیگری که برای پشتیبانی از کاربردهای عمومی و یا خاصی مورد نیاز است، ارسال می‌شود. با توجه به اینکه اشیاء به صورت پیوسته در زمان‌های مختلف در دسترس هستند و در تمامی زمان‌ها اطلاعات آنها در اختیار است، کل مسیر اشیاء متحرک در دسترس می‌باشد و می‌توان بر روی کل داده‌های مسیرهای اشیاء متحرک تجزیه و تحلیل انجام داد.

3- روش پیشنهادی

در این مقاله یک روش گمنام‌سازی برای داده‌های مسیر براساس مدل حریم خصوصی k-گمنامی و l-گوناگونی ارائه شده است. برای این منظور نقشه محیط به صورت یک گرید به قطعاتی مستطیلی تقسیم می‌شود، که هر یک از مناطق دارای یک شناسه هستند بعلاوه یک واحد زمانی داریم که در هر واحد زمانی مکان کاربرهای فعال در آن بازه به جمع‌کننده داده ارسال می‌شود. و از آنجایی که داده‌های یک پنجره لغزان به طول T قرار است با هم گمنام و منتشر شود. بنابراین مکان کاربران در T واحد زمانی را داریم. الگوریتم گمنام‌سازی پیشنهادی برای داده‌های یک پنجره شامل 3 مرحله‌ی زیر است:

3-1- تبدیل مسیر اصلی به رشته بیتی به طول m

از آنجایی که داده‌های مسیر در زمان دارای ابعاد بالایی است و پردازش بر روی آن سخت است در ابتدای فرآیند برای راحتی گروه‌بندی و یافتن مسیرهای مشابه هم، مسیر کاربر i در طول پنجره را به رشته‌های بیتی به طول m (B_i) تبدیل می‌کنیم. برای اینکار مجموعه داده‌های مسیر اصلی به صورت دوتایی‌های پشت سر هم از مکان و زمان درخواهند آمد. به عنوان مثال: داده مسیر به صورت $\{L_{6,1}, L_{7,2}, L_{8,3}, L_{11,4}\}$ برای پنجره‌ای به طول 4 به صورت دوتایی‌های پشت سر هم به این صورت $\{(6,7),(7,8),(8,11)\}$در خواهد آمد و به ازای هر دوتایی درون مسیر با توجه به جدول درهم‌سازی، یک عدد بین 0 تا m-1 نسبت داده می‌شود. برای مثال: h(6,7)=31, h(7,8)=40, h(8,11)=49 به ازای اعداد خروجی تابع درهم‌سازی بیت‌های متناظر با آنها را در رشته بیتی متناظر با مسیر کاربر را 1 می‌کنیم. در مثال فوق برای این مسیر،بیت‌های 31,40,49 رشته باینری مسیر یک می‌شود. در ادامه از این رشته بیتی برای پیدا کردن شباهت بین مسیرها و گروه‌بندی استفاده خواهیم کرد.

اگر طول پنجره را W در نظر بگیریم و تعداد مسیرها برای گمنام‌سازی n باشد، زمان اجرای این مرحله به ازای هر مسیر O(W) می‌باشد که طول پنجره لغزان W ثابت است، پیچیدگی زمانی این مرحله برای همه مسیرها O(n) می‌باشد.

3-2- محاسبه معیار شباهت دو مسیر

اگر چه الگوریتم‌های متعددی برای k-گمنامی برای مجموعه داده‌های جدولی و رابطه‌ای ارائه شده است، اما این روشها برای گروه‌بندی داده‌های مسیر که دارای ابعاد بسیار بالایی هستند مناسب نیستند. اکثر الگوریتم های خوشه بندی سعی می‌کنند تا جهت افزایش سودمندی داده‌های گمنام شده داده‌های شبیه به هم را در یک خوشه به اندازه k یا بیشتر قرار دهند و به ازای هر خوشه یک مسیر تعمیم یافته منتشر کنند. به این ترتیب هر مسیر منتشر شده

از حداقل k-1 مسیر دیگر قابل تمیز و شناسایی نیست. اما چالش اصلی در داده‌های مسیر تعیین یک معیار مناسب برای فاصله بین مسیرها و میزان شباهت آنها است. از آنجا که هدف ما کاهش اتلاف داده‌ها به دلیل تعمیم و گروه‌بندی هست، پیدا کردن گروه‌بندی بهینه با حداقل کاهش اطلاعات دارای زمان اجرای نمایی است. به همین دلیل در این مقاله روشی حریصانه براساس معیار فاصله رشته‌های بیتی تولید شده از مسیرها برای گروه‌بندی مسیرها ارائه شده است.

در مرحله قبل هر مسیر کاربر به یک رشته بیتی به طول m نگاشت شد. اکنون می‌خواهیم تا این رشته‌های بیتی را گروه‌بندی کنیم به گونه‌ای که ترازی از رشته‌های بیتی را به گونه‌ای پیدا کنیم که فاصله ویرایش کامل جفت رشته‌ها به حداقل برسد[15]. در این روش با استفاده از همینگ وزندار [14] عیاری برای میزان شباهت مسیرها به هم و گروه‌بندی ارائه می‌شود.

معیار فاصله همینگ وزندار دو رشته بیتی X و Y به طول m را به صورت زیر محاسبه می‌کنیم:

$$WH(X,Y) = \sum_{k=1}^{m} |X_k - Y_k| * 2^k \quad (1)$$

این معیار مکانهایی از رشته بیتی (در واقع زوج مکانهای متوالی مسیر) که با هم متفاوت بوده‌اند را به صورت وزندار براساس اینکه در کدام بیت با هم متفاوت بوده‌اند محاسبه می‌کند. محاسبه WH برای هر زوج مسیر O(m) زمان نیاز دارد از آنجائیکه m یک عدد ثابت است پیچیدگی زمانی آن O(1) است.

سپس برای هر مسیر در پنجره مورد نظر، مجموع فاصله همینگ وزندار آن را با سایر مسیرها به عنوان امتیاز P آن مسیر به صورت زیر محاسبه می‌شود:

$$P_i = \sum_{j=1}^{n} WH(B_i, B_j) \quad (2)$$

امتیازات به دست آمده برای هر مسیر مشخص می‌کند این مسیر تا چه میزان از مسیرهای دیگر متفاوت است. یا به عبارتی مسیرهایی که دارای P یکسانی هستند یعنی به نسبت یکسانی با نسبت به بقیه رشته‌ها تمایز داشته‌اند و در واقع به هم شبیه‌تر هستند. بنابراین هر چه معیار P دو رشته بیتی به هم نزدیک تر باشد میزان شباهت آن دو به هم بیشتر است. و می‌توانند در گروه‌بندی در یک گروه گمنام‌شده قرار بگیرند.

محاسبه معیار P برای یک کاربر برابر O(n) می‌باشد و محاسبه آن برای تمام کاربران $O(n^2)$ می‌باشد.

3-3- گروه‌بندی

بنابر آنچه که در مورد معیار P گفته شد، معیار حریصانه ما برای گروه‌بندی امتیاز P در نظر گرفته می‌شود. و ابتدا تمام مسیرها براساس معیار P مرتب می‌شود. و بر اساس ترتیب ایجاد شده، k مسیر اول را در خوشه اول قرار می‌دهیم تا به نیازمندی k–گمنامی برسیم. سپس برای رسیدن به l–گوناگونی در صفت حساس مسیر، اگر تعداد صفات حساس متمایز گروه ایجاد شده کمتر از l است به همان ترتیب به مسیرهای گروه اضافه می‌کنیم تا نیازمندی مورد نظر برآورده شود. سپس اعضای یک گروه گمنام‌سازی مشخص می‌شود و همین روند برای ساخت سایر گروه‌ها ادامه پیدا می‌کند تا همه مسیرها گروه‌بندی شود ممکن است در گروه آخر تعداد مسیرهای باقی‌مانده کمتر از k باشد بنابراین مسیرهای باقی‌مانده را به آخرین گروه ایجاد شده اضافه می‌کنیم تا نیازمندی‌های حریم خصوصی برای همه مسیرها تامین شود. هر چه اندازه گروه‌های ایجاد شده کمتر باشد میزان کاهش اطلاعات در هنگام تعمیم مسیرهای یک گروه کمتر خواهد شد.

پیچیدگی زمانی این مرحله برابر O(nlogn) برای مرتب‌سازی بعلاوه O(n) برای گروه‌بندی می‌باشد.

3-4- ساخت مسیر تعمیم یافته هر گروه

ابتدا برای هر گروه یک رشته بیتی نماینده که با Head نشان داده می‌شود، ساخته می‌شود. اگر تعداد یکهای بیت jام، از رشته‌های بیتی متناظر با اعضای گروه G از $|G|/2$ بیشتر باشد، آنگاه مقدار بیت Head[j] برابر یک می‌شود در ~~غیر~~غیر اینصورت صفر باقی می‌ماند. بعد از انجام این فرایند یک رشته Head گروه مورد نظر ساخته می‌شود. چون طول ~~داده باینری~~رشته بیتی m است پیچیدگی زمانی برای یک گروه O(|G|m) می‌باشد اگر اندازه گروه‌ها O(k) باشد پیچیدگی زمانی O(km) می‌شود..

$$Head_G[j] = \begin{cases} 1 & \sum_{\forall e \in G} B_e[j] \geq \frac{k}{2} \\ 0 & else \end{cases} \quad (3)$$

با توجه به رشته بیتی Head تولید شده برای هر گروه، فاصله همینگ وزندار $WH(B_i, Head_g)$ آن را نسبت به تمام اعضای گروه اندازه گرفته. عضوی از گروه که دارای کمترین مقدار فاصله باشد، شباهت بیشتری به سایر مسیرهای آن گروه داشته است و به عنوان مسیر برگزیده آن گروه انتخاب می‌شود و به عنوان مسیر گمنام شده برای تمام کاربران آن گروه به همراه صفت حساس آن برای آن پنجره منتشر می‌شود. به دلیل اینکه تعداد اعضای هر گروه در حدود k است. پیچیدگی زمانی آن حدود O(k) می‌باشد.

پیچیدگی زمانی این مرحله برای کل گروه‌ها برابر O(mn) می‌شود که از آنجائیکه m مقدار ثابتی است O(n) در نظر گرفته می‌شود.

4- نتایج شبیه‌سازی

روش پیشنهادی با روش ITSA تحت پارامترهای یکسانی مقایسه شده است.تمامی اجراها با سیستم با پردازشگر Intel(R) Core(TM) i7-3630QM CPU @ 2.40GHz و با 6 گیگ رم اجرا شده است.داده اصلی ما یک داده جریانی مسیر1000 تایی است که در بازه زمانی 60 واحد زمانی جمع‌آوری شده است.

پارامترهای روش پیشنهادی k حداقل اندازه گروه‌های گمنام‌سازی، L نیازمندی لازم برای جلوگیری از کشف صفت حساس و W اندازه پنجره داده های جریانی می‌باشد. در روش ITSA نیز پارامتر K و W به همین شکل است و یک پارامتر Len دارد که در آن می‌توان تنظیم کرد که طول مسیری که به عنوان مسیر گمنام شده برای هر گروه منتشر می‌شود چقدر است. از آنجائیکه در روش پیشنهادی ما طول مسیر گمنام شده هر پنجره با طول پنجره یکسان است بنابراین در مقایسه‌های انجام شده برای ITSA مقدار Len برابر با W در نظر گرفته شده است. بعلاوه با توجه به اینکه ITSA از

کشف صفت جلوگیری نمی‌کند، روش پیشنهادی فقط با L=2 با ITSA مقایسه شده است.

4-1- زمان پردازش الگوریتم گمنام‌سازی

در روش پیشنهادی چون داده‌ها به صورت رشته‌های بیتی پردازش می‌شوند لذا زمان پردازش کمتری دارد و طبق تحلیل زمان‌اجرا که در هر بخش ارائه شد کل پیچیدگی زمانی الگوریتم پیشنهادی $O(n^2)$ می‌باشد.

شکل (1) : مقایسه زمان الگوریتم پیشنهادی با روش ITSA)، زمان پردازش برای پارامترهای متفاوت (اندازه پنجره w ، k–گمنامی) را نشان می‌دهد. مشاهده می‌شود که با افزایش طول پنجره زمان اجرا افزایش می‌یابد. و در همین شکل زمان اجرای روش پیشنهادی با ITSA مقایسه شده است. همانطور که

در

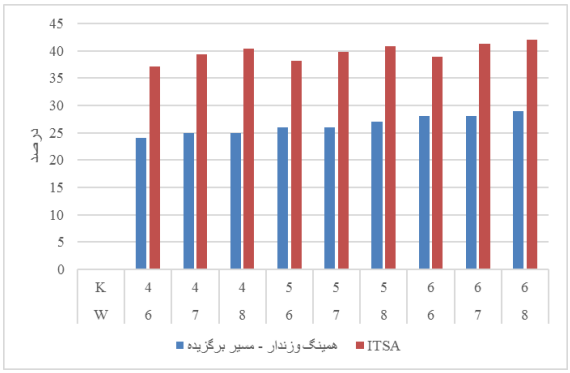
شکل (1) : مقایسه زمان الگوریتم پیشنهادی با روش ITSA1) مشاهده می-شود، زمان اجرای الگوریتم پیشنهادی در همه حالات کمتر از روش ITSA می‌باشد.

شکل (1) : مقایسه زمان الگوریتم پیشنهادی با روش ITSA

4-2- مقایسه میزان اتلاف اطلاعات

برای ارزیابی میزان حفظ سودمندی داده در داده گمنام شده، میانگین میزان انحراف مسیرهای گمنام‌شده به مسیر اصلی همه کاربران در هر پنجره محاسبه می‌شود. میزان انحراف یک مسیر نسبت به مسیر گمنام شده برابر فاصله همینگ دو رشته بیتی متناظر با هر مسیر (تعداد بیت‌های متفاوت) تقسیم بر طول پنجره می‌باشد. که در هر پنجره میانگین آن برای همه مسیرها و در نهایت میانگین آن برای کل پنجره‌ها محاسبه می‌شود.

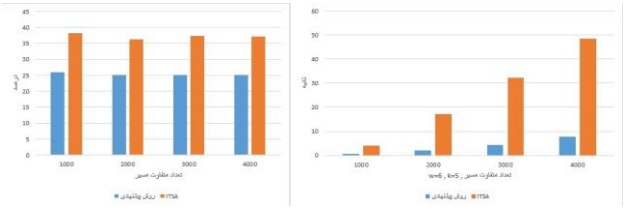
شکل (2) میانگین درصد انحرافات و یا اتلاف اطلاعات داده گمنام شده، روش پیشنهادی و ITSA را با پارامترهای متفاوت نشان می‌دهد. همانطور که نشان داده شده است در همه حالات میزان انحراف و اتلاف اطلاعات روش پیشنهادی کمتر است. و در هر دو روش با افزایش k و W به دلیل افزایش اندازه گروه‌ها و طول مسیر، اتلاف اطلاعات بیشتر می‌شود.



شکل (2): مقایسه انحراف بین الگوریتم پیشنهادی و روش ITSA

4-3- مقایسه مقیاس‌پذیری

شکل (5) زمان اجرا و میزان اتلاف اطلاعات را برای روش پیشنهادی و ITSA بر روی مجموعه داده با اندازه‌های متفاوت 1000، 2000، 3000 و 4000 تایی از مسیرها، نشان می‌دهد. همانطور که مشاهده می‌شود، زمان‌اجرای روش پیشنهادی نسبت به رشد اندازه مجموعه داده نسبت به روش ITSA رشد کمتری دارد. میزان اتلاف اطلاعات هر دو روش برای مجموعه‌های متفاوت تقریبا ثابت است اما در روش پیشنهادی درصد اتلاف کمتر است.



(ب) (الف)

شکل (5) : مقایسه (الف)زمان اجرا و (ب) درصد اتلاف اطلاعات برای مجموعه دادهایی با اندازه‌های مختلف

5- نتیجه گیری

به دلیل پیشرفت در تکنولوژی فناوری‌های همراه،داده‌های مکانی، زمانی پیوسته تولید می‌شوند و نیاز به پردازش روی این اطلاعات احساس می‌شود. در این مقاله، یک رویکرد جدید برای گمنام‌سازی داده‌های جریانی مسیر به صورت بیتی پیشنهاد شده است ، که داده‌های مسیر همزمان گمنام و منتشر می‌شود.

در روش حریم خصوصی پیشنهادی داده‌های مسیر با سرعت و حجم بالا پردازش شدند و از ویژگی حساس مسیرها در برابر حملات متخاصم جلوگیری می‌کند. عملکرد روش پیشنهادی روی داده واقعی دارای زمان بسیار

کمی است و در برابر حملات در برابر متخاصم از روش‌های موجود بهتر عمل می‌کند. در ادامه این روش می‌توان جهت حفظ سودمندی بیشتر داده، روشهایی دیگری برای انتخاب مسیری که به عنوان مسیر گمنام شده هر گروه منتشر می‌شود، پیشنهاد داد.

مراجع

[1] H. Korala, D. Georgakopoulos, P. P. Jayaraman, and A. Yavari, "A Time-Sensitive IoT Data Analysis Framework," in *Proceedings of the 54th Hawaii International Conference on System Sciences*, 2021, p. 7185.

[2] Q. Zhang, Y. Zhang, C. Li, C. Yan, Y. Duan, and H. Wang, "Sport Location-based User Clustering with Privacy-preservation in Wireless IoT-driven Healthcare," *IEEE Access*, 2021.

[3] Y. Zhang, J. Pan, L. Qi, and Q. He, "Privacy-preserving quality prediction for edge-based IoT services," *Future Generation Computer Systems*, vol. 114, pp. 336-348, 2021.

[4] J. Cao, Q. Li, W. Tu, Q. Gao, R. Cao, and C. Zhong, "Resolving urban mobility networks from individual travel graphs using massive-scale mobile phone tracking data," *Cities*, vol. 110, p. 103077, 2021.

[5] M. T. Niles, L. A. Schimanski, E. C. McKiernan, and J. P. Alperin, "Why we publish where we do: Faculty publishing values and their relationship to review, promotion and tenure expectations," *PLoS One*, vol. 15, no. 3, p. e0228914, 2020.

[6] T. Saito ,S. Nakamura, T. Enokido, and M. Takizawa, "A topic-based publish/subscribe system in a fog computing model for the IoT," in *Conference on Complex, Intelligent, and Software Intensive Systems*, 2020: Springer, pp. 12-21.

[7] F. Bonchi, "Privacy preserving publication of moving object data," in *Privacy in Location-Based Applications*: Springer, 2009, pp. 190-215.

[8] M. Terrovitis and N. Mamoulis, "Privacy Preservation in the Publication of Trajectories," in *MDM*, 2008, vol. 8, pp. 65-72.

[9] R. Yarovoy, F. Bonchi, L. V. Lakshmanan, and W. H. Wang, "Anonymizing moving objects: How to hide a mob in a crowd?," in *Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology*, 2009: ACM, pp. 72-83.

[10] K. Wang, Y. Xu, R. C.-W. Wong, and A. W.-C. Fu, "Anonymizing temporal data," in *2010 IEEE International Conference on Data Mining*, 2010: IEEE, pp. 1109-1114.

[11] X. Xiao and Y. Tao, "M-invariance: towards privacy preserving re-publication of dynamic datasets," in *Proceedings of the 2007 ACM SIGMOD international conference on Management of data*, 2007: ACM, pp. 689-700.

[12] F. Li, J. Sun, S. Papadimitriou, G. A. Mihaila, and I. Stanoi, "Hiding in the crowd: Privacy preservation on evolving streams through correlation tracking," in *2007 IEEE 23rd International Conference on Data Engineering*, 2007: IEEE, pp. 686-695.

[13] C. Dwork, M. Naor, T. Pitassi, G. N. Rothblum, and S. Yekhanin, "Pan-Private Streaming Algorithms," in *ICS*, 2010, pp. 66-80.

[14] B. Zhou, Y. Han, J. Pei, B. Jiang, Y. Tao, and Y. Jia, "Continuous privacy preserving publishing of data streams," in *Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology*, 2009: ACM, pp. 648-659.

[15] L. Gan *et al.*, "Preparation, characterization and functional properties of a novel exopolysaccharide produced by the halophilic strain Halomonas saliphila LCB169T," *International journal of biological macromolecules*, vol. 156, pp. 372-380, 2020.