



ارائه یک پروتکل جدید برای انتشار داده‌ها با حفظ حریم خصوصی در محیط‌های توزیع شده مبتنی بر مدل‌های احتمالاتی

الیاس مسیبی^۱، رضا ابراهیمی آتانی^{۲*}

دانش آموخته کارشناسی ارشد گروه مهندسی کامپیوتر دانشگاه گیلان، رشت، ایران^۱

دانشیار گروه مهندسی کامپیوتر دانشگاه گیلان، رشت، ایران^{۲*}

چکیده

امروزه بسیاری از خدمات در بخش خصوصی و دولتی به صورت الکترونیکی ارائه می‌شوند. اطلاعات ایجاد شده به وسیله این خدمات به طور معمول شامل رکوردهایی با اطلاعات حساس و خصوصی افرادی است که از این خدمات استفاده می‌کنند. سازمان‌های ارائه دهنده خدمات، موظفانند با ارائه خدمات امن و مورد اعتماد از نقض حریم خصوصی افراد جلوگیری کنند؛ از سوی دیگر اطلاعات ذخیره شده با حجم بالا منبع مناسبی برای کشف دانش به شمار می‌روند و با انتشار و اشتراک اطلاعات می‌توان به این مهم دست یافت که ممکن است به ایجاد چالش امنیتی و نقض حریم خصوصی منجر شود. راه کارهای ارائه شده در انتشار با حفظ محرمانگی داده‌های توزیع شده بیشتر با استفاده از روش‌های شخص سوم مورد اعتماد و محاسبات امن چندگانه همراه‌اند. این روش‌ها شامل مشکلات و چالش‌های متعدد امنیتی از قبیل ارتباط، هماهنگی و تبانی شرکت کنندگان در اشتراک داده، نبود شخص سوم قابل اعتماد، چالش‌های امنیتی پروتکل‌های محاسبات امن چندگانه و همچنین حملات درون سازمانی مشارکت کنندگان است. در این مقاله علاوه بر مرور روش‌های حفظ محرمانگی و حریم خصوصی در داده‌های توزیع شده و بررسی نقاط ضعف و قدرت آن‌ها، طرح جدیدی بر پایه مدل احتمالاتی برای حفظ حریم خصوصی در حوزه توزیع شده ارائه می‌شود. نتایج به دست آمده از ارزیابی امنیتی نشان می‌دهد روش پیشنهادی، بدون نیاز به شخص سوم مورد اعتماد و یا استفاده از محاسبات امن چندگانه، علاوه بر مقاومت در برابر حملات خارجی، در مقابل حملات داخلی که توسط مشارکت کنندگان پی‌ریزی می‌شود نیز مقاومت مناسبی دارد؛ به نحوی که در تمام حالات کمتر از یک درصد از رکوردهای منتشر شده نقض می‌شوند؛ همچنین نتایج حاصل از دقت طبقه بندی نشان می‌دهد، در این روش با توجه به اشتراک داده، دقت داده‌ها نسبت به حالتی که تنها یک فراهم کننده قصد انتشار داده را دارد افزایش می‌یابد؛ اما از سوی دیگر استفاده از روش پیشنهادی افزایش سربار پردازشی را به دنبال دارد؛ به طوری که این روش در تمامی حالات نسبت به روش پایه کندتر عمل می‌کند.

واژگان کلیدی: اشتراک داده، امنیت، انتشار داده، محرمانگی، حفظ حریم خصوصی، داده کاوی.

A Novel Privacy-Preserving Distributed Data Publishing Protocol Based on Probabilistic Models

Elyas Mosayebi¹, Reza Ebrahimi Atani^{2*}

M.Sc. in Computer Engineering, University of Guilan, Rasht, Iran¹

Associated Professor, Department of Computer Engineering, University of Guilan, Rasht, Iran^{2*}

Abstract

In the era of digital transformation, government agencies and corporations increasingly rely on electronic services, generating vast volumes of sensitive data stored in distributed databases. While these records hold immense potential for knowledge discovery through data mining, their publication or sharing raises critical privacy concerns, particularly when sensitive individual information is at risk. Traditional Privacy-Preserving Distributed Data Publishing (PPDDP) methods rely heavily on Trusted Third-Party (TTP) intermediaries and Secure Multi-Party Computation (SMC), which introduce systemic vulnerabilities such as communication bottlenecks, synchronization failures, insider attacks, and inherent distrust in centralized entities. In healthcare analytics, hospitals leverage patient data to enhance diagnostic precision, optimize clinical workflows, and advance preventive and precision medicine. Yet, reliance on siloed datasets from individual institutions often restricts model generalizability and impedes comprehensive insights into health outcomes. Patient health is a multidimensional construct influenced not only by genetic and biological factors but also by behavioral patterns and socio-environmental determinants. Cross-institutional collaboration integrating diverse

* Corresponding author

* نویسنده عهده دار مکاتبات

datasets from geographically distributed sources is essential to develop robust analytical models. However, such collaboration raises critical privacy concerns, as centralized aggregation of sensitive data risks exposure to breaches or misuse. Our probabilistic framework for privacy-preserving distributed data publishing directly addresses this challenge. By eliminating dependencies on trusted third parties and secure multi-party computation, our approach enables secure, decentralized integration of heterogeneous healthcare data. Through uncertainty-aware probabilistic anonymization and adaptive noise injection, the framework ensures compliance with stringent privacy regulations (e.g., GDPR, CPRA, HIPAA) while preserving the analytical utility required for accurate, actionable health outcome predictions. This balance of *utility and privacy* empowers researchers to harness the full potential of distributed datasets without compromising individual confidentiality, ultimately fostering innovation in precision medicine and population health management. This paper introduces a novel probabilistic framework for privacy preservation in distributed environments, eliminating dependencies on TTP and SMC. Unlike existing approaches, this method leverages uncertainty-aware probabilistic models to dynamically anonymize and perturb data across distributed nodes while preserving global data utility. First a survey of privacy preservation data publishing methods is presented in this paper and then we discuss about pros and cons of the techniques. After this we present the model and its implementation details. The results obtained by security evaluations shows that the presented method will balance out the privacy security and the accuracy of distributed data better, using the probability model without needing a Trusted Third-Party and Secure Multi-party Computation.

Keywords: Data Mining, Data Publishing, Data sharing, Privacy Preserving, Security.

۱- مقدمه

امروزه نقش فناوری و خدمات الکترونیکی در زندگی روزمره کلیه افراد جامعه هر روز پررنگ تر می شود؛ تا جایی که کمابیش بیشتر خدمات بخش عمومی که در طول روز از آن بهره گرفته می شود، به وسیله سرویس های دولت یا بخش خصوصی به صورت الکترونیکی به شهروندان عرضه می شود؛ از همین رو بسیاری از شرکت ها و سازمان های ارائه دهنده این سرویس ها، اطلاعات مربوط به خدمات مشتریان خود را ذخیره می کنند. ناگزیر، این اطلاعات می تواند شامل داده هایی باشد که محتوای داده بخشی از حریم خصوصی مشتریان باشد؛ به همین دلیل پاسداری و صیانت از این اطلاعات بسیار حائز اهمیت است. از سوی دیگر، اطلاعاتی که به مرور زمان جمع آوری می شوند؛ علاوه بر اینکه حکم تاریخچه را دارند، در صورت انتشار و یا برون سپاری می تواند منبع بسیار مناسبی برای کشف دانش [۱] به حساب آید. بدیهی است که اگر داده ها بدون هیچ گونه پردازشی انتشار داده شوند، حریم خصوصی مشتریان نقض می شود. انتشار با حفظ حریم خصوصی داده [۲ و ۵] (PPDP)^۱، راه کاری برای حل این مشکل است [۱]. در روش های PPDP سعی می شود داده ها به نحوی انتشار یابند که حریم خصوصی افراد نقض نشود.

اطلاعاتی که در قالب رکورد به وسیله شرکت هایی مانند بیمه، ثبت احوال، بیمارستان، بانک و غیره جمع آوری می شوند حاوی اطلاعات خصوصی و حساس احاد جامعه است. انتشار این داده ها باید تضمین کند که طی مراحل کشف دانش و یا هرگونه بررسی دیگر، اطلاعات حساس افراد فاش نشود؛ زیرا بر اساس اعلامیه جهانی حقوق بشر نباید به حریم خصوصی افراد خدشه ای وارد شود [۲]. مشکلات امنیتی که برای شرکت های Netflix و AOL اتفاق افتاد، مثالی از نقض حریم خصوصی در انتشار داده ها هستند [۳، ۴].

^۱ Privacy Preserving Data Publishing (PPDP)

همان طور که در بخش قبل بیان شد PPDP راه کاری برای حل این چالش است. دو رکن مهم در به کارگیری PPDP، حفظ حریم خصوصی و سودمندی داده ها است [۵]؛ به نحوی که با برقراری امنیت حریم خصوصی افراد، نتایج حاصل از کشف دانش نیز ارزش قابل قبولی داشته باشند. به طور معمول در دنیای واقعی شرکت ها به دلایل مختلف برای مثال رسیدن به سودمندی بیشتر یا یک پارچه سازی داده ها، اقدام به انتشار اشتراکی آن ها می کنند؛ به عنوان یک مثال کاربردی می توان به چند شرکت بیمه که برای ساخت سامانه کشف تقلب [۳]، داده های خود را به صورت اشتراکی انتشار می دهند، اشاره کرد. در این حالت چالش های جدیدی برای انتشار اشتراکی به وجود می آید. استفاده از روش های مرسوم PPDP به نحوی که برای حالت اشتراکی مناسب باشند یکی از این چالش هاست؛ از طرف دیگر ارتباط و هماهنگی شرکت کنندگان [۴] در اشتراک داده و همچنین محدودیت هایی که اشتراک داده برای شرکت کنندگان ایجاد می کند نیز از مسائل مهم این مبحث است. در بسیاری از پژوهش ها از شخص سوم مورد اعتماد [۵] برای عملیات گمنام سازی استفاده می شود. در مجموع برای حالتی که چندین شرکت یا سازمان قصد انتشار داده اشتراکی را دارند به دلیل ذات توزیعی انتشار، چالش های جدیدی برای عملیات انتشار به وجود می آید که حل آن ها هدف اصلی مقاله پیش رو است.

در این مقاله مشکلات موجود در انتشار با حفظ محرمانگی داده های توزیع شده (به اختصار PPDDP)^۲ مورد بررسی قرار می گیرد و سپس روش های موجود در این حوزه واریسی شده و به نقاط قوت و ضعف آن ها اشاره می شود.

^۲ Privacy Preserving Distributed Data Publishing (PPDDP)

و بهینه‌شدن آن روی زیرساخت‌های غیرمتمرکز مثل اینترنت اشیا و محاسبات لبه، به‌کارگیری آن در مقیاس سازمانی یا بین سازمانی که ساختار دادگان هم به‌صورت عمودی^۶ و هم افقی^۷ توسعه یافته‌اند و به‌طور اصولی هم ساختار نیستند، مناسب نیست[۴۶].

در این مقاله با درنظرگرفتن ساختار ارتباطات سازمانی و یا بین سازمانی، هدف پیاده‌سازی الگوریتمی برای انتشار با حفظ محرمانگی و حریم خصوصی داده‌های توزیع شده است که مبتنی بر روش‌های احتمالاتی باشد. این روش بر پایه قوانین احتمالاتی و با درنظرگرفتن اطلاعات آماری مجموعه داده موردنظر، عملیات گمنام‌سازی را انجام می‌دهد. روش پیشنهادی با درنظرگرفتن حفظ حریم خصوصی مالکین داده‌ها، ارتباط بین مجموعه‌های متعلق به شرکت‌کنندگان را به‌نحوی برقرار می‌کند که بتوانند عملیات گمنام‌سازی را بدون نیاز به TTP یا SMC انجام دهند.

در ادامه مقاله، ساختار نگارشی آن بر این اساس ارائه می‌شود: در بخش دوم مشخصات و ویژگی‌های داده‌های قابل انتشار، مدل‌ها و روش‌های حفظ محرمانگی و حریم خصوصی مختصر معرفی می‌شوند. در ادامه و در بخش سوم، انتشار با حفظ محرمانگی در داده‌های توزیع شده مورد بررسی قرار خواهد گرفت و چالش‌ها و حملات ممکن در PPDDP برشمرده می‌شوند. در بخش چهارم پیشینه این پژوهش ارائه می‌شود و راه‌کارهای کلی برای انتشار داده در حالت توزیعی مورد بررسی قرار می‌گیرد. در بخش پنجم، روشی برای انتشار داده‌های توزیع شده به‌نحوی که توازن مناسبی بین حفظ حریم خصوصی و دقت داده‌های منتشرشده برقرار شود، ارائه می‌شود؛ سپس پیاده‌سازی الگوریتم و نتایج حاصل از ارزیابی روش پیشنهادی مورد بررسی قرار می‌گیرد. سرانجام در بخش ششم، مقاله جمع‌بندی و نتیجه‌گیری می‌شود.

۲- مدل‌های محرمانگی و حملات در انتشار توزیع شده داده‌ها

کمابیش تمام مؤسسات و شرکت‌ها و یا حتی دولت‌ها، اطلاعات دیجیتالی مربوط به حوزه کاری خود را جمع‌آوری می‌کنند. این داده‌ها علاوه بر اینکه به‌عنوان تاریخچه به‌کار می‌آیند، برای عملیات کشف دانش نیز بسیار مناسب‌اند و دارندۀ داده^۸ می‌تواند از نتایج به‌دست‌آمده از داده‌کاوی برای پیشبرد اهداف مؤسسه یا شرکت استفاده کند، اما در بسیاری از مواقع به دلایل مختلفی مانند نبود کارشناس حوزه داده‌کاوی در شرکت موردنظر و یا اشتراک داده به‌منظور

بسیاری از روش‌های ارائه‌شده با استفاده از شخص سوم مورداعتماد (TTP)^۱ یا پروتکل‌های محاسبات امن چندگانه (SMC)^۲ قابل اجرا هستند. استفاده از TTP و SMC باعث بروز محدودیت و مشکلاتی برای شرکت‌کنندگان در انتشار داده می‌شود. هدف اصلی این مقاله ارائه روشی است که شرکت‌کنندگان در انتشار داده، بدون نیاز به TTP و SMC، اقدام به انتشار داده‌ها به‌صورت اشتراکی و با توازن مناسب میان امنیت و دقت داده‌ها کنند؛ همچنین در روش پیشنهادی، شرکت‌کنندگان در انتشار داده در حین گمنام‌سازی به‌صورت مستقل از هم عمل می‌کنند و وابستگی آن‌ها کاهش می‌یابد.

یکی از راه‌حل‌های اخیر برای حل چالش توزیع داده‌های غیرمتمرکز با حفظ حریم خصوصی استفاده از یادگیری فدرال^۳ است که برای رفع محدودیت‌های یادگیری ماشین متمرکز در سال ۲۰۱۷ به‌وسیله گوگل ارائه شد [۵۰, ۴۹, ۴۵]. این مدل یادگیری غیرمتمرکز با آموزش مدل‌ها به‌صورت محلی روی داده‌های توزیع شده، حریم خصوصی و امنیت را بر اساس ساختار معماری خود ارتقا می‌دهد؛ باین‌حال برای مقابله با مسائل امنیتی فزاینده، دولت‌ها مقرراتی برای حفاظت از داده‌ها ارائه کرده‌اند؛ از جمله مقررات عمومی حفاظت از داده (GDPR)^۴ در اروپا، قانون حقوق حریم خصوصی کالیفرنیا (CPRA)^۵ در ایالات متحده، و قانون حفاظت از داده‌های شخصی در سنگاپور. بر اساس گزارش‌های امنیتی ۴۹ درصد از سازمان‌ها در ایالات متحده دست‌کم یک نقض داده را در دوران کاری تجربه کرده‌اند. مورد مشابهی نیز برای گوگل پیش آمد که به‌دلیل نقض داده‌ها توسط GDPR مبلغ ۵۷ میلیون دلار جریمه و این جریمه در مارس ۲۰۲۰ اعمال شد [۴۴].

یادگیری فدرال پاسخ مناسبی برای برون‌سپاری داده‌ها در محیط‌های توزیع شده است. در این فرایند یادگیری، یک مدل با استفاده از چندین کاربر ساخته می‌شود و این کاربرها تمام داده‌ها را در طول فرایند آموزش به‌صورت محلی نگه می‌دارند؛ درنتیجه، این رویکرد از منابع داده متنوع برای بهبود امنیت مدل و حفظ حریم خصوصی داده‌ها استفاده می‌کند؛ بنابراین، پتانسیل یادگیری فدرال در کاربردهای واقعی مانند صفحه‌کلیدهای پیش‌بینی‌کننده در گوشی‌های هوشمند، مراقبت‌های بهداشتی [۷]، کشاورزی، امور مالی، شهرهای هوشمند، حمل‌ونقل و مدیریت انرژی مورد توجه قرار گرفته است؛ باین‌حال ساختار پیچیده این شیوه یادگیری

¹ Trusted Third Party (TTP)

² Secure Multi Party Computation (SMC)

³ Federated Learning

⁴ General Data Protection Regulation (GDPR)

⁵ California Privacy Rights Act (CPRA)

⁶ Vertical

⁷ Horizontal

⁸ Data Holder

کشف دانش بهتر و مطمئن تر با شرکت‌های دیگر، دارنده داده مجبور می‌شود مجموعه داده^۱ را برای عملیات داده‌کاوی برون‌سپاری کند.

بیشتر افراد خواستار حفظ حریم خصوصی خودشان و نگران عواقب ناشی از اشتراک اطلاعات شخصی هستند. در این حالت منتشرکننده داده به عنوان سازمان مورد اعتماد مالکان رکورد، باید تغییراتی روی مجموعه داده برای جلوگیری از نقض حریم خصوصی افراد اعمال کند و سپس آن را نشر دهد تا به دست دریافت کننده داده برسد.

در برخی مواقع داده‌ها برای یک عمل داده‌کاوی معین مثل دسته‌بندی، خوشه‌بندی و یا قوانین انجمنی آماده می‌شوند[۴۷]. به عملیات انتشار داده برای یک عمل داده‌کاوی خاص در یک محیط غیرمطمئن، حفظ حریم خصوصی در داده‌کاوی یا به اختصار^۲ PPDM گفته می‌شود[۸]. در این حالت امکان دارد دریافت کننده از یک عمل داده‌کاوی دیگر برای دسترسی به اطلاعات حساس استفاده کند[۵] یا مجموعه داده برای چندین دریافت کننده فرستاده شود. در این صورت تضمینی برای حفظ حریم خصوصی داده‌ها وجود ندارد.

به همین منظور راه کار دیگری برای انتشار مجموعه داده بیان شد که مجموعه داده به صورت کلی و در حالت عام برای دریافت کنندگان پردازش و گمنام سازی می‌شود. در این حالت داده برای هر عمل داده‌کاوی پردازش و منتشر می‌شود[۵]. PPDP به طور کلی گمنامی داده‌ها را با مخفی کردن شناسه‌های مالک رکورد انجام می‌دهد؛ درحالی که PPDM مستقیم سعی در مخفی کردن داده‌های حساس دارد. به تعبیری دیگر PPDP در سطح داده حفظ حریم خصوصی را برآورده می‌کند و PPDM در سطح پردازش این کار را انجام می‌دهد[۱].

۲-۱- انتشار با حفظ حریم خصوصی داده

PPDP از دو فاز جمع‌آوری داده و انتشار داده تشکیل شده است. در مرحله نخست منتشرکننده داده که به طور معمول یک سازمان مستقل است، اطلاعات را از کاربران واقعی جمع‌آوری می‌کند. در این حالت فرض بر این است که منتشرکننده داده نسبت به اطلاعات حساس افراد مسئول است و به عبارتی برای صاحبان رکورد، شخصی مورد اعتماد به حساب می‌آید، اما کسانی که داده‌های انتشار یافته را در اختیار می‌گیرند مورد اعتماد نیستند و اطلاعات حساس و شخصی افراد باید از آن‌ها مخفی نگاه داشته شود؛ بنابراین در مرحله بعد بر اساس مدل‌های حفظ حریم خصوصی، عملیاتی بر روی داده‌ها اعمال می‌شود و در نهایت خروجی در اختیار دریافت کنندگان داده قرار می‌گیرد. گفتنی است که هرچه میزان حفظ حریم خصوصی بیشتر مدنظر باشد، ممکن است

دقت^۳ نتایج به دست آمده از داده‌ها پایین تر بیاید و برعکس [۹،۱]؛ بنابراین عمل انتشار با حفظ محرمانگی داده باید تعادلی بین دقت نتایج و حفظ محرمانگی حریم خصوصی برقرار کند.

ارکان مختلفی در انتشار با حفظ محرمانگی وجود دارند که از آن جمله می‌توان به مجموعه داده، رکوردهای مجموعه داده، صفات مربوط به هر رکورد و همچنین روش‌هایی که برای حفظ محرمانگی مورد استفاده قرار می‌گیرند، اشاره کرد. در ادامه به بررسی این ارکان می‌پردازیم.

انواع مجموعه داده

داده‌های ذخیره شده به وسیله سازمان‌ها به دو صورت ساختاریافته و بدون ساختار تقسیم می‌شوند[۵]. داده‌های ساختاریافته داده‌هایی هستند که دارای شمای ثابتی برای تمام رکوردها باشند؛ به بیان دیگر مجموعه داده^۴ برای تمام رکوردها دارای خصوصیت‌های^۵ یکسان است[۱]؛ برای مثال می‌توان به مجموعه اطلاعات ذخیره شده برای بیماران یک بیمارستان اشاره کرد؛ از سوی دیگر داده‌های بدون ساختار، داده‌هایی هستند که خصوصیات رکوردهای آن با هم متفاوت باشد[۱۰].

صفات در مجموعه داده

به طور کلی صفات یا شناسه‌های مجموعه داده D را می‌توان در چهار دسته زیر طبقه‌بندی کرد[۵]:

شناسه صریح^۶: مجموعه صفاتی که به صراحت برای شناسایی مالک رکورد مورد استفاده قرار می‌گیرند؛ مانند کد ملی اشخاص.

شبه شناسه^۷: ترکیبی از صفات که هر یک به تنهایی برای شناسایی صاحب رکورد کافی نیستند؛ اما با به کارگیری ترکیبی از آن‌ها می‌توان مالک رکورد را تشخیص داد. مانند ترکیب کدپستی و تاریخ تولد و جنسیت.

صفات حساس^۸: صفاتی که صاحب رکورد مایل نیست عموم از آن‌ها اطلاع داشته باشند، مانند: نوع بیماری و میزان حقوق. این صفات برای داده‌کاوی و تجزیه و تحلیل آماری مورد استفاده قرار می‌گیرند.

صفات غیر حساس^۹: صفاتی که در هیچ کدام از دسته‌های پیشین جای نمی‌گیرند و در عملیات پردازش داده برای حفظ محرمانگی مورد استفاده نیستند.

در عملیات حفظ حریم خصوصی نخستین گام فراهم کننده داده، حذف شناسه‌های صریح از مجموعه داده است، اما تجربه‌های پیشین نشان می‌دهد که این کار

³ Accuracy

⁴ Data Set

⁵ Attributes

⁶ Explicit Identifier

⁷ Quasi Identifier

⁸ Sensitive Attributes

⁹ Non-Sensitive Attributes

¹ Data Set

² Privacy Preserving Data Mining

به‌تنهایی ضامن حفظ حریم خصوصی و اطلاعات حساس نیست[۱]. روش‌های مختلفی برای حفظ حریم خصوصی داده وجود دارد که در ادامه به دسته‌بندی آن‌ها می‌پردازیم.

۲-۲- روش‌های حفظ حریم خصوصی

روش‌های مختلفی برای انتشار با حفظ حریم خصوصی داده ارائه شده‌است که در تمام این روش‌ها هدف تغییر دادن داده‌ها به‌نحوی است که ارتباط بین اطلاعات حساس و افراد واقعی مشخص نباشد، اما به‌طور کلی می‌توان این روش‌ها را در چهار دسته طبقه‌بندی کرد[۵].

روش تصادفی

در روش تصادفی^۱ با اضافه کردن نوفه به داده اصلی از فاش شدن اطلاعات حساس جلوگیری می‌شود. این روش بیشتر برای مخفی کردن داده‌های عددی^۲ مورد استفاده قرار می‌گیرد. مشکل این نوع روش تقریبی بودن آن و همچنین از دست دادن داده‌های مفید به مقدار زیاد است[۵]. دقت و صحت نتیجه در این روش به میزان بزرگی تابعی که به‌عنوان نویز به داده اصلی اضافه می‌شود بستگی دارد[۱۱].

روش جایگزینی

در روش جایگزینی^۳ داده، ویژگی‌های حساس با مقداری دیگر جایگزین می‌شوند به‌نحوی که رابطه آماری بین داده‌ها تغییری نکند[۱۲]؛ در این حالت در سطح رکورد نمی‌توان به داده‌ها اطمینان کرد؛ یعنی در سطح مجموعه داده اطلاعات درستی به دست می‌آید؛ اما رکوردها ممکن است مربوط به افراد حقیقی نباشند. از این روش می‌توان برای ویژگی‌های عددی و طبقه‌بندی شده^۴ استفاده کرد[۱].

روش رمزنگاری

روش رمزنگاری^۵ برای داده‌های توزیع شده در محاسبات امن چندگانه (SMC) مورد استفاده قرار می‌گیرد[۵]. در این روش چند شریک با ورودی‌های خصوصی متعلق به خود، مایل به اجرای یک تابع خاص مانند یک عمل داده‌کاوی به‌صورت اشتراکی هستند. بر طبق تعریف، زمانی عملیات به‌صورت امن صورت می‌گیرد که هیچ‌کدام از شرکا چیزی بیشتر از خروجی موردانتظار درک نکنند[۱۳].

روش گمنام‌سازی

گمنام‌سازی^۶ رایج‌ترین روش برای حفظ حریم خصوصی در انتشار داده به‌شمار می‌آید. ایده اصلی روش‌های مبتنی بر این روش، همسان کردن رکوردهای مجموعه داده است؛ به‌نحوی که رکورد موردنظر برای شخص مهاجم^۷ به راحتی قابل تشخیص نباشد[۱]؛ به بیان دیگر، این روش رکوردها را

در گروه‌های هم‌ارزی^۸ قرار می‌دهد به‌صورتی که هر رکورد با رکوردهای هم‌گروه خود دارای شبه شناسه‌های مشابه باشند و بر این اساس از یکدیگر قابل تفکیک نباشند. عمل‌گرهایی که در ساخت گروه هم‌ارزی استفاده می‌شوند عمل‌گرهای گمنامی نام دارند. عمل‌گرهای گمنامی مختلفی برای گمنام‌سازی در مجموعه داده‌ها استفاده می‌شود، اما پرکاربردترین آن‌ها، عمومی‌سازی^۹ و فرونشانی^{۱۰} هستند.

عمل‌گر عمومی‌سازی: در عمل‌گر عمومی‌سازی مقدار یک صفت به بازه بزرگ‌تر تعمیم داده می‌شود؛ برای مثال مقدار روز تولد با ماه تولد جایگزین می‌شود و در مرحله بعد می‌توان ماه تولد را با سال تولد جایگزین کرد. عمومی‌سازی هر صفت به کمک درخت طبقه‌بندی^{۱۱} آن انجام می‌گیرد. تشکیل درخت طبقه‌بندی برای هر صفت توسط یک شخص خبره صورت می‌گیرد که دانش کاملی در حوزه مقادیر آن صفت دارد[۱۴].

عمل‌گر فرونشانی: عمل‌گر فرونشانی از عمومی‌سازی سخت‌گیرانه‌تر است و مقدار صفت را در خروجی حذف می‌کند. این عمل‌گر نیاز به درخت طبقه‌بندی ندارد؛ زیرا مقدار صفت یک رکورد را مستقیماً از برگ به ریشه درخت می‌نشانند و با داشتن ریشه هر صفت که (null؟ یا هر چیزی) می‌تواند باشد، این عمل‌گر قابل استفاده است، اما باید توجه داشت که استفاده نادرست از این عمل‌گر موجب کاهش شدید کیفیت مجموعه داده می‌شود[۱۴].

۲-۳- مدل‌های حفظ حریم خصوصی

اصولاً پس از انجام عملیات PPDP دانش شخص مهاجم در مورد اطلاعات افراد، پیش و پس از رؤیت مجموعه منتشر شده، نباید تفاوتی داشته باشد؛ به بیان ساده‌تر نباید دانش جدیدی بر دانسته‌های فرد مهاجم پس از رؤیت مجموعه داده منتشر شده اضافه شود، اما برآورده کردن این سطح از امنیت غیرممکن است. با در نظر گرفتن اطلاعاتی که ممکن است از قبل در اختیار مهاجم باشد این سطح از حریم خصوصی نقض خواهد شد (دانش پیش‌زمینه مهاجم) [۱]؛ به همین دلیل در بیشتر کارهایی که در حوزه حرمانگی داده انجام می‌گیرد از مدل‌های تعدیل‌یافته‌تری استفاده می‌شود که هر کدام می‌توانند در مقابل دسته‌ای از حملات از داده‌ها حفاظت کنند. این مدل‌ها به دو دسته اصلی مدل‌های پیوندهی^{۱۲} و مدل‌های احتمالاتی^{۱۳} تقسیم می‌شوند[۱۶، ۱۰].

⁸ Equivalence Group

⁹ Generalization

¹⁰ Suppression

¹¹ Taxonomy Tree

¹² Linkage Model

¹³ Probabilistic Model

¹ Randomization Method

² Numerical Attributes

³ Data Swapping Method

⁴ Categorical Attributes

⁵ Cryptographic Method

⁶ Anonymization Method

⁷ Adversary

مدل پیونددهی یا مدل نحوی^۱ [۱۷] برای مقابله با حمله پیونددهی^۲ طراحی شده است. در حمله پیونددهی مهاجم با بهره‌گیری از مجموعه داده منتشر شده و همچنین صفت، رکورد یا جدولی که از اطلاعات دانش عمومی یا پیش‌زمینه^۳ مهاجم فراهم می‌شود، اقدام به کشف اطلاعات شخصی افراد می‌کند یا به بیان بهتر به اطلاعاتی که از پیش داشته می‌افزاید. مدل پیونددهی خود به سه دسته تقسیم می‌شود [۱۰] که در ادامه به صورت مختصر هر کدام را توضیح می‌دهیم.

الف. مدل پیونددهی رکورد: در حملات پیونددهی رکورد مهاجم شبه شناسه‌های قربانی را با رکوردهای انتشار یافته تطبیق می‌دهد. بعد از انجام این کار دسته‌ای کوچک از رکوردها باقی می‌ماند که ممکن است متعلق به قربانی باشد. پیونددهی رکورد زمانی با موفقیت انجام می‌شود که مهاجم بتواند رکورد مربوط به قربانی را به طور یکتا شناسایی کند. روش پایه برای مقابله با این حملات k-anonymity است [۱]. این روش با استفاده از عمل‌گرهای گمنام‌سازی که پیش‌تر بیان شد؛ مانند فرونشانی و عمومی‌سازی، شبه شناسه‌های هر رکورد از جدول را دست‌کم با k-1 رکورد دیگر یکسان می‌سازد؛ به عبارتی دیگر باعث می‌شود در جدول به‌ازای هر مقدار از شبه شناسه‌ها، k رکورد وجود داشته باشد. در این حالت مهاجم در بیشترین حد با درجه اطمینان $1/k$ می‌تواند رکورد قربانی را شناسایی کند [۱۸].

ب. مدل پیونددهی صفت: در حمله پیونددهی صفت به‌دست آوردن صفت حساس قربانی، مهاجم را به هدف خود می‌رساند و نیازی به شناسایی رکورد نیست [۱]؛ برای مثال اگر مهاجم پس از تطبیق شبه شناسه‌ها به یک دسته هم‌ارزی رسید که صفت حساس تمام رکوردهای موجود در آن یکسان است، دیگر نیازی به تشخیص رکورد قربانی وجود ندارد؛ زیرا هم‌اکنون مقدار صفت حساس وی را به‌دست آورده است. یکی از روش‌های مورد استفاده برای مقابله با حمله پیونددهی صفت، روش l-diversity است [۱۹].

ج. مدل پیونددهی جدول: در مدل پیونددهی صفت و پیونددهی رکورد، فرض بر این بود که مهاجم از وجود رکورد قربانی در مجموعه داده منتشر شده اطمینان دارد. اما در مدل پیونددهی جدول این فرض برقرار نیست و اطمینان از وجود یا نبود رکورد قربانی در مجموعه داده توسط مهاجم، خود به‌عنوان نقض محرمانگی تلقی می‌شود؛ به عبارت دیگر مهاجم با دانستن اینکه رکورد مورد نظرش در مجموعه داده منتشر شده وجود دارد یا خیر به هدف خود رسیده است [۱]. روش δ -Presence برای مقابله با این نوع حمله ارائه شده است. در روش δ -Presence مهاجم کمتر از $\delta\%$ از وجود رکورد قربانی در جدول منتشر شده اطمینان دارد، احتمال

اینکه بتواند پیونددهی رکورد یا صفت را با موفقیت انجام دهد حداکثر $\delta\%$ است؛ بنابراین این روش برای مقابله با حملات پیونددهی رکورد و صفت نیز کارایی دارد. در محاسبه احتمال δ فرض می‌شود منتشرکننده داده و مهاجم به یک جدول خارجی مشترک دسترسی دارند. عیب عمده روش δ -Presence این است که چنین فرضی در دنیای واقعی زیاد محتمل نیست [۱].

• مدل احتمالاتی

در مدل‌های احتمالاتی یا معنایی^۴ سعی می‌شود تغییر اطمینان مهاجم قبل از مشاهده داده منتشر شده به مقدار کوچکی محدود شود [۱]. شخص مهاجم پیش از رؤیت داده‌ها، اطمینان مشخصی نسبت به دانش و مفروضات در مورد هدف خود دارد. پس از در اختیار گرفتن مجموعه داده این اطمینان نباید بیشتر شود؛ برای مثال اگر مهاجم در ابتدا با اطمینان ۱۰ درصد صفت حساس قربانی را بداند و پس از در اختیار گرفتن جدول منتشر شده اطمینان وی به ۸۰ درصد افزایش یابد این مدل نقض شده و امنیت مجموعه داده منتشر شده زیر سؤال می‌رود. حریم خصوصی تفاضلی معروف‌ترین روشی است که بر اساس مدل احتمالاتی ارائه شده است [۱]. این مدل بیان می‌کند که حذف یا اضافه کردن یک رکورد، نباید تأثیر چندانی بر نتایج به‌دست آمده از تجزیه و تحلیل مجموعه داده داشته باشد [۲۰]. به بیان ساده‌تر در این مدل انتشار یا عدم انتشار رکورد که متعلق به قربانی است، نباید تأثیری در دانش مهاجم داشته باشد.

۲-۴- انتشار با حفظ حریم خصوصی در داده‌های توزیع شده

در دنیای واقعی ممکن است تعداد منتشرکنندگان بیش از یک سازمان یا ارگان باشد. در برخی مواقع چندین منتشرکننده برای رسیدن به هدفی خاص اقدام به همکاری می‌کنند و یا اینکه داده‌های خود را در اختیار دریافت‌کننده یا دریافت‌کنندگان دیگر قرار می‌دهند؛ به‌عنوان مثال کاربردی می‌توان به همکاری دولت‌ها^۵ (OGP) اشاره کرد که در آن دولت‌ها برای رسیدن به اهداف مشترک و کشف دانش جدید اقدام به اشتراک اطلاعات می‌کنند [۱۳]، اما از سوی دیگر نگران استفاده سو از داده‌های حساس خود هستند. در حالت انتشار داده به‌صورت اشتراکی نیز باید مجموعه داده نهایی از لحاظ حریم خصوصی افراد ایمن باشد.

در حالت توزیع شده که چندین فراهم‌کننده داده قصد انتشار داده‌های مربوط به خود را دارند، انتشار با حفظ محرمانگی در داده‌های توزیع شده یا به‌اختصار PPDDP اتفاق می‌افتد. در این حالت به‌دلیل ذات توزیعی، چالش‌های جدید برای حفظ حریم خصوصی وجود دارد.

^۱ Syntactic

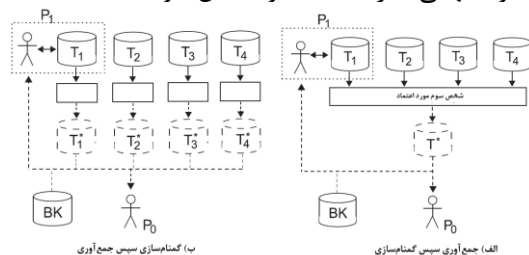
^۲ Linkage Attack

^۳ Background Knowledge

^۴ Semantic

^۵ Open Government Partnership

در روش دوم که «ابتدا گمنام‌سازی سپس جمع‌آوری» نام دارد، ابتدا هرکدام از منتشرکننده‌ها، مجموعه داده متناظر با خود را با استفاده از روش‌های گمنام‌سازی بیان شده در بخش پیشین، گمنام کرده و پس از آن مجموعه‌های گمنام‌شده با هم ادغام می‌شود و مورد عملیات داده‌کاوی یا تجزیه و تحلیل قرار می‌گیرند [۱]. این روش در شکل (۱) بخش (ب) نمایش داده شده است. در این روش اگر تمام منتشرکننده‌ها دارای یک نوع مجموعه با خصیصه‌ها و دامنه‌های یکسان باشند هیچ مشکل خاصی از لحاظ دقت نتایج پیش نمی‌آید، اما اگر مجموعه داده‌ها یکسان نباشد نتیجه به دست آمده از تجزیه و تحلیل آماری یا داده‌کاوی به شدت تضعیف می‌شود [۲۲]؛ همچنین در این حالت اگر از روش‌های چندپردازشی امن (SMC) برای ارتباط فراهم‌کنندگان در طول زمان گمنام‌سازی استفاده نشود، امنیت داده‌ها به دلیل حملاتی که ممکن است توسط هر یک از فراهم‌کنندگان به مجموعه نهایی تدارک دیده شود، نقض شود [۱۷].



شکل (۱)- سیاست‌های حفظ حریم خصوصی داده‌های

توزیع شده [۲۲]

(Figure-1): Privacy Preservation Policies for Distributed Data [22]

• سناریوهای حمله در داده‌های توزیع شده
به‌طور کلی برای حالتی که چند منتشرکننده، داده‌های خود را گردآوری کرده و برای عملیات داده‌کاوی به اشتراک می‌گذارند، سه سناریوی حمله وجود دارد [۲۲]:

حمله‌ای که با استفاده از داده‌های خارجی، توسط گیرنده بر روی مجموعه داده منتشرشده صورت می‌گیرد: در این حالت دریافت‌کننده داده با استفاده از دانش پیش‌زمینه خود و به‌کارگیری حملاتی مانند حمله پیوندهای، اقدام به نقض حریم خصوصی در رکوردهای مجموعه داده منتشرشده می‌کند؛ برای مثال فرض کنیم P_0 با استفاده از مجموعه داده منتشرشده و دانش پیش‌زمینه مربوط به خود اقدام به حمله پیوندهای کرده و به داده‌های حساس مالکین رکورد در هر یک از مجموعه داده‌های اولیه دسترسی پیدا کند. برای جلوگیری از چنین حملاتی می‌توان از روش‌های گمنام‌سازی که برای داده‌های غیر توزیع شده طراحی شده‌اند مانند k -anonymity، l -diversity، (n, t) -Closeness، δ -Presence و غیره استفاده کرد [۲۲].

حمله منتشرکنندگان داده‌ها به مجموعه‌های یکدیگر با استفاده از داده‌های تشکیل شده میانی: در این حالت

روش‌های مختلفی برای انتشار با حفظ محرمانگی داده‌های یک سازمان ارائه شده است، اما در برخی شرایط نیاز به انتشار داده با تشریک مساعی بین چند سازمان به وجود می‌آید که در این حالت علاوه بر تهدیدهایی که در شرایط عادی متوجه حریم خصوصی افراد است، حمله‌های جدیدی به دلیل توزیع بودن داده‌ها به وجود می‌آید؛ برای مثال در حوزه داده‌های بیمارستانی، چند بیمارستان اقدام به انتشار یا به اشتراک‌گذاری داده‌های خود برای کشف دانش پزشکی جدید می‌کنند. فرض بر این است که داده‌های مربوط به هر بیمارستان با روش‌های گمنام‌سازی که در بخش پیشین بیان شد، از لحاظ حریم خصوصی امن شده باشند؛ سپس داده‌های بیمارستان‌های مختلف با هم جمع شده و مجموعه داده گمنام‌شده نهایی را تشکیل می‌دهند. در این حالت ممکن است بیمارستانی با در اختیار داشتن مجموعه نهایی و داده‌های مربوط به خود، اقدام به نقض حریم خصوصی داده‌های بیمارستانی دیگر کند. این شرایط حادث خواهد شد، اگر چند بیمارستان برای این کار با هم تبانی کنند. در این حالت از روش‌های حفظ محرمانگی در داده‌های توزیع شده استفاده می‌شود.

برای برقراری حفظ حریم خصوصی در انتشار داده‌ها به صورت توزیع شده بر حسب معمول از دو سیاست کلی استفاده می‌شود:

• سیاست‌های حفظ حریم خصوصی در حالت توزیعی

به‌طور کلی وقتی چندین منتشرکننده وجود دارد از دو نوع سیاست یا راه‌کار کلی می‌توان برای جلوگیری از نقض حریم خصوصی بهره گرفت [۱، ۲۲]. شکل (۱) این دو راه‌کار را به صورت کلی نشان می‌دهد. «ابتدا جمع‌آوری و بعد گمنام‌سازی» روش اولی است که در برخی از کارها مورد استفاده قرار می‌گیرد. این روش زمانی بدون مشکل امنیتی خواهد بود که تمام منتشرکننده‌ها به هم اعتماد داشته باشند؛ زیرا اطلاعات حساس بلافاصله بعد از مرحله جمع‌آوری یا ادغام در اختیار تمام شرکت‌کننده‌ها قرار دارد و یا این‌که راه‌کار موردنظر با روش شخص سوم مورد اعتماد ادغام شود [۲۲]. شکل (۱) قسمت «الف»، به این روش اشاره دارد. مجموعه‌های داده T_1 تا T_4 متعلق به منتشرکننده‌های P_1 تا P_4 هستند. ابتدا داده‌ها توسط شخص سوم مورد اعتماد که می‌تواند یکی از منتشرکننده‌ها باشد جمع‌آوری شده و سپس گمنام‌سازی می‌شود و نتیجه پردازش آماری یا داده‌کاوی در اختیار تمام منتشرکننده‌ها یا هر شخص دیگری مانند P_0 قرار می‌گیرد، اما پیدا کردن شخص سومی که قابل اعتماد همه فراهم‌کنندگان باشد به‌طور معمول امکان‌پذیر نیست [۱۷]. به همین دلیل این روش کمتر مورد استفاده قرار می‌گیرد.

^۱ Integrate-then-Anonymize

^۲ Anonymize-then-Integrate

فرض می‌شود که منتشرکننده‌های داده نسبت به هم، نیمه‌صادق‌اند^۱ (صادق اما کنجکاو) و با استفاده از SMC داده‌های میانی را در اختیار هم قرار می‌دهند. در این حالت منتشرکننده مهاجم، با استفاده از داده‌های میانی و همچنین مجموعه متعلق به خود می‌تواند به اطلاعات حساس در مورد مجموعه منتشرکنندگان دیگر دسترسی پیدا کند. این مشکل به‌طور معمول به دلیل ضعف پروتکل‌های SMC به‌وجود می‌آید.

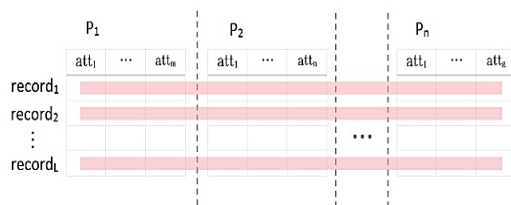
حمله منتشرکنندگان داده‌ها به مجموعه‌های یکدیگر با استفاده از مجموعه داده منتشرشده نهایی و مجموعه داده متعلق به هر منتشرکننده: هرکدام از منتشرکننده‌ها مانند P_1 به مجموعه نهایی T^* دسترسی دارد و با به‌کارگیری اطلاعات مجموعه متعلق به خود (T_1) می‌تواند این نوع حمله را سازمان‌دهی کند. در این حالت مهاجم P_1 چیزی بیشتر از مهاجم P_0 در اختیار دارد. حال اگر چند منتشرکننده دیگر مانند P_2 و P_3 در این نوع حمله برای به‌دست‌آوردن اطلاعات حساس P_4 با P_1 تباری کنند این حمله شدیدتر خواهد شد. این حملات به حملات داخلی^۲ معروف‌اند[۲۲].

برای حل مشکل حملات بیان‌شده در قسمت پیش، برخی از روش‌ها با فرض عمودی (رکوردها یکی هستند، اما صفات متفاوت) یا افقی بودن (صفات یکی هستند، اما رکوردها متفاوت) مجموعه‌ها و به‌کارگیری روش‌های گمنام‌سازی برای یک مجموعه، و همچنین به‌کارگیری SMC، روش‌هایی برای حل چالش‌های PPDDP ارائه داده‌اند.

• انواع توزیع برای انتشار داده‌ها

در دنیای واقعی مثال‌های زیادی از نرم‌افزارها و سامانه‌هایی وجود دارد که یک منتشرکننده یا دارنده مجموعه داده، بیشتر اوقات تمام داده‌ها را در اختیار ندارد و نیاز دارد تا با دیگر منتشرکنندگان تشریک‌مساعی کند؛ برای مثال بانک‌های عضو شبکه سراسری شتاب برای ساخت یک نرم‌افزار تشخیص نفوذ به اطلاعات کارت‌های عابر بانک نیاز دارند. در این حالت بانک‌ها داده‌های کارت‌های الکترونیکی خود را که حاوی اطلاعات شخصی کاربران است در اختیار یکدیگر یا یک شرکت سوم قرار می‌دهند، اما بانک‌ها به دلیل امنیتی و حریم خصوصی نمی‌خواهند اطلاعات خام در اختیار دیگر بانک‌ها قرار گیرد. مجموعه‌های داده که بین منتشرکنندگان به اشتراک گذاشته می‌شوند یا به‌صورت افقی جمع‌آوری می‌شوند، مانند مثال کارت الکترونیکی بانک‌ها که بیشتر دارای یک مجموعه از صفات هستند و یا به‌صورت عمودی جمع‌آوری می‌شوند که مجموعه‌ها، صفات‌های یکسان ندارند، اما رکوردهای یکسان دارند. در ادامه به بررسی هرکدام از این دو نوع گردآوری داده می‌پردازیم.

توزیع عمودی: توزیع عمودی در شکل (۲) نشان داده شده‌است؛ همان‌طور که در شکل مشخص است، چندین دارنده داده مانند P_1 ، P_2 تا P_k به‌ترتیب دارای مجموعه صفات $\{att_1, \dots, att_n\}$ ، $\{att_1, \dots, att_n\}$ و $\{att_1, \dots, att_n\}$ هستند؛ از طرف دیگر این مجموعه‌ها رکوردهای هم‌سان دارند و ارتباط بین مجموعه‌ها روی یک رکورد خاص با شناسه خاص برای مثال att_1 که در تمام مجموعه‌ها تکرار شده‌است، برقرار می‌شود[۱]. دلیل اصلی اشتراک داده بین چند منتشرکننده در توزیع عمودی، کامل‌تر شدن مجموعه صفات و بالا رفتن دقت عمل داده‌کاوی یا تجزیه و تحلیل آماری است. به‌طور معمول هیچ‌کدام از شرکت‌کننده‌ها در اشتراک داده به‌صورت توزیع عمودی، نمی‌توانند به‌تنهایی و بدون داشتن مجموعه دیگر اعضا اقدام به عمل داده‌کاوی یا انتشار داده کنند. نمونه واقعی از انتشار با حفظ محرمانگی در داده‌های توزیع‌شده برنامه‌های Data Mashup هستند. Mashup نرم‌افزارهای تحت وب هستند که اطلاعات و سرویس‌ها را از بیش از یک منبع جمع‌آوری کرده و در یک برنامه ارائه می‌دهند[۲۳].



(شکل ۲-): توزیع عمودی داده‌ها
(Figure-2): Vertical Data Distribution

توزیع افقی: در توزیع افقی، صفات برای تمام دارندگان مجموعه‌های داده یکی است، اما این مجموعه‌ها رکوردهای متفاوت دارند[۱]. این نوع از توزیع در شکل (۳) به تصویر کشیده شده‌است. در این شکل دارندگان داده P_1 ، P_2 تا P_k دارای صفات یکسان $\{att_1, \dots, att_n\}$ هستند؛ هرکدام از این منتشرکننده‌ها می‌توانند رکوردهای مربوط به خود را داشته باشند.

در روش توزیع عمودی، صفات مربوط به یک رکورد خاص، در بین چندین منتشرکننده تقسیم شده‌است و به همین دلیل منتشرکنندگان از اشتراک داده استفاده می‌کنند، اما این موضوع در روش افقی صادق نیست و سؤال اساسی که به ذهن خطور می‌کند این است که چرا در این روش، نیاز به اشتراک مجموعه‌های داده وجود دارد. هدف اصلی بالا بردن دقت عمل کشف دانش با افزایش داده‌ها است و با اشتراک داده، اطلاعات مربوط در آن حوزه کامل‌تر می‌شود[۱۳]. به‌عنوان نمونه‌ای از توزیع افقی، تمام سازمان‌های مراقبت بهداشتی در ایالات متحده آمریکا از قوانین HIPAA^۳ برای اشتراک داده‌های گمنام‌شده خود استفاده می‌کنند[۲۴].

^۳ Health Insurance Portability and Accountability Act

^۱ Semi-Honest

^۲ Insider Attack

همچنین به بیان ریاضی اگر بررسی‌های مهاجم به وسیله تابع تصادفی F مدل شود، برای هر مجموعه داده T_1 و T_2 که در وجود یک رکورد اختلاف دارند، باید رابطه (۱) برقرار باشد:

$$\forall S \in \text{Range}(F) \left(\left| \ln \frac{P[F(T_1) = S]}{P[F(T_2) = S]} \right| \leq \epsilon \right) \quad (1)$$

که در آن S صفت حساس و ϵ مقداری است که هرچه کوچک‌تر انتخاب شود، میزان تأثیر یک رکورد بر روی نتایج تجزیه و تحلیل را کاهش می‌دهد. برای برآورده کردن رابطه مدل حریم خصوصی تفاضلی، نوبت زیادی به داده‌های مجموعه اضافه می‌شود؛ زیرا این مدل تمایزی بین صفات حساس و دیگر صفات قائل نمی‌شود. از طرف دیگر این روش تضمینی برای محافظت در برابر حملات پیونددهی ندارد [۵۱ و ۴۸ و ۵۱].

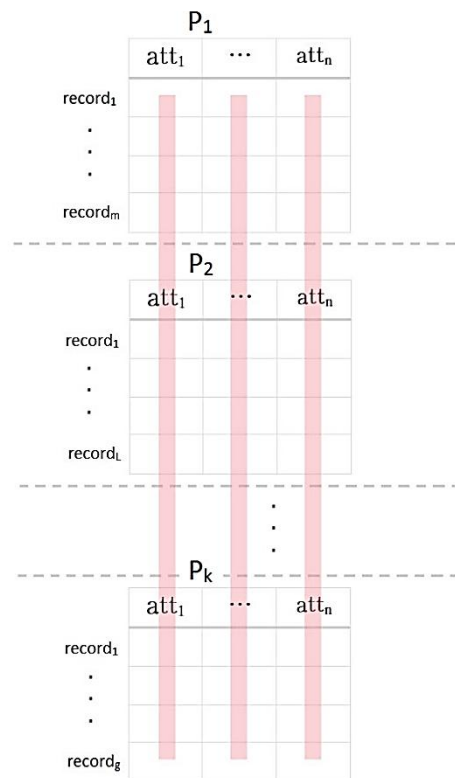
در [۶] روش تعدیل یافته‌تری نسبت به حریم خصوصی تفاضلی و بر پایه مدل احتمالاتی ارائه شده است. این روش به نحوی اطمینان مهاجم برای درک صفت حساس یک رکورد بعد از دیدن مجموعه داده را مدل کرده و آن را محدود می‌کند. روش ارائه شده در [۶] به عنوان مدل حفظ حریم خصوصی در مقاله پیش رو استفاده شده است؛ به همین دلیل در ادامه به بررسی کامل این روش می‌پردازیم و آن را به اختصار روش GF_PPDP می‌نامیم. روش GF_PPDP بر مبنای دو مفهوم اطمینان مورد انتظار^۱ و اطمینان مشاهده شده^۲ بنا شده است. مهاجم پس از مشاهده مجموعه داده منتشر شده، با جستجوی شبه‌شناسه‌های قربانی، سعی می‌کند صفت حساس وی را به دست آورد؛ با این کار، تعدادی از رکوردها شامل رکورد قربانی در یک گروه هم‌ارزی قرار می‌گیرند که شبه‌شناسه‌های یکسانی دارند، اما ممکن است صفات حساس متفاوتی داشته باشند. احتمال اینکه مهاجم صفت حساس قربانی را از این گروه هم‌ارزی تشخیص دهد، اطمینان مشاهده شده قربانی می‌نامند؛ به بیان دیگر اطمینان مشاهده شده بیان می‌کند که مهاجم با چه احتمالی می‌تواند صفت حساس یک رکورد را به قربانی نسبت دهد؛ به دلیل این که در این روش در ابتدا از رکوردها با نرخ β نمونه‌برداری می‌شود، باید احتمال گزینش رکورد قربانی در مرحله نمونه‌برداری نیز مدنظر قرار بگیرد. همان‌طور که رابطه (۲) نشان می‌دهد از ضرب این دو احتمال، اطمینان مشاهده شده رکورد r به دست می‌آید.

$$\text{ObservedConfidence}(r) = \beta \times \frac{\#S(r)}{|E(r)|} \quad (2)$$

که $|E(r)|$ اندازه گروه هم‌ارزی و $\#S(r)$ تعداد تکرار صفت حساس رکورد r در گروه هم‌ارزی است؛ هرچه اطمینان مشاهده شده که به وسیله فرمول بالا محاسبه می‌شود مقدار کمتری باشد، مهاجم نیز با اطمینان کمتری می‌تواند بین صفات حساس و شناسه‌های صریح رابطه برقرار

¹ Expected Confidence

² Observed Confidence



(شکل - ۳): توزیع افقی داده‌ها
(Figure-3): Horizontal Data Distribution

۳- پیشینه پژوهش

در این بخش پژوهش‌هایی که در حوزه انتشار با حفظ محرمانگی و حریم خصوصی داده‌های توزیع شده انجام شده است، مورد بررسی قرار می‌گیرد. تمام روش‌های ارائه شده، بر پایه روش‌های حفظ محرمانگی در حالت غیر توزیعی بنا شده‌اند و به نحوی در این حالت روش‌های PPDP به همراه یکی از دو سیاست حفظ محرمانگی در حالت توزیعی که در بخش پیشین بیان شد، مورد استفاده قرار می‌گیرند.

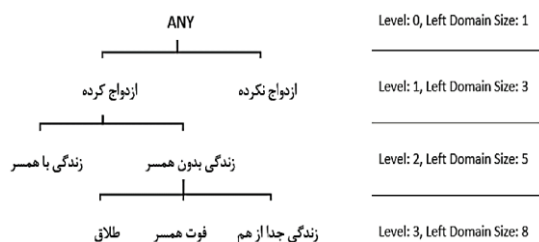
• انتشار با حفظ حریم خصوصی در حالت غیر توزیع شده

برای حالتی که تنها یک فراهم‌کننده داده قصد انتشار مجموعه داده خود را دارد، روش‌های مختلفی برای حفظ محرمانگی ارائه شده است. مانند روش k -anonymity که در بخش پیش به آن اشاره شد [۲۷، ۲۶، ۱۸]؛ همچنین پس از آن روش‌هایی برای بهبود k -anonymity ارائه شدند [۱] که از جمله آن‌ها می‌توان به δ -Presence، l -diversity و t -closeness اشاره کرد [۳۰، ۲۹، ۲۸، ۱۹].

روش‌هایی نیز بر پایه مدل احتمالاتی ارائه شده که مهم‌ترین آن‌ها حریم خصوصی تفاضلی است. ایده این مفهوم ابتدا در [۳۱] معرفی شد و سپس در [۲۰] به حریم خصوصی تفاضلی تبدیل شد. این مدل بیان می‌کند که حذف یا اضافه کردن یک رکورد، نباید تأثیر چندانی بر نتایج به دست آمده از تجزیه و تحلیل مجموعه داده داشته باشد؛ به بیان ساده‌تر در این مدل انتشار یا عدم انتشار رکورد که متعلق به قربانی است، نباید تأثیری در دانش مهاجم داشته باشد؛

$$\delta = \|r - \hat{r}\| = \frac{1}{d} \sum_{i=1}^d \left(1 - \frac{\text{CurrentLevel of } \hat{q}_i}{\text{Maximal level in } q_i} \right) \quad (6)$$

که در این رابطه r' معادل گمنام‌شده رکورد r است. اگر معیار تخریب برابر با «۰» باشد به معنی این است که تمام شبه‌شناسه‌های رکورد مقدار اصلی خود را حفظ کرده‌اند و اگر برابر با «۱» باشد، یعنی تمام شبه‌شناسه‌ها به ریشه درخت طبقه‌بندی جایگزین شده‌اند؛ بنابراین هر چه این مقدار به‌ازای یک رکورد به صفر نزدیک‌تر باشد، رکورد موردنظر برای عملیات داده‌کاوی مناسب‌تر است.



(شکل ۴-): معیار LDF در درخت طبقه‌بندی وضعیت تأهل [۲۵]
(Figure-4): LDF Metric in the Marital Status Taxonomy Tree [25]

• انتشار با حفظ حریم خصوصی در حالت توزیع‌شده

همان‌طور که در بخش پیشین بیان شد روش‌های مختلفی برای حالتی که تنها یک فراهم‌کننده انتشار داده را برعهده دارد وجود دارد، اما در عمل به‌طور معمول فراهم‌کنندگان مجبور به اشتراک داده می‌شوند. در این حالت فراهم‌کننده باید عملیات گمنام‌سازی و انتشار مجموعه‌داده متعلق به خود را با توجه به مجموعه دیگر فراهم‌کنندگان انجام دهد. در بیشتر روش‌های انتشار با حفظ حریم خصوصی در حالت غیر توزیع‌شده، با فرض سطح دانش پیش‌زمینه مهاجم، روشی برای حفظ محرمانگی در مقابل حمله‌های خاص (برای مثال حملات پیونددهی) ارائه می‌دهند، اما در حالت توزیع‌شده، مدل کردن دانش پیش‌زمینه مهاجم با در نظر گرفتن شرکت‌کنندگان (فراهم‌کنندگان) دیگر، کار دشواری است و این کار در صورت تبانی شرکت‌کنندگان با یکدیگر یا با مهاجم، دشوارتر نیز خواهد شد [۱۷].

بسیاری از پژوهش‌ها در حوزه توزیع‌شده بر مبنای روش معروف k-anonymity بنا شده‌است. در [۳۲] از روش k-anonymity برای گمنامی داده‌هایی که به‌صورت عمودی بین شرکت‌کنندگان پخش شده، استفاده می‌شود. ارائه‌دهندگان این روش با استفاده از روش محاسبه امن چندگانه پروتکلی ارائه داده‌اند تا کلید نامزد پیونددهی مجموعه‌داده شرکت‌کنندگان به نحوی غیرقابل‌شناسایی گمنام شود. SMC در بسیاری از روش‌ها برای ارتباط امن بین فراهم‌کنندگان استفاده می‌شود. هدف اصلی SMC ارائه پروتکلی برای اجرای یک تابع یا عمل، بر روی داده‌های مشترک در محیط توزیع شده‌است؛ به‌نحوی که حریم خصوصی داده‌ها حفظ شود و همچنین خروجی تابع صحیح باشد [۳۳]؛ به عبارت دیگر حفظ

کند و در نتیجه سطح محرمانگی مجموعه‌داده بالاتر می‌رود؛ بنابراین باید مقدار آن را به سقف معینی محدود کرد. بر اساس مدل ارزیابی ارائه‌شده در GF_PPDP، اطمینان مشاهده‌شده یک رکورد نباید از احتمال دیده‌شدن آن رکورد در مجموعه‌داده‌ای با اندازه یکسان که به‌صورت تصادفی از فضای نمونه رکوردها ایجاد شده‌است (D_0) بیشتر باشد.

این احتمال اطمینان مورد انتظار نامیده می‌شود. با فرض اینکه مجموعه‌داده شامل n رکورد باشد، احتمال وجود رکورد r در D_0 برابر با احتمال مکمل «۰» پیروزی در n آزمون با احتمال پیروزی $\Pr(r)$ است و با استفاده از توزیع دوجمله‌ای طبق رابطه (۳) محاسبه می‌شود:

$$\begin{aligned} \text{ExpectedConfidence}(r) &= 1 - f(0; n, \Pr(r)) \\ &= 1 - (1 - \Pr(r))^n \end{aligned} \quad (3)$$

که $\Pr(r)$ با فرض مستقل بودن صفات از یکدیگر به‌وسیله ضرب احتمال شبه‌شناسه‌ها در احتمال صفت حساس به‌دست می‌آید:

$$\Pr(r) = \left(\prod_i \Pr(q_i) \right) \times \Pr(s) \quad (4)$$

با توجه به روابط (۲) و (۳)، برای تأمین حریم خصوصی طبق قاعده ذکرشده، به‌ازای هر رکورد موجود در مجموعه‌داده، باید رابطه (۵) برقرار باشد:

$$\begin{aligned} \text{ObservedConfidence}(r) & \\ \leq \text{ExpectedConfidence}(r) \end{aligned} \quad (5)$$

الگوریتم ارائه‌شده در روش GF_PPDP به‌ازای هر رکورد رابطه (۵) را تشکیل داده و اگر این رابطه برای رکورد یادشده برقرار نباشد، با استفاده از معیاری به نام $\text{LeftDomainSize(LDF)}$ یک شبه‌شناسه را انتخاب کرده و با اعمال عمل‌گرهای گمنام‌سازی، رابطه (۵) را برای رکورد برقرار می‌کند. LDF به‌ازای تمام شبه‌شناسه‌ها محاسبه و شبه‌شناسه‌ای که بیشترین مقدار را داشته باشد، برای عملیات عمومی‌سازی انتخاب می‌شود. LDF برای شبه‌شناسه‌ای با مقدار v ، تعداد گره‌هایی است که در درخت طبقه‌بندی در سطح برابر یا بالاتر از v قرار دارد؛ برای مثال در شکل (۴) معیار LDF در درخت طبقه‌بندی وضعیت تأهل نمایش داده شده‌است.

برای این‌که رکوردها پس از استفاده چندباره از عمل‌گرهای فرونشانی و عمومی‌سازی، کیفیت مناسبی برای عملیات داده‌کاوی داشته باشند GF_PPDP از معیار تخریب δ^1 استفاده کرده و رکوردهایی که میزان تخریب آن‌ها از حد آستانه δ بیشتر باشد از مجموعه خروجی حذف می‌شوند. میزان تخریب برای رکورد r با شبه‌شناسه‌های q_1 تا q_n از رابطه (۶) به‌دست می‌آید.

¹ Distortion

حریم خصوصی اقتضا می‌کند اطلاعاتی بیشتر از آن چیزی که باید دانسته شود فاش نشود؛ یعنی افراد شرکت‌کننده و همچنین کاربران تابع خروجی نباید چیزی به‌جز نتیجه تابع خروجی را درک کنند [۳۴].

در [۲۳] راه‌حلی برای حفظ محرمانگی در Data Mashup ارائه شده‌است. راه‌حل پیشنهادی باید هم حافظ حریم خصوصی مالکان داده باشد و هم این‌که کاربر نهایی نتواند تشخیص دهد نتایج به‌دست‌آمده متعلق به کدام فراهم‌کننده داده است. این روش برای مجموعه داده‌هایی که به‌صورت عمودی توزیع شده‌اند، مفهوم محدودیت‌های گمنام‌سازی برای پیوندهای^۱ را مطرح کرد که به مجموعه‌ای از دوتایی‌های شبیه‌شناسه QID و مقدار K به‌شکل $\{ \langle QID_1, k_1 \rangle, \dots, \langle QID_p, k_p \rangle \}$ گفته می‌شود. در طی مراحل گمنام‌سازی محدودیت‌های بالا باید ارضا شوند. برای گمنام‌سازی از روش TDS^۲ استفاده می‌شود که برای هر صفت شبیه‌شناسه از عمل‌گر ویژه‌سازی^۳ برای رسیدن به k-anonymity استفاده می‌شود [۳۵]. در فرایند گمنامی Mashup هر فراهم‌کننده داده یا همان پارتی، یک رونوشت از مجموعه Tg که بیان‌کننده مجموعه نهایی در آن لحظه است و مجموعه مربوط به خودش را دارد و در هر مرحله از اجراء پارتی‌ها با هم همکاری می‌کنند تا یک عمل ویژه‌سازی را با توجه به محدودیت‌های تعریف‌شده و با استفاده از ردوبدل کردن اطلاعات انجام دهند. این روند تا رسیدن به مجموعه نهایی انجام می‌شود.

برخی از کارها برای مجموعه‌هایی که به‌صورت افقی توزیع شده‌اند ارائه شده‌است؛ برای مثال در [۱۳] پروتکل غیرمتمرکزی برای اشتراک‌گذاری این نوع مجموعه داده‌ها ارائه شده‌است؛ از آنجایی که فراهم‌کنندگان داده ترجیح می‌دهند تعلق رکوردهای منتشرشده نهایی به آن‌ها مشخص نشود، در این روش مفهوم جدیدی به نام l-site-diversity برای حفظ حریم خصوصی فراهم‌کنندگان داده معرفی شد؛ به این ترتیب که شبیه‌شناسه‌هایی که به‌همراه دانش پیش‌زمینه موجب تشخیص تعلق رکورد به فراهم‌کننده می‌شود را انتخاب می‌کنند و تا حدی عمومی‌سازی انجام می‌دهند که تعداد تنوع فراهم‌کنندگان داده در یک گروه هم‌ارزی برابر با مقدار l باشد. این روش از SMC برای ارتباط بین مجموعه داده‌ها و گمنام‌سازی آن‌ها استفاده می‌کند. فراهم‌کنندگان داده با استفاده از یک پروتکل توزیع‌شده مجموعه داده گمنام‌شده مجازی^۴ می‌سازند؛ به بیان دیگر مجموعه داده‌های محلی متعلق به فراهم‌کنندگان بدون تغییر باقی می‌مانند و با استفاده از عمل‌گر امن پیوندهای، تحت پروتکل امن، تجمع می‌شوند و می‌توانند منتشر شوند و یا مورد پرس‌وجو قرار گیرند.

روش ارائه‌شده در [۳۶] برای حالتی مناسب است که چند فراهم‌کننده داده قصد دارند داده‌های خود را که به‌صورت افقی توزیع شده‌است در اختیار یک داده‌کاو^۵ قرار دهند. این روش برای مقابله با حمله پیوندهای صفت، پروتکلی بر پایه روش گمنام‌سازی k-anonymity ارائه داده‌است.

روشی منعطف برای حفظ حریم خصوصی در مجموعه داده‌هایی که به‌صورت افقی یا عمودی توزیع شده‌اند در [۳۷] ارائه شده‌است. این روش برخلاف روش‌های بیان‌شده در بالا به الگوریتم خاصی برای گمنام‌سازی وابسته نیست و با استفاده از SMC و به‌کارگیری یکی از الگوریتم‌های k-anonymity, l-diversity, t-closeness و δ -presence اقدام به گمنام‌سازی مجموعه‌ها می‌کند.

در تمام روش‌های گمنام‌سازی مجموعه‌های توزیع‌شده‌ای که تا به اینجا بیان شدند، فرض بر این بود که فراهم‌کنندگان نیمه‌صادق هستند و برای رسیدن به هدف نهایی، پروتکل SMC را به‌درستی دنبال می‌کنند [۳۸]، اما در برخی موارد خود فراهم‌کنندگان ممکن است در نقش مهاجم ظاهر شوند و حمله داخلی را به مجموعه داده نهایی تدارک ببینند [۲۲]؛ اگر چند فراهم‌کننده با هم تبانی کنند این وضعیت دشوارتر نیز خواهد شد.

در [۲۲] روشی برای مقابله با حملات داخلی ارائه شده‌است. در این کار مفهومی به نام m-privacy معرفی کرده‌است که با بررسی کردن محدودیت‌های وضع‌شده برای مجموعه نهایی، آن را در مقابل مجموعه تبانی فراهم‌کنندگان متخاصم (m-adversary)، ایمن می‌دارد. m-privacy با توجه به محدودیت C که شرط یا شرایطی است که برای انتشار رکورد یا مجموعه‌ای از رکوردها باید ارضا شود (مثل مقدار k در k-anonymity یا l در l-diversity)، تعریف می‌شود. بر طبق تعریف برای n فراهم‌کننده داده که مجموعه رکوردهای آن‌ها T را تشکیل می‌دهند، A سازوکار گمنام‌سازی، I مجموعه‌ای از m فراهم‌کننده متخاصم ($m \leq n-1$) و T_I مجموعه متعلق به I، A(T) شرایط m-privacy را نسبت به C ارضا می‌کند اگر و فقط اگر به‌ازای هر فوق مجموعه گمنام‌شده $A(T')$ از مجموعه غیر متخاصم (متمم T_I) شرط C را ارضا کند. به بیان ریاضی داریم:

$$\forall I \subset P, |I| = m, \forall T' : T \setminus T_I \subseteq T' \\ \subseteq T, C(A(T')) = true \quad (V)$$

در رابطه بالا $T \setminus T_I$ بیان‌کننده متمم مجموعه T_I است. اگر n فراهم‌کننده $(P_1, P_2, \dots, P_n)$ داشته باشیم، حالت‌های مختلف برای m فراهم‌کنندگان متخاصم، که m می‌تواند از «۰» تا «n-۱» باشد، مانند شکل (۵) است. این شکل برای حالتی که چهار فراهم‌کننده وجود دارد، رسم شده‌است. در هر ردیف از شکل به‌ازای هر m تعداد حالت‌ها برابر با ترکیب m از n است؛ برای مثال در سطح m برابر با «۱»، هر کدام از چهار فراهم‌کننده

^۵ Data Miner

^۱ Join Anonymity Requirement

^۲ Top-Down Specialization

^۳ Specialization

^۴ Virtual Anonymized Database

(جدول ۱-): مقایسه روش‌های حفظ محرمانگی در داده‌های

توزیع شده

(Table-1): Comparison of Privacy Preservation Methods in Distributed Data

مدل محرمانگی	نوع داده	مقایسه با حملات داخلی	توضیحات
[۳۲] پیونددهی	عمودی	خیر	استفاده از مدل محرمانگی k-anonymity
[۲۳] پیونددهی	عمودی	خیر	به کارگیری k-anonymity و روش ویژگی‌سازی
[۱۳] پیونددهی	افقی	خیر	محرمانگی در سطح فراهم‌کنندگان و ارائه مفهوم l-site-diversity
[۳۶] پیونددهی	افقی	خیر	استفاده از مدل محرمانگی k-anonymity
[۳۷] پیونددهی	عمودی-افقی	خیر	قابلیت استفاده از مدل‌های مختلف مانند k-anonymity, l-diversity, t-closeness و غیره
[۲۲] پیونددهی-احتمالاتی	افقی	بلی	قابلیت به کارگیری مدل‌های مختلف و محافظت در مقابل حملات داخلی

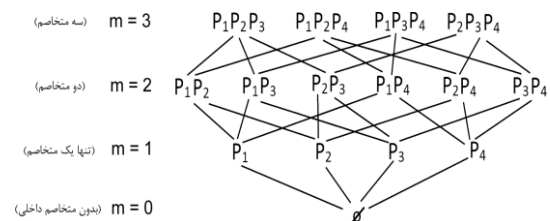
۴- طرح الگوریتم پیشنهادی

مدل حفظ حریم خصوصی مورد استفاده در روش پیشنهادی برگرفته‌شده از روش ارائه‌شده در [۶] است. این مدل با استفاده از دو مفهوم اطمینان مشاهده‌شده و اطمینان مورد انتظار و برقراری رابطه بین این دو مفهوم برای هر رکورد، محرمانگی داده‌ها را در سطح رکورد تضمین می‌کند. اطمینان مورد انتظار برابر با احتمالی است که مهاجم پیش از مشاهده مجموعه داده انتشار یافته برای دیدن رکورد در مجموعه داده‌ای با اندازه یکسان که به صورت تصادفی از فضای نمونه رکوردها ایجاد شده است (D_0)، قائل می‌شود؛ رابطه (۳) و از طرف دیگر اطمینان مشاهده‌شده یک رکورد، بیان‌کننده این است که مهاجم پس از مشاهده مجموعه منتشرشده با چه احتمالی می‌تواند صفت حساس آن رکورد را به قربانی نسبت دهد؛ رابطه (۲)؛ همچنین بر طبق مدل GF_PPDP که در رابطه (۵) بیان شده است، اطمینان مشاهده‌شده برای یک رکورد نباید بیشتر از مقدار اطمینان مورد انتظار برای آن رکورد باشد.

مدل GF_PPDP برای حالتی مورد استفاده قرار می‌گیرد که تنها یک فراهم‌کننده قصد انتشار مجموعه داده را دارد. در حالتی که چندین فراهم‌کننده در انتشار به صورت توزیع شده شرکت دارند روابط بیان‌شده در مدل GF_PPDP باید تغییر کند؛ زیرا در رابطه (۳) باید احتمال کل مجموعه‌ها را در نظر گرفت و همچنین در رابطه (۲) مقدار اطمینان مشاهده‌شده به‌ازای هر رکورد، از گروه‌های هم‌ارزی متعلق به هر فراهم‌کننده به دست می‌آید، که باید گروه‌های هم‌ارزی تمام مجموعه‌ها در نظر گرفته شود. بر اساس مدل GF_PPDP هر

می‌تواند به عنوان فراهم‌کننده متخاصم در نظر گرفته شوند. برای رسیدن به امنیت در سطح m باید تمام حالت‌های کوچک‌تر و مساوی برابر با آن m به خصوص را چک کرد؛ برای مثال اگر m برابر با دو در نظر گرفته شود، باید حالت‌های m از مجموعه $\{0, 1, 2\}$ چک شود؛ که اگر تعداد فراهم‌کنندگان زیاد باشد، بررسی تمام حالات دشوار می‌شود. بر اساس [۲۲] تعداد کل حالات از مرتبه $O(2^n n^{-1/2}) = O(n/2)$ است که بررسی کردن تمام حالت‌ها زمان اجرایی بالایی را خواستار است. البته در مقاله m-privacy با توجه به خاصیت EG Monotonic که برای برخی از محدودیت‌ها مانند k-anonymity یا l-diversity برقرار است و همچنین انواع هرس، تعداد کل حالاتی که باید چک شود کاهش می‌یابد.

پس از بررسی کردن تمام حالت‌ها توسط الگوریتم، اگر شرط C برای رکورد یا مجموعه‌ای از رکوردها برقرار نباشد از مجموعه قابل انتشار نهایی حذف می‌شوند و یا می‌توان عملیات گمنامی را بر رکورد مربوطه اعمال کرد تا محدودیت C را ارضا کند. الگوریتم m-privacy را می‌توان به وسیله TTP یا پروتکل SMC که در [۲۲] ارائه شده اجرا کرد.



(شکل ۵-): حالت‌های مختلف برای m فراهم‌کننده متخاصم [۲۲]
(Figure-5): Different Scenarios for m Malicious Providers [22]

یکی از مشکلات روش m-privacy تعیین مقدار یا مقادیر برای C است. باید تمام فراهم‌کنندگان بر روی یک مقدار خاص توافق کنند (برای مثال $k=20$ برای k-anonymity). حال آن که ممکن است با توجه به حجم مجموعه‌های متعلق به فراهم‌کنندگان، مقدار توافقی برای برخی از آن‌ها مناسب و برای برخی دیگر نامناسب باشد؛ از طرف دیگر فراهم‌کنندگان باید بر روی تعداد فراهم‌کنندگان متخاصم به توافق برسند (m).

با توجه به مشکلات TTP اجرای الگوریتم m-privacy بر روی آن غیر قابل اعتماد است. از طرف دیگر استفاده از SMC نیازمند بار پردازشی بالایی است [۳۸] و چنانچه شرط C از نوع EG Monotonic نباشد بار پردازشی افزایش پیدا می‌کند؛ همچنین در اجرای SMC باید تمام فراهم‌کنندگان صادق باشند و به روند پروتکل احترام بگذارند.

جدول (۱) خلاصه‌ای از ویژگی‌های روش‌های معرفی شده در این بخش را نشان می‌دهد. در ستون توضیحات ویژگی‌های هر کدام از روش‌ها بیان شده است. نکته مهمی که در جدول ذکر نشده، وابستگی تمام روش‌ها به پروتکل SMC و یا به کارگیری TTP است.

هوشمندانه، باید کاربرد داده‌های انتشار یافته، مشخص شود تا بر اساس آن روشی مناسب برای نمونه‌برداری برگزید. از آنجایی که در PPDP هدف انتشار داده برای عملیات مختلف داده‌کاوی است و حالت عام در نظر گرفته می‌شود، برای روش پیشنهادی این مقاله از روش نمونه‌برداری خاصی استفاده نمی‌شود.

با توجه به رابطه (۳) یکی دیگر از متغیرهایی که باعث تغییر اطمینان مورد انتظار می‌شود، $Pr(r)$ یا همان احتمال رکورد است. در حالتی که تنها یک فراهم‌کننده وجود دارد تنها یک مقدار برای این متغیر وجود دارد و در عمل نمی‌توان از آن برای تغییر اطمینان مورد انتظار استفاده کرد، اما در این مقاله که انتشار برای حالت توزیع شده مدنظر است با توجه به چندمقداری بودن این متغیر می‌توان از آن برای کاهش اطمینان مورد انتظار استفاده کرد؛ به همین منظور در رابطه (۳) مقدار $Pr(r)$ کاهش داده می‌شود تا مقدار نهایی اطمینان مورد انتظار کاهش یابد.

با توجه به مرحله نشست، برای هر رکورد به‌ازای هر صفت شبه‌شناسه یا صفت حساس، n مقدار برای احتمال رخداد آن صفت در اختیار است که n تعداد فراهم‌کنندگان را مشخص می‌کند؛ بنابراین برای محاسبه رابطه (۴)، به‌ازای هر صفت حساس یا شبه‌شناسه، از n احتمال یادشده مقدار کمتر مورد استفاده قرار می‌گیرد تا $Pr(r)$ کاهش یابد. با کاهش مقدار $Pr(r)$ مقدار اطمینان مورد انتظار نیز کاهش می‌یابد. با این روش اطمینان مورد انتظار برای هر رکورد با در نظر گرفتن خصوصیات مجموعه تمام فراهم‌کنندگان محاسبه می‌شود؛ برای مثال فرض می‌شود رکورد r با مشخصات $\{سن=23$ ، جنسیت=زن، بیماری=سرطان $\}$ ، متعلق به فراهم‌کننده P_1 از مجموعه فراهم‌کنندگان $\{P_1, P_2, P_3\}$ باشد. جدول (۲) احتمال مربوط به هر فراهم‌کننده را برای رکورد r نشان می‌دهد.

(جدول-۲): احتمال مقادیر رکورد r در مجموعه هر فراهم‌کننده
(Table-2): Probability of Record r 's Values in Each Provider's Dataset

	$Pr[۲۰-۳۰]$	$Pr[زن]$	$Pr[سرطان]$
P_1	۰.۱۵	۰.۴	۰.۰۲۱
P_2	۰.۲	۰.۵	۰.۰۲۵
P_3	۰.۲	۰.۵۵	۰.۰۱۸
P	۰.۱۸	۰.۴۸	۰.۰۲۱

با توجه به رابطه (۴) و در نظر گرفتن کمترین مقدار بین احتمال‌ها Pr برابر با مقدار زیر می‌شود.

$$Pr = \min\{0.15, 0.2, 0.2\} \times \min\{0.4, 0.5, 0.55\} \times \min\{0.021, 0.025, 0.018\} = 0.00108$$

چه احتمال صفت‌های شبه‌شناسه به دنیای واقعی نزدیک‌تر باشد و از مجموعه بزرگ‌تری انتخاب شود احتمال خطرپذیری حرمانگی کاهش می‌یابد [۶]؛ بنابراین در روش پیشنهادی برای رابطه (۴) احتمال $Pr(s)$ با توجه به احتمال صفت‌های شبه‌شناسه و صفت حساس به‌ازای تمام مجموعه‌های فراهم‌کنندگان انتخاب می‌شود.

برای این کار در گام نخست و پیش از عملیات گمنام‌سازی که نام آن مرحله نشست است، فراهم‌کنندگان مقادیر احتمال صفت‌های شبه‌شناسه و صفت حساس و همچنین اندازه مجموعه داده متعلق به خود را در اختیار هم قرار می‌دهند. این عمل حریم خصوصی هیچ‌یک از فراهم‌کنندگان را نقض نمی‌کند و فراهم‌کنندگان می‌توانند مجموعه فرضی D_0 را با توجه به آمار کل مجموعه‌ها محاسبه کنند؛ همچنین اگر تعداد شرکت‌کنندگان در این عمل بیش از دو فراهم‌کننده باشد، حتی تشخیص ارتباط بین یک مجموعه از احتمالات و یک فراهم‌کننده دشوار می‌شود.

در ادامه روشی برای محاسبه اطمینان مشاهده شده و اطمینان مورد انتظار در مدل پیشنهادی این مقاله ارائه می‌شود؛ در واقع روش پیشنهادی برای محاسبه اطمینان مورد انتظار، یک روش حریصانه^۱ است که از بین چند روش حریصانه دیگر انتخاب شد. معیار انتخاب، مقایسه روش‌ها بر مبنای درصد تعداد رکوردهای نقض شده به کل رکوردها و همچنین سودمندی داده‌ها در عملیات داده‌کاوی بود. پس از معرفی روش پیشنهادی، در جدولی این روش با روش‌های دیگر مقایسه می‌شود.

• محاسبه اطمینان مورد انتظار

برقراری امنیت رکوردها نسبت به حملات خارجی و همچنین حملات داخلی که توسط فراهم‌کنندگان پیاده‌سازی می‌شود، هدف اصلی طرح پیشنهادی است؛ بنابراین باید اطمینان مورد انتظار فراهم‌کنندگان دیگر را نسبت به رخداد رکورد مورد نظر کاهش داد. در GF_PPDP برای کاهش اطمینان مورد انتظار از نمونه‌برداری رکوردها استفاده شده است؛ برای مثال رکوردها با نرخ نود درصد نمونه‌برداری می‌شوند تا مقدار n در رابطه (۳) کاهش یابد و در نتیجه مقدار اطمینان مورد انتظار نیز کاهش یابد، اما نرخ نمونه‌برداری را نمی‌توان آنچنان کاهش داد؛ زیرا تعداد رکوردهای خروجی کاهش می‌یابد. این واقعیت که در نرخ پایین برای مثال سی درصد تنها سی درصد از کل رکوردها امکان انتشار می‌یابند انکار نشدنی است.

در روش پیشنهادی نمی‌توان نرخ نمونه‌برداری را آنچنان کاهش داد، تنها می‌توان با به‌کارگیری روش‌های نمونه‌برداری هوشمندانه رکوردهای بهتری را برای انتشار انتخاب کرد؛ از طرف دیگر برای استفاده از نمونه‌برداری

^۱ Greedy

با فرض اینکه تعداد رکوردهای متعلق به مجموعه P_1 برابر با پانصد باشد و جاگذاری Pr در رابطه (۳) اطمینان مشاهده شده برای رکورد r به دست می آید.

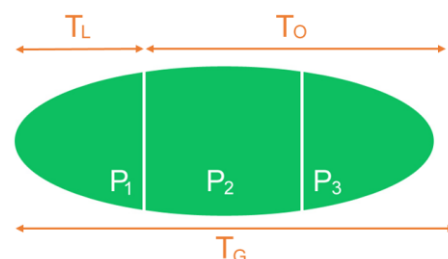
$$EC(r) = 1 - (1 - 0.00108)^{500} = 0.41$$

• محاسبه اطمینان مشاهده شده

همان طور که در شکل (۶) مشخص است، اگر مجموعه متعلق به تمام فراهم کنندگان T_G ، مجموعه متعلق به هر فراهم کننده T_L و مجموعه فراهم کنندگان دیگر به غیر از فراهم کننده مورد نظر T_O نامیده شود و همچنین رابطه (۲) با توجه به ثابت بودن β ، به صورت ساده شده $OC_G = \frac{S_G}{E_G} = \frac{S_L + S_O}{E_L + E_O}$ در نظر گرفته شود، آنگاه مقادیر S_L و E_L از مجموعه T_L در دسترس است (هر فراهم کننده به آمار مجموعه متعلق به خود دسترسی دارد) و مقدار تقریبی E_O را نیز می توان از ضرب احتمال صفت های شبه شناسه به دست آمده از مرحله نشست به صورت رابطه (۸) محاسبه کرد:

$$E_O \cong \left| \left(\left(\prod_i Pr(q_i) \right) \times N \right) - E_L \right| \quad (8)$$

در رابطه بالا مقدار Pr از احتمال مربوط به تمام فراهم کنندگان به دست می آید و همچنین N تعداد کل رکوردهای متعلق به تمام فراهم کنندگان را مشخص می کند. در رابطه ساده شده OC_G مقادیر E_L و S_L در دسترس اند، اما مقدار S_O برای فراهم کننده ای که در حال حاضر عملیات گمنام سازی را انجام می دهد قابل محاسبه نیست. S_O بیان کننده تعداد رکوردهایی از گروه های هم ارزی متعلق به مجموعه T_O است که مقدار صفت حساس آن ها برابر با رکورد مورد نظر است؛ بنابراین مقدار آن در دسترس نیست، اما می تواند مقادیر بین صفر تا E_O را اختیار کند. به ازای تمام مقادیر این بازه OC_G را محاسبه می کنیم و شرط رابطه (۵) را برای آن ها بررسی می کنیم. با این کار تعداد حالاتی که رابطه حریم خصوصی مدل گمنام سازی را ارضا می کنند را به دست آورده و اگر از حد آستانه γ بیشتر بود، رکورد مورد نظر را به عنوان رکورد قابل انتشار در مجموعه نهایی قرار می دهیم. در این تعریف γ آستانه انتشار توزیعی نامیده می شود و می تواند توسط هر فراهم کننده مقداردهی شود.



(شکل -۶): رابطه مجموعه های متعلق به هر فراهم کننده با

مجموعه فراهم کنندگان دیگر

(Figure-6): Relationship Between Datasets of Each Provider and Other Providers

برای مثال برای محاسبه اطمینان مشاهده شده رکورد r با مشخصات {سن ۲۳، جنسیت=زن، بیماری=سرطان} مربوط به P_1 از جدول (۲)، فرض می کنیم تعداد رکوردهایی که با رکورد r هم ارز هستند در مجموعه متعلق به فراهم کننده P_1 برابر با سی ($E_L=40$)، تعداد رکوردهایی از این گروه هم ارزی که مقدار صفت بیماری آن ها برابر با «سرطان» است نیز برابر ۱۷ باشد ($S_L=17$) باشد؛ همچنین فرض می کنیم مقدار N در رابطه (۸) که تعداد رکوردهای مربوط به فراهم کنندگان P_1 ، P_2 و P_3 است، برابر با هزار و پانصد باشد. در این صورت مقدار E_O با استفاده از رابطه (۸) به صورت زیر محاسبه می شود.

$$E_O = [(0.18 \times 0.48 \times 1500) - 40] = 89$$

بنابراین S_O نیز می تواند مقادیر صفر تا ۸۹ را اختیار کند. به ازای تمام این مقادیر اطمینان مشاهده شده OC_G را محاسبه کرده و با توجه به شرط محرمانگی با مقدار اطمینان مورد انتظار به دست آمده (۰.۴۱) مقایسه می شود. نسبت تعداد حالاتی که شرط محرمانگی را ارضا می کنند به کل حالت ها، اگر از حد آستانه انتشار توزیعی (۰.۷) بیشتر بود، رکورد مورد نظر به عنوان یک رکورد مناسب به مجموعه قابل انتشار اضافه می شود؛ در غیر این صورت عملیات گمنام سازی بر روی این رکورد انجام می گیرد.

اگر مجموعه احتمال شبه شناسه ها و صفت حساس متعلق به هر فراهم کننده را با Pr نمایش دهیم، در این صورت به تعداد فراهم کنندگان مجموعه احتمال وجود دارد $\{Pr_1, Pr_2, \dots, Pr_n\}$. زیرنویس احتمال ها نشان دهنده تعلق آن به فراهم کننده متناظر است؛ همچنین از دید هر فراهم کننده مجموعه احتمال متعلق به خودش با Pr_L و مجموعه احتمال مربوط به تمام فراهم کنندگان که از میانگین احتمالات به دست می آید، با Pr_G نشان داده می شود. در این مقاله پنج روش برای به دست آوردن اطمینان مورد انتظار با هم مقایسه شدند که شرح روش ها به صورت زیر است:

روش نخست: این روش اصلی مقاله است؛ به طوری که در رابطه (۴) به ازای هر صفت شبه شناسه و صفت حساس از کوچک ترین احتمال استفاده شود $\text{Min}(Pr_1, Pr_2, \dots, Pr_n)$.

روش دوم: به ازای هر صفت شبه شناسه و صفت حساس تنها از احتمال کل و احتمال مربوط به همان فراهم کننده استفاده شود و با مقایسه این دو احتمال، عدد کوچک تر برگزیده شود $\text{Min}(Pr_G, Pr_L)$.

روش سوم: استفاده از مجموعه احتمال کل Pr_G .
روش چهارم: استفاده از بزرگ ترین احتمال بین مجموعه احتمال های مربوط به فراهم کنندگان $\text{Max}(Pr_1, Pr_2, \dots, Pr_n)$.

روش پنجم: در این روش ابتدا به ازای هر فراهم کننده یک اطمینان مورد انتظار با استفاده از مجموعه احتمال مربوط به

همان فراهم‌کننده به‌دست می‌آید و سپس از بین این اطمینان‌ها کمترین انتخاب می‌شود $\text{Min}(EC_1, EC_2, \dots, EC_n)$.

جدول (۳) مقایسه‌ای بر مبنای محرمانگی و سودمندی داده‌های منتشرشده به‌ازای این پنج روش را نشان می‌دهد. در این جدول درصد تعداد رکوردهایی که پس از انتشار رابطه محرمانگی را نقض می‌کنند، به کل رکوردها، به‌عنوان معیار محرمانگی در نظر گرفته شده‌است. معیار محرمانگی برای دو حالت حملات داخلی و خارجی در جدول لیست شده‌است؛ همچنین دقت طبقه‌بندی داده‌های منتشرشده برای چهار روش طبقه‌بندی به‌عنوان معیار سودمندی داده‌ها در نظر گرفته شده‌است. گفتنی است که نتایج مقایسه به‌ازای چهل‌هزار رکورد به‌دست آمده‌است.

(جدول-۳): مقایسه روش‌های محاسبه اطمینان مورد انتظار
(Table-3): Comparison of Expected Confidence Calculation Methods

روش	محرمانگی در حملات خارجی	محرمانگی در حملات داخلی	سودمندی داده			
			J48	Naive Bayes	PAR T	Logistic
روش نخست	۰.۰۸۴۲	۰.۵۱۹۴	۸۳.۷۷	۸۱.۰۱	۸۱.۲۷	۸۰.۱۰
روش دوم	۰.۴۳۸۵	۱.۱۳۲۰	۸۴.۶۲	۸۳.۲۱	۸۱.۸۶	۸۴.۰۴
روش سوم	۰.۵۰۴۳	۸۸۳۳۶	۸۴.۳۶	۸۳.۱۴	۸۱.۳۰	۸۳.۲۷
روش چهارم	۱۵.۹۳۵۸	۴۴.۱۵۹۸	۸۴.۵۵	۸۲.۳۸	۸۰.۲۳	۷۷.۵۰
روش پنجم	۴۸۳۷۶	۲۵۶۴۸۷	۸۴.۶۸	۸۲.۵۳	۸۲.۰۸	۸۱.۶۸

روش نخست در معیار حفظ حریم خصوصی به‌طرز محسوسی بهتر از روش‌های دیگر عمل کرده‌است، اما در معیار سودمندی داده‌ها، روش نخست به‌صورت میانگین ضعیف‌تر از روش‌های دیگر است، اما این اختلاف آن‌چنان زیاد نیست و می‌توان روش نخست را به‌عنوان روش حریصانه برتر انتخاب کرد.

روند انتشار مجموعه‌داده

همانگی فراهم‌کنندگان در طی روند گمنام‌سازی، یکی از مشکلات انتشار مجموعه‌داده در حالت توزیع‌شده است. در برخی از روش‌ها در روند گمنام‌سازی، برای هر رکورد باید بین تمام فراهم‌کنندگان توافق حاصل شود، برای مثال در [۲۳] باید بر روی صفتی که در هر مرحله عمومی‌سازی می‌شود تصمیم‌گیری شود که همانگی لازم به‌وسیله پروتکل‌های SMC انجام می‌شود.

یکی از خصوصیات روش پیشنهادی وابسته‌نبودن فراهم‌کنندگان به یکدیگر در طی مراحل گمنام‌سازی است. این نداشتن وابستگی از مجموعه احتمال‌هایی که از مرحله نشست حاصل می‌شود و همچنین تخمین اطمینان

موردانتظار فراهم‌کنندگان دیگر، ناشی می‌شود. با توجه به این موضوع در ادامه روند انتشار مجموعه‌داده با استفاده از روش پیشنهادی در حالت توزیع‌شده را به‌صورت فهرست نشان می‌دهیم:

- با توجه به توزیع افقی داده‌ها، فراهم‌کنندگان بر روی صفات شبه‌شناسه و صفت حساس توافق می‌کنند.
- فراهم‌کنندگان می‌توانند بر روی درخت طبقه‌بندی مربوط به صفات شبه‌شناسه به توافق برسند. این کار باعث جلوگیری از به‌وجودآمدن دامنه متفاوت در مقادیر شبه‌شناسه‌ها می‌شود [۳۹].
- هرکدام از فراهم‌کنندگان احتمال مربوط به صفات شبه‌شناسه و حساس در مجموعه فرضی برابر با مجموعه متعلق به خود (D_0) در تعریف اطمینان مورد انتظار را محاسبه می‌کنند.
- احتمال‌های محاسبه‌شده در مرحله پیش بین فراهم‌کنندگان توزیع می‌شود (مرحله نشست). این کار به امنیت مجموعه‌ها خللی وارد نمی‌کند.
- فراهم‌کنندگان به‌صورت مستقل با استفاده از الگوریتم پیشنهادی اقدام به گمنام‌سازی مجموعه متعلق به خود می‌کنند.
- در مرحله آخر مجموعه‌داده گمنام‌شده فراهم‌کنندگان، جمع‌آوری شده و به‌صورت یک مجموعه واحد منتشر می‌شود.

الگوریتم گمنام‌سازی

الگوریتم گمنام‌سازی استفاده‌شده در این مقاله دقیق برابر با الگوریتم ارائه‌شده در روش GF_PPDP [۶] است؛ با این تفاوت که برای محاسبه اطمینان مورد انتظار و اطمینان مشاهده‌شده از روش‌های بیان‌شده در بخش پیشین استفاده می‌شود.

الگوریتم محاسبه اطمینان مورد انتظار در شبه‌کد (۱) آورده شده‌است. ورودی‌های الگوریتم به‌ترتیب برابرند با رکوردی که باید اطمینان مورد انتظار برای آن محاسبه شود، اندازه مجموعه‌داده‌ای که رکورد r متعلق به آن است و آرایه‌ای از احتمال‌ها که از مرحله نشست به‌دست آمده‌است.

در خط یک الگوریتم مقدار احتمال برابر با یک در نظر گرفته می‌شود؛ سپس در خطوط دو تا چهار به‌ازای تمام صفات‌های شبه‌شناسه و حساس، مقدار احتمالی که کمترین مقدار را از مجموعه احتمالات مرحله نشست دارد، انتخاب شده و در مقدار متغیر Probability ضرب می‌شود. این عمل به‌وسیله تابع $\text{getMinimumProbabilityFrom_N_prvider}$ انجام می‌شود و در آخر مقدار نهایی اطمینان مورد انتظار طبق رابطه (۳) محاسبه و مقدار آن برمی‌گردد.

CalcExpectedConfidence ()**Input:** r, n, array of N probability for QIDs and S value**Output:** EC

```

1: Probability ← 1
2: FOR each atr in QIDs and S DO
3: Probability ← Probability *
   getMinimumProbabilityFrom_N_provider(atr, r)
4: END FOR
5: RETURN (1 – Probability) ^ n

```

CalcObservedConfidence ()**Input:** r, EQGroups, EC, total, γ **Output:** OC

```

1: EL ← | EQGroups |
2: SL ← getSensitive(EQGroups, r)
3: EG ← 1
4: FOR each atr in QIDs DO
5: EG ← EG * nodeProbability(atr, r)
6: END FOR
7: EG ← EG * total
8: EO ← EG – EL
9: IF EO > 0 THEN
10: OCG ← 1, Satisfy ← 0
11: FOR i from 0 to EO DO
12: OCG ←  $\beta * ((SL + i) / (EL + EO))$ 
13: IF OCG < EC THEN
14: Satisfy ← Satisfy + 1
15: END IF
16: END FOR
17: IF (satisfy / (EO + 1)) <  $\gamma$  THEN
18: RETURN 1
19: END IF
20: END IF
21: RETURN  $\beta * (S\_L / E\_L)$ 

```

تعداد حالاتی که شرط محرمانگی را ارضا می کنند در متغیر Satisfy قرار می گیرد. در خط هفده تا نوزده اگر نسبت حالت هایی که شرط محرمانگی را ارضا می کنند به تعداد کل حالات از حد آستانه γ کمتر باشد مقدار «یک» برگردانده می شود تا شرط رابطه (۵) در الگوریتم گمنام سازی GF_PPDP نقض شود و عمل عمومی سازی بر روی رکورد صورت گیرد؛ همچنین اگر تعداد گروه های هم ارزی مربوط به دیگر فراهم کنندگان صفر باشد، در خط ۲۱ مقدار محاسبه شده اطمینان مشاهده شده بر اساس رابطه (۲) برگردانده می شود.

• پیاده سازی

برای پیاده سازی روش پیشنهادی و مقایسه آن با روش های دیگر از زبان جاوا که یکی از زبان های شناخته شده و محبوب است، استفاده شد. برای عملیات مربوط به مجموعه ها و داده کاوی از کتابخانه وکا بهره گرفته شد [۴۰]؛ وکا مجموعه ای از الگوریتم ها و ابزارهای پیش پردازش و ارزیابی در حوزه یادگیری ماشین است.

برای اجرای الگوریتم های گمنام سازی ابتدا باید مجموعه داده با فرمت csv. در اختیار برنامه قرار گیرد؛ سپس عملیات پیش پردازش داده شامل تشخیص صفات شبه شناسه و صفت حساس انجام می شود. طبق قرارداد صفات های شبه شناسه به پسوند #QI نام گذاری شده اند و همچنین آخرین صفت از مجموعه نیز به عنوان صفت حساس S در نظر گرفته می شود. مجموعه داده با توجه به نوع آزمایش به چند مجموعه کوچک تر (Party) تقسیم می شود و عملیات گمنام سازی با توجه به روش پیشنهادی بر روی آن ها صورت می گیرد و مجموعه نهایی قابل انتشار را تولید می کنند. این مجموعه با فرمت arff. که فرمت استاندارد وکا است ذخیره می شود. فایل مورد نظر با توجه به شامل بودن نام و خصوصیات صفات برای عملیات داده کاوی مناسب است. نتایج آماری حاصل از اجرای برنامه شامل دقت طبقه بندی^۱ و زمان اجرا در فایل با نام results.txt ذخیره می شود.

• آزمایش و ارزیابی نتایج

این بخش را با ارائه آزمون هایی جهت مقایسه عملکرد روش پیشنهادی (که DGF_PPDP^۲ نام گذاری می شود) آغاز کرده و سپس نتایج حاصل شده مورد بررسی قرار خواهد گرفت؛ همان طور که در فصول پیشین بیان شد PPDP علم ایجاد توازن بین محرمانگی و دقت داده های منتشر شده است؛ به همین دلیل باید الگوریتم پیشنهادی را از این دو دیدگاه مورد ارزیابی قرار داد. در شکل (۷) روند ارزیابی مربوط به آزمایش نشان داده شده است. مجموعه ابتدایی به دو مجموعه آموزش و آزمون تقسیم شده (نسبت یک به ده برای مجموعه آزمون)؛ سپس مجموعه آموزش با حجم برابر بین چهار

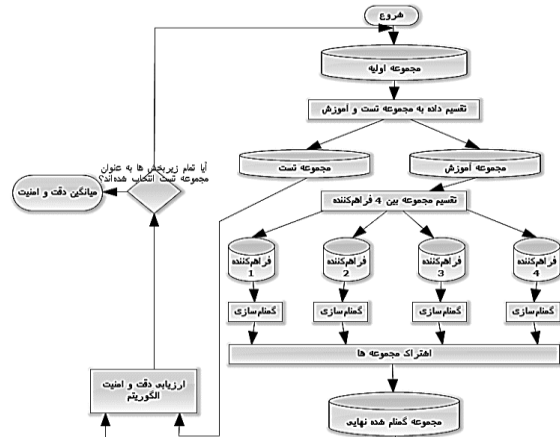
الگوریتم محاسبه اطمینان مشاهده شده در شبه کد (۲) آورده شده است؛ همان طور که در شبه کد مشخص است، ورودی های این الگوریتم به ترتیب از چپ به راست؛ رکورد r، گروه های هم ارز با رکورد مورد نظر، مقدار محاسبه شده برای اطمینان مورد انتظار، تعداد کل رکوردهای مربوط به تمام فراهم کنندگان و حد آستانه انتشار توزیعی است؛ همچنین خروجی الگوریتم مقدار اطمینان مشاهده شده به ازای رکورد r است. تابع getSensitive در خط دو تعداد رکوردهای موجود در گروه هم ارزی، که صفت حساس آن ها برابر با رکورد r است را برمی گرداند. در خط چهار تا هشت تعداد گروه های هم ارزی متعلق به فراهم کنندگان دیگر محاسبه می شود. تابع nodeProbability احتمال صفت شبه شناسه ها QIDs را برمی گرداند.

در خطوط نه تا شانزده به ازای تمام مقادیری که So می تواند اختیار کند، اطمینان مشاهده شده محاسبه می شود و

^۱ Classification^۲ Distributed General Framework for Privacy Preserving Data Publishing

فراهم‌کننده تقسیم می‌شود و هر فراهم‌کننده بر اساس مدل محرمانگی (GF_PPDP یا DGF_PPDP) داده‌های مربوط به خود را گمنام می‌کند، پس از آن داده‌های گمنام‌شده تجمیع می‌شوند و مجموعه قابل انتشار نهایی را تولید می‌کنند. این مجموعه به همراه مجموعه آزمون، از لحاظ دقت طبقه‌بندی مورد ارزیابی قرار می‌گیرند؛ همچنین مجموعه‌های تولیدشده توسط هر کدام از فراهم‌کنندگان با استفاده از مکانیسم m -privacy [۲۲] که در بخش پیش تشریح شد، از لحاظ محرمانگی مورد ارزیابی قرار می‌گیرند. فرایند بالا تا زمانی که تمام ده قسمت مجموعه ابتدایی به عنوان آزمون در نظر گرفته شوند ادامه پیدا می‌کند. و در انتها میانگین نتایج به دست آمده مورد استفاده قرار می‌گیرد.

فراهم‌کننده تقسیم می‌شود و هر فراهم‌کننده بر اساس مدل محرمانگی (GF_PPDP یا DGF_PPDP) داده‌های مربوط به خود را گمنام می‌کند، پس از آن داده‌های گمنام‌شده تجمیع می‌شوند و مجموعه قابل انتشار نهایی را تولید می‌کنند. این مجموعه به همراه مجموعه آزمون، از لحاظ دقت طبقه‌بندی مورد ارزیابی قرار می‌گیرند؛ همچنین مجموعه‌های تولیدشده توسط هر کدام از فراهم‌کنندگان با استفاده از مکانیسم m -privacy [۲۲] که در بخش پیش تشریح شد، از لحاظ محرمانگی مورد ارزیابی قرار می‌گیرند. فرایند بالا تا زمانی که تمام ده قسمت مجموعه ابتدایی به عنوان آزمون در نظر گرفته شوند ادامه پیدا می‌کند. و در انتها میانگین نتایج به دست آمده مورد استفاده قرار می‌گیرد.



(شکل ۷): روند ارزیابی الگوریتم برای دقت و امنیت
(Figure-7): Algorithm Evaluation Process for Accuracy and Security

مجموعه داده استفاده شده در تمام آزمون‌های این مقاله Adult [۴۱] نام دارد که در بسیاری از پژوهش‌ها [۳۵،۲۲،۶] مورد استفاده قرار گرفته است. این مجموعه داده شامل ۴۸۸۴۲ رکورد و پانزده صفت است که در جدول (۴) مشخصات این صفات ذکر شده است. صفت Income یا درآمد سالیانه، به عنوان صفت حساس و هفت صفت (age, workclass, education, occupation, race, sex و native-country) به عنوان صفات شبه‌شناسه مورد استفاده قرار گرفتند. در تمامی آزمون‌های صورت گرفته برای عملیات عمومی سازی از درخت طبقه‌بندی^۱ مطابق با [۴۲] استفاده شد، نرخ نمونه‌برداری از مجموعه‌ها برابر «نود درصد»، حد تخریب مربوط به مدل GF_PPDP برابر «۰.۶» و حد آستانه انتشار توزیعی مربوط به مدل پیشنهادی برابر «۰.۱» انتخاب شده‌اند. با توجه به اینکه حد آستانه انتشار توزیعی موازنه‌ای بین دقت و محرمانگی ایجاد می‌کند، در آزمایش‌ها مقدار ۰.۱ اختیار شد تا به دقت بالاتری برسیم. در کارهایی که امنیت بیشتری مدنظر است می‌توان این مقدار را افزایش داد.

• امنیت و حفظ حریم خصوصی داده‌ها

در این آزمایش حفظ حریم خصوصی داده‌های منتشرشده به وسیله روش پیشنهادی (DGF_PPDP) با روش

(جدول ۴): مشخصات صفات در مجموعه Adult

(Table-4): Attributes Specification in the Adult Dataset

صفت	نوع	تعداد/دامنه تغییرات	صفت	نوع	تعداد/دامنه تغییرات
Age	پیوسته	۱۷-۹۰	Race	گسسته	۵
Work-class	گسسته	۸	Sex	گسسته	۲
Final-weight	پیوسته	-۱۳۴۹۲ تا ۱۴۹۰۴۰۰	Capital-gain	پیوسته	۰-۹۹۹۹۹۰
Education	گسسته	۱۶	Capital-loss	پیوسته	۰-۴۳۵۶
Education-num	پیوسته	۱-۱۶	Hours-per-week	پیوسته	۱-۹۹
Marital-status	گسسته	۷	Native-country	گسسته	۴۰
Occupation	گسسته	۱۴	Income	گسسته	۲
Relationship	گسسته	۶			

جدول (۵) نتایج حاصل از نقض حریم خصوصی در مجموعه نهایی برای دو الگوریتم GF_PPDP و DGF_PPDP را به ازای اندازه مجموعه داده نشان می‌دهد. تعداد و درصد به صورت میانگین به ازای رکوردهای نقض شده برای حمله داخلی و خارجی در جدول مشخص شده است. اندازه مجموعه داده بر اساس اندازه مجموعه آموزش مشخص شده است؛ برای مثال عدد ۴۰۵۰ مربوط به مجموعه‌ای با اندازه پنج هزار رکورد است که ده درصد آن به عنوان مجموعه آزمون انتخاب و باقی آن با نرخ «نود درصد» نمونه‌برداری شده‌اند. برای درک بهتر نتایج امنیت مجموعه منتشرشده به ازای دو الگوریتم یادشده، شکل (۸) نمودار حفظ حریم خصوصی در مقابل حملات خارجی را نشان می‌دهد. در این شکل درصد رکوردهای نقض شده به کل رکوردهای انتشار یافته، برای مجموعه داده با اندازه‌های مختلف به تصویر کشیده شده است. رنگ سبز مربوط به روش پیشنهادی و رنگ قرمز مربوط به روش GF_PPDP است. در تمامی

² The direct m -privacy verification algorithm

¹ Taxonomy Tree

حالات رکوردهای نقض شده برای روش پیشنهادی کمتر از روش GF_PPDP است. نکته قابل توجه در شکل این است که کمابیش در تمامی حالات درصد رکوردهای نقض شده برای روش پیشنهادی برابر صفر است.

شکل (۹) نیز نمودار درصد رکوردهای نقض شده در مقابله با حملات داخلی را نشان می دهد. در این حالت روش پیشنهادی بهتر از روش GF_PPDP عمل کرده است و در تمام حالت ها درصد رکوردهای نقض شده از ۰.۸ درصد تجاوز نکرده است. تفاوت نرخ رکوردهای نقض شده بین دو روش به این دلیل است که الگوریتم GF_PPDP بدون در نظر گرفتن مشخصات فراهم کنندگان دیگر اقدام به گمنام سازی می کند، اما روش پیشنهادی با در نظر گرفتن احتمال شبه شناسه ها در مجموعه های فراهم کنندگان

دیگر، اطمینان مورد انتظار فراهم کنندگان دیگر برای دیدن یک رکود را کاهش می دهد. جدول (۶) میانگین رکوردهای نقض شده در هر سطح m از حمله داخلی را نمایش می دهد. نمودار مورد نظر به ازای تعداد چهل هزار رکود که بین چهار فراهم کننده تقسیم شده، رسم شده است. به ازای تمام حالاتی که در شکل (۵) نشان داده شده است رکوردها از نظر محرمانگی چک شده و تعداد رکوردهایی که محرمانگی را نقض کردند محاسبه شده است. بر طبق روش m-privacy اگر رکوردی در سطح m نقض شود به طور قطع در سطح m-1 نیز نقض خواهد شد [۲۲]؛ لذا در جدول (۶) تعداد رکوردهای نقض شده در هر سطح با مجموع تعداد رکوردهای سطوح پایین تر تجمیع شده است.

(جدول ۵): نتایج رکوردهای نقض شده به ازای حملات داخلی و خارجی
(Table-5): Violated Records Count for Internal and External Attacks

DGF PPDP				GF PPDP				
خارجی		داخلی		خارجی		داخلی		اندازه مجموعه داده آموزش
درصد رکورد نقض شده	تعداد رکورد نقض شده	درصد رکورد نقض شده	تعداد رکورد نقض شده	درصد رکورد نقض شده	تعداد رکورد نقض شده	درصد رکورد نقض شده	تعداد رکورد نقض شده	
۰.۰	۰.۰	۰.۷۰۳	۲۸.۵	۶.۷۴۵	۲۷۳.۲	۲۳.۸۹۸	۹۶۷.۹	
۰.۰	۰.۰	۰.۲۹۶	۲۴.۰	۹.۳۱۱	۷۵۴.۲	۲۶.۳۹۰	۲۱۳۸.۱	۸۱۰۰
۰.۰	۰.۰	۰.۳۴۸	۵۶.۴	۶.۳۳۲	۱۲۹۴.۱	۲۱.۰۰۱	۳۴۰۲.۲	۱۶۲۰۰
۰.۰≈	۰.۱	۰.۳۳۲	۸۰.۷	۴.۱۷۶	۱۵۳۸.۷	۱۶.۸۶۹	۴۰۹۹.۲	۲۴۳۰۰
۰.۰۸۵	۲۷.۷	۰.۵۲۱	۱۶۸.۹	۴.۱۷۶	۱۳۵۳.۳	۱۱.۱۱۰	۳۵۹۹.۹	۳۲۴۰۰

(Figure-9): Privacy Against Internal Attacks

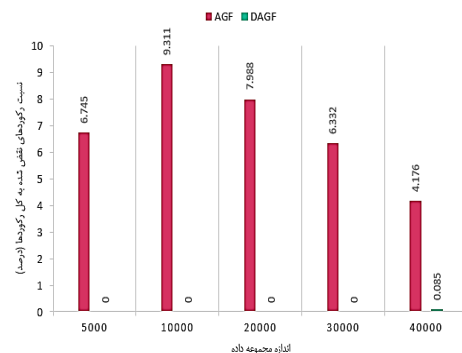
(جدول ۶): میانگین رکوردهای نقض شده در حملات داخلی به ازای هر سطح

(Table-6): Average Violated Records in Internal Attacks per Level

M=۳	m=۲	m=۱	
۶۷.۹+۳۵۴۷.۳=۳۶۱۵.۲	۱۲۱۰.۲+۲۳۳۷.۱=۳۵۴۷.۳	۲۳۳۷.۱	GF_PPDP
۴۸.۹+۱۱۹۰.۴=۱۶۸.۳	۴۶+۷۳.۴=۱۱۹.۴	۷۳.۴	DGF_PPDP

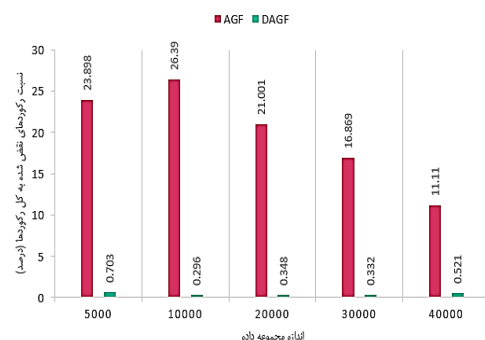
• دقت داده های انتشار یافته

برآورده کردن شاخص حفظ حریم خصوصی و امنیت در داده های منتشر شده بسیار حائز اهمیت است، اما از طرف دیگر باید داده های نهایی سودمندی لازم را جهت عملیات داده کاوی حفظ کنند؛ از این رو در این بخش نتایج دقت طبقه بندی داده های منتشر شده در حالت های مختلف مورد ارزیابی قرار می گیرد. روند ارزیابی به صورت چارت ارائه شده در شکل (۷) است. در این حالت برای ارزیابی تأثیر گمنام سازی بر دقت طبقه بندی از الگوریتم های J48، NaiveBayes و Logistic استفاده شده و همچنین برای مقایسه بهتر، دقت طبقه بندی بر روی داده های خام



(شکل ۸): محرمانگی در مقابل حمله خارجی

(Figure-8): Privacy Against External Attacks



(شکل ۹): محرمانگی در مقابل حمله داخلی

در دو نمودار پیش برای روشن‌تر شدن سودمندی اشتراک داده، دقت در دو حالت انفرادی و اشتراکی مورد ارزیابی قرار می‌گیرد. در حالت نخست یعنی حالتی که تنها یک فراهم‌کننده قصد انتشار داده را دارد، مجموعه متعلق به هر فراهم‌کننده با روش GF_PPDP گمنام‌سازی می‌شود و در حالت دوم مجموعه داده فراهم‌کنندگان با استفاده از روش DGF_PPDP گمنام‌سازی و به اشتراک گذاشته می‌شود. در این آزمون مشخص می‌کنیم که اشتراک داده چه تأثیری بر روی دقت داده‌ها دارد. شکل (۱۲) تفاوت دو حالت یادشده را به‌ازای تعداد فراهم‌کنندگان ($p = \{2, 3, 4, 5\}$) نشان می‌دهد. گفتنی است که اندازه مجموعه داده هر بخش حدود نه‌هزار رکورد است و تعداد شبه‌شناسه در این آزمون برابر با نه است. اگر دقت طبقه‌بندی روش GF_PPDP را A_1 بنامیم و دقت طبقه‌بندی مربوط به روش DGF_PPDP را A_2 بنامیم، هر میله از نمودار شکل (۱۲) مقدار $(A_2 - A_1)$ را نشان می‌دهد. در جدولی که زیر نمودار (۱۳) قرار دارد، عدد اول مربوط به هر ستون، درصد دقت طبقه‌بندی در حالت اشتراکی (A_2) و عدد دوم این مقدار در حالت انفرادی (A_1) به‌ازای یکی از فراهم‌کنندگان را نشان می‌دهد. تفاوت این دو مقدار در نمودار به‌ازای الگوریتم‌های طبقه‌بندی رسم شده‌است؛ همان‌طور که در شکل مشخص است، برای تمام الگوریتم‌های طبقه‌بندی با افزایش تعداد فراهم‌کننده (Party) تفاوت دقت طبقه افزایش پیدا کرده‌است؛ به‌عبارت دیگر اشتراک داده و استفاده از روش پیشنهادی به‌نفع فراهم‌کنندگان بوده‌است. باینکه در الگوریتم پیشنهادی عمل‌گرهای گمنامی مانند عمومی‌سازی بیشتر استفاده می‌شود، اما اشتراک داده باعث شده‌است در مجموع الگوریتم DGF_PPDP سودمندی بیشتری نسبت به الگوریتم GF_PPDP داشته باشد.

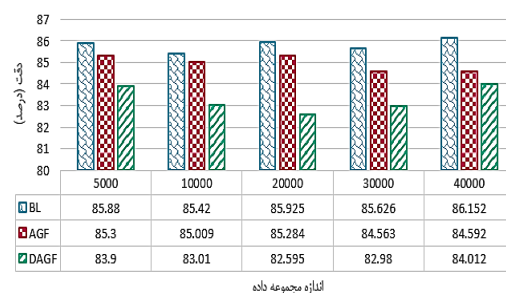
• زمان اجرا

یکی دیگر از پارامترها در ارزیابی الگوریتم‌ها زمان اجرا است. سازوکار گمنام‌سازی الگوریتم پیشنهادی با الگوریتم GF_PPDP تنها در محاسبه اطمینان مشاهده‌شده و اطمینان مورد انتظار متفاوت است. در محاسبه اطمینان مورد انتظار همان‌طور که شبه‌کد (۱) نشان می‌دهد، به‌ازای هر شبه‌شناسه حلقه‌ای به تعداد آمار فراهم‌کنندگان وجود دارد؛ بنابراین بدیهی است با افزایش تعداد فراهم‌کنندگان زمان اجرا نیز افزایش پیدا می‌کند؛ همچنین در روند محاسبه اطمینان مشاهده‌شده (شبه‌کد (۲)) نیز دو حلقه اضافه نسبت به الگوریتم GF_PPDP وجود دارد. حلقه نخست به‌تعداد شبه‌شناسه‌ها برای محاسبه اطمینان مورد انتظار و حلقه دیگر به‌تعداد رکوردهایی که صفت حساس مشابه با رکورد r دارند (S_0 در شبه‌کد (۲))؛ بنابراین در محاسبه اطمینان مورد انتظار نیز الگوریتم پیشنهادی کندتر از روش GF_PPDP عمل می‌کند.

(گمنام‌نشده) نیز اندازه‌گیری شده‌است که در نمودار آن را با نام BL نمایش می‌دهیم.

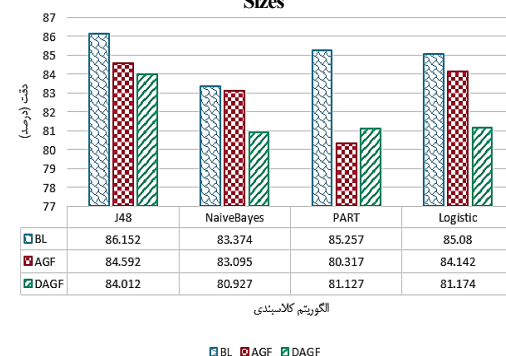
شکل (۱۰) نمودار درصد دقت طبقه‌بندی الگوریتم J48، به‌ازای مجموعه داده با اندازه‌های مختلف برای سه روش یادشده در بالا را نشان می‌دهد. واضح است که روش‌های GF_PPDP و DGF_PPDP به‌دلیل استفاده از عمل‌گرهای عمومی‌سازی دقت طبقه‌بندی کمتری نسبت به حالت پایه دارند؛ همچنین روش پیشنهادی برای برقراری امنیت بیشتر عملیات گمنام‌سازی مانند عمل‌گر عمومی‌سازی را تا سطح بیشتری انجام می‌دهد و از این‌رو دقت داده‌های منتشرشده با الگوریتم DGF_PPDP کمی پایین‌تر از روش GF_PPDP است؛ همچنین برای ارزیابی بهتر سودمندی داده‌ها، شکل (۱۱) دقت طبقه‌بندی روش GF_PPDP، DGF_PPDP و روش پایه را به‌ازای الگوریتم‌های مختلف طبقه‌بندی در مجموعه داده اولیه با اندازه چهار هزار رکورد نشان می‌دهد؛ همان‌طور که در شکل مشخص است دقت طبقه‌بندی برای الگوریتم‌های مختلف متفاوت است، اما بر حسب معمول روش GF_PPDP از دقت بهتری نسبت به DGF_PPDP برخوردار است و فقط در حالت الگوریتم PART این مقدار به نفع روش DGF_PPDP است، اما به‌طور کلی با توجه به امنیت داده‌های روش DGF_PPDP، تفاوت دقت طبقه‌بندی روش پیشنهادی و روش پایه مقدار قابل قبولی دارد.

الگوریتم J48



(شکل -۱۰): دقت طبقه‌بندی به‌ازای مجموعه با اندازه‌های مختلف

(Figure-10): Classification Accuracy for Datasets of Different Sizes



(شکل -۱۱): دقت طبقه‌بندی به‌ازای الگوریتم‌های مختلف طبقه‌بندی

(Figure-11): Classification Accuracy for Different Classification Algorithms

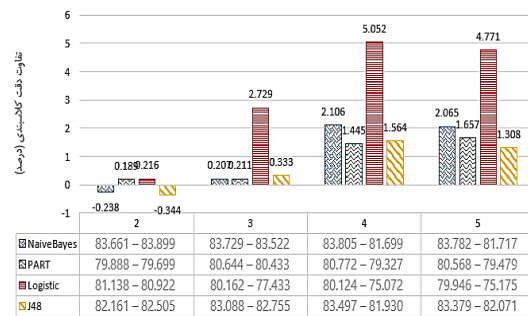
است. نتایج پیاده‌سازی طرح روی مجموعه‌داده‌گان استاندارد نشان می‌دهد که این روش توازن مناسبی میان امنیت و کارایی داده‌های منتشرشده نهایی برقرار می‌کند. مهم‌ترین نوآوری در این مقاله ارائه روشی است که بدون نیاز به TTP و SMC محرمانگی را در انتشار داده‌های توزیع‌شده برقرار کند.

یکی دیگر از ویژگی‌های مهم این روش استفاده از مدل احتمالاتی برای عملیات گمنام‌سازی است؛ درحالی‌که بیشتر روش‌های توزیع‌شده از مدل‌های پیونددهی برای حفظ حریم خصوصی استفاده می‌کنند. در روش پیشنهادی با تخمین اطمینان مهاجم (ممکن است مهاجم از مجموعه فرام‌کنندگان باشد) پیش و پس از دیدن داده‌های منتشرشده، رکوردهای مجموعه به‌نحوی گمنام‌سازی می‌شوند که در مقابل حملات خارجی و داخلی مقاوم باشد. روش ارائه‌شده در این مقاله می‌تواند مورد استفاده شرکت‌هایی قرار گیرد که در یک حوزه مشترک فعالیت دارند؛ مانند بیمارستان‌ها و بانک‌ها که داده‌های ذخیره‌شده آن‌ها شمای یکسانی دارند؛ همچنین این روش برای زیرمجموعه‌های شرکت‌های بزرگ مانند بیمه، که قصد جمع‌آوری اطلاعات و انتشار آن‌ها را دارند می‌تواند مورد استفاده قرار گیرد.

6-References

۶-مراجع

- [1] A. Majeed and S. Lee, (2021) "Anonymization Techniques for Privacy Preserving Data Publishing: A Comprehensive Survey," in *IEEE Access*, vol. 9, pp. 8512-8545, doi: 10.1109/ACCESS.2020.3045700.
- [2] "Universal Declaration of human Rights", [Online]. Available: <http://www.un.org/en/documents/udhr>
- [3] J. Bennett and S. Lanning, (2007), "The Netflix Prize", *Proceeding of the KDD Cup Workshop*, pp. 3-6
- [4] Barbaro, M., Zeller, T., Hansell, S, (2006). "A face is exposed for AOL searcher no. 4417749", *New York Times*, 9(2008), 8For.
- [5] Kiran, P., Kavya, N. P, (2012). "A Survey on Methods, Attacks and Metric for Privacy Preserving Data Publishing", *International Journal of Computer Applications*, Vol. 53, No. 18.
- [6] Sattar, A. S., Li, J., Ding, X., Liu, J., Vincent, M. (2013), "A general framework for privacy preserving data publishing", *Knowledge-Based Systems*, Vol. 54, pp. 276-287.
- [7] Parsa, Mojtaba, (2009), "Privacy and Confidentiality in Medicine and its Various AspectsK", *Journal of Medical Ethics and History of Medicine*, Vol. 2, No 4, pp. 1-14.
- [7] پارسا، مجتبی (۱۳۸۸)، «حریم خصوصی و رازداری در پزشکی و جنبه‌های مختلف آن»، *مجله ایرانی اخلاق و تاریخ پزشکی*، جلد ۲، شماره ۴، صص ۱-۱۴.

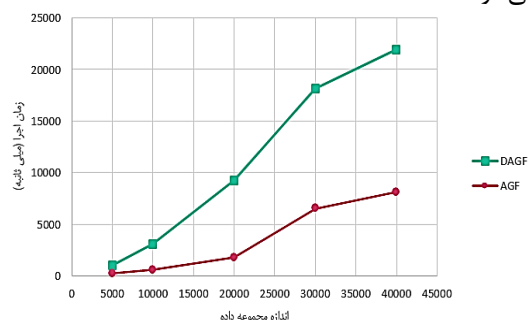


(شکل - ۱۲): مقایسه دقت طبقه‌بندی در حالت انفرادی و

اشتراکی به‌ازای تعداد فراهم‌کنندگان

(Figure-12): Comparison of Classification Accuracy in Individual vs. Collaborative Modes for Different Provider Counts

برای ارزیابی تفاوت یادشده در اجرای دو روش، میانگین زمان اجرای الگوریتم‌ها به‌ازای مجموعه‌داده‌ها با اندازه مختلف محاسبه و در شکل (۱۳) نمایش داده شده‌است. دو الگوریتم بر روی رایانه‌ای با پردازنده Intel core i5-4210U (2.4 GHz) و حافظه اصلی اجرا شدند؛ همان‌طور که در شکل مشخص است، با افزایش اندازه مجموعه‌داده زمان اجرای هر دو الگوریتم افزایش می‌یابد؛ همچنین اختلاف زمان اجرای دو الگوریتم نیز با افزایش تعداد رکوردها بیشتر می‌شود.



(شکل - ۱۳): میانگین زمان اجرا در گمنام‌سازی داده‌ها با اندازه

مختلف

(Figure-13): Average Execution Time for Anonymizing Datasets of Different Sizes

۵- جمع‌بندی و نتیجه‌گیری

در این مقاله ابتدا مفاهیم و مدل‌های مورد نیاز در روش‌های انتشار با حفظ حریم خصوصی در داده‌های توزیع‌شده مورد بررسی قرار گرفت؛ سپس مروری جامع روی روش‌های مطرح این حوزه انجام شد و یک طرح جدید برای حفظ محرمانگی در داده‌های توزیع‌شده مبتنی بر مدل‌های احتمالاتی ارائه شد که دارای ویژگی استقلال زمانی شرکت‌کنندگان در انتشار داده از لحاظ زمان عملیات گمنام‌سازی، عدم نیاز به توافق برای تعیین سطح معینی از محرمانگی، عدم نیاز به شخص سوم مورد اعتماد، عدم نیاز به پروتکل‌های محاسبات امن چندگانه در روند گمنام‌سازی است و در برابر حملات داخلی مقاوم

- Knowledge Discovery from Data (TKDD)*, Vol. 1, No 1.
- [20] C. Dwork, (2006), "Differential Privacy," in *Automata, Languages and Programming*, Springer Berlin Heidelberg, pp. 1-12.
- [21] Sweeney, L., (2002), "k-anonymity: A model for protecting privacy", *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, Vol. 10, No. 05, pp. 557-570.
- [22] Goryczka, S., Xiong, L., Fung, B. C., (2014), "m-Privacy for Collaborative Data Publishing", *Ieee Transactions On Knowledge And Data Engineering*, Vol. 26, No. 10, pp. 2520-2533.
- [23] Mohammed, N., Fung, B., Wang, K., Hung, P. C. (2009, March), "Privacy-preserving data mashup", In *Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology* (pp. 228-239). ACM.
- [24] Office for Civil Rights, H. H. S. (2002), "Standards for privacy of individually identifiable health information. Final rule", *Federal Register*, Vol. 67, No. 157, pp. 53181.
- [۲۵] صادق پور، مهدی، (۱۳۹۴)، «حفظ حرمانگی در انتشار داده‌ها به وسیله گمنام‌سازی دسته‌ای»، پایان‌نامه کارشناسی ارشد مهندسی کامپیوتر، گیلان: دانشگاه گیلان.
- [25] Sadeghpour, Mehdi, (2015), "Preserving Confidentiality in Data Publication through Batch Anonymization", *Master's Thesis in Computer Engineering*, Gilan: University of Gilan.
- [26] W Surapon Riyana, Noppamas Riyana, and Srikul Nanthachumphu, (2021), "Privacy Preservation Techniques for Sequential Data Releasing", In *Proceedings of the 12th International Conference on Advances in Information Technology (IAIT '21)*, Association for Computing Machinery, New York, NY, USA, Article 24, 1-9. <https://doi.org/10.1145/3468784.3470468>.
- [27] Samarati, P. (2001), "Protecting respondents identities in microdata release", *IEEE transactions on Knowledge and Data Engineering*, Vol. 13, No. 6, pp. 1010-1027.
- [28] Li, N., Li, T., Venkatasubramanian, S. (2010), "Closeness: A new privacy measure for data publishing", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 22, No. 7, pp. 943-956.
- [29] Silva de Garcia, P., Oliveira, M., & Brohman, K. (2020), "Knowledge sharing, hiding and hoarding: how are they related?", *Knowledge Management Research & Practice*, 20(3), 339-351. <https://doi.org/10.1080/14778238.2020>.
- [30] W. Ren, K. Ghazinour and X. Lian, (2023) "kt-Safety: Graph Release via k-Anonymity and t-Closeness," in *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 9, pp. 9102-9113, doi: 10.1109/TKDE.2022.3221333.
- [8] Aggarwal, C. C., Philip, S. Y. (2008), "A general survey of privacy-preserving data mining models and algorithms", In *Privacy-preserving data mining*, Springer US, pp. 11-52.
- [9] Manta, A, (2013), "Literature survey on privacy preserving mechanisms for data publishing", *MSc. Thesis on Department of Intelligence Systems*, Faculty EEMCS, Delft University of Technology.
- [10] Benjamin, C. M., Fung, M., Wang, K. E., Chen, R., Yu, P. S. (2010), "Privacy-preserving data publishing: A survey of recent developments", *ACM Computing Surveys*, Vol. 42, No 4, pp. 141-153.
- [11] Sun, Chang; Ippel, Lianne; Dekker, Andre; Dumontier, Michel; van Soest, Johan (2021), *A systematic review on privacy-preserving distributed data mining*, IOS Press, Data Science 4 (2021) 121-150, doi: 10.3233/DS-210036.
- [12] Tânia Carvalho, Nuno Moniz, Pedro Faria, and Luís Antunes. (2023), *Survey on Privacy-Preserving Techniques for Microdata Publication*, ACM Comput, Surv. 55, 14s, Article 309 (December 2023), 42 pages. <https://doi.org/10.1145/3588765>
- [13] Jurczyk, P., Xiong, L. (2009, July), "Distributed anonymization: Achieving privacy for both data subjects and data providers", In *IFIP Annual Conference on Data and Applications Security and Privacy* (pp. 191-207). Springer Berlin Heidelberg.
- [14] Fung, B. C., Wang, K., Philip, S. Y. (2007), "Anonymizing classification data for privacy preservation", *IEEE transactions on knowledge and data engineering*, Vol. 19, No 5.
- [15] D Joseph Ficek, Wei Wang, Henian Chen, Getachew Dagne, Ellen Daley, (2021), "Differential privacy in health research: A scoping review", *Journal of the American Medical Informatics Association*, Vol. 28, Issue 10, Pages 2269-2276, <https://doi.org/10.1093/jamia/ocab135>.
- [16] Clifton, C., Tassa, T. (2013, April), "On syntactic anonymity and differential privacy", In *Data Engineering Workshops (ICDEW)*, 2013 IEEE 29th International Conference on (pp. 88-93). IEEE.
- [17] Slawomir A. Goryczka, (2014), "Secure and Privacy-Preserving Distributed Data Release", *PhD. Thesis on Computer Science*, Emory University.
- [18] A. Majeed and S. O. Hwang, (2024) "Differential Privacy and k-Anonymity-Based Privacy Preserving Data Publishing Scheme With Minimal Loss of Statistical Information," in *IEEE Transactions on Computational Social Systems*, vol. 11, no. 3, pp. 3753-3765, doi: 10.1109/TCSS.2023.3320141.
- [19] Machanavajjhala, A., Kifer, D., Gehrke, J., Venkatasubramanian, M. (2007), "l-diversity: Privacy beyond k-anonymity", *ACM Transactions on*

- [42] "Taxonomy trees of the Adult data set", [Online]. Available: <http://ddm.cs.sfu.ca/dmsoft/privacy/products/adultHierarchy.txt>. [Accessed 5 October 2020].
- [43] مسیعی، الیاس. (۱۳۹۵)، «تحلیل امنیتی روش‌های انتشار با حفظ محرمانگی در داده‌های توزیع‌شده»، پایان‌نامه کارشناسی/ارشد مهندسی کامپیوتر، گیلان: دانشگاه گیلان.
- [44] Mosayebi, Elyas (2016), "Security Analysis of Privacy-Preserving Publication Methods in Distributed Data", *Master's Thesis in Computer Engineering, Gilan: University of Gilan*.
- [44] Kim, Pauline and Bodie, Matthew T., (2021), "Artificial Intelligence and the Challenges of Workplace Discrimination and Privacy (September 2021)", *35 ABA Journal of Labor and Employment Law* 289, Washington University in St. Louis Legal Studies Research Paper No. 21-09-02, Saint Louis U, Legal Studies Research Paper No. 2021-26, Available at SSRN: <https://ssrn.com/abstract=392906>
- [45] McMahan Brendan, Moore Eider, Ramage Daniel, Hampson Seth, Arcas Blaise Aguera y (2017), "Communication-efficient learning of deep networks from decentralized data", *In: Artificial intelligence and statistics*, pp 1273–1282. PMLR.
- [46] Latif, N., Ma, W. & Ahmad, H.B, (2025), "Advancements in securing federated learning with IDS: a comprehensive review of neural networks and feature engineering techniques for malicious client detection", *Artif Intell Rev*, 58, 91 <https://doi.org/10.1007/s10462-024-11082-w>.
- [47] مرادی لالمی، علی، شاه بهرامی، اسداله، ابراهیمی آتانی، رضا، علی دوست نیا، مهران (۱۳۹۵)، «خوشه‌بندی فراابتکاری اسناد فارسی اکس‌ام‌ال مبتنی بر شباهت ساختاری و محتوایی»، نشریه پردازش علائم و داده‌ها، ۲ (۲۸)، صص ۱۱-۲۳.
- [47] Moradi A, Shahbahrami A, Ebrahimi Atani R, Alidoust Nia M (2016), "Persian XML Documents Metaheuristic Clustering Based on Structure and Content Similarity", *Journals of signal and Processing Data*; 13 (2): 11-23
- [48] ابراهیمی آتانی، رضا، صادقیور، مهدی (۱۳۹۷)، «ارایه یک روش جدید انتشار داده‌ها با حفظ محرمانگی با هدف بهبود دقت طبقه‌بندی روی داده‌های گمنام»، نشریه پردازش علائم و داده‌ها، ۳ (۳۷)، صص ۳۱-۴۶.
- [48] Ebrahimi Atani R, Sadeghpour M (2018), "A New Privacy Preserving Data Publishing Technique Conserving Accuracy of Classification on Anonymized Data", *Journals of signal and Processing Data*; 15 (3): 31-46.
- [49] S. Ameri and R. E. Atani, (2024), "A Novel Decentralized Privacy Preserving Federated Learning Model for
- [31] Dwork, C., McSherry, F., Nissim, K., Smith, A. (2006, March), "Calibrating noise to sensitivity in private data analysis", *In Theory of Cryptography Conference* (pp. 265-284), Springer Berlin Heidelberg.
- [32] Jiang, W., Clifton, C. (2006), "A secure distributed framework for achieving k-anonymity", *The VLDB Journal—The International Journal on Very Large Data Bases*, Vol. 15, No. 4, pp. 316-333.
- [33] Hewage, U.H.W.A., Sinha, R. & Naeem, M.A. (2023), "Privacy-preserving data (stream) mining techniques and their impact on data mining accuracy: a systematic literature review", *Artif Intell Rev* 56, 10427–10464, doi:10.1007/s10462-023-10425-3.
- [34] Jun Liu, Yuan Tian, Yu Zhou, Yang Xiao, Nirwan Ansari, (2020) "Privacy preserving distributed data mining based on secure multi-party computation", *Computer Communications*, Volume 153, Pages 208-216, ISSN 0140-3664, doi:10.1016/j.comcom.2020.02.014.
- [35] FDjordje Slijepčević, Maximilian Henzl, Lukas Daniel Klausner, Tobias Dam, Peter Kieseberg, Matthias Zeppelzauer, (2021), "k-Anonymity in practice: How generalisation and suppression affect machine learning classifiers", *Computers & Security*, Vol. 111, 2021, 102488, doi: 10.1016/j.cose.2021.102488.
- [36] Zhong, S., Yang, Z., Wright, R. N. (2005, June), "Privacy-enhancing k-anonymization of customer data", *In Proceedings of the twenty-fourth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems* (pp. 139-147). ACM.
- [37] Kohlmayer, F., Prasser, F., Eckert, C., Kuhn, K. A. (2014), "A flexible approach to distributed data anonymization", *Journal of biomedical informatics*, Vol. 50, pp. 62-76.
- [38] Nergiz, M. E., Cicek, E., Pedersen, T., Saygin, Y. (2012), "A look-ahead approach to secure multiparty protocols", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 24, No. 7, pp. 1170-1185.
- [39] Sakthivel, S. and Vinotha, N. (2023), "An Intellectual Optimization of K-anonymity Model for Efficient Privacy Preservation in Cloud Platform". *Journal of Intelligent & Fuzzy Systems*, Vol. 45, no. 1, pp. 1497-1512, doi: 10.3233/JIFS-223509.
- [40] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I. H. (2009), "The WEKA data mining software: an update", *ACM SIGKDD explorations newsletter*, Vol. 11, No. 1, pp. 10-18.
- [41] "UCI Machine learning repository: Adult Data Set", [Online]. Available: <http://archive.ics.uci.edu/ml/datasets/Adult>. [Accessed 5 September 2020]

Healthcare Application," 15th International Conference on Information and Knowledge Technology (IKT), Isfahan, Iran, Islamic Republic of, 2024, pp. 115-120, doi: 10.1109/IKT65497.2024.10892736.

- [50] K. Mohammadi and R. E. Atani, (2024), "Sigma: A Secure Federated Network Gaming Platform," 2024 15th International Conference on Information and Knowledge Technology (IKT), Isfahan, Iran, Islamic Republic of, pp. 222-227, doi: 10.1109/IKT65497.2024.10892800.

[۵۱] ابراهیمی آتانی، رضا، صادقی‌پور، مهدی (۱۳۹۵).

مروری بر روش‌های حفظ حریم خصوصی در انتشار داده‌ها، امنیت فضای تولید و تبادل اطلاعات (منادی)، ۵ (۲)، صص ۴۹-۶۲.

- [51] Sadeghpour M, Ebrahimi Atani R. (2017), "An overview of privacy preserving Data Publishing Techniques", *Biannual Journal Monadi for Cyberspace Security (AFTA)*, Vol. 5(2), pages: 49-62.



الیا س مسی بی مدرک کارشناسی و

کارشناسی ارشد خود را در رشته مهندسی کامپیوتر - نرم افزار از گروه مهندسی کامپیوتر دانشگاه گیلان به ترتیب در

سال‌های ۱۳۹۲ و ۱۳۹۵ دریافت کرد. ایشان در سمت مدیر تیم‌های توسعه در شرکت‌های حوزه فناوری اطلاعات و هوش مصنوعی مشغول به فعالیت و زمینه‌های پژوهشی ایشان در حوزه‌های یادگیری ماشین، مهندسی نرم افزار و هوش مصنوعی است.

نشانی رایانامه ایشان عبارت است از:

elyas.mosayebi@gmail.com



رضا ابراهیمی آتانی دانشیار گروه

مهندسی کامپیوتر دانشکده فنی و مهندسی دانشگاه گیلان هستند. ایشان مدرک دکترای خود را در سال ۱۳۸۹ در رشته مهندسی الکترونیک از دانشگاه علم

و صنعت دریافت کرد و در حال حاضر عضو پیوسته انجمن رمز ایران و انجمن‌های IACR و IEEE هستند. از ایشان تاکنون چهار عنوان کتاب و بیش از یکصد و پنجاه مقاله در مجلات و کنفرانس‌های ملی و بین‌المللی به چاپ رسیده است. زمینه‌های پژوهشی ایشان طراحی، پیاده‌سازی و تحلیل الگوریتم‌ها و پروتکل‌های رمزنگاری، امنیت و حفظ حریم خصوصی داده‌ها در سامانه‌های محاسباتی و طراحی و توسعه شبکه‌های کامپیوتری است.

نشانی رایانامه ایشان عبارت است از:

rebrahimi@guilan.ac.ir

