نشریه علمی "پدافند الکترونیکی و سایبری"
سال نهم، شماره ۱، بهار ۱۴۰۰، ص ۴۲-۲۹

علمی – پژوهشی

انتشار پایگاه‌های داده مسیر حرکت با ضمانت حریم خصوصی تفاضلی

فاطمه دلدار[۱]، مهدی آبادی[۲*]

۱- دکتری، ۲- استادیار، گروه کامپیوتر، دانشکده مهندسی برق و کامپیوتر، دانشگاه تربیت مدرس، تهران، ایران

(دریافت: ۱۳۹۸/۱۲/۲۴، پذیرش: ۱۳۹۹/۰۵/۱۵)

چکیده

در سال‌های اخیر، سازوکارهای متعددی برای اجرای پرس‌وجوهای آماری با ضمانت حریم خصوصی تفاضلی روی پایگاه‌های داده مسیر حرکت پیشنهاد شده است. هدف اغلب این سازوکارها پاسخ به پرس‌وجوهای آماری بدون انتشار مسیرهای حرکت اشیا متحرک است. در این مقاله، یک سازوکار حریم خصوصی تفاضلی جدید به نام DP-STDR پیشنهاد می‌شود که با حفظ سودمندی‌های فضایی و زمانی، مسیرهای حرکت مصنوعی را با ضمانت حریم خصوصی تفاضلی و برای اهداف تحلیل داده منتشر می‌کند. DP-STDR برخی ویژگی‌های اصلی فضایی، زمانی و آماری مسیرهای حرکت واقعی را حفظ کرده و ساختار درختی جدیدی را با ضمانت حریم خصوصی تفاضلی برای نگهداری محتمل‌ترین مسیرهای موجود با طول‌ها و نقاط شروع مختلف تعریف می‌کند. از این ساختار درختی برای تولید مسیرهای حرکت مصنوعی استفاده می‌شود. آزمایش‌های انجام‌شده نشان می‌دهند که DP-STDR در مقایسه با کارهای مرتبط پیشین، سودمندی پاسخ پرس‌وجوها را افزایش داده و ویژگی‌های فضایی، زمانی و آماری مسیرهای حرکت واقعی را بهتر حفظ می‌کند.

کلیدواژه‌ها: حریم خصوصی تفاضلی، انتشار پایگاه داده مسیر حرکت، درخت مسیر نویزی، الگوی مسیر حرکت

۱- مقدمه

با رشد سریع و روزافزون دستگاه‌های هوشمند، محبوبیت خدمات و برنامه‌های کاربردی مکان‌محور رو به افزایش است و این محبوبیت منجر به رشد سریع پایگاه‌های داده مسیر حرکت شده است. هر مسیر حرکت تاریخچه حرکت یک شیء متحرک را در طول زمان نشان می‌دهد. امروزه تمایل زیادی به استفاده از پایگاه‌های داده مسیر حرکت برای اهدافی از قبیل طراحی شهری، کنترل ترافیک و حمل‌ونقل شهری وجود دارد. برای مثال، مسئولان حمل‌ونقل می‌توانند از این پایگاه‌های داده برای طراحی بهتر سامانه‌های حمل‌ونقل و بهینه‌سازی مصرف منابع استفاده کنند. علی‌رغم کاربرد فراوان پایگاه‌های داده مسیر حرکت، نگرانی‌های حریم خصوصی زیادی در استفاده از آن‌ها وجود دارد که این نگرانی‌ها اغلب به دلیل این واقعیت است که برخی اطلاعات شخصی و حرفه‌ای حساس در مورد یک شیء متحرک را می‌توان با اطلاع از حضور آن شیء در مکان‌های خاص به دست آورد.

در روش‌های سنتی، اغلب برای حفظ حریم خصوصی هر شی متحرک در هر نقطه از مسیر حرکت وی آشفتگی ایجاد می‌شود. این روش‌های سنتی معمولاً برای حفظ ویژگی‌های فضایی و زمانی مسیرهای حرکت کافی نیستند. از طرف دیگر، حذف شناسه‌های کاربری یا سایر اطلاعات هویتی از پایگاه‌های داده مسیر حرکت به تنهایی نمی‌تواند مانع از شناسایی دقیق اشیاء متحرک شود [۱].

حریم خصوصی تفاضلی (DP[۲]) [۳-۲] یکی از قوی‌ترین سازوکارهای موجود برای حفظ حریم خصوصی است. ایده حریم خصوصی تفاضلی این است که مستقل از وجود یا عدم وجود یک رکورد خاص در یک پایگاه داده، نتایج یکسانی از آن پایگاه داده گرفته شود. به‌طور خاص، حریم خصوصی تفاضلی اطمینان می‌دهد که احتمال این‌که یک پرس‌وجوی آماری نتیجه مشخصی را روی یک پایگاه داده به دست آورد، تقریباً برابر با حالتی است که یک رکورد خاص به پایگاه داده افزوده شده یا از آن حذف شود. در ابتدا، اغلب کارهای انجام‌شده روی حریم خصوصی تفاضلی بر پاسخ‌گویی به پرس‌وجوهای آماری متمرکز بودند [۶-۲]. هرچند، در برخی از کارهای اخیر، از حریم خصوصی تفاضلی در سناریوهای انتشار داده استفاده شده است [۹-۷].

در سال‌های گذشته، چندین سازوکار حریم خصوصی تفاضلی برای پاسخ‌گویی به پرس‌وجوهای آماری روی پایگاه‌های داده مسیر حرکت پیشنهاد شده است [۱۳-۱۰]. اما همان طور که

* رایانامه نویسنده مسئول: abadi@modares.ac.ir

[2] Differential privacy

اشاره شد، بیشتر آن‌ها یک نسخه با ضمانت حریم خصوصی تفاضلی (نسخه مصنوعی) از پایگاه داده مسیر حرکت را منتشر نمی‌کنند. اگرچه سازوکارهای اندکی برای حل این مشکل ارائه شده است [۱۵-۱۴]، اما این سازوکارها نمی‌توانند به‌درستی ویژگی‌های فضایی، زمانی و آماری مسیرهای حرکت واقعی را حفظ کنند. در این مقاله، این خط پژوهشی با ارائه یک سازوکار حریم خصوصی تفاضلی جدید ادامه داده می‌شود. در این سازوکار، ابتدا برخی از ویژگی‌های اصلی پایگاه داده مسیر حرکت، مانند تعداد مسیرهای حرکت، تعداد نقاط در هر مسیر حرکت و الگوهای حرکت مسیرهای حرکت با ضمانت حریم خصوصی تفاضلی استخراج می‌شوند. سپس تعدادی درخت که به‌اصطلاح درخت‌های مسیر نویزی نامیده می‌شوند، برای نگهداری محتمل‌ترین مسیرهای موجود با طول‌ها و نقاط شروع مختلف ساخته می‌شود. در نهایت، با استفاده از درخت‌های مسیر نویزی و با توجه به ویژگی‌های فضایی، زمانی و آماری که با ضمانت حریم خصوصی تفاضلی از مسیرهای حرکت واقعی به دست آمده‌اند، پایگاه داده مسیر حرکت مصنوعی به‌صورت کارا تولید می‌شود. در ادامه، مهم‌ترین نوآوری‌های این مقاله مرور می‌شود:

- یک سازوکار حریم خصوصی تفاضلی به نام DP-STDR برای تولید پایگاه‌های داده مسیر حرکت مصنوعی پیشنهاد می‌شود که ویژگی‌های فضایی و زمانی مسیرهای حرکت واقعی را حفظ می‌کند.
- ساختار درختی جدیدی به نام درخت مسیر نویزی ارائه می‌شود که هدف آن نگهداری محتمل‌ترین مسیرهای موجود با طول‌ها و نقاط شروع مختلف و با ضمانت حریم خصوصی تفاضلی است.
- با تولید مسیرهای حرکت مصنوعی به یک روش پایین به بالا، پایگاه داده مسیر حرکت مصنوعی با ضمانت حریم خصوصی تفاضلی به‌صورت مؤثر و کارا تولید می‌شود. هر مسیر حرکت مصنوعی با الحاق محتمل‌ترین مسیرها به یکدیگر تولید می‌شود.
- آزمایش‌های انجام‌شده نشان می‌دهند که DP-STDR در مقایسه با کارهای مرتبط پیشین، سودمندی پاسخ پرس‌وجوها را افزایش داده و ویژگی‌های فضایی، زمانی و آماری مسیرهای حرکت واقعی را بهتر حفظ می‌کند.

ادامه مقاله به‌صورت زیر سازمان‌دهی شده است. در بخش ۲ پیشینه پژوهش مرور می‌شود. در بخش ۳ مفاهیم پایه و در بخش ۴ جزئیات DP-STDR شرح داده می‌شود. در بخش ۵ نتایج آزمایش‌ها گزارش شده و در بخش ۶ نتیجه‌گیری انجام می‌شود.

۲- پیشینه پژوهش

حریم خصوصی تفاضلی از زمان معرفی آن در سال ۲۰۰۶ میلادی [۲] به طیف وسیعی از کارهای تحلیل داده اعمال شده است [۲۰-۱۶]. در این بخش، سازوکارهای پیشنهادی در پژوهش‌های پیشین برای ضمانت حریم خصوصی تفاضلی در پایگاه‌های داده مسیر حرکت مرور می‌شود.

چن[1] و همکاران [۱۰] برای اولین بار یک الگوریتم حریم خصوصی تفاضلی مستقل از داده پیشنهاد کردند که یک درخت پیشوندی نویزی را روی پایگاه داده مسیر حرکت می‌سازد. در این درخت پیشوندی، مسیرهای حرکت با پیشوند یکسان به شاخه یکسانی گروه‌بندی می‌شوند. هرچند، با رشد درخت پیشوندی نویزی، تعداد مسیرهای حرکت که در یک شاخه قرار می‌گیرند به‌سرعت کاهش پیدا می‌کند که منجر به کاهش سودمندی می‌شود. برای حل این مشکل، چن و همکاران [۲۱] در کار بعدی خود مدلی ارائه دادند که اطلاعات لازم از یک پایگاه داده ترتیبی را در قالب مجموعه‌ای از ان-گرم‌ها[2] با طول متغیر استخراج کرده و برای کاهش بزرگی نویز افزوده‌شده از یک درخت اکتشاف مبتنی بر نظریه مارکوف استفاده می‌کند. اما این مدل از این مشکل رنج می‌برد که با افزایش تعداد مکان‌ها، اندازه درخت اکتشاف به‌صورت نمایی رشد پیدا می‌کند و بنابراین برای دامنه‌های فضایی با تعداد مکان‌های زیاد مقیاس‌پذیر نیست. هی[3] و همکاران [۱۱] سامانه‌ای به نام DPT برای انتشار مسیرهای حرکت کاربران با ضمانت حریم خصوصی تفاضلی پیشنهاد کردند. در این سامانه، با هدف ضبط سرعت‌های مختلف حرکت کاربران، دامنه فضایی با ریزدانگی‌های مختلف گسسته‌سازی شده و برای هر ریزدانگی یک درخت پیشوندی نگهداری می‌شود. ونگ[4] و همکاران [۲۲] سامانه‌ای به نام RTCP برای انتشار و درجه‌بندی مسیرهای حرکت با ضمانت حریم خصوصی تفاضلی پیشنهاد کردند. در این سامانه، با ساخت درخت‌های پیشوندی نویزی، راهکاری برای انتشار مسیرهای حرکت کالیبره‌شده نویزی با ضمانت حریم خصوصی تفاضلی ارائه می‌شود. همه روش‌های فوق از ساختارهای درختی برای نمایش پایگاه‌های داده مسیر حرکت استفاده می‌کنند که موجب می‌شود نویز افزوده‌شده به گره‌های دارای مقادیر کوچک منجر به خطای نسبی بزرگ شود. علاوه بر این، استفاده از ساختارهای درختی برای نمایش پایگاه‌های داده مسیر حرکت معمولاً منجر به سربارهای فضایی و زمانی بالا می‌شود.

[1] Chen
[2] N-grams
[3] He
[4] Wang

لی[1] و همکاران [۲۳] سازوکاری برای انتشار مسیرهای حرکت با ضمانت حریم خصوصی تفاضلی پیشنهاد کردند که از یک الگوریتم تولید نویز محدود و یک الگوریتم ادغام مسیر استفاده می‌کند. ونگ و همکاران [۱۲] سازوکاری به نام DP-PSP را برای انتشار داده‌های آماری با ضمانت حریم خصوصی تفاضلی در جریان‌های مسیر حرکت زمان واقعی ارائه دادند. در این سازوکار، نقاط اصلی حساس شناسایی شده و هر مکان در یک جریان مسیر حرکت به نزدیک‌ترین نقطه اصلی خود کالیبره می‌شود. همچنین، این سازوکار به کاربران اجازه می‌دهد که برای بهینه‌سازی بودجه حریم خصوصی خود، توزیع پویای آن را مشخص کنند.

گورسوی[2] و همکاران [۱۴] روشی آگاه از سودمندی به نام AdaTrace برای انتشار مسیرهای حرکت با ضمانت حریم خصوصی تفاضلی ارائه دادند. در این روش، استخراج ویژگی‌ها، یادگیری و تزریق نویز با استفاده از پایگاه داده‌ای از مسیرهای حرکت واقعی انجام می‌شود. سپس مسیرهای حرکت مصنوعی با ضمانت حریم خصوصی تفاضلی به‌گونه‌ای تولید می‌شوند که در برابر حملات استنتاج مقاوم بوده و سودمندی‌های فضایی و آماری را حفظ کنند. همچنین، آن‌ها روشی به نام DP-Star [۱۵] ارائه دادند که مشابه با AdaTrace است با این تفاوت که از یک الگوریتم هنجارسازی[3] برای کوتاه‌کردن مسیرهای حرکت خام اولیه با استفاده از نقاط نماینده آن‌ها استفاده می‌کند. هر دو روش فوق، برخی از ویژگی‌های مفید مسیرهای حرکت (مانند تعداد نقاط و الگوهای حرکت) را هنگام تولید مسیرهای حرکت مصنوعی به‌درستی در نظر نمی‌گیرند و بنابراین، نمی‌توانند برخی از ویژگی‌های فضایی و زمانی مسیرهای حرکت واقعی را حفظ کنند.

دلدار و آبادی [۱۳] یک سازوکار حریم خصوصی تفاضلی به نام PDP-SAG ارائه دادند که تعمیم ویژگی حساس را با حریم خصوصی تفاضلی شخصی‌سازی‌شده به‌صورت یکپارچه ترکیب می‌کند. هدف از این کار تضمین سطوح مختلف حریم خصوصی تفاضلی برای اشیای متحرک در آن دسته از پایگاه‌های داده مسیر حرکت است که ویژگی‌های حساس غیرفضایی-زمانی نیز دارند.

۳- مفاهیم پایه

در این بخش، تعاریف و مفاهیم پایه مورداستفاده در این مقاله شرح داده می‌شوند.

۳-۱- حریم خصوصی تفاضلی

حریم خصوصی تفاضلی یکی از قوی‌ترین تعاریف حریم خصوصی برای انتشار داده‌های آماری است. ایده حریم خصوصی تفاضلی این است که مستقل از وجود یا عدم وجود یک رکورد خاص در یک پایگاه داده، نتایج آماری تقریباً یکسانی از آن پایگاه داده گرفته شود. بنابراین، حتی اگر نتایج پرس‌وجوها روی دو پایگاه داده همسایه وجود داشته باشد، نمی‌توان هیچ‌گونه اطلاعاتی را درباره یک رکورد خاص به دست آورد. در ادامه، مفاهیم مرتبط با حریم خصوصی تفاضلی تعریف می‌شوند.

تعریف ۱ (پایگاه‌های داده همسایه). دو پایگاه داده مجزای $\mathcal{D}_1$ و $\mathcal{D}_2$ همسایه نامیده شده و با $\mathcal{D}_1 \sim \mathcal{D}_2$ نمایش داده می‌شود، اگر و فقط اگر یکی از آن‌ها را بتوان با افزودن یا حذف یک رکورد خاص از دیگری به دست آورد.

تعریف ۲ (ضمانت حریم خصوصی تفاضلی ε). الگوریتم تصادفی $\mathcal{A}$ دارای ضمانت حریم خصوصی تفاضلی ε است اگر و فقط اگر برای هر دو پایگاه داده همسایه $\mathcal{D}_1$ و $\mathcal{D}_2$ و هر زیرمجموعه O از خروجی‌های ممکن $\mathcal{A}$ رابطه زیر برقرار باشد:

$$\Pr[\mathcal{A}(\mathcal{D}_1) \in O] \leq \exp(\varepsilon) \times \Pr[\mathcal{A}(\mathcal{D}_2) \in O] \quad (۱)$$

که ε قدرت ضمانت حریم خصوصی را تعیین کرده و بودجه حریم خصوصی نامیده می‌شود.

توجه داشته باشید که $\varepsilon > 0$ پارامتری است که مالک داده می‌تواند آن را وابسته به میزان حریم خصوصی موردنیاز انتخاب کند. بدیهی است که مقادیر کوچک‌تر ε منجر به افشای اطلاعات کمتری می‌شوند.

یک روش مرسوم و متداول به‌منظور دستیابی به ضمانت حریم خصوصی تفاضلی ε برای پرس‌وجوهای آماری، استفاده از سازوکار لاپلاس[4] [۲۴] برای افزودن نویز تصادفی به پاسخ پرس‌وجوها است. در این سازوکار، نویز تصادفی با توزیع لاپلاس تولید شده و به نتایج پرس‌وجوها افزوده می‌شود. بزرگی نویز به بودجه حریم خصوصی ε و حساسیت تابع پرس‌وجو وابسته است. حساسیت تابع پرس‌وجو کران تغییرات ممکن در پاسخ پرس‌وجوها بر روی هر دو پایگاه داده همسایه را نشان می‌دهد.

تعریف ۳ (حساسیت). فرض کنید f یک تابع پرس‌وجو باشد که هر پایگاه داده دلخواه را به یک بردار d بعدی از اعداد حقیقی نگاشت می‌کند. حساسیت تابع f که با σ_f نمایش داده می‌شود، به‌صورت زیر تعریف می‌شود:

$$\sigma_f = \max_{\mathcal{D}_1 \sim \mathcal{D}_2} \|f(\mathcal{D}_1) - f(\mathcal{D}_2)\|_1 \quad (۲)$$

[1] Li
[2] Gursoy
[3] Normalization
[4] Laplace

که $\|\cdot\|_1$ نرم[1] L1 یک بردار است.

فرض کنید پایگاه داده $\mathcal{D}$، تابع پرس‌وجوی f و بودجه حریم خصوصی ε داده شده باشد، سازوکار لاپلاس نویز تصادفی را به پاسخ پرس‌وجوها می‌افزاید که از توزیع لاپلاس با میانگین صفر و پارامتر مقیاس σ_f/ε تولید می‌شود. واریانس پاسخ‌های نویزی با واریانس سازوکار لاپلاس برابر است:

$$\mathrm{Var}(\mathrm{Lap}(\sigma_f/\varepsilon)) = 2\sigma_f^2/\varepsilon^2 \qquad (۳)$$

که $\mathrm{Lap}(\lambda)$ متغیر تصادفی لاپلاس با تابع چگالی احتمال $h_\lambda(z) = \frac{1}{2\lambda}\exp(-|z|/\lambda)$ است. مقادیر بزرگ‌تر σ_f یا مقادیر کوچک‌تر ε منجر به افزودن نویز تصادفی بیشتری به پاسخ پرس‌وجوها شده و در نتیجه حریم خصوصی قوی‌تری را ضمانت می‌کنند.

تعریف ۴ (سازوکار لاپلاس). فرض کنید σ_f حساسیت تابع پرس‌وجوی f باشد. الگوریتم تصادفی $\mathcal{A}$ حریم خصوصی تفاضلی ε را برای پایگاه داده $\mathcal{D}$ ضمانت می‌کند اگر و فقط اگر

$$\mathcal{A}(\mathcal{D}) = f(\mathcal{D}) + \mathrm{Lap}(\sigma_f/\varepsilon). \qquad (۴)$$

از سازوکار لاپلاس نمی‌توان برای پرس‌وجوهایی استفاده کرد که خروجی‌های دسته‌ای یا گسسته دارند. سازوکار نمایی[2] [۲۵] کلی‌تر از سازوکار لاپلاس بوده و به همه انواع پرس‌وجوها قابل‌اعمال است. این سازوکار از یک تابع امتیازدهی استفاده می‌کند که پایگاه داده مسیر حرکت $\mathcal{D}$ و مقدار گسسته r را گرفته و یک امتیاز با مقدار حقیقی به r نسبت می‌دهد تا کیفیت آن را کمی‌سازی کند.

تعریف ۵ (سازوکار نمایی). فرض کنید q یک تابع امتیازدهی دلخواه باشد. الگوریتم تصادفی $\mathcal{A}$ که خروجی گسسته r را برای پایگاه داده $\mathcal{D}$ با احتمالی متناسب با $\exp(\varepsilon q(\mathcal{D}, r)/2\sigma_q)$ برمی‌گرداند، حریم خصوصی تفاضلی ε را ضمانت می‌کند که σ_q حساسیت تابع q است و به‌صورت زیر تعریف می‌شود:

$$\sigma_q = \max_{r, \mathcal{D}_1 \sim \mathcal{D}_2} \|q(\mathcal{D}_1, r) - q(\mathcal{D}_2, r)\|_1. \qquad (۵)$$

۳-۲- پایگاه داده مسیر حرکت

هر پایگاه داده مسیر حرکت تغییرات گسسته یا پیوسته اشیاء متحرک را در یک دامنه فضایی ذخیره و مدیریت می‌کند.

فرض کنید $\mathcal{L} = \{L_1, L_2, \dots, L_{|\mathcal{L}|}\}$ مجموعه‌ای متناهی شامل مکان‌های موجود در یک دامنه فضایی دلخواه باشد. هر مسیر حرکت دلخواه $T = \langle X_1, X_2, \dots, X_{|T|}\rangle$ یک توالی از نقاط یا مکان‌های $X_i \in \mathcal{L}$ است که سابقه یا تاریخچه حرکت یک شیء متحرک را نشان می‌دهد. تعداد نقاط مسیر حرکت T با $|T|$ نمایش داده شده و طول آن مسیر حرکت نامیده می‌شود. هر پایگاه داده مسیر حرکت $\mathcal{D} = \{T_1, T_2, \dots, T_{|\mathcal{D}|}\}$ شامل زیرمجموعه‌ای از مسیرهای حرکت اشیا متحرک است. به اولین نقطه از هر مسیر حرکت نقطه شروع و به زیرمسیری که با حذف نقطه شروع به دست می‌آید دنباله آن مسیر حرکت گفته می‌شود.

۴- سازوکار DP-STDR

در این بخش، سازوکار DP-STDR برای انتشار پایگاه‌های داده مسیر حرکت با ضمانت حریم خصوصی تفاضلی معرفی می‌شود. سازوکار پیشنهادی شامل سه مرحله اصلی است. در مرحله اول که هدف از آن حفظ ویژگی‌های آماری پایگاه داده مسیر حرکت واقعی است، ابتدا دامنه فضایی پیوسته به مجموعه‌ای متناهی از سلول‌ها گسسته شده و یک بافت‌نگار[3] نویزی از سلول‌های ابتدایی ابتدایی مسیرهای حرکت واقعی ساخته می‌شود. سپس مسیرهای حرکت واقعی بر اساس سلول‌های ابتدایی خود به دسته‌های مختلفی تقسیم شده و میانه نویزی طول‌های مسیرهای حرکت در هر دسته محاسبه می‌شود. در مرحله دوم که هدف از آن حفظ الگوهای حرکت مسیرهای حرکت واقعی است، ماتریس هزینه انتقال تولید شده و با استفاده از آن درخت‌های مسیر نویزی ایجاد می‌شوند. در نهایت، در مرحله سوم، با استفاده از اطلاعات به‌دست‌آمده از دو مرحله قبل، پایگاه داده مسیر حرکت مصنوعی با ضمانت حریم خصوصی تفاضلی تولید می‌شود. در ادامه این مقاله، پایگاه‌های داده مسیر حرکت واقعی و مصنوعی به‌ترتیب با $\mathcal{D}$ و $\widehat{\mathcal{D}}$ نمایش داده می‌شوند.

۴-۱- حفظ ویژگی‌های آماری پایگاه داده مسیر حرکت واقعی

در این مرحله، ابتدا دامنه فضایی پیوسته به مجموعه متناهی $\mathcal{C}$ از سلول‌ها گسسته شده و نقاط هر مسیر حرکت در $\mathcal{D}$ به سلول‌های متناظرشان در $\mathcal{C}$ نگاشت می‌شوند. سپس یک بافت‌نگار از تمامی سلول‌های ابتدایی مسیرهای حرکت در $\mathcal{D}$ ایجاد می‌شود. منظور از سلول ابتدایی یک مسیر حرکت سلولی است که نقطه شروع آن مسیر حرکت را پوشش می‌دهد. در ادامه، نویز لاپلاس با پارامتر مقیاس $1/\varepsilon_1$ به‌صورت مستقل به هر بازه از این بافت‌نگار افزوده می‌شود. این کار یک بافت‌نگار نویزی را نتیجه می‌دهد که به آن بافت‌نگار نویزی سلول‌های ابتدایی گفته می‌شود.

در ادامه، برای هر سلول $C_i \in \mathcal{C}$ میانه طول‌های آن دسته از

[1] Norm
[2] Exponential
[3] Histogram

مسیرهای حرکت که با این سلول شروع می‌شوند محاسبه می‌شود. سپس میانه‌های به‌دست‌آمده با ضمانت حریم خصوصی تفاضلی (با استفاده از ε_2 به‌عنوان پارامتر حریم خصوصی) نویزی می‌شوند. از میانه‌های نویزی برای حفظ ویژگی طول مسیر هنگام تولید مسیرهای حرکت مصنوعی استفاده می‌شود. هر میانه نویزی را می‌توان با افزودن نویز لاپلاس به مقدار واقعی آن محاسبه کرد. اما تابع میانه برای طول‌های مسیرهای حرکت حساسیت بالایی دارد (حساسیت تابع میانه برای یک توالی از اعداد برابر با تفاضل بین بزرگ‌ترین و کوچک‌ترین عدد موجود در آن توالی است) و به این دلیل که نویز لاپلاس رابطه مستقیمی با حساسیت دارد (به تعریف ۴ مراجعه شود)، بنابراین افزودن نویز لاپلاس به میانه می‌تواند سودمندی را به‌شدت کاهش دهد. بنابراین در این مقاله، مشابه با کارهای پیشین [۲۶، ۲۷]، از سازوکار نمایی (تعریف ۵) به‌جای سازوکار لاپلاس برای محاسبه میانه‌های نویزی با ضمانت حریم خصوصی تفاضلی استفاده می‌شود.

از روش زیر برای محاسبه میانه‌های نویزی با استفاده از سازوکار نمایی استفاده می‌شود. برای هر سلول $C_i \in \mathcal{C}$، ابتدا توالی طول‌های آن دسته از مسیرهای حرکت که نقطه شروع آن‌ها در این سلول قرار دارد با L نمایش داده می‌شود. سپس L به‌صورت غیرنزولی مرتب شده و طول $l \in L$ با احتمالی متناسب با $\exp(\varepsilon_2 s(l, \theta_{C_i})/2)$ به‌عنوان میانه نویزی $\hat{\theta}_{C_i}$ انتخاب می‌شود که θ_{C_i} میانه واقعی طول‌های مسیرهای حرکت در L و $s(l, \theta_{C_i})$ امتیاز l نسبت به θ_{C_i} است که به‌صورت زیر تعریف می‌شود:

$$s(l, \theta_{C_i}) = -|r(l) - r(\theta_{C_i})| \qquad (۶)$$

که تابع r رتبه[1] طول هر مسیر حرکت در L را برمی‌گرداند. ویژگی تابع امتیاز s این است که اگر طول یک مسیر حرکت به θ_{C_i} نزدیک باشد، رتبه آن مسیر حرکت به رتبه θ_{C_i} نزدیک خواهد بود. بنابراین، امتیاز طول هر مسیر حرکت باید با منفی تفاضل مطلق رتبه آن مسیر حرکت با رتبه θ_{C_i} متناسب باشد. این ویژگی موجب می‌شود که طول‌هایی که به θ_{C_i} نزدیک‌تر هستند با احتمال بالاتری به‌عنوان میانه نویزی انتخاب شوند.

۴-۲- حفظ الگوهای حرکت پایگاه داده مسیر حرکت واقعی

در این مرحله، ابتدا ماتریس هزینه انتقال $\mathbf{C} = (c_{ij})_{m\times m}$ با ضمانت حریم خصوصی تفاضلی (با استفاده از ε_3 به‌عنوان پارامتر حریم خصوصی) تولید می‌شود که m تعداد سلول‌های $\mathcal{C}$ است. هدف از این ماتریس، حفظ الگوهای حرکت مسیرهای حرکت در $\mathcal{D}$ است. هر عنصر $c_{ij} \in \mathbf{C}$ هزینه انتقال از سلول $C_i \in \mathcal{C}$ به سلول $C_j \in \mathcal{C}$ را ذخیره می‌کند:

$$c_{ij} = -\log p_{ij} \qquad (۷)$$

که p_{ij} احتمال انتقال با ضمانت حریم خصوصی تفاضلی (احتمال انتقال نویزی) از C_i به C_j است.

برای محاسبه احتمال‌های انتقال نویزی، ابتدا ماتریس فراوانی هنجاریافته[2] $\mathbf{F} = (f_{ij})_{m\times m}$ روی $\mathcal{D}$ ایجاد می‌شود که در آن سطرها و ستون‌ها به‌صورت یکتا با سلول‌های $\mathcal{C}$ برچسب زده می‌شوند. هر عنصر دلخواه $f_{ij} \in \mathbf{F}$ فراوانی هنجاریافته زیرمسیر $\langle C_i, C_j\rangle$ در $\mathcal{D}$ را ذخیره می‌کند.

تعریف ۶ (فراوانی هنجاریافته). فرض کنید $\mathcal{C}$ مجموعه تمام سلول‌ها در یک دامنه فضایی دلخواه باشد. فراوانی هنجاریافته زیرمسیر $\langle C_i, C_j\rangle$ در پایگاه داده مسیر حرکت گسسته $\mathcal{D}$ به‌صورت زیر تعریف می‌شود:

$$\bar{c}_{\mathcal{D}}(\langle C_i, C_j\rangle) = \sum_{T\in\mathcal{D}} \frac{c_T(\langle C_i, C_j\rangle)}{|T| - 1} \qquad (۸)$$

که T یک مسیر حرکت دلخواه در $\mathcal{D}$ و $c_T(\langle C_i, C_j\rangle)$ فراوانی زیرمسیر $\langle C_i, C_j\rangle$ در T است.

مثال ۱. پایگاه داده مسیر حرکت جدول (۱) را در نظر بگیرید که بر روی یک دامنه فضایی گسسته با سلول‌های C_1، C_2، C_3 و C_4 تولید شده است. فراوانی هنجاریافته زیرمسیرهای $\langle C_1, C_2\rangle$ و $\langle C_2, C_4\rangle$ به‌صورت زیر محاسبه می‌شود:

$$\bar{c}_{\mathcal{D}}(\langle C_1, C_2\rangle) = \frac{1}{4} + \frac{1}{5} = 0.45,$$

$$\bar{c}_{\mathcal{D}}(\langle C_2, C_4\rangle) = \frac{1}{4} + \frac{1}{5} + \frac{1}{4} + \frac{1}{3} = 1.03.$$

جدول (۱): مثالی از یک پایگاه داده مسیر حرکت

شناسه	مسیر حرکت
۱	$\langle C_1, C_4, C_4\rangle$
۲	$\langle C_2, C_1, C_2, C_2, C_4\rangle$
۳	$\langle C_1, C_2, C_4, C_4, C_3, C_2\rangle$
۴	$\langle C_2, C_1, C_3, C_2, C_3, C_2\rangle$
۵	$\langle C_4, C_2, C_3, C_2\rangle$
۶	$\langle C_1, C_3, C_2, C_4, C_2\rangle$
۷	$\langle C_2, C_2, C_4, C_3\rangle$

در ادامه، به عناصر $\mathbf{F}$ نویز لاپلاس با پارامتر مقیاس $1/\varepsilon_3$ افزوده می‌شود. سپس احتمال‌های انتقال نویزی با هنجارسازی سطرهای $\mathbf{F}$ محاسبه می‌شوند به‌طوری که مجموع عناصر هر سطر برابر با ۱ شود:

[1] Rank

[2] Normalized frequency

$$p_{ij} = \frac{\hat{f}_{ij}}{\sum_j \hat{f}_{ij}} \quad (۹)$$

که $\hat{f}_{ij}$ نسخه نویزی از عنصر $f_{ij} \in \mathbf{F}$ است.

پس از تولید ماتریس هزینه انتقال، برای هر سلول از دامنه فضایی، یک درخت مسیر نویزی با حداکثر ارتفاع h_{max} ساخته می‌شود تا محتمل‌ترین مسیرهای موجود با طول‌های متفاوت (حداکثر تا $h_{max}+1$) که از آن سلول شروع می‌شوند را نگهداری کند.

تعریف ۷ (درخت مسیر نویزی). فرض کنید $\mathcal{C}$ مجموعه‌ای شامل m سلول یک دامنه فضایی دلخواه باشد. درخت مسیر نویزی برای هر سلول $C_i \in \mathcal{C}$ با سه‌تایی $\Phi_{C_i} = (V, E, \vartheta)$ نمایش داده می‌شود که V مجموعه گره‌ها، E مجموعه یال‌ها و $\vartheta: V \to \mathbb{R}_{\geq 0}$ تابع هزینه است. هر سطح از Φ_{C_i} حداکثر شامل m گره است که هرکدام از آن‌ها به‌صورت یکتا با یک سلول از $\mathcal{C}$ برچسب زده می‌شود. به این سلول به‌عنوان سلول متناظر با گره رجوع می‌شود. ریشه Φ_{C_i} در سطح صفر است و با C_i برچسب زده می‌شود. تابع هزینه برای هر گره $v \in V$ در سطح l، هزینه محتمل‌ترین مسیر با طول $l+1$ (در صورتی که چنین مسیری وجود داشته باشد) از C_i به سلول متناظر با v است.

از یک الگوریتم برنامه‌نویسی پویای کارا برای ساخت درخت مسیر نویزی Φ_{C_i} برای هر سلول دلخواه $C_i \in \mathcal{C}$ استفاده می‌شود (الگوریتم ۱). در این الگوریتم، ابتدا ریشه r از Φ_{C_i} در سطح صفر ایجاد شده و با C_i برچسب زده می‌شود (هزینه این گره به صفر تنظیم می‌شود) (خطوط ۱ و ۲). سپس سایر گره‌های Φ_{C_i} به‌ترتیب اول سطح ایجاد می‌شوند. این کار تا زمانی انجام می‌شود که درخت به سطح h_{max} برسد (خطوط ۳ تا ۱۲). برای هر سطح از درخت، m گره به‌صورت بالقوه در نظر گرفته می‌شود. هر گره بالقوه v در این سطح با یک سلول یکتا از $\mathcal{C}$ برچسب زده شده (خط ۵) و هزینه $\vartheta(v)$ برای آن گره به‌صورت زیر محاسبه می‌شود (خط ۶):

$$\vartheta(v) = \min_{u \in L_V(v)} \vartheta(u) + c(u, v) \quad (۱۰)$$

که $L_V(v)$ مجموعه تمام گره‌ها در سطح قبلی v و $c(u, v)$ هزینه انتقال از سلول متناظر با u به سلول متناظر با v است. همچنین، والد v به‌صورت زیر تعیین می‌شود:

$$\eta(v) = \arg\min_{u \in L_V(v)} \vartheta(u) + c(u, v). \quad (۱۱)$$

اگر $\vartheta(v) \neq \infty$ باشد، حداقل یک مسیر از C_i به سلول متناظر با v وجود دارد. در این حالت، گره v به مجموعه $V(\Phi_{C_i})$ و یال $(\eta(v), v)$ به مجموعه $E(\Phi_{C_i})$ افزوده می‌شود (خطوط ۷ تا ۱۰).

الگوریتم (۱): ساخت درخت مسیر نویزی

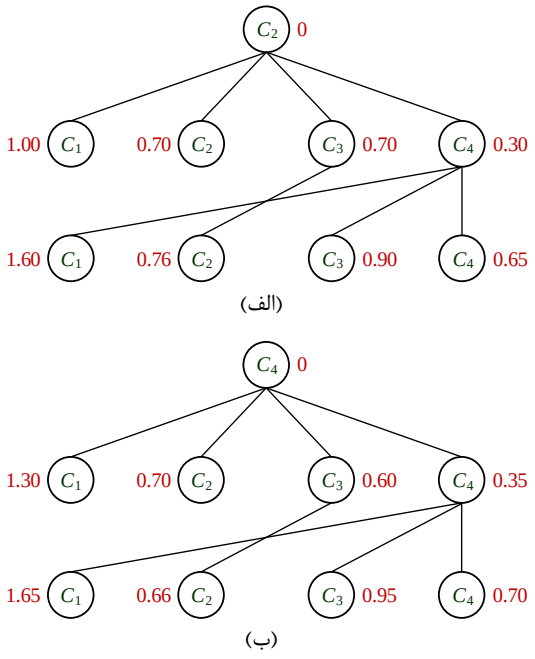
Input:
C_i: A domain cell
Output:
Φ_{C_i}: A noisy path tree

1: Create the root r of Φ_{C_i} at level 0
2: Label r by C_i and set its cost to 0
3: **for** each level of Φ_{C_i} from 1 up to h_{max} **do**
4: **for** each domain cell $C_j \in \mathcal{C}$ **do**
5: Create a node v and label it by C_j
6: Compute the cost $\vartheta(v)$ using (10)
7: **if** $\vartheta(v) \neq \infty$ **then**
8: Compute the parent $\eta(v)$ using (11)
9: Add v to $V(\Phi_{C_i})$ and $(\eta(v), v)$ to $E(\Phi_{C_i})$
10: **end if**
11: **end for**
12: **end for**

مثال ۲. فرض کنید ماتریس هزینه انتقال برای پایگاه داده مسیر حرکت جدول (۱) به‌صورت زیر باشد:

$$\mathbf{C} = \begin{bmatrix} 1.30 & 0.52 & 0.46 & 0.52 \\ 1.00 & 0.70 & 0.70 & 0.30 \\ 1.22 & 0.06 & 1.40 & 1.30 \\ 1.30 & 0.70 & 0.60 & 0.35 \end{bmatrix}.$$

در شکل (۱) درخت‌های مسیر نویزی Φ_{C_2} و Φ_{C_4} با تنظیم حداکثر ارتفاع h_{max} به ۲ نمایش داده شده است. سلول متناظر با هر گره در درون و هزینه در بیرون آن گره قرار داده شده است.



شکل (۱): درخت‌های مسیر نویزی با تنظیم حداکثر ارتفاع h_{max} به ۲. (الف) Φ_{C_2}، (ب) Φ_{C_4}.

از Φ_{C_i} در مرحله بعد برای یافتن محتمل‌ترین مسیر شروع‌شده از C_i با یک طول مشخص استفاده می‌شود. در واقع،

برای یافتن محتمل‌ترین مسیر با طول $l+1$ که از C_i شروع می‌شود، از ریشه Φ_{C_i} یعنی r شروع کرده و مسیری دنبال می‌شود که به گره‌ای با کمترین هزینه در سطح l از Φ_{C_i} می‌رسد. به چنین مسیری با عنوان محتمل‌ترین مسیر کامل با طول $l+1$ رجوع می‌شود.

۴-۳- تولید پایگاه داده مسیر حرکت مصنوعی با ضمانت حریم خصوصی تفاضلی

در این مرحله، پایگاه داده مسیر حرکت مصنوعی $\widehat{\mathcal{D}}$ با ضمانت حریم خصوصی تفاضلی تولید می‌شود. فرض کنید l_{max} حداکثر طول ممکن برای مسیرهای حرکت باشد. ابتدا برای هر سلول $C_i \in \mathcal{C}$، گام‌های زیر به تعداد دفعاتی مساوی با مقدار بازه متناظر با C_i در بافت‌نگار نویزی سلول‌های ابتدایی تکرار می‌شوند. در هر تکرار، نمونه l از توزیع نمایی با پارامتر $\ln(2)/\hat{\theta}_{C_i}$ استخراج می‌شود (l به l_{max} محدود می‌شود) که $\hat{\theta}_{C_i}$ میانه نویزی طول‌های آن دسته از مسیرهای حرکت است که نقطه شروع آن‌ها در C_i قرار دارد. سپس مسیر حرکت مصنوعی $\hat{T}$ با طول l و با شروع از C_i تولید می‌شود. برای این منظور، ابتدا $\hat{T}$ به $\langle C_i \rangle$ مقداردهی می‌شود. سپس به‌صورت تکراری و با شروع از سلول انتهایی E از $\hat{T}$، محتمل‌ترین مسیر کامل با طول $h+1$ در درخت مسیر نویزی Φ_E پیدا شده و دنباله آن به $\hat{T}$ اضافه می‌شود که h ارتفاع Φ_E است. این کار آن‌قدر تکرار می‌شود که اختلاف بین l و $|\hat{T}|$ کوچک‌تر از یا مساوی با h شود. در این هنگام، محتمل‌ترین مسیر کامل با طول $l-|\hat{T}|+1$ در Φ_E پیدا شده و دنباله آن به $\hat{T}$ اضافه می‌شود. در نهایت، هر سلول از $\hat{T}$ با یک مکان (طول جغرافیایی/عرض جغرافیایی) از داخل آن سلول جایگزین شده (برای مثال، طول جغرافیایی/عرض جغرافیایی مرکز آن سلول) و $\hat{T}$ به $\widehat{\mathcal{D}}$ اضافه می‌شود.

مثال ۳. درخت‌های مسیر نویزی شکل ۱ را در نظر بگیرید. برای تولید مسیر حرکت مصنوعی $\hat{T}$ با طول $l=4$ که از C_4 شروع می‌شود، ابتدا $\hat{T}$ به $\langle C_4 \rangle$ مقداردهی می‌شود. سپس محتمل‌ترین مسیر کامل با طول ۳ در Φ_{C_4} پیدا شده و دنباله آن به $\hat{T}$ اضافه می‌شود که $\hat{T}=\langle C_4, C_3, C_2 \rangle$ را نتیجه می‌دهد. در ادامه، محتمل‌ترین مسیر کامل با طول ۲ در Φ_{C_2} پیدا شده و دنباله آن به $\hat{T}$ اضافه می‌شود که $\hat{T}=\langle C_4, C_3, C_2, C_4 \rangle$ را نتیجه می‌دهد.

۴-۴- تحلیل پیچیدگی زمانی

در ادامه، پیچیدگی زمانی سازوکار DP-STDR مورد تحلیل قرار می‌گیرد. همان طور که اشاره شد، DP-STDR شامل سه مرحله اصلی است. در مرحله اول، ابتدا پایگاه داده مسیر حرکت $\mathcal{D}$ یک‌بار پیمایش شده و بافت‌نگار نویزی سلول‌های ابتدایی ایجاد می‌شود که مرتبه زمانی آن $O(|\mathcal{D}|)$ است. سپس میانه نویزی طول‌های مسیرهای حرکت واقعی که در هر سلول شروع می‌شوند محاسبه می‌شود که این کار زمانی متناسب با تعداد این مسیرهای حرکت را مصرف می‌کند. بنابراین، مرتبه زمانی این مرحله در کل $O(|\mathcal{D}|)$ است. در مرحله دوم، ابتدا ماتریس فراوانی هنجاریافته در زمان $O(l_{max} \times |\mathcal{D}|)$ ایجاد شده و ماتریس هزینه انتقال در زمان $O(m^2)$ تولید می‌شود که m تعداد سلول‌های $\mathcal{C}$ و $\mathcal{C}$ مجموعه سلول‌های ممکن است. بنابراین ایجاد این دو ماتریس دارای مرتبه زمانی $O(l_{max} \times |\mathcal{D}| + m^2)$ است که l_{max} بیشترین طول ممکن برای یک مسیر حرکت است. سپس درخت‌های مسیر نویزی ساخته می‌شوند. به این دلیل که تعداد گره‌ها در هر درخت مسیر نویزی حداکثر برابر با $h_{max} \times m$ بوده و محاسبه هزینه برای هر گره از مرتبه زمانی $O(m)$ است، ساخت تمامی m درخت مسیر نویزی دارای مرتبه زمانی $O(h_{max} \times m^3)$ است. در مرحله سوم، مسیرهای حرکت مصنوعی منتشر می‌شوند. از آنجایی که طول هر مسیر حرکت مصنوعی حداکثر برابر با l_{max} است و تعداد مسیرهای حرکت مصنوعی تقریباً مشابه با تعداد مسیرهای حرکت واقعی است، مرتبه زمانی این مرحله $O(l_{max} \times |\mathcal{D}|)$ است. بنابراین، پیچیدگی زمانی کلی DP-STDR برابر با $O(l_{max} \times |\mathcal{D}| + h_{max} \times m^3)$ است. با توجه به این واقعیت که معمولاً h_{max} به یک مقدار کوچک تنظیم می‌شود، پیچیدگی زمانی کلی به $O(l_{max} \times |\mathcal{D}| + m^3)$ تبدیل می‌شود.

۵- ارزیابی

در این بخش، با انجام آزمایش‌های مختلف کارایی سازوکار پیشنهادی DP-STDR مورد ارزیابی قرار می‌گیرد.

۵-۱- تنظیمات آزمایش‌ها

در آزمایش‌ها از دو مجموعه داده مسیر حرکت زیر استفاده می‌شود:

- مجموعه داده Geolife [۲۸]. این مجموعه داده شامل مسیرهای حرکت ۱۸۲ کاربر در طول یک دوره پنج‌ساله (از آوریل ۲۰۰۷ تا اوت ۲۰۱۲ میلادی) است و نه‌تنها شامل مسیرهای حرکت روزمره‌ای مانند رفتن به خانه و رفتن به محل کار است، بلکه شامل طیف گسترده‌ای از سایر حرکات متنوع در فضای باز، از جمله برخی فعالیت‌های تفریحی و ورزشی مانند خرید، گشت‌وگذار، پیاده‌روی و دوچرخه‌سواری است. اغلب این مسیرهای حرکت مربوط به شهر پکن در چین هستند. با انجام یک پیش‌پردازش ساده، مسیرهای حرکت پرت حذف‌شده و ۱۷۰۰۰ مسیر حرکت

در یک محدوده جغرافیایی خاص از شهر پکن انتخاب می‌شوند.

- مجموعه داده Taxi [۱۱]. این مجموعه داده شامل تقریباً ۱۷۰۰۰۰۰ مسیر حرکت از ۸۶۰۲ تاکسی در شهر پکن است که در ماه مه سال ۲۰۰۹ میلادی ضبط شده است.

در آزمایش‌های انجام‌شده، هنگام تولید پایگاه داده مسیر حرکت مصنوعی، دامنه فضایی به‌صورت یکنواخت به ۱۰۲۴ سلول تقسیم شده و حداکثر ارتفاع درخت‌های مسیر نویزی به ۳ تنظیم می‌شود. همچنین، بودجه حریم خصوصی ε به‌صورت مساوی به سه بودجه حریم خصوصی محلی ε_1، ε_2 و ε_3 تقسیم می‌شود.

به‌دلیل اینکه سازوکارهای لاپلاس و نمایی احتمالی هستند، هر آزمایش پنج بار تکرار شده و میانگین نتایج گزارش می‌شود. برای انجام مقایسه دقیق‌تر، هنگام محاسبه سنجه‌های ارزیابی، دو سناریوی مختلف در نظر گرفته می‌شود: (۱) سناریوی ریزدانه که در آن دامنه فضایی پیوسته به‌صورت یکنواخت به ۱۰۲۴ سلول تقسیم می‌شود و (۲) سناریوی درشت‌دانه که در آن دامنه فضایی پیوسته به‌صورت یکنواخت به ۳۶ سلول تقسیم می‌شود.

۵-۲- سنجه‌های ارزیابی

منظور از سودمندی میزان نزدیکی خروجی‌های نویزی انتشاریافته به مقادیر واقعی آن‌ها است. در ادامه این بخش، از سنجه‌های متفاوتی برای ارزیابی سودمندی‌های فضایی و زمانی پایگاه‌های داده مسیر حرکت مصنوعی تولیدشده توسط DP-STDR استفاده می‌شود.

خطای پرس‌وجوهای شمارشی

کیفیت پاسخ‌های نویزی به پرس‌وجوهای شمارشی با خطای نسبی آن‌ها نسبت به پاسخ‌های واقعی اندازه‌گیری می‌شود. فرض کنید Q یک پرس‌جوی شمارشی (مانند فراوانی یک زیرمسیر مفروض را بازیابی کنید) باشد. خطای نسبی پاسخ نویزی به Q به‌صورت زیر محاسبه می‌شود:

$$\mathcal{E}(Q)=\frac{\left|c_{\mathcal{D}}(Q)-c_{\widehat{\mathcal{D}}}(Q)\right|}{\max\left\{c_{\mathcal{D}}(Q),\delta\right\}}\times 100 \qquad (۱۲)$$

که $c_{\mathcal{D}}(Q)$ و $c_{\widehat{\mathcal{D}}}(Q)$ به‌ترتیب پاسخ‌های واقعی و نویزی به Q هستند (هنگامی که Q به‌ترتیب بر روی $\mathcal{D}$ و $\widehat{\mathcal{D}}$ اجرا می‌شود) و δ کرانی است که برای کاهش تأثیر پرس‌وجوهای شمارشی با پاسخ‌های واقعی خیلی کوچک بر خطای نسبی استفاده می‌شود. در این مقاله، مشابه با کارهای پیشین [۲۱، ۲۹]، δ به یک‌دهم درصد از تعداد کل مسیرهای حرکت در $\mathcal{D}$ تنظیم می‌شود.

همبستگی رتبه مکان‌ها

ضریب همبستگی رتبه کندال[1] یک سنجه آماری است که برای اندازه‌گیری همبستگی رتبه بین دو کمیت دلخواه استفاده می‌شود. در صورتی که مشاهدات بین دو متغیر رتبه مشابهی داشته باشند، ضریب همبستگی رتبه کندال بین آن دو متغیر زیاد و در غیر این صورت کم خواهد بود. از این ضریب همبستگی می‌توان برای محاسبه همبستگی رتبه مکان‌ها در $\mathcal{D}$ و $\widehat{\mathcal{D}}$ استفاده کرد. فرض کنید $\mathcal{C}$ مجموعه همه سلول‌ها در یک دامنه فضایی گسسته باشد. زوج سلول‌های $C_i, C_j \in \mathcal{C}$ موافق نامیده می‌شوند اگر یکی از شرایط زیر برقرار باشد:

$$\begin{aligned} &c_{\mathcal{D}}(C_i) > c_{\mathcal{D}}(C_j) \text{ and } c_{\widehat{\mathcal{D}}}(C_i) > c_{\widehat{\mathcal{D}}}(C_j),\\ &c_{\mathcal{D}}(C_i) < c_{\mathcal{D}}(C_j) \text{ and } c_{\widehat{\mathcal{D}}}(C_i) < c_{\widehat{\mathcal{D}}}(C_j) \end{aligned} \qquad (۱۳)$$

که $c_{\mathcal{D}}(\cdot)$ و $c_{\widehat{\mathcal{D}}}(\cdot)$ به‌ترتیب فراوانی یک سلول مفروض در $\mathcal{D}$ و $\widehat{\mathcal{D}}$ هستند. زوج سلول‌های $C_i, C_j \in \mathcal{C}$ مخالف نامیده می‌شوند اگر یکی از شرایط زیر برقرار باشد:

$$\begin{aligned} &c_{\mathcal{D}}(C_i) > c_{\mathcal{D}}(C_j) \text{ and } c_{\widehat{\mathcal{D}}}(C_i) < c_{\widehat{\mathcal{D}}}(C_j),\\ &c_{\mathcal{D}}(C_i) < c_{\mathcal{D}}(C_j) \text{ and } c_{\widehat{\mathcal{D}}}(C_i) > c_{\widehat{\mathcal{D}}}(C_j). \end{aligned} \qquad (۱۴)$$

با توجه به تعاریف فوق، همبستگی رتبه مکان‌ها به‌صورت زیر محاسبه می‌شود:

$$\tau_s=\frac{2}{m(m-1)}\left(\eta_s(\mathcal{D},\widehat{\mathcal{D}})-\tilde{\eta}_s(\mathcal{D},\widehat{\mathcal{D}})\right) \qquad (۱۵)$$

که $\eta_s(\mathcal{D},\widehat{\mathcal{D}})$ و $\tilde{\eta}_s(\mathcal{D},\widehat{\mathcal{D}})$ به‌ترتیب تعداد زوج سلول‌های موافق و مخالف را نمایش می‌دهند و m تعداد سلول‌های $\mathcal{C}$ است.

همبستگی رتبه الگوهای پرتکرار

فرض کنید $\mathcal{F}$ مجموعه‌ای از پرتکرارترین الگوها در $\mathcal{D}$ باشد. در این سنجه، از ضریب همبستگی رتبه کندال برای محاسبه همبستگی رتبه الگوهای پرتکرار استفاده می‌شود. در واقع، این سنجه شباهت‌ها و تفاوت‌های بین رتبه الگوهای پرتکرار از $\mathcal{F}$ در $\mathcal{D}$ و $\widehat{\mathcal{D}}$ را اندازه‌گیری می‌کند. همبستگی رتبه الگوهای پرتکرار به‌صورت زیر محاسبه می‌شود:

$$\tau_t=\frac{2}{k(k-1)}\left(\eta_t(\mathcal{D},\widehat{\mathcal{D}})-\tilde{\eta}_t(\mathcal{D},\widehat{\mathcal{D}})\right) \qquad (۱۶)$$

که $\eta_t(\mathcal{D},\widehat{\mathcal{D}})$ و $\tilde{\eta}_t(\mathcal{D},\widehat{\mathcal{D}})$ به‌ترتیب تعداد زوج الگوهای پرتکرار موافق و مخالف را نمایش می‌دهند و k تعداد الگوهای پرتکرار $\mathcal{F}$ است.

خطای سفر

نقاط شروع و انتهایی دو ویژگی مهم هر مسیر حرکت هستند.

[1] Kendall

این نقاط یک سفر خاص را نشان می‌دهند (برای مثال، سفر از خانه به محل کار یا سفر با تاکسی) و برای بسیاری از کارهای تحلیل داده از قبیل طراحی شهری و تحلیل بر اساس تقاضای مسافر دارای اهمیت هستند. از این سنجه برای اندازه‌گیری میزان اطلاعات حفظ‌شده در مورد نقاط شروع و انتهایی مسیرهای حرکت واقعی در مسیرهای حرکت مصنوعی استفاده می‌شود. در این مقاله، برای محاسبه خطای سفر، ابتدا نقاط شروع و انتهایی مسیرهای حرکت به سلول‌های دامنه فضایی گسسته نگاشت می‌شوند. سپس واگرایی جنسن-شانون[1] بین توزیع سفر (توزیع تمام زوج سلول‌های ابتدایی و انتهایی ممکن) در $\mathcal{D}$ و توزیع سفر در $\widehat{\mathcal{D}}$ محاسبه می‌شود.

خطای طول

طول هر مسیر حرکت برابر با تعداد کل نقاط آن مسیر حرکت است. از این سنجه برای اندازه‌گیری میزان اطلاعات حفظ‌شده در مورد طول مسیرهای حرکت واقعی در مسیرهای حرکت مصنوعی استفاده می‌شود. خطای طول مسیر حرکت با واگرایی جنسن-شانون بین توزیع طول مسیرهای حرکت واقعی در $\mathcal{D}$ و توزیع طول مسیرهای حرکت مصنوعی در $\widehat{\mathcal{D}}$ محاسبه می‌شود.

۵-۳- مقایسه

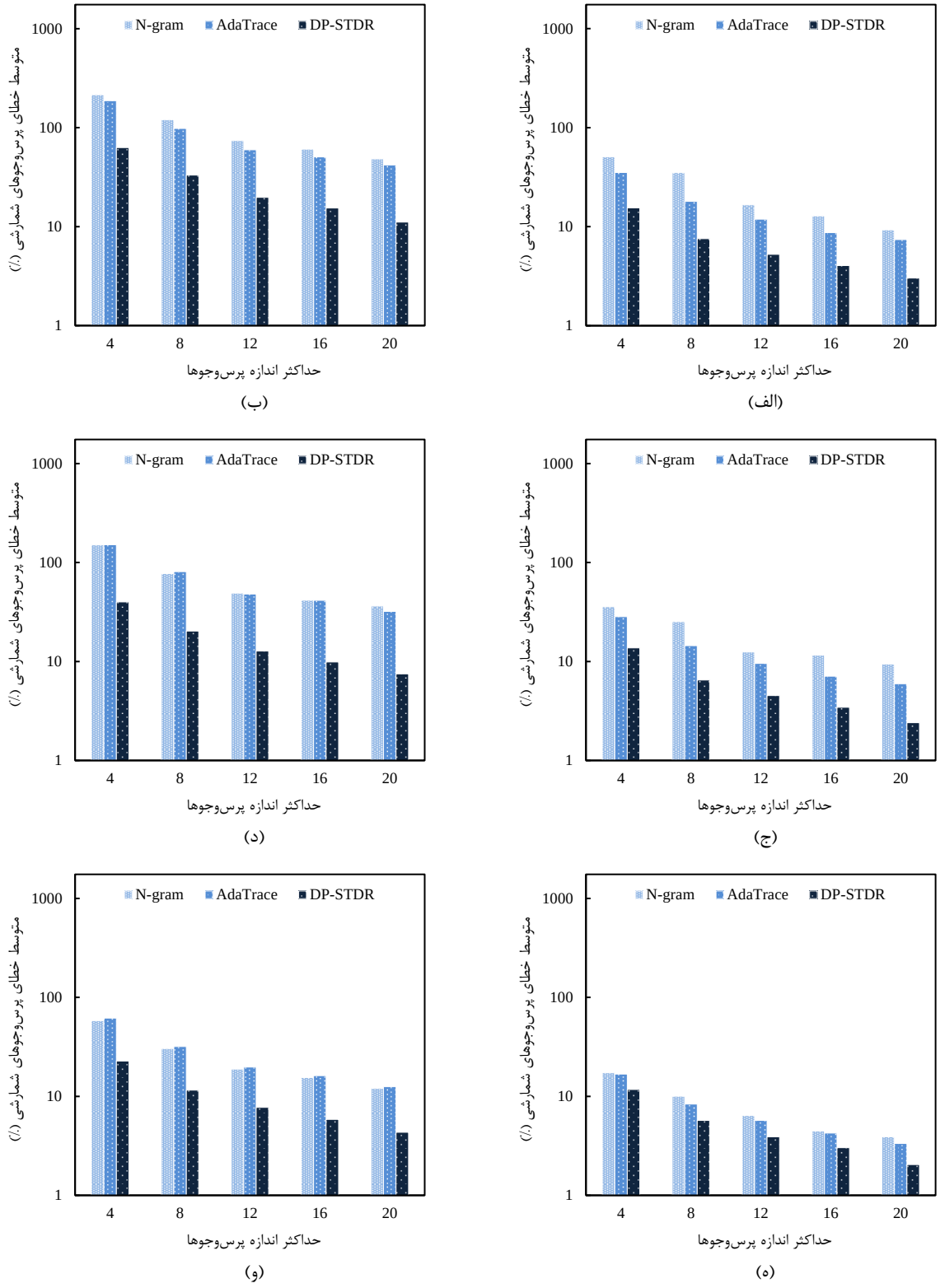
در ادامه، سودمندی‌های فضایی و زمانی DP-STDR با روش‌های N-gram [۲۱] و AdaTrace [۱۴] مقایسه می‌شود. برای N-gram، دامنه فضایی پیوسته به‌صورت یکنواخت به ۶۴ سلول برای مجموعه داده Geolife و به ۱۰۲۴ سلول برای مجموعه داده Taxi تقسیم می‌شود که در بیشتر سنجه‌ها به نتایج بهتری از سایر ریزدانگی‌ها منجر می‌شود.

ابتدا، سنجه خطای پرس‌وجوهای شمارشی مورد مقایسه قرار می‌گیرد. به این منظور، پنج مجموعه پرس‌وجوی شمارشی متفاوت روی هر مجموعه داده مسیر حرکت تولید می‌شود. حداکثر اندازه پرس‌وجوها در هر مجموعه داده متفاوت است (یعنی ۴، ۸، ۱۲، ۱۶ و ۲۰) و هر مجموعه داده شامل ۱۰۰۰۰ پرس‌وجوی شمارشی است که به‌صورت تصادفی تولید می‌شوند. هر مکان در یک پرس‌وجوی شمارشی به‌صورت یکنواخت از مجموعه سلول‌های دامنه فضایی گسسته انتخاب می‌شود. در شکل‌های (۲ و ۳) متوسط خطای پرس‌وجوهای شمارشی در DP-STDR با N-gram و AdaTrace مقایسه می‌شود. این مقایسه برای مجموعه پرس‌وجوهای شمارشی متفاوت و مقادیر مختلف ε روی دو مجموعه داده Geolife و Taxi انجام می‌شود. در اینجا، برای نمایش بهتر نتایج با توجه به اختلاف زیاد آن‌ها با یکدیگر، از مقیاس لگاریتمی به‌جای مقیاس خطی استفاده می‌شود. با توجه به شکل‌ها مشاهده می‌شود که متوسط خطای پرس‌وجوهای شمارشی DP-STDR اغلب بهتر از N-gram و AdaTrace است. دلیل بالابودن خطای پرس‌وجوهای شمارشی در N-gram، به‌ویژه برای مقادیر کوچک ε این است که N-gram از یک درخت اکتشاف برای پاسخ به پرس‌وجوهای شمارشی استفاده می‌کند. این درخت اکتشاف الگوهای حرکت مسیرهای حرکت واقعی را نگهداری می‌کند، اما بودجه حریم خصوصی بین سطوح مختلف آن تقسیم می‌شود که منجر به خطای بالای پرس‌وجوهای شمارشی برای مقادیر کوچک ε می‌شود (به‌ویژه برای مجموعه‌های داده‌ای مانند Geolife که در آن‌ها فراوانی مسیرهای حرکت واقعی کوچک است). در برخی موارد از سناریوی ریزدانه، N-gram می‌تواند برای مجموعه‌های داده بزرگ مانند Taxi (که در آن‌ها مسیرهای حرکت واقعی معمولاً فراوانی‌های بزرگ دارند) و مقادیر بزرگ ε (که منجر به افزودن مقدار نویز کوچکی به فراوانی‌ها می‌شود) بهتر عمل کند که این برتری به‌دلیل نگه‌داشتن الگوهای حرکت مسیرهای حرکت واقعی است. هرچند، دو نکته را باید در نظر گرفت: (۱) N-gram سربارهای فضایی و زمانی بالایی دارد و (۲) سودمندی خوب برای بودجه‌های حریم خصوصی کوچک از سودمندی خوب برای بودجه‌های حریم خصوصی بزرگ اهمیت بیشتری دارد. علاوه بر این، AdaTrace سودمندی خوبی در پاسخ به پرس‌وجوهای شمارشی ندارد که در نتیجه این واقعیت است که به‌صورت کامل روابط زمانی بین نقاط را حفظ نمی‌کند. همچنین، مشاهده می‌شود که متوسط خطای پرس‌وجوهای شمارشی در سناریوی درشت‌دانه بیشتر از سناریوی ریزدانه است. دلیل این امر این است که در سناریوی درشت‌دانه، پرس‌وجوهای شمارشی بر روی سلول‌های بزرگ‌تر تولید می‌شوند و بنابراین هنگام تولید پایگاه داده مسیر حرکت مصنوعی، خطای اعمال‌شده به سلول‌های کوچک با هم جمع می‌شوند.

در ادامه، سایر سنجه‌ها مورد ارزیابی قرار می‌گیرند. لازم به ذکر است که سنجه‌های همبستگی رتبه مکان‌ها و همبستگی رتبه الگوهای پرتکرار مقداری بین ۱- و ۱ دارند که مقادیر نزدیک‌تر به ۱ سودمندی بهتری را نشان می‌دهند. همچنین، سنجه‌های خطای سفر و خطای طول مقداری بین ۰ و ۱ دارند که مقادیر پایین‌تر سودمندی بهتری را نشان می‌دهند. برای محاسبه همبستگی رتبه الگوهای پرتکرار، ۵۰ الگوی پرتکرار مسیرهای حرکت واقعی در نظر گرفته می‌شوند. در جدول‌های (۲ و ۳) نتایج N-gram، AdaTrace و DP-STDR بر روی دو مجموعه داده Geolife و Taxi و برای مقادیر مختلف ε گزارش شده است. همان طور که مشاهده می‌شود، در اغلب موارد DP-STDR نتایج بهتری از N-gram و AdaTrace دارد. دلیل این

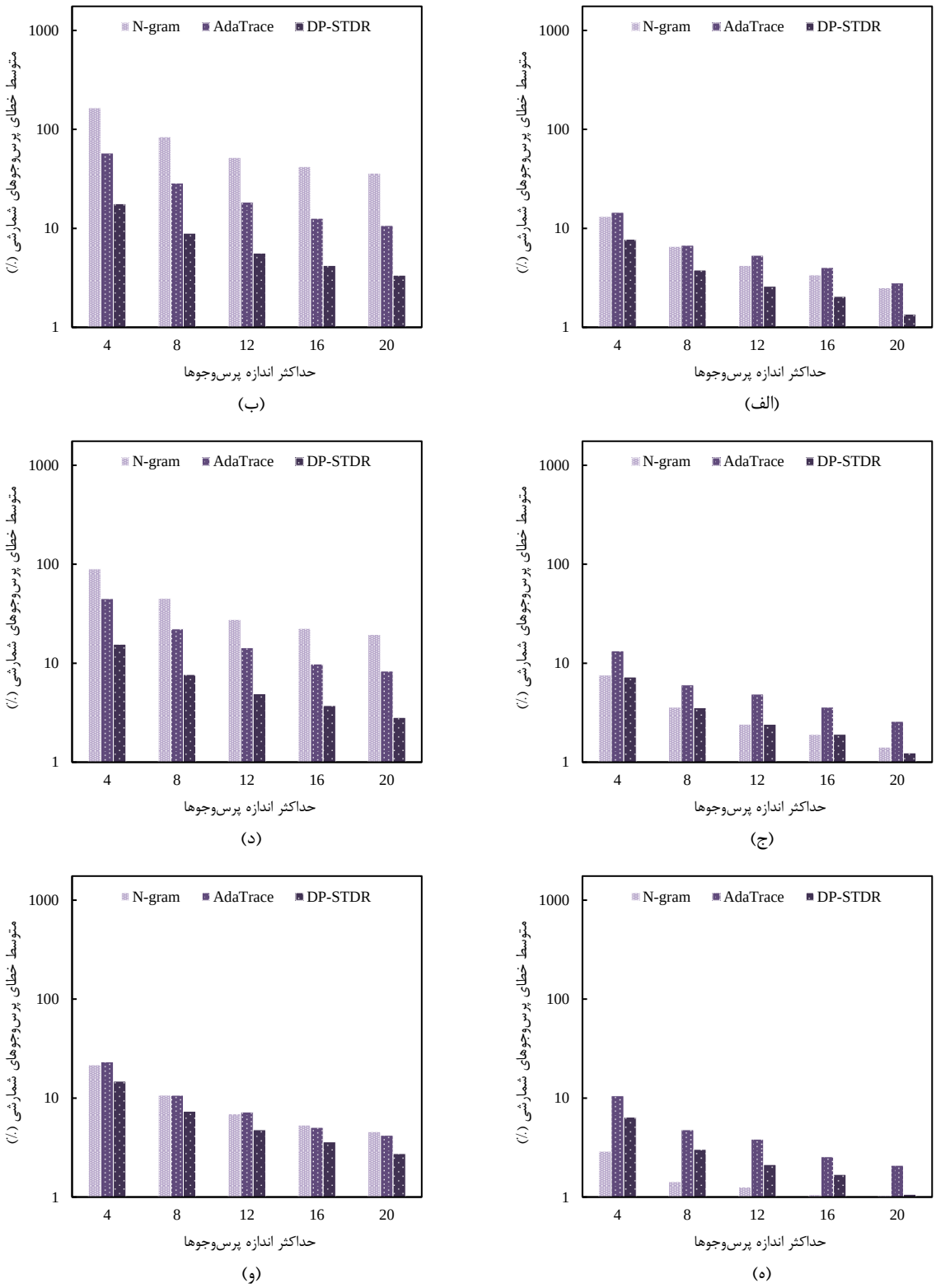
[1] Jensen–Shannon

امر این است که DP-STDR مسیرهای حرکت مصنوعی را به‌صورت پایین به بالا و با الحاق محتمل‌ترین مسیرها تولید می‌کند که موجب می‌شود الگوهای حرکت مسیرهای حرکت و فراوانی آن‌ها در پایگاه داده مسیر حرکت واقعی را به‌صورت کارا حفظ کند.



شکل (۲): متوسط خطای پرس‌وجوهای شمارشی N-gram، AdaTrace و DP-STDR روی مجموعه داده Geolife برای مجموعه پرس‌وجوهای شمارشی متفاوت و مقادیر مختلف ε. (الف) ریزدانه، $\varepsilon = 0.05$. (ب) درشت‌دانه، $\varepsilon = 0.05$. (ج) ریزدانه، $\varepsilon = 0.1$. (د) درشت‌دانه، $\varepsilon = 0.1$. (ه) ریزدانه، $\varepsilon = 0.5$. (و)

درشت‌دانه، $\varepsilon = 0.5$.



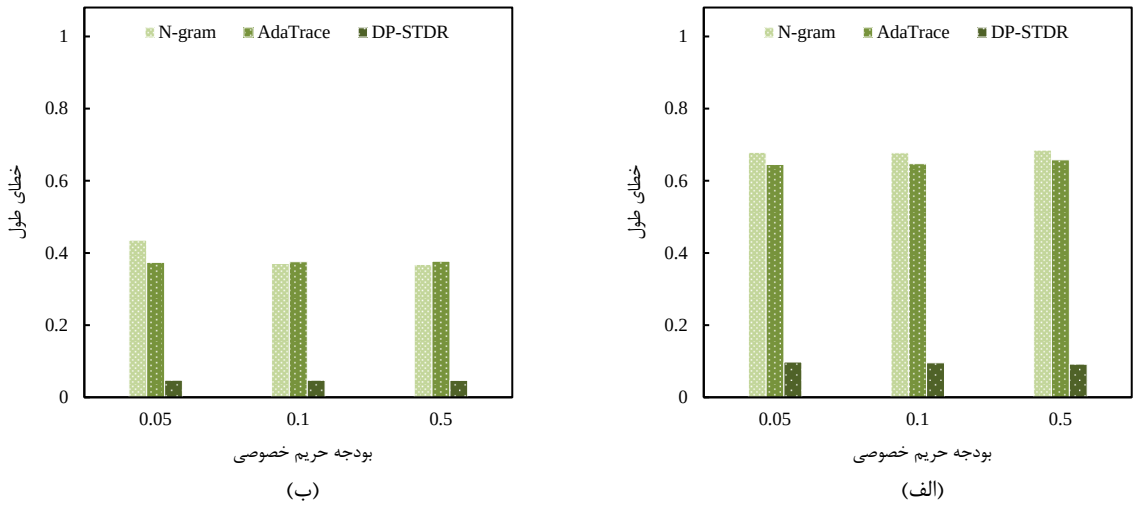
شکل (۳): متوسط خطای پرس‌وجوهای شمارشی N-gram، AdaTrace و DP-STDR روی مجموعه داده Taxi برای مجموعه پرس‌وجوهای شمارشی متفاوت و مقادیر مختلف ε. (الف) ریزدانه، $\varepsilon = 0.05$. (ب) درشت‌دانه، $\varepsilon = 0.05$. (ج) ریزدانه، $\varepsilon = 0.1$. (د) درشت‌دانه، $\varepsilon = 0.1$. (ه) ریزدانه، $\varepsilon = 0.5$. (و) درشت‌دانه، $\varepsilon = 0.5$.

جدول (۲): همبستگی رتبه مکان‌ها، همبستگی رتبه الگوهای پرتکرار و خطای سفر N-gram، AdaTrace و DP-STDR روی مجموعه داده Geolife برای مقادیر مختلف ε

سناریو	سنجه ارزیابی	N-gram			AdaTrace			DP-STDR		
		ε			ε			ε		
		۰/۰۵	۰/۱	۰/۵	۰/۰۵	۰/۱	۰/۵	۰/۰۵	۰/۱	۰/۵
ریزدانه	همبستگی رتبه مکان‌ها	۰/۰۱	۰/۰۰	۰/۰۱	**۰/۲۵**	**۰/۲۸**	**۰/۳۷**	۰/۱۷	۰/۱۸	۰/۲۳
	همبستگی رتبه الگوهای پرتکرار	۰/۴۸	۰/۳۷	۰/۱۲	۰/۰۱	۰/۰۱	۰/۰۲	**۰/۹۵**	**۰/۹۹**	**۱/۰۰**
	خطای سفر	۰/۴۲	۰/۴۰	۰/۴۱	۰/۶۷	۰/۶۶	۰/۶۱	**۰/۳۳**	**۰/۳۰**	**۰/۲۶**
درشت‌دانه	همبستگی رتبه مکان‌ها	۰/۳۵	۰/۳۳	۰/۴۹	۰/۴۴	۰/۵۰	۰/۵۹	**۰/۴۷**	**۰/۵۱**	**۰/۷۰**
	همبستگی رتبه الگوهای پرتکرار	۰/۳۴	**۰/۴۶**	۰/۰۶	۰/۰۳	۰/۰۳	۰/۰۲	**۰/۳۸**	۰/۴۰	**۰/۴۱**
	خطای سفر	۰/۲۸	۰/۲۸	۰/۲۴	۰/۵۲	۰/۴۸	۰/۳۴	**۰/۱۶**	**۰/۱۲**	**۰/۱۲**

جدول (۳): همبستگی رتبه مکان‌ها، همبستگی رتبه الگوهای پرتکرار و خطای سفر N-gram، AdaTrace و DP-STDR روی مجموعه داده Taxi برای مقادیر مختلف ε

سناریو	سنجه ارزیابی	N-gram			AdaTrace			DP-STDR		
		ε			ε			ε		
		۰/۰۵	۰/۱	۰/۵	۰/۰۵	۰/۱	۰/۵	۰/۰۵	۰/۱	۰/۵
ریزدانه	همبستگی رتبه مکان‌ها	۰/۵۰	۰/۵۵	۰/۶۶	۰/۵۳	۰/۵۷	۰/۶۲	**۰/۶۵**	**۰/۶۷**	**۰/۷۰**
	همبستگی رتبه الگوهای پرتکرار	**۰/۵۷**	**۰/۶۶**	**۰/۷۳**	۰/۱۵	۰/۱۷	۰/۲۰	۰/۴۳	۰/۴۴	۰/۴۶
	خطای سفر	**۰/۲۹**	**۰/۲۴**	**۰/۲۸**	۰/۳۴	۰/۳۳	۰/۲۷	۰/۴۹	۰/۴۸	۰/۴۷
درشت‌دانه	همبستگی رتبه مکان‌ها	۰/۷۲	۰/۷۶	۰/۸۵	۰/۶۶	۰/۷۳	۰/۸۳	**۰/۸۷**	**۰/۸۸**	**۰/۸۸**
	همبستگی رتبه الگوهای پرتکرار	۰/۶۸	۰/۷۱	۰/۶۲	۰/۳۷	۰/۳۶	۰/۳۵	**۰/۷۰**	**۰/۷۵**	**۰/۸۴**
	خطای سفر	**۰/۰۸**	**۰/۰۷**	**۰/۰۸**	۰/۲۱	۰/۱۹	۰/۱۰	۰/۱۶	۰/۱۷	۰/۱۹



شکل (۴): خطای طول N-gram، AdaTrace و DP-STDR روی دو مجموعه داده Geolife و Taxi برای مقادیر مختلف ε. (الف) Geolife. (ب) Taxi.

در نهایت، خطای طول مورد ارزیابی قرار می‌گیرد. لازم به ذکر است که سنجه خطای طول به گسسته‌سازی دامنه فضایی پیوسته وابسته نیست (برای اطلاعات بیشتر به تعریف این سنجه مراجعه شود) و بنابراین مقادیر یکسانی را برای سناریوهای ریزدانه و درشت‌دانه نتیجه می‌دهد. در شکل (۴) خطای طول N-gram، AdaTrace و DP-STDR روی دو مجموعه داده Geolife و Taxi و برای مقادیر مختلف ε نمایش داده شده است. همان طور که مشاهده می‌شود، DP-STDR خطای طول کمتری از N-gram و

AdaTrace دارد. دلیل این امر این است که DP-STDR تلاش می‌کند تا با نگه‌داشتن میانه نویزی طول‌های مسیرهای حرکت واقعی که در هر سلول شروع می‌شوند، توزیع طول مسیرهای حرکت را حفظ کند. لازم به ذکر است که توزیع طول مسیرهای حرکت یکی از مهم‌ترین ویژگی‌های یک پایگاه داده مسیر حرکت است. همان طور که در این شکل مشاهده می‌شود، N-gram و AdaTrace نمی‌توانند این توزیع را به‌خوبی حفظ کنند.

۶- نتیجه‌گیری

در این مقاله، سازوکاری به نام DP-STDR برای انتشار پایگاه‌های داده مسیر حرکت با ضمانت حریم خصوصی تفاضلی و با هدف حفظ سودمندی‌های فضایی و زمانی پیشنهاد شده است. در این سازوکار، ابتدا برخی ویژگی‌های مفید پایگاه داده مسیر حرکت واقعی، از قبیل توزیع نقاط شروع، الگوهای حرکت و توزیع طول مسیرهای حرکت با ضمانت حریم خصوصی تفاضلی استخراج می‌شود. سپس تعدادی درخت مسیر نویزی برای نگهداری محتمل‌ترین مسیرهای موجود با طول‌ها و سلول‌های ابتدایی مختلف ساخته می‌شود. در نهایت، مسیرهای حرکت مصنوعی با ضمانت حریم خصوصی تفاضلی و با توجه این ویژگی‌های فضایی، زمانی و آماری تولید می‌شوند. هر مسیر حرکت مصنوعی به‌صورت پایین به بالا و با الحاق محتمل‌ترین مسیرها در درخت‌های مسیر نویزی تولید می‌شود.

آزمایش‌های انجام‌شده روی دو مجموعه داده مسیر حرکت با استفاده از چندین سنجه ارزیابی فضایی و زمانی نشان می‌دهند که در مقایسه با کارهای مرتبط پیشین، DP-STDR سودمندی پاسخ پرس‌وجوها را افزایش داده و بسیاری از ویژگی‌های فضایی، زمانی و آماری مسیرهای حرکت واقعی را بهتر حفظ می‌کند.

۷- مراجع

[1] B. C. M. Fung, K. Wang, R. Chen, and P. S. Yu, "Privacy-preserving data publishing: A survey of recent developments," ACM Comput. Surv., vol. 42, no. 4, pp. 1–14, Jun. 2010.

[2] C. Dwork, "Differential privacy," In Automata, Languages and Programming (M. Bugliesi, B. Preneel, V. Sassone, and I. Wegener, eds.), Lecture Notes in Computer Science, pp. 1–12, Berlin, Heidelberg, Germany: Springer, 2006.

[3] T. Zhu, G. Li, W. Zhou, and P. S. Yu, "Differentially private data publishing and analysis: A survey," IEEE Trans. Knowl. Data Eng., vol. 29, no. 8, pp. 1619–1638, Aug. 2017.

[4] N. Niknami, M. Abadi, and F. Deldar, "SpatialPDP: A personalized differentially private mechanism for range counting queries over spatial databases," In Proceedings of the 2014 4th International Conference on Computer and Knowledge Engineering, Mashhad, Iran, pp. 709–715, Oct. 2014.

[5] F. Deldar and M. Abadi, "PLDP-TD: Personalized-location differentially private data analysis on trajectory databases," Pervasive Mob. Comput., vol. 49, pp. 1–22, Sep. 2018.

[6] G. Cormode, T. Kulkarni, and D. Srivastava, "Answering range queries under local differential privacy," Proc. VLDB Endow., vol. 12, no. 10, pp. 1126–1138, Jun. 2019.

[7] K. Al-Hussaeni, B. C. M. Fung, F. Iqbal, J. Liu, and P. C. K. Hung, "Differentially private multidimensional data publishing," Knowl. Inf. Syst., vol. 56, no. 3, pp. 717–752, Sep. 2018.

[8] C. Piao, Y. Shi, J. Yan, C. Zhang, and L. Liu, "Privacy-preserving governmental data publishing: A fog-computing-based differential privacy approach," Future Gener. Comput. Syst., vol. 90, pp. 158–174, Jan. 2019.

[9] Z. Zheng, T. Wang, J. Wen, S. Mumtaz, A. K. Bashir, and S. H. Chauhdary, "Differentially private high-dimensional data publication in Internet of Things," IEEE Internet Things J., vol. 7, no. 4, pp. 1–10, Apr. 2020.

[10] R. Chen, B. C. M. Fung, B. C. Desai, and N. M. Sossou, "Differentially private transit data publication: A case study on the Montreal transportation system," In Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Beijing, China, pp. 213–221, Aug. 2012.

[11] X. He, G. Cormode, A. Machanavajjhala, C. M. Procopiuc, and D. Srivastava, "DPT: Differentially private trajectory synthesis using hierarchical reference systems," Proc. VLDB Endow., vol. 8, pp. 1154–1165, Jul. 2015.

[12] S. Wang, R. Sinnott, and S. Nepal, "Privacy-protected statistics publication over social media user trajectory streams," Future Gener. Comput. Syst., vol. 87, pp. 792–802, Oct. 2018.

[13] F. Deldar and M. Abadi, "PDP-SAG: Personalized privacy protection in moving objects databases by combining differential privacy and sensitive attribute generalization," IEEE Access, vol. 7, pp. 85887–85902, Jun. 2019.

[14] M. E. Gursoy, L. Liu, S. Truex, L. Yu, and W. Wei, "Utility-aware synthesis of differentially private and attack-resilient location traces," In Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security, Toronto, Canada, pp. 196–211, Jan. 2018.

[15] M. E. Gursoy, L. Liu, S. Truex, and L. Yu, "Differentially private and utility preserving publication of trajectory data," IEEE Trans. Mob. Comput., vol. 18, no. 10, pp. 2315–2329, Oct. 2019.

[16] N. Holohan, D. J. Leith, and O. Mason, "Differential privacy in metric spaces: Numerical, categorical and functional data under the one roof," Inf. Sci., vol. 305, pp. 256–268, Jun. 2015.

[17] J. Zhang, X. Xiao, and X. Xie, "PrivTree: A differentially private algorithm for hierarchical decompositions," In Proceedings of the 2016 ACM SIGMOD International Conference on Management of Data, San Francisco, CA, USA, pp. 155–170, Jun. 2016.

[18] C. Xu, J. Ren, Y. Zhang, Z. Qin, and K. Ren, "DPPro: Differentially private high-dimensional data release via random projection," IEEE Trans. Inf. Forensics Secur., vol. 12, no. 12, pp. 3081–3093, Dec. 2017.

[19] G. Cormode, S. Jha, T. Kulkarni, N. Li, D. Srivastava, and T. Wang, "Privacy at scale: Local differential privacy in practice," In Proceedings of the 2018 ACM SIGMOD International Conference on Management of Data, Houston, TX, USA, pp. 1655–1658, May 2018.

[20] F. Deldar and M. Abadi, "Differentially private count queries over personalized-location trajectory databases,"

Annual IEEE Symposium on Foundations of Computer Science, Providence, RI, USA, pp. 94–103, Oct. 2007.

[26] Z. Jorgensen, T. Yu, and G. Cormode, "Conservative or liberal? personalized differential privacy," In Proceedings of the 2015 IEEE 31st International Conference on Data Engineering, Seoul, South Korea, pp. 1023–1034, Apr. 2015.

[27] N. Kohli and P. Laskowski, "Epsilon voting: Mechanism design for parameter selection in differential privacy," In Proceedings of the 2018 IEEE Symposium on Privacy-Aware Computing, Washington, DC, USA, pp. 19–30, Sep. 2018.

[28] Y. Zheng, L. Zhang, X. Xie, and W.-Y. Ma, "Mining interesting locations and travel sequences from GPS trajectories," In Proceedings of the 18th International Conference on World Wide Web, Madrid, Spain, pp. 791–800, Apr. 2009.

[29] W. Qardaji, W. Yang, and N. Li, "Differentially private grids for geospatial data," In Proceedings of the 2013 IEEE 29th International Conference on Data Engineering, Brisbane, Australia, pp. 757–768, Apr. 2013.

Data Brief, vol. 20, pp. 1510–1514, Oct. 2018.

[21] R. Chen, G. Acs, and C. Castelluccia, "Differentially private sequential data publication via variable-length n-grams," In Proceedings of the 2012 ACM SIGSAC Conference on Computer and Communications Security, Raleigh, NC, USA, pp. 638–649, Oct. 2012.

[22] S. Wang and R. O. Sinnott, "Protecting personal trajectories of social media users through differential privacy," Comput. Secur., vol. 67, pp. 142–163, Jun. 2017.

[23] M. Li, L. Zhu, Z. Zhang, and R. Xu, "Achieving differential privacy of trajectory data publishing in participatory sensing," Inf. Sci., vol. 400, pp. 1–13, Aug. 2017.

[24] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," In Theory of Cryptography (S. Halevi and T. Rabin, eds.), Lecture Notes in Computer Science, pp. 265–284, Berlin, Heidelberg, Germany: Springer, 2006.

[25] F. McSherry and K. Talwar, "Mechanism design via differential privacy," In Proceedings of the 2007 48th

Journal of Electronical & Cyber Defence
Vol. 9, No. 1, 2021, Serial No. 33

Trajectory Database Release with Differential Privacy Guarantee

F. Deldar, M. Abadi*

* Tarbiat Modares University

(Received: 14/03/2020, Accepted: 05/08/2020)

ABSTRACT

Over the last years, several differentially private mechanisms have been proposed to answer statistical queries over trajectory databases. However, most of these mechanisms aim to answer statistical queries without releasing trajectories. In this paper, we present DP-STDR; a new differentially private mechanism that releases synthetic trajectories for data analysis purposes while preserving spatial and temporal utilities. DP-STDR keeps some main spatial, temporal, and statistical properties of original trajectories and defines a new differentially private tree structure to keep the most probable paths with different lengths and different starting points. This tree structure is used to generate synthetic trajectories. Our experiments show that DP-STDR enhances the utility of query answers and better preserves the main spatial, temporal, and statistical properties of original trajectories in comparison to prior related work.

Keywords: Differential Privacy, Trajectory Database Release, Noisy Path Tree, Trajectory Pattern

* Corresponding Author Email: abadi@modares.ac.ir