ارائۀ یک پروتکل جدید برای انتشار داده‌ها با حفظ حریم خصوصی در محیط‌های توزیع‌شدۀ مبتنی بر مدل‌های احتمالاتی

الیاس مسیبی[1]، رضا ابراهیمی آتانی[2*]

دانش‌آموختۀ کارشناسی‌ارشد گروه مهندسی کامپیوتر دانشگاه گیلان، رشت، ایران [1]

دانشیار گروه مهندسی کامپیوتر دانشگاه گیلان، رشت، ایران [2*]

چکیده

امروزه بسیاری از خدمات در بخش خصوصی و دولتی به‌صورت الکترونیکی ارائه می‌شوند. اطلاعات ایجادشده به‌وسیلۀ این خدمات به‌طورمعمول شامل رکوردهایی با اطلاعات حساس و خصوصی افرادی است که از این خدمات استفاده می‌کنند. سازمان‌های ارائه‌دهندۀ خدمات، موظف‌اند با ارائۀ خدمات امن و مورد اعتماد از نقض حریم خصوصی افراد جلوگیری کنند؛ از سوی دیگر اطلاعات ذخیره‌شده با حجم بالا منبع مناسبی برای کشف دانش به‌شمار 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. راه‌کارهای ارائه‌شده در انتشار با حفظ محرمانگی داده‌های توزیع‌شده 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 شرکت‌کنندگان در اشتراک داده، نبود شخص سوم قابل اعتماد، چالش‌های امنیتی پروتکل‌های محاسبات امن چندگانه و همچنین حملات درون‌سازمانی مشارکت‌کنندگان است. در این مقاله علاوه‌بر مرور روش‌های حفظ محرمانگی و حریم خصوصی در داده‌های توزیع‌شده و بررسی نقاط ضعف و قدرت آن‌ها، طرح جدیدی بر پایۀ مدل احتمالاتی برای حفظ حریم خصوصی در حوزۀ توزیع‌شده ارائه می‌شود. نتایج به‌دست‌آمده از ارزیابی امنیتی نشان می‌دهد روش پیشنهادی، بدون نیاز به شخص سوم مورد اعتماد و یا استفاده از محاسبات امن چندگانه، علاوه‌بر مقاومت در برابر حملات خارجی، در مقابل حملات داخلی که توسط مشارکت‌کنندگان پی‌ریزی می‌شود نیز مقاومت مناسبی دارد؛ به‌نحوی که در تمام حالات کمتر از یک درصد از رکوردهای منتشرشده نقض می‌شوند؛ همچنین نتایج حاصل از دقت طبقه‌بندی نشان می‌دهد، در این روش با توجه به اشتراک داده، دقت داده‌ها نسبت به حالتی که تنها یک فراهم‌کننده قصد انتشار داده را دارد افزایش می‌یابد؛ اما از سوی دیگر استفاده از روش پیشنهادی افزایش سربار پردازشی را به‌دنبال دارد؛ به‌طوری که این روش در تمامی حالات نسبت به روش پایه کندتر عمل می‌کند.

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

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

Elyas Mosayebi[1], Reza Ebrahimi Atani[2*]

M.Sc. in Computer Engineering, University of Guilan, Rasht, Iran[1]

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

* نویسندۀ عهده‌دار مکاتبات

* Corresponding author

• تاریخ ارسال مقاله: ۱۴۰۴/۱/۱۹ • تاریخ پذیرش: ۱۴۰۴/۴/۳۰ • تاریخ انتشار: ۱۴۰۴/۹/۲۸ • نوع مطالعه: پژوهشی

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 prose 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 کلیهٔ افراد جامعه هر روز پررنگ‌تر می‌شود؛ تا جایی که کمابیش بیشتر خدمات بخش عمومی که در طول روز از آن بهره گرفته می‌شود، به‌وسیلهٔ سرویس‌های دولت یا بخش خصوصی به‌صورت الکترونیکی به شهروندان عرضه می‌شود؛ از همین رو بسیاری از شرکت‌ها و سازمان‌های ارائه‌دهندهٔ این سرویس‌ها، اطلاعات مربوط به خدمات مشتریان خود را ذخیره می‌کنند. ناگزیر، این اطلاعات می‌تواند شامل داده‌هایی باشد که محتوای دادهٔ بخشی از حریم خصوصی مشتریان باشد؛ به همین دلیل پاسداری و صیانت از این اطلاعات بسیار حائز اهمیت است. از سوی دیگر، اطلاعاتی که به‌مرور زمان جمع‌آوری می‌شوند؛ علاوه‌بر اینکه حکم تاریخچه را دارند، در صورت انتشار و یا برون‌سپاری می‌تواند منبع بسیار مناسبی برای کشف دانش [۱] به‌حساب آید. بدیهی است که اگر داده‌ها بدون هیچ‌گونه پردازشی انتشار داده شوند، حریم خصوصی مشتریان نقض می‌شود. انتشار با حفظ حریم خصوصی داده [۲و۵۱] (PPDP)[1]، راه‌کاری برای حل این مشکل است[۱]. در روش‌های PPDP سعی می‌شود داده‌ها به‌نحوی انتشار یابند که حریم خصوصی افراد نقض نشود.

اطلاعاتی که در قالب رکورد به‌وسیلهٔ شرکت‌هایی مانند بیمه، ثبت احوال، بیمارستان، بانک و غیره جمع‌آوری می‌شوند حاوی اطلاعات خصوصی و حساس آحاد جامعه است. انتشار این داده‌ها باید تضمین کند که طی مراحل کشف دانش و یا هرگونه بررسی دیگر، اطلاعات حساس افراد فاش نشود؛ زیرا بر اساس اعلامیهٔ جهانی حقوق بشر نباید به حریم خصوصی افراد خدشه‌ای وارد شود[۲]. مشکلات امنیتی که برای شرکت‌های Netflix و AOL اتفاق افتاد، مثالی از نقض حریم خصوصی در انتشار داده‌ها هستند[۴,۳].

همان‌طور که در بخش قبل بیان شد PPDP راه‌کاری برای حل این چالش است. دو رکن مهم در به‌کارگیری PPDP، حفظ حریم خصوصی و سودمندی داده‌ها است[۵]؛ به‌نحوی که با برقراری امنیت حریم خصوصی افراد، نتایج حاصل از کشف دانش نیز ارزش قابل قبولی داشته باشند.

به‌طورمعمول در دنیای واقعی شرکت‌ها به‌دلایل مختلف برای مثال رسیدن به سودمندی بیشتر یا یک‌پارچه‌سازی داده‌ها، اقدام به انتشار اشتراکی آن‌ها می‌کنند؛ به‌عنوان یک مثال کاربردی می‌توان به چند شرکت بیمه که برای ساخت سامانهٔ کشف تقلب[۳]، داده‌های خود را به‌صورت اشتراکی انتشار می‌دهند، اشاره کرد. در این حالت چالش‌های جدیدی برای انتشار اشتراکی به‌وجود می‌آید. استفاده از روش‌های مرسوم PPDP به‌نحوی که برای حالت اشتراکی مناسب باشند یکی از این چالش‌هاست؛ از طرف دیگر ارتباط و هماهنگی شرکت‌کنندگان [۴] در اشتراک داده و همچنین محدودیت‌هایی که اشتراک داده برای شرکت‌کنندگان ایجاد می‌کند نیز از مسائل مهم این مبحث است. در بسیاری از پژوهش‌ها از شخص سوم مورد اعتماد [۵] برای عملیات گمنام‌سازی استفاده می‌شود. در مجموع برای حالتی که چندین شرکت یا سازمان قصد انتشار دادهٔ اشتراکی را دارند به‌دلیل ذات توزیعی انتشار، چالش‌های جدیدی برای عملیات انتشار به‌وجود می‌آید که حل آن‌ها هدف اصلی مقالهٔ پیشِ رو است.

در این مقاله مشکلات موجود در *انتشار با حفظ محرمانگی داده‌های توزیع‌شده* (به‌اختصار PPDDP)[2] مورد بررسی قرار می‌گیرد و سپس روش‌های موجود در این حوزه وارسی شده و به نقاط قوت و ضعف آن‌ها اشاره می‌شود.

[1] Privacy Preserving Data Publishing (PPDP)

[2] Privacy Preserving Distributed Data Publishing (PPDDP)

بسیاری از روش‌های ارائه‌شده با استفاده از شخص سوم مورداعتماد (TTP)[1] یا پروتکل‌های محاسبات امن چندگانه (SMC)[2] قابل‌اجرا هستند. استفاده از TTP و SMC باعث بروز محدودیت و مشکلاتی برای شرکت‌کنندگان در انتشار داده می‌شود. هدف اصلی این مقاله ارائهٔ 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[3] است که برای رفع محدودیت‌های یادگیری ماشین متمرکز در سال ۲۰۱۷ به‌وسیلهٔ گوگل ارائه شد[۵۰,۴۹,۴۵]. این مدل یادگیری غیرمتمرکز با آموزش مدل‌ها به‌صورت محلی روی داده‌های توزیع‌شده، حریم خصوصی و امنیت را بر اساس ساختار معماری خود ارتقا می‌دهد؛ بااین‌حال برای مقابله با مسائل امنیتی فزاینده، دولت‌ها مقرراتی برای حفاظت از داده‌ها ارائه کرده‌اند؛ از جمله مقررات عمومی حفاظت از داده(GDPR)[4] در اروپا، قانون حقوق حریم خصوصی کالیفرنیا (CPRA)[5] 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 توسطGDPR مبلغ ۵۷ میلیون دلار جریمه و این جریمه در مارس ۲۰۲۰ اعمال شد[۴۴].

یادگیری فدرال پاسخ مناسبی برای برون‌سپاری داده‌ها در محیط‌های توزیع شده‌است. 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 دادهٔ 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 گرفته‌است؛ بااین‌حال ساختار پیچیدهٔ این شیوهٔ یادگیری 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[6] و هم افقی[7] 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 توزیع‌شده است که مبتنی بر روش‌های احتمالاتی باشد. این روش بر پایهٔ قوانین احتمالاتی و با درنظرگرفتن اطلاعات آماری مجموعه‌دادهٔ 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 ادامهٔ مقاله، ساختار نگارشی آن بر این اساس ارائه می‌شود: در بخش دوم مشخصات و ویژگی‌های داده‌های قابل‌انتشار، مدل‌ها و روش‌های حفظ محرمانگی و حریم خصوصی مختصر معرفی می‌شوند. در ادامه و در بخش سوم، انتشار با حفظ محرمانگی در داده‌های توزیع‌شده مورد بررسی قرار خواهد گرفت و چالش‌ها و حملات ممکن در PPDDP برشمرده می‌شوند. در بخش چهارم پیشینهٔ این پژوهش ارائه می‌شود و راه‌کارهای کلی برای انتشار داده در حالت توزیعی مورد بررسی قرار می‌گیرد. در بخش پنجم، روشی برای انتشار داده‌های توزیع‌شده 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 توزیع‌شدهٔ داده‌ها

کمابیش تمام مؤسسات و شرکت‌ها و یا حتی دولت‌ها، اطلاعات دیجیتالی مربوط به حوزهٔ کاری خود را جمع‌آوری می‌کنند. این داده‌ها علاوه‌بر اینکه به‌عنوان تاریخچه به‌کار می‌آیند، برای عملیات کشف دانش نیز بسیار مناسب‌اند و دارندهٔ داده[8] 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 حوزهٔ داده‌کاوی در شرکت موردنظر و یا اشتراک داده به‌منظور

[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 مطمئن‌تر با شرکت‌های دیگر، دارندهٔ داده مجبور می‌شود مجموعه‌داده[1] 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 منتشرکنندهٔ داده به‌عنوان سازمان مورد اعتماد مالکان رکورد، باید تغییراتی روی مجموعه‌داده برای جلوگیری از نقض حریم خصوصی افراد اعمال کند و سپس آن را نشر دهد تا به‌دست دریافت‌کنندهٔ 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 به‌اختصار[2] PPDM گفته می‌شود[۸]. در این حالت امکان دارد دریافت‌کننده از یک عمل داده‌کاوی دیگر برای دسترسی به اطلاعات حساس استفاده کند[۵] یا مجموعه‌داده برای چندین دریافت‌کننده فرستاده شود. در این صورت تضمینی برای حفظ حریم خصوصی داده‌ها وجود ندارد.

به همین منظور راه‌کار دیگری برای انتشار مجموعه‌داده بیان شد که مجموعه‌داده به‌صورت کلی و در حالت عام برای دریافت‌کنندگان پردازش و گمنام‌سازی می‌شود. در این حالت داده برای هر عمل داده‌کاوی پردازش و منتشر می‌شود[۵]. PPDP به‌طور کلی گمنامی داده‌ها را با مخفی‌کردن شناسه‌های مالک رکورد انجام می‌دهد؛ درحالی‌که PPDM مستقیم سعی در مخفی‌کردن 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 شده‌است. در مرحلهٔ نخست منتشرکنندهٔ داده که به‌طور معمول یک سازمان مستقل است، اطلاعات را از کاربران واقعی جمع‌آوری می‌کند. در این حالت فرض بر این است که منتشرکنندهٔ داده نسبت به اطلاعات حساس افراد مسئول است و به‌عبارتی برای صاحبان رکورد، شخصی مورد اعتماد به‌حساب می‌آید، اما کسانی که داده‌های انتشاریافته را در اختیار می‌گیرند مورد اعتماد نیستند و اطلاعات حساس و شخصی افراد باید از آن‌ها مخفی نگاه داشته شود؛ بنابراین در مرحلهٔ بعد بر اساس مدل‌های حفظ حریم خصوصی، عملیاتی بر روی داده‌ها اعمال می‌شود و درنهایت خروجی در اختیار دریافت‌کنندگان داده قرار می‌گیرد. گفتنی است که هرچه میزان حفظ حریم خصوصی بیشتر مدنظر باشد، ممکن است دقت[3] نتایج به‌دست‌آمده از داده‌ها پایین‌تر 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 مجموعه‌داده، رکوردهای مجموعه‌داده، صفات مربوط به هر رکورد و همچنین روش‌هایی که برای حفظ محرمانگی مورداستفاده قرار می‌گیرند، اشاره کرد. در ادامه به بررسی این ارکان می‌پردازیم.

انواع مجموعه‌داده

داده‌های ذخیره‌شده به‌وسیلهٔ 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 مجموعه‌داده[4] برای تمام رکوردها دارای خصوصیت‌های[5] یکسان است[۱]؛ برای مثال می‌توان به مجموعه اطلاعات ذخیره‌شده 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 مجموعه‌داده

به‌طورکلی صفات یا شناسه‌های مجموعه‌دادهٔ D را می‌توان در چهار دستهٔ زیر طبقه‌بندی کرد[۵]:

شناسهٔ صریح[6]: مجموعه صفاتی که به‌صراحت برای شناسایی مالک رکورد مورداستفاده قرار می‌گیرند؛ مانند کد ملی اشخاص.

شبه شناسه[7]: ترکیبی از صفات که هر یک به‌تنهایی برای شناسایی صاحب رکورد کافی نیستند؛ اما با به‌کارگیری ترکیبی از آن‌ها می‌توان مالک رکورد را تشخیص داد. مانند ترکیب کدپستی و تاریخ تولد و جنسیت.

صفات حساس[8]: صفاتی که صاحب رکورد مایل نیست عموم از آن‌ها اطلاع داشته باشند، مانند: نوع بیماری و میزان حقوق. این صفات برای داده‌کاوی و تجزیه‌وتحلیل آماری مورد استفاده قرار می‌گیرند.

صفات غیرحساس[9]: صفاتی که در هیچ‌کدام از دسته‌های پیشین جای نمی‌گیرند و در عملیات پردازش داده برای حفظ محرمانگی مورد استفاده نیستند.

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

[1] Data Set
[2] Privacy Preserving Data Mining
[3] Accuracy
[4] Data Set
[5] Attributes
[6] Explicit Identifier
[7] Quasi Identifier
[8] Sensitive Attributes
[9] Non-Sensitive Attributes

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 تغییردادن داده‌ها به‌نحوی است که ارتباط بین اطلاعات حساس و افراد واقعی مشخص نباشد، اما به‌طورکلی می‌توان این روش‌ها را در چهار دسته طبقه‌بندی کرد[۵].

روش تصادفی

در روش تصادفی[1] با اضافه‌کردن نوفه به دادهٔ اصلی از فاش‌شدن اطلاعات حساس جلوگیری می‌شود. این روش بیشتر برای مخفی‌کردن داده‌های عددی[2] مورداستفاده قرار می‌گیرد. مشکل این نوع روش تقریبی‌بودن آن و همچنین ازدست‌دادن داده‌های مفید به مقدار زیاد است[۵]. دقت و صحت نتیجه در این روش به میزان بزرگی تابعی که به‌عنوان نویز به دادهٔ اصلی اضافه می‌شود بستگی دارد[۱۱].

روش جایگزینی

در روش جایگزینی[3] داده، ویژگی‌های حساس با مقداری دیگر جایگزین می‌شوند به‌نحوی که رابطهٔ آماری بین داده‌ها تغییری نکند[۱۲]؛ در این حالت در سطح رکورد نمی‌توان به داده‌ها اطمینان کرد؛ یعنی در سطح مجموعه‌داده اطلاعات درستی به‌دست می‌آید؛ اما رکوردها ممکن است مربوط به افراد حقیقی نباشند. از این روش می‌توان برای ویژگی‌های عددی و طبقه‌بندی‌شده[4] استفاده کرد[۱].

روش رمزنگاری

روش رمزنگاری[5] برای داده‌های توزیع‌شده در محاسبات امن چندگانه (SMC) مورداستفاده 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[6] رایج‌ترین روش برای حفظ حریم خصوصی در انتشار داده به‌شمار می‌آید. ایدهٔ اصلی روش‌های مبتنی بر این روش، همسان‌کردن رکوردهای مجموعه‌داده است؛ به‌نحوی که رکورد موردنظر برای شخص مهاجم[7] به‌راحتی قابل‌تشخیص نباشد[۱]؛ به بیان دیگر، این روش رکوردها را در گروه‌های هم‌ارزی[8] 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 هم ارزی استفاده می‌شوند عمل‌گرهای گمنامی نام دارند. عمل‌گرهای گمنامی مختلفی برای گمنام‌سازی در مجموعه‌داده‌ها استفاده می‌شود، اما پرکاربردترین آن‌ها، عمومی‌سازی[9] و فرونشانی[10] 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 مرحلهٔ بعد می‌توان ماه تولد را با سال تولد جایگزین کرد. عمومی‌سازی هر صفت به کمک درخت طبقه‌بندی[11] آن انجام می‌گیرد. تشکیل درخت طبقه‌بندی برای هر صفت توسط یک شخص خبره صورت می‌گیرد که دانش کاملی در حوزهٔ 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 مستقیم از برگ به ریشهٔ درخت می‌نشاند و با داشتن ریشهٔ هر صفت که (null،؟ یا هر چیزی) می‌تواند باشد، این عمل‌گر قابل‌استفاده است، اما باید توجه داشت که استفادهٔ نادرست از این عمل‌گر موجب کاهش شدید کیفیت مجموعه‌داده 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 مجموعهٔ منتشرشده، نباید تفاوتی داشته باشد؛ به بیان ساده‌تر نباید دانش جدیدی بر دانسته‌های فرد مهاجم پس از رؤیت مجموعه‌دادهٔ منتشرشده اضافه شود، اما برآورده‌کردن این سطح از امنیت غیرممکن است. با درنظرگرفتن اطلاعاتی که ممکن است از قبل در اختیار مهاجم باشد این سطح از حریم خصوصی نقض خواهد شد (دانش پیش‌زمینهٔ مهاجم) [۱]؛ به همین دلیل در بیشتر کارهایی که در حوزهٔ محرمانگی داده انجام می‌گیرد از مدل‌های تعدیل‌یافته‌تری استفاده می‌شود که هرکدام می‌توانند در مقابل دسته‌ای از حملات از داده‌ها حفاظت کنند. این مدل‌ها به دو دستهٔ اصلی مدل‌های پیونددهی[12] و مدل‌های احتمالاتی[13] تقسیم می‌شوند[۱۶،۱۰].

[1] Randomization Method
[2] Numerical Attributes
[3] Data Swapping Method
[4] Categorical Attributes
[5] Cryptographic Method
[6] Anonymization Method
[7] Adversary

[8] Equivalence Group
[9] Generalization
[10] Suppression
[11] Taxonomy Tree
[12] Linkage Model
[13] Probabilistic Model

مدل پیونددهی

مدل پیونددهی یا مدل نحوی[1] [۱۷] برای مقابله با حملهٔ پیونددهی[2] طراحی شده‌است. در حملهٔ پیونددهی مهاجم با بهره‌گیری از مجموعه‌دادهٔ منتشرشده و همچنین صفت، رکورد یا جدولی که از اطلاعات دانش عمومی یا پیش‌زمینهٔ[3] 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 به‌صورت مختصر هرکدام را توضیح می‌دهیم.

الف. مدل پیونددهی رکورد: در حملات پیونددهی رکورد مهاجم شبه شناسه‌های قربانی را با رکوردهای انتشاریافته 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 به‌طور 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 درجهٔ اطمینان 1/k می‌تواند رکورد قربانی را شناسایی کند [۱۸].

ب. مدل پیونددهی صفت: در حملهٔ پیونددهی صفت به‌دست‌آوردن 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 دستهٔ 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 آورده‌است. یکی از روش‌های مورداستفاده برای مقابله با حملهٔ پیونددهی صفت، روش l-diversity است[۱۹].

ج. مدل پیونددهی جدول: در مدل پیونددهی صفت و پیونددهی رکود، فرض بر این بود که مهاجم از وجود رکود قربانی در مجموعه‌دادهٔ منتشرشده اطمینان دارد. اما در مدل پیونددهی جدول این فرض برقرار نیست و اطمینان از وجود یا نبود رکورد قربانی در مجموعه‌داده توسط مهاجم، خود به‌عنوان نقض محرمانگی تلقی می‌شود؛ به‌عبارت‌دیگر مهاجم با دانستن اینکه رکورد موردنظرش در مجموعه‌دادهٔ منتشرشده وجود دارد یا خیر به هدف خود رسیده‌است[۱]. روش δ-Presence برای مقابله با این نوع حمله ارائه شده‌است. در روش δ-Presence مهاجم کمتر از δ٪ از وجود رکورد قربانی در جدول منتشرشده اطمینان دارد، احتمال اینکه بتواند پیونددهی رکورد یا صفت را با موفقیت انجام دهد حداکثر δ٪ است؛ بنابراین این روش برای مقابله با حملات پیونددهی رکورد و صفت نیز کارایی دارد. در محاسبهٔ احتمال δ فرض می‌شود منتشرکنندهٔ داده و مهاجم به یک جدول خارجی مشترک دسترسی دارند. عیب عمدهٔ روش δ-Presence این است که چنین فرضی در دنیای واقعی زیاد محتمل نیست[۱].

- **مدل احتمالاتی**

در مدل‌های احتمالاتی یا معنایی[4] سعی می‌شود تغییر اطمینان مهاجم قبل از مشاهدهٔ دادهٔ منتشرشده 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 دراختیارگرفتن مجموعه‌داده این اطمینان نباید بیشتر شود؛ برای مثال اگر مهاجم در ابتدا با اطمینان ۱۰درصد صفت حساس قربانی را بداند و پس از دراختیارگرفتن جدول منتشرشده اطمینان وی به ۸۰درصد افزایش یابد این مدل نقض شده و امنیت مجموعه‌دادهٔ منتشرشده زیر سؤال می‌رود. حریم خصوصی تفاضلی معروف‌ترین روشی است که بر اساس مدل احتمالاتی ارائه شده‌است[۱]. این مدل بیان می‌کند که حذف یا اضافه‌کردن یک رکورد، نباید تأثیر چندانی بر نتایج به‌دست‌آمده از تجزیه‌وتحلیل مجموعه‌داده 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 توزیع‌شده

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؛ به‌عنوان‌مثال کاربردی می‌توان به همکاری دولت‌ها[5] (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 استفادهٔ سو از داده‌های حساس خود هستند. در حالت انتشار داده به‌صورت اشتراکی نیز باید مجموعه‌دادهٔ نهایی از لحاظ حریم خصوصی افراد ایمن باشد.

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

[1] Syntactic
[2] Linkage Attack
[3] Background Knowledge
[4] Semantic
[5] Open Government Partnership

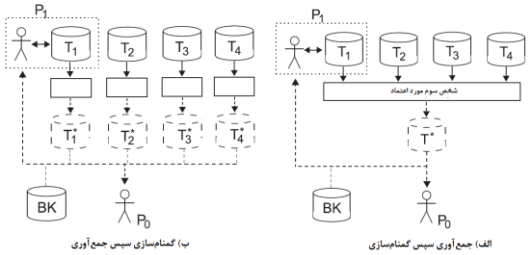
روش‌های مختلفی برای انتشار با حفظ محرمانگی داده‌های یک سازمان ارائه شده‌است، اما در برخی شرایط نیاز به انتشار داده با تشریک مساعی بین چند سازمان به‌وجود می‌آید که در این حالت علاوه‌بر تهدیدهایی که در شرایط عادی متوجه حریم خصوصی افراد است، حمله‌های جدیدی به‌دلیل توزیعی‌بودن داده‌ها به‌وجود می‌آید؛ برای مثال در حوزهٔ داده‌های بیمارستانی، چند بیمارستان اقدام به انتشار یا به‌اشتراک‌گذاری 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 مجموعه‌دادهٔ گمنام‌شدهٔ نهایی را تشکیل می‌دهند. در این حالت ممکن است بیمارستانی با دراختیارداشتن مجموعهٔ نهایی و داده‌های مربوط به خود، اقدام به نقض حریم خصوصی داده‌های بیمارستانی دیگر کند. این شرایط حادتر خواهد شد، اگر چند بیمارستان برای این کار با هم تبانی کنند. در این حالت از روش‌های حفظ محرمانگی در داده‌های توزیع‌شده استفاده می‌شود.

برای برقراری حفظ حریم خصوصی در انتشار داده‌ها به‌صورت توزیع‌شده بر حسب‌معمول از دو سیاست کلی استفاده می‌شود:

- **سیاست‌های حفظ حریم خصوصی در حالت توزیعی**

به‌طورکلی 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 به‌صورت کلی نشان می‌دهد. «ابتدا جمع‌آوری و بعد گمنام‌سازی[1]» روش اولی است که در برخی از کارها مورداستفاده قرار می‌گیرد. این روش زمانی بدون مشکل امنیتی خواهد بود که تمام منتشرکننده‌ها به هم اعتماد داشته باشند؛ زیرا اطلاعات حساس بلافاصله بعد از مرحلهٔ 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 دادهٔ T_1 تا T_4 متعلق به منتشرکننده‌های P_1 تا P_4 هستند. ابتدا داده‌ها توسط شخص سوم مورداعتماد که می‌تواند یکی از منتشرکننده‌ها باشد جمع‌آوری شده و سپس گمنام‌سازی می‌شود و نتیجهٔ پردازش آماری یا داده‌کاوی در اختیار تمام منتشرکننده‌ها یا هر شخص دیگری مانند P_0 قرار می‌گیرد، اما پیداکردن شخص سومی که قابل اعتماد همهٔ فراهم‌کنندگان باشد به‌طورمعمول امکان‌پذیر نیست[۱۷]. به همین دلیل این روش کمتر مورد استفاده قرار می‌گیرد.

در روش دوم که «ابتدا گمنام‌سازی سپس جمع‌آوری[2]» نام دارد، ابتدا هرکدام از منتشرکننده‌ها، مجموعه‌دادهٔ متناظر با خود را با استفاده از روش‌های گمنام‌سازی بیان‌شده در بخش پیشین، گمنام کرده و پس از آن مجموعه‌های گمنام‌شده با هم ادغام می‌شود و مورد عملیات داده‌کاوی یا تجزیه‌وتحلیل قرار می‌گیرند[۱]. این روش در شکل (۱) بخش (ب) نمایش داده شده‌است. 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 مجموعه‌داده‌ها یکسان نباشد نتیجهٔ به‌دست‌آمده از تجزیه‌وتحلیل آماری یا داده‌کاوی به‌شدت تضعیف می‌شود[۲۲]؛ همچنین در این حالت اگر از روش‌های چندپردازشی امن (SMC) برای ارتباط فراهم‌کنندگان در طول زمان گمنام‌سازی استفاده نشود، امنیت داده‌ها به‌دلیل حملاتی که ممکن است توسط هر یک از فراهم‌کنندگان به مجموعهٔ نهایی تدارک دیده شود، نقض شود[۱۷].



(شکل -۱): سیاست‌های حفظ حریم خصوصی داده‌های توزیع‌شده[۲۲]

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

- **سناریوهای حمله در داده‌های توزیع‌شده**

به‌طورکلی 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 مجموعه‌دادهٔ منتشرشده صورت می‌گیرد: در این حالت دریافت‌کنندهٔ داده با استفاده از دانش پیش‌زمینهٔ خود و به‌کارگیری حملاتی مانند حملهٔ پیونددهی، اقدام به نقض حریم خصوصی در رکوردهای مجموعه‌دادهٔ منتشرشده می‌کند؛ برای مثال فرض کنیم P_0 با استفاده از مجموعه‌دادهٔ منتشرشده و دانش پیش‌زمینهٔ مربوط به خود اقدام به حملهٔ پیونددهی 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 توزیع‌شده طراحی شده‌اند مانند k-anonymity، l-diversity، (n,t)-Closeness، δ-Presence و غیره استفاده کرد[۲۲].

حملهٔ منتشرکنندگان داده‌ها به مجموعه‌های یکدیگر با استفاده از داده‌های تشکیل‌شدهٔ میانی: در این حالت

[1] Integrate-then-Anonymize

[2] Anonymize-then-Integrate

فرض می‌شود که منتشرکننده‌های داده نسبت به هم، نیمه‌صادق‌اند[1] (صادق اما کنجکاو) و با استفاده از SMC داده‌های میانی را در اختیار هم قرار می‌دهند. در این حالت منتشرکنندهٔ مهاجم، با استفاده از داده‌های میانی و همچنین مجموعهٔ متعلق به خود می‌تواند به اطلاعات حساس در مورد مجموعهٔ منتشرکنندگان دیگر دسترسی پیدا کند. این مشکل به‌طورمعمول به‌دلیل ضعف پروتکل‌های SMC به‌وجود می‌آید.

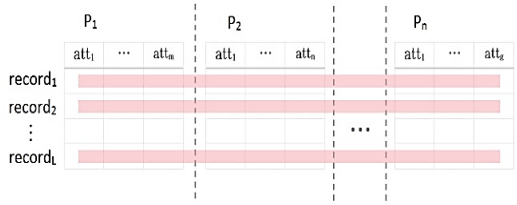
حملهٔ منتشرکنندگان داده‌ها به مجموعه‌های یکدیگر با استفاده از مجموعه‌دادهٔ منتشرشدهٔ نهایی و مجموعه‌دادهٔ متعلق به هر منتشرکننده: هرکدام از منتشرکننده‌ها مانند P_1 به مجموعهٔ نهایی T^* دسترسی دارد و با به‌کارگیری اطلاعات مجموعهٔ متعلق به خود (T_1) می‌تواند این نوع حمله را سازمان‌دهی کند. در این حالت مهاجم P_1 چیزی بیشتر از مهاجم P_0 در اختیار دارد. حال اگر چند منتشرکنندهٔ دیگر مانند P_2 و P_3 در این نوع حمله برای به‌دست‌آوردن اطلاعات حساس P_4 با P_1 تبانی کنند این حمله شدیدتر خواهد شد. این حملات به حملات داخلی[2] معروف‌اند[۲۲].

برای حل مشکل حملات بیان‌شده در قسمت پیش، برخی از روش‌ها با فرض عمودی (رکوردها یکی هستند، اما صفت‌ها متفاوت) یا افقی‌بودن (صفت‌ها 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 دارندهٔ مجموعه‌داده، بیشتر اوقات تمام داده‌ها را در اختیار ندارد و نیاز دارد تا با دیگر منتشرکنندگان تشریک‌مساعی کند؛ برای مثال بانک‌های عضو شبکهٔ 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 به‌دلیل 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، صفت‌های 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 دارندهٔ داده مانند P_1، P_2 تا P_k به‌ترتیب دارای مجموعه صفات $\{att_1,\ldots,att_m\}$، $\{att_1,\ldots,att_n\}$ و $\{att_1,\ldots,att_g\}$ هستند؛ از طرف دیگر این مجموعه‌ها رکوردهای هم‌سان دارند و ارتباط بین مجموعه‌ها روی یک رکورد خاص با شناسهٔ خاص برای مثال att_1 که در تمام مجموعه‌ها تکرار شده‌است، برقرار می‌شود[۱]. دلیل اصلی اشتراک داده بین چند منتشرکننده در توزیع عمودی، کامل‌ترشدن مجموعه صفات و بالارفتن دقت عمل داده‌کاوی یا تجزیه‌وتحلیل آماری است. به‌طورمعمول هیچ‌کدام از شرکت‌کننده‌ها در اشتراک داده به‌صورت توزیع عمودی، نمی‌توانند به‌تنهایی و بدون داشتن مجموعهٔ دیگر اعضا اقدام به عمل داده‌کاوی یا انتشار داده کنند. نمونهٔ 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].

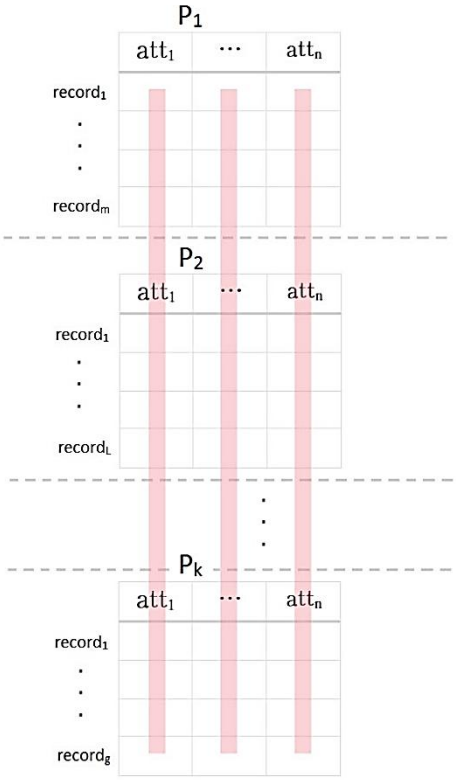


(شکل -۲): توزیع عمودی داده‌ها
(Figure-2): Vertical Data Distribution

توزیع افقی: در توزیع افقی، صفت‌ها 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 دادهٔ P_1، P_2 تا P_k دارای صفات یکسان $\{att_1,\ldots,att_n\}$ 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 بالابردن 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 ایالات‌متحده آمریکا از قوانین HIPAA[3] برای اشتراک داده‌های گمنام‌شدهٔ خود استفاده می‌کنند[۲۴].

[1] Semi-Honest
[2] Insider Attack
[3] Health Insurance Portability and Accountability Act



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

۳- پیشینهٔ پژوهش

در این بخش پژوهش‌هایی که در حوزهٔ انتشار با حفظ محرمانگی و حریم خصوصی داده‌های توزیع‌شده انجام شده‌است، مورد بررسی قرار می‌گیرد. تمام روش‌های ارائه‌شده، بر پایهٔ روش‌های حفظ محرمانگی در حالت غیر توزیعی بناشده‌اند و به نحوی در این حالت روش‌های PPDP به‌همراه یکی از دو سیاست حفظ محرمانگی در حالت توزیعی که در بخش پیشین بیان شد، مورداستفاده قرار می‌گیرند.

• انتشار با حفظ حریم خصوصی در حالت غیر توزیع‌شده

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

روش‌هایی نیز بر پایهٔ مدل احتمالاتی ارائه شده که مهم‌ترین آن‌ها حریم خصوصی تفاضلی است. ایدهٔ این مفهوم ابتدا در [۳۱] معرفی شد و سپس در [۲۰] به حریم خصوصی تفاضلی تبدیل شد. این مدل بیان می‌کند که حذف یا اضافه‌کردن یک رکورد، نباید تأثیر چندانی بر نتایج به‌دست‌آمده از تجزیه‌وتحلیل مجموعه‌داده 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؛ همچنین به بیان ریاضی اگر بررسی‌های مهاجم به‌وسیلهٔ تابع تصادفی F مدل شود، برای هر مجموعه‌دادهٔ T_1 و T_2 که در وجود یک رکورد اختلاف دارند، باید رابطهٔ (۱) برقرار باشد:

$$\forall S \in Range(F)\left(\left|\ln\frac{P[F(T_1)=S]}{P[F(T_2)=S]}\right| \le \epsilon\right) \qquad (۱)$$

که در آن S صفت حساس و ϵ مقداری است که هرچه کوچک‌تر انتخاب شود، میزان تأثیر یک رکورد بر روی نتایج تجزیه‌وتحلیل را کاهش می‌دهد. برای برآورده‌کردن رابطهٔ مدل حریم خصوصی تفاضلی، نوفهٔ زیادی به داده‌های مجموعه اضافه می‌شود؛ زیرا این مدل تمایزی بین صفات حساس و دیگر صفت‌ها قائل نمی‌شود. از طرف دیگر این روش تضمینی برای محافظت در برابر حملات پیونددهی ندارد[۲۵و۴۸و۵۱].

در [۶] روش تعدیل‌یافته‌تری نسبت به حریم خصوصی تفاضلی و بر پایهٔ مدل احتمالاتی ارائه شده‌است. این روش به نحوی اطمینان مهاجم برای درک صفت حساس یک رکورد بعد از دیدن مجموعه‌داده را مدل کرده و آن را محدود می‌کند. روش ارائه‌شده در [۶] به‌عنوان مدل حفظ حریم خصوصی در مقالهٔ پیش‌رو استفاده شده‌است؛ به همین دلیل در ادامه به بررسی کامل این روش می‌پردازیم و آن را به‌اختصار روش GF_PPDP می‌نامیم. روش GF_PPDP بر مبنای دو مفهوم اطمینان مورد انتظار[1] و اطمینان مشاهده‌شده[2] بنا شده‌است. مهاجم پس از مشاهدهٔ مجموعه‌دادهٔ منتشرشده، با جست‌وجوی شبه‌شناسه‌های قربانی، سعی می‌کند صفت حساس وی را به‌دست 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 مشاهده‌شدهٔ قربانی می‌نامند؛ به بیان دیگر اطمینان مشاهده‌شده بیان می‌کند که مهاجم با چه احتمالی می‌تواند صفت حساس یک رکورد را به قربانی نسبت دهد؛ به‌دلیل این‌که در این روش در ابتدا از رکوردها با نرخ β نمونه‌برداری می‌شود، باید احتمال گزینش رکورد قربانی در مرحلهٔ نمونه‌برداری نیز مدنظر قرار بگیرد. همان‌طور که رابطهٔ (۲) نشان می‌دهد از ضرب این دو احتمال، اطمینان مشاهده‌شدهٔ رکورد r به‌دست می‌آید.

$$ObservedConfidence(r) = \beta \times \frac{\#S(r)}{|E(r)|} \qquad (۲)$$

که $|E(r)|$ اندازهٔ گروه هم‌ارزی و $\#S(r)$ تعداد تکرار صفت حساس رکورد r در گروه هم‌ارزی است؛ هرچه اطمینان مشاهده‌شده که به‌وسیلهٔ 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

کند و درنتیجه سطح محرمانگی مجموعه‌داده بالاتر می‌رود؛ بنابراین باید مقدار آن را به سقف معینی محدود کرد. بر اساس مدل ارزیابی ارائه‌شده در GF_PPDP، اطمینان مشاهده‌شدهٔ یک رکورد نباید از احتمال دیده‌شدن آن رکورد در مجموعه‌داده‌ای با اندازهٔ یکسان که به‌صورت تصادفی از فضای نمونهٔ 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 رابطهٔ (۳) محاسبه می‌شود:

$$ExpectedConfidence(r) = 1 - f(0; n, \Pr(r)) = 1 - (1 - \Pr(r))^n \quad (۳)$$

که Pr(r) با فرض مستقل‌بودن صفات از یکدیگر به‌وسیلهٔ ضرب احتمال شبه‌شناسه‌ها در احتمال صفت حساس به‌دست می‌آید:

$$\Pr(r) = (\prod_i \Pr(q_i)) \times \Pr(s) \quad (۴)$$

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

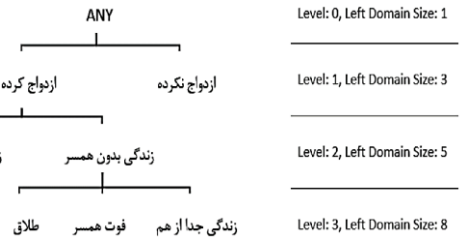
$$\text{ObservedConfidence}(r) \leq \text{ExpectedConfidence}(r) \quad (۵)$$

الگوریتم ارائه‌شده در روش GF_PPDP به‌ازای هر رکورد رابطهٔ (۵) را تشکیل داده و اگر این رابطه برای رکورد یادشده برقرار نباشد، با استفاده از معیاری به نام LeftDomainSize(LDF) یک شبه‌شناسه را انتخاب کرده و با اِعمال عمل‌گرهای گمنام‌سازی، رابطهٔ (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 استفادهٔ 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 مجموعهٔ خروجی حذف می‌شوند. میزان تخریب برای رکورد r با شبه‌شناسه‌های q_1 تا q_d از رابطه (۶) به‌دست می‌آید.

$$\delta = \|r - \acute{r}\| = \frac{1}{d}\sum_{i=}^{d}\left(1 - \frac{CurrentLevel\ of\ \acute{q}_i}{Maximal\ level\ in\ q_i}\right) \quad (۶)$$

که در این رابطه r' معادل گمنام‌شدهٔ 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 ریشهٔ 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 فراهم‌کنندهٔ 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 مجموعهٔ دیگر فراهم‌کنندگان انجام دهد. در بیشتر روش‌های انتشار با حفظ حریم خصوصی در حالت غیر توزیع‌شده، با فرض سطح دانش پیش‌زمینهٔ مهاجم، روشی برای حفظ محرمانگی در مقابل حمله‌های خاص (برای مثال حملات پیونددهی) ارائه می‌دهند، اما در حالت توزیع‌شده، مدل‌کردن دانش پیش‌زمینهٔ مهاجم با درنظرگرفتن 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 حوزهٔ 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 محاسبهٔ امن چندگانه پروتکلی ارائه داده‌اند تا کلید نامزد پیونددهی مجموعه‌دادهٔ شرکت‌کنندگان به نحوی غیرقابل‌شناسایی گمنام شود. SMC در بسیاری از روش‌ها برای ارتباط امن بین فراهم‌کنندگان استفاده می‌شود. هدف اصلی SMC ارائهٔ 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

[1] Distortion

حریم خصوصی اقتضا می‌کند اطلاعی بیشتر از آن چیزی که باید دانسته شود فاش نشود؛ یعنی افراد شرکت‌کننده و همچنین کاربران تابع خروجی نباید چیزی به‌جز نتیجهٔ 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 فراهم‌کنندهٔ داده است. این روش برای مجموعه‌داده‌هایی که به‌صورت عمودی توزیع شده‌اند، مفهوم محدودیت‌های گمنام‌سازی برای پیونددهی[1] را مطرح کرد که به مجموعه‌ای از دوتایی صفت‌های شبه‌شناسه QID و مقدار K به‌شکل $\{<QID_1,k_1>,\ldots,<QID_p,k_p>\}$ گفته می‌شود. در طی مراحل گمنام‌سازی محدودیت‌های بالا باید ارضا شوند. برای گمنام‌سازی از روش TDS[2] استفاده می‌شود که برای هر صفت شبه‌شناسه از عمل‌گر ویژه‌سازی[3] برای رسیدن به k-anonymity استفاده می‌شود[۳۵]. در فرایند گمنامی Mashup هر فراهم‌کنندهٔ داده یا همان پارتی، یک رونوشت از مجموعهٔ Tg که بیان‌کنندهٔ مجموعهٔ نهایی در آن لحظه است و مجموعهٔ مربوط به خودش را دارد و در هر مرحله از اجرا، پارتی‌ها با هم همکاری می‌کنند تا یک عمل ویژه‌سازی را با توجه به محدودیت‌های تعریف‌شده و با استفاده از ردوبدل‌کردن اطلاعات انجام دهند. این روند تا رسیدن به مجموعهٔ نهایی انجام می‌شود.

برخی از کارها برای مجموعه‌هایی که به‌صورت افقی توزیع‌شده‌اند ارائه شده‌است؛ برای مثال در [۱۳] پروتکل غیرمتمرکزی برای اشتراک‌گذاری این نوع مجموعه‌داده‌ها ارائه شده‌است؛ ازآنجایی که فراهم‌کنندگان داده ترجیح می‌دهند تعلق رکوردهای منتشرشدهٔ 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 شبه‌شناسه‌هایی 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 مجموعه‌داده‌ها و گمنام‌سازی آن‌ها استفاده می‌کند. فراهم‌کنندگان داده با استفاده از یک پروتکل توزیع‌شده مجموعه‌دادهٔ گمنام‌شدهٔ مجازی[4] می‌سازند؛ به بیان دیگر مجموعه‌داده‌های محلی متعلق به فراهم‌کنندگان بدون تغییر باقی می‌مانند و با استفاده از عمل‌گر امن پیونددهی، تحت پروتکل امن، تجمیع می‌شوند و می‌توانند منتشر شوند و یا مورد پرس‌وجو قرار گیرند.

روش ارائه‌شده در [۳۶] برای حالتی مناسب است که چند فراهم‌کنندهٔ داده قصد دارند داده‌های خود را که به‌صورت افقی توزیع‌شده‌است در اختیار یک داده‌کاو[5] قرار دهند. این روش برای مقابله با حملهٔ پیونددهی صفت، پروتکلی بر پایهٔ روش گمنام‌سازی k-anonymity ارائه داده‌است.

روشی منعطف برای حفظ حریم خصوصی در مجموعه‌داده‌هایی که به‌صورت افقی یا عمودی توزیع‌شده‌اند 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 حملهٔ داخلی را به مجموعه‌دادهٔ 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 بررسی‌کردن محدودیت‌های وضع‌شده برای مجموعهٔ نهایی، آن را در مقابل مجموعهٔ تبانی فراهم‌کنندگان متخاصم (m-adversary)، ایمن می‌دارد. m-privacy با توجه به محدودیت C که C 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-dieversity)، تعریف می‌شود. بر طبق تعریف برای n فراهم‌کنندهٔ داده که مجموعه رکوردهای آن‌ها T را تشکیل می‌دهند، A سازوکار گمنام‌سازی، I مجموعه‌ای از m فراهم‌کنندهٔ متخاصم ($m \leq n\text{-}1$) و T_I مجموعهٔ متعلق به I، A(T) شرایط m-privacy را نسبت به C ارضا می‌کند اگر و فقط اگر به‌ازای هر فوق‌مجموعهٔ گمنام‌شده A(T′) از مجموعهٔ غیر متخاصم (متمم T_I) شرط C را ارضا کند. به بیان ریاضی داریم:

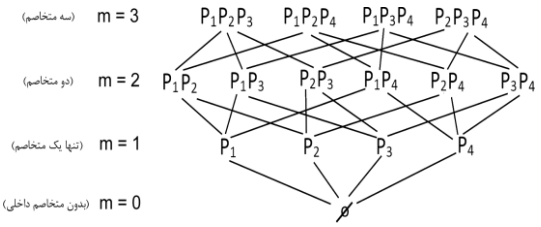
$$\forall I \subset P, |I| = m, \forall \acute{T} : T \backslash T_I \subseteq \acute{T} \subseteq T, C\left(A\left(\acute{T}\right)\right) = true \quad (۷)$$

در رابطهٔ بالا $T\backslash T_I$ بیان‌کنندهٔ متمم مجموعه T_I است. اگر n فراهم‌کنندهٔ $(P_1,P_2,\ldots,P_n)$ داشته باشیم، حالت‌های مختلف برای m فراهم‌کنندگان متخاصم، که m می‌تواند از «0» تا «n-1» 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»، هرکدام از چهار فراهم‌کننده

[1] Join Anonymity Requierment
[2] Top-Down Specialization
[3] Specialization
[4] Virtual Anonymized Database
[5] Data Miner

می‌توانند به‌عنوان فراهم‌کنندهٔ متخاصم در نظر گرفته شوند. برای رسیدن به امنیت در سطح m باید تمام حالت‌های کوچک‌تر و مساوی برابر با آن m به‌خصوص 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 $\binom{n}{n/2} = O(2^n n^{-1/2})$ است که بررسی‌کردن تمام حالت‌ها زمان اجرایی بالایی را خواستار است. البته در مقالهٔ 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 بررسی‌کردن تمام حالت‌ها توسط الگوریتم، اگر شرط C برای رکورد یا مجموعه‌ای از رکوردها برقرار نباشند از مجموعهٔ قابل‌انتشار 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 به‌وسیلهٔ TTP یا پروتکل SMC که در [۲۲] ارائه‌شده اجرا کرد.



(شکل -۵): حالت‌های مختلف برای m فراهم‌کنندهٔ 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 آن‌که 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 معرفی‌شده در این بخش را نشان می‌دهد. در ستون توضیحات ویژگی‌های هرکدام از روش‌ها بیان شده‌است. نکتهٔ مهمی که در جدول ذکر نشده، وابستگی تمام روش‌ها به پروتکل SMC و یا به‌کارگیری TTP است.

(جدول-۱): مقایسهٔ روش‌های حفظ محرمانگی در داده‌های توزیع‌شده

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

	مدل محرمانگی	نوع داده	مقابله با حملات داخلی	توضیحات
[۳۲]	پیونددهی	عمودی	خیر	استفاده از مدل محرمانگی k-anonymity
[۲۳]	پیونددهی	عمودی	خیر	به‌کارگیری k-anonimity و روش ویژه‌سازی
[۱۳]	پیونددهی	افقی	خیر	محرمانگی در سطح فراهم‌کنندگان و ارائهٔ 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 به‌کارگیری 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 مشاهدهٔ مجموعه‌دادهٔ انتشاریافته برای دیدن رکورد در مجموعه‌داده‌ای با اندازهٔ یکسان که به‌صورت تصادفی از فضای نمونه رکوردها ایجاد شده‌است (D_0)، قائل می‌شود؛ رابطهٔ (۳) و از طرف دیگر اطمینان مشاهده‌شدهٔ یک رکورد، بیان‌کنندهٔ این است که مهاجم پس از مشاهدهٔ مجموعهٔ منتشرشده با چه احتمالی می‌تواند صفت حساس آن رکورد را به قربانی نسبت دهد؛ رابطهٔ (۲)؛ همچنین بر طبق مدل GF_PPDP که در رابطهٔ (۵) بیان شده‌است، 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 مجموعه‌داده را دارد. در حالتی که چندین فراهم‌کننده در انتشار به‌صورت توزیع‌شده شرکت دارند روابط بیان‌شده در مدل GF_PPDP باید تغییر کند؛ زیرا در رابطهٔ (۳) باید احتمال کل مجموعه‌ها را در نظر گرفت و همچنین در رابطهٔ (۲) مقدار اطمینان مشاهده‌شده به‌ازای هر رکورد، از گروه‌های هم‌ارزی متعلق به هر فراهم‌کننده به‌دست می‌آید، که باید گروه‌های هم‌ارزی تمام مجموعه‌ها در نظر گرفته شود. بر اساس مدل GF_PPDP هر

چه احتمال صفت‌های شبه‌شناسه به دنیای واقعی نزدیک‌تر باشد و از مجموعهٔ بزرگ‌تری انتخاب شود احتمال خطرپذیری محرمانگی کاهش می‌یابد[۶]؛ بنابراین در روش پیشنهادی برای رابطهٔ (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 **مرحلهٔ نشست** است، فراهم‌کنندگان مقادیر احتمال صفت‌های شبه‌شناسه و صفت حساس و همچنین اندازهٔ مجموعه‌دادهٔ متعلق به خود را در اختیار هم قرار می‌دهند. این عمل حریم خصوصی هیچ‌یک از فراهم‌کنندگان را نقض نمی‌کند و فراهم‌کنندگان می‌توانند مجموعهٔ 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 محاسبهٔ اطمینان مشاهده‌شده و اطمینان مورد انتظار در مدل پیشنهادی این مقاله ارائه می‌شود؛ درواقع روش پیشنهادی برای محاسبهٔ اطمینان مورد انتظار، یک روش حریصانه[1] است که از بین چند روش حریصانهٔ دیگر انتخاب شد. معیار انتخاب، مقایسهٔ روش‌ها بر مبنای درصد تعداد رکوردهای نقض‌شده 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.

• محاسبهٔ 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 شده‌است؛ برای مثال رکوردها با نرخ نود درصد نمونه‌برداری می‌شوند تا مقدار n در رابطهٔ (۳) کاهش یابد و درنتیجه 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 هوشمندانه، باید کاربرد داده‌های انتشاریافته، 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 رابطهٔ (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 درعمل نمی‌توان از آن برای تغییر اطمینان مورد انتظار استفاده کرد، اما در این مقاله که انتشار برای حالت توزیع‌شده مدنظر است با توجه به چندمقداری‌بودن این متغیر می‌توان از آن برای کاهش اطمینان مورد انتظار استفاده کرد؛ به همین منظور در رابطهٔ (۳) مقدار Pr(r) کاهش داده می‌شود تا مقدار نهایی اطمینان مورد انتظار کاهش یابد.

با توجه به مرحلهٔ 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 محاسبهٔ رابطهٔ (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 درنظرگرفتن خصوصیات مجموعهٔ تمام فراهم‌کنندگان محاسبه می‌شود؛ برای مثال فرض می‌شود رکورد r با مشخصات {سن=23، جنسیت=زن، بیماری=سرطان}، متعلق به فراهم‌کنندهٔ P_1 از مجموعهٔ 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 مجموعهٔ 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 میانگین	۰.۱۸	۰.۴۸	۰.۰۲۱

با توجه به رابطهٔ (۴) و درنظرگرفتن کمترین مقدار بین احتمال‌ها Pr برابر با مقدار زیر می‌شود.

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

[1] Greedy

با فرض اینکه تعداد رکوردهای متعلق به مجموعهٔ P_1 برابر با پانصد باشد و جاگذاری Pr در رابطهٔ (۳) اطمینان مشاهده‌شده برای رکورد r به‌دست می‌آید.

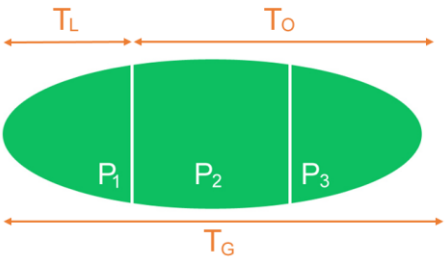
$$EC(r) = 1 - (1 - 0.00108)^{500} = 0.41$$

- **محاسبهٔ اطمینان مشاهده‌شده**

همان‌طور که در شکل (۶) مشخص است، اگر مجموعهٔ متعلق به تمام فراهم‌کنندگان T_G، مجموعهٔ متعلق به هر فراهم‌کننده T_L و مجموعهٔ فراهم‌کنندگان دیگر به‌غیر از فراهم‌کنندهٔ مورد نظر T_O نامیده شود و همچنین رابطهٔ (۲) با توجه به ثابت‌بودن β، به‌صورت ساده‌شده $OC_G = \frac{S_G}{E_G} = \frac{S_L+S_O}{E_L+E_O}$ درنظر گرفته شود، آنگاه مقادیر S_L و E_L از مجموعهٔ T_L در دسترس است (هر فراهم‌کننده به آمار مجموعهٔ متعلق به خود دسترسی دارد) و مقدار تقریبی E_O را نیز می‌توان از ضرب احتمال صفت‌های شبه‌شناسهٔ به‌دست‌آمده از مرحلهٔ نشست به‌صورت رابطهٔ (۸) محاسبه کرد:

$$E_O \cong \left| \left(\left(\prod_i \Pr(q_i) \right) \times N \right) - E_L \right| \qquad (۸)$$

در رابطهٔ بالا مقدار Pr از احتمال مربوط به تمام فراهم‌کنندگان به‌دست می‌آید و همچنین N تعداد کل رکوردهای متعلق به تمام فراهم‌کنندگان را مشخص می‌کند. در رابطهٔ ساده‌شدهٔ OC_G مقادیر E_L، E_O و S_L در دسترس‌اند، اما مقدار S_O برای فراهم‌کننده‌ای که در حال حاضر عملیات گمنام‌سازی را انجام می‌دهد قابل محاسبه نیست. S_O بیان‌کنندهٔ تعداد رکوردهایی از گروه‌های هم‌ارزی متعلق به مجموعهٔ T_O است که مقدار صفت حساس آن‌ها برابر با رکورد موردنظر است؛ بنابراین مقدار آن در دسترس نیست، اما می‌تواند مقادیر بین صفر تا E_O را اختیار کند. به‌ازای تمام مقادیر این بازه OC_G را محاسبه می‌کنیم و شرط رابطهٔ (۵) را برای آن‌ها بررسی می‌کنیم. با این کار تعداد حالاتی که رابطهٔ حریم خصوصی مدل گمنام‌سازی را ارضا می‌کنند را به‌دست آورده و اگر از حد آستانهٔ γ بیشتر بود، رکورد موردنظر را به‌عنوان رکورد قابل‌انتشار در مجموعهٔ نهایی قرار می‌دهیم. در این تعریف γ آستانه انتشار توزیعی نامیده می‌شود و می‌تواند توسط هر فراهم‌کننده مقداردهی شود.



(شکل -۶): رابطهٔ مجموعه‌های متعلق به هر فراهم‌کننده با مجموعهٔ فراهم‌کنندگان دیگر

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

برای مثال برای محاسبهٔ اطمینان مشاهده‌شدهٔ رکورد r با مشخصات {سن ۲۳، جنسیت=زن، بیماری=سرطان} مربوط به P_1 از جدول (۲)، فرض می‌کنیم تعداد رکوردهایی که با رکورد r هم‌ارز هستند در مجموعهٔ متعلق به فراهم‌کنندهٔ P_1 برابر با سی (E_L=40)، تعداد رکوردهایی از این گروه هم‌ارزی که مقدار صفت بیماری آن‌ها برابر با «سرطان» است نیز برابر ۱۷ باشد (S_L=17) باشد؛ همچنین فرض می‌کنیم مقدار N در رابطهٔ (۸) که تعداد رکوردهای مربوط به فراهم‌کنندگان P_1، P_2 و P_3 است، برابر با هزار و پانصد باشد. در این صورت مقدار E_O با استفاده از رابطهٔ (۸) به‌صورت زیر محاسبه می‌شود.

$$E_O = \lfloor (0.18 \times 0.48 \times 1500) - 40 \rfloor = 89$$

بنابراین S_O نیز می‌تواند مقادیر صفر تا ۸۹ را اختیار کند. به‌ازای تمام این مقادیر اطمینان مشاهده‌شده OC_G را محاسبه کرده و با توجه به شرط محرمانگی با مقدار اطمینان مورد انتظار به‌دست‌آمده (0.41) مقایسه می‌شود. نسبت تعداد حالاتی که شرط محرمانگی را ارضا می‌کنند به کل حالت‌ها، اگر از حد آستانهٔ انتشار توزیعی (γ) بیشتر بود، رکورد مورد نظر به‌عنوان یک رکورد مناسب به مجموعهٔ قابل انتشار اضافه می‌شود؛ در غیر این صورت عملیات گمنام‌سازی بر روی این رکورد انجام می‌گیرد.

اگر مجموعه احتمال شبه‌شناسه‌ها و صفت حساس متعلق به هر فراهم‌کننده را با Pr نمایش دهیم، در این صورت به تعداد فراهم‌کنندگان مجموعه احتمال وجود دارد $\{Pr_1, Pr_2, \ldots, Pr_n\}$. زیرنویس احتمال‌ها نشان‌دهندهٔ تعلق آن به فراهم‌کنندهٔ متناظر است؛ همچنین از دید هر فراهم‌کننده مجموعه احتمال متعلق به خودش با Pr_L و مجموعه احتمال مربوط به تمام فراهم‌کنندگان که از میانگین احتمالات به‌دست می‌آید، با Pr_G نشان داده می‌شود. در این مقاله پنج روش برای به‌دست‌آوردن اطمینان مورد انتظار با هم مقایسه شدند که شرح روش‌ها به‌صورت زیر است:

روش نخست: این روش روش اصلی مقاله است؛ به‌طوری که در رابطهٔ (۴) به‌ازای هر صفت شبه‌شناسه و صفت حساس از کوچک‌ترین احتمال استفاده شود $\text{Min}(Pr_1, Pr_2, \ldots Pr_n)$.

روش دوم: به‌ازای هر صفت شبه‌شناسه و صفت حساس تنها از احتمال کل و احتمال مربوط به همان فراهم‌کننده استفاده شود و با مقایسهٔ این دو احتمال، عدد کوچک‌تر برگزیده شود $\text{Min}(Pr_G, Pr_L)$.

روش سوم: استفاده از مجموعه احتمال کل Pr_G.

روش چهارم: استفاده از بزرگ‌ترین احتمال بین مجموعه احتمال‌های مربوط به فراهم‌کنندگان $\text{Max}(Pr_1, Pr_2, \ldots Pr_n)$.

روش پنجم: در این روش ابتدا به‌ازای هر فراهم‌کننده یک اطمینان مورد انتظار با استفاده از مجموعه احتمال مربوط به

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

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 رابطهٔ محرمانگی را نقض می‌کنند، به کل رکوردها، به‌عنوان معیار محرمانگی درنظر 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): مقایسهٔ روش‌های محاسبهٔ 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	PART	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	۸.۸۳۳۶	۸۴.۳۶	۸۳.۱۴	۸۱.۳۰	۸۳.۲۷
روش چهارم	۱۵.۹۳۵۸	۴۴.۱۵۹۸	۸۴.۵۵	۸۲.۳۸	۸۰.۲۳	۷۷.۵۰
روش پنجم	۴.۸۳۷۶	۲۵.۶۴۸۷	۸۴۶۸	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 حریصانهٔ برتر انتخاب کرد.

روند انتشار مجموعه‌داده

هماهنگی 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 هماهنگی لازم به‌وسیلهٔ 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 مرحلهٔ 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 دامنهٔ متفاوت در مقادیر شبه‌شناسه‌ها می‌شود[۳۹].
- هرکدام از فراهم‌کنندگان احتمال مربوط به صفات شبه‌شناسه و حساس در مجموعهٔ فرضی برابر با مجموعهٔ متعلق به خود (D_0 در تعریف اطمینان مورد انتظار) را محاسبه می‌کنند.
- احتمال‌های محاسبه‌شده در مرحلهٔ پیش بین فراهم‌کنندگان توزیع می‌شود (مرحلهٔ نشست). این کار به امنیت مجموعه‌ها خللی وارد نمی‌کند.
- فراهم‌کنندگان به‌صورت مستقل با استفاده از الگوریتم پیشنهادی اقدام به گمنام‌سازی مجموعهٔ متعلق به خود می‌کنند.
- در مرحلهٔ آخر مجموعه‌دادهٔ گمنام‌شده فراهم‌کنندگان، جمع‌آوری شده و به‌صورت یک مجموعهٔ 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 محاسبهٔ اطمینان مورد انتظار و اطمینان مشاهده‌شده از روش‌های بیان‌شده در بخش پیشین استفاده می‌شود.

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

در خط یک الگوریتم مقدار احتمال برابر با یک در نظر گرفته می‌شود؛ سپس در خطوط دو تا چهار به‌ازای تمام صفت‌های شبه‌شناسه و حساس، مقدار احتمالی که کمترین مقدار را از مجموعه احتمالات مرحلهٔ نشست دارد، انتخاب شده و در مقدار متغیر Probability ضرب می‌شود. این عمل به‌وسیلهٔ تابع getMinimumProbabilityFrom_N_privider انجام می‌شود و در آخر مقدار نهایی اطمینان مورد انتظار طبق رابطهٔ (۳) محاسبه و مقدار آن برمی‌گردد.

(شبه‌کد -۱): الگوریتم محاسبهٔ اطمینان مورد انتظار

(Pseudocodes-1) : Expected Confidence Calculation Algorithm

```
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_privider(atr, r)
 4: END FOR
 5: RETURN (1 – Probability) ^ n
```

(شبه‌کد-۲): الگوریتم محاسبهٔ اطمینان مشاهده‌شده

(Pseudocodes-2): Observed Confidence Calculation Algorithm

```
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 ← β * ((SL + i) / (EL + EO))
 13:  IF OCG < EC THEN
 14:  Satisfy ← Satisfy + 1
 15:  END IF
 16:  END FOR
 17:  IF (satisfy / (EO + 1)) < γ THEN
 18:  RETURN 1
 19:  END IF
 20: END IF
 21: RETURN β * (S_L / E_L)
```

الگوریتم محاسبهٔ اطمینان مشاهده‌شده در شبه‌کد (۲) آورده شده‌است؛ همان‌طور که در شبه‌کد مشخص است، ورودی‌های این الگوریتم به‌ترتیب از چپ به راست؛ رکورد r، گروه‌های هم‌ارز با رکورد موردنظر، مقدار محاسبه‌شده برای اطمینان موردانتظار، تعداد کل رکوردهای مربوط به تمام فراهم‌کنندگان و حد آستانهٔ انتشار توزیعی است؛ همچنین خروجی الگوریتم مقدار اطمینان مشاهده‌شده به‌ازای رکورد r است. تابع getSensitive در خط دو تعداد رکوردهای موجود در گروه هم‌ارزی، که صفت حساس آن‌ها برابر با رکورد r است را برمی‌گرداند. در خط چهار تا هشت تعداد گروه‌های هم‌ارزی متعلق به فراهم‌کنندگان دیگر محاسبه می‌شود. تابع nodeProbability احتمال صفت شبه‌شناسه‌ها QIDs را برمی‌گرداند.

در خطوط نُه تا شانزده به‌ازای تمام مقادیری که S_O می‌تواند اختیار کند، اطمینان مشاهده‌شده محاسبه می‌شود و تعداد حالاتی که شرط محرمانگی را ارضا می‌کنند در متغیر Satisfy قرار می‌گیرد. در خط هفده تا نوزده اگر نسبت حالت‌هایی که شرط محرمانگی را ارضا می‌کنند به تعداد کل حالات از حد آستانهٔ γ کمتر باشد مقدار «یک» برگردانده می‌شود تا شرط رابطهٔ (۵) در الگوریتم گمنام‌سازی GF_PPDP نقض شود و عمل عمومی‌سازی بر روی رکورد صورت گیرد؛ همچنین اگر تعداد گروه‌های هم‌ارزی مربوط به دیگر فراهم‌کنندگان صفر باشد، در خط ۲۱ مقدار محاسبه‌شدهٔ اطمینان مشاهده‌شده بر اساس رابطهٔ (۲) برگردانده می‌شود.

- **پیاده‌سازی**

برای پیاده‌سازی روش پیشنهادی و مقایسهٔ آن با روش‌های دیگر از زبان جاوا که یکی از زبان‌های شناخته‌شده و محبوب است، استفاده شد. برای عملیات مربوط به مجموعه‌ها و داده‌کاوی از کتابخانهٔ وکا بهره گرفته شد[۴۰]؛ وکا مجموعه‌ای از الگوریتم‌ها و ابزارهای پیش‌پردازش و ارزیابی در حوزهٔ یادگیری ماشین است.

برای اجرای الگوریتم‌های گمنام‌سازی ابتدا باید مجموعه‌داده با فرمت csv. در اختیار برنامه قرار گیرد؛ سپس عملیات پیش‌پردازش داده شامل تشخیص صفات شبه‌شناسه و صفت حساس انجام می‌شود. طبق قرارداد صفت‌های شبه‌شناسه به پسوند #QI نام‌گذاری شده‌اند و همچنین آخرین صفت از مجموعه نیز به‌عنوان صفت حساس S در نظر گرفته می‌شود. مجموعه‌داده با توجه به نوع آزمایش به چند مجموعهٔ کوچک‌تر (Party) تقسیم می‌شود و عملیات گمنام‌سازی با توجه به روش پیشنهادی بر روی آن‌ها صورت می‌گیرد و مجموعهٔ نهایی قابل‌انتشار را تولید می‌کنند. این مجموعه با فرمت arff. که فرمت استاندارد وکا است ذخیره می‌شود. فایل مورد نظر با توجه به شامل‌بودن نام و خصوصیات صفات برای عملیات داده‌کاوی مناسب است. نتایج آماری حاصل از اجرای برنامه شامل دقت طبقه‌بندی[1] و زمان اجرا در فایل با نام results.txt ذخیره می‌شود.

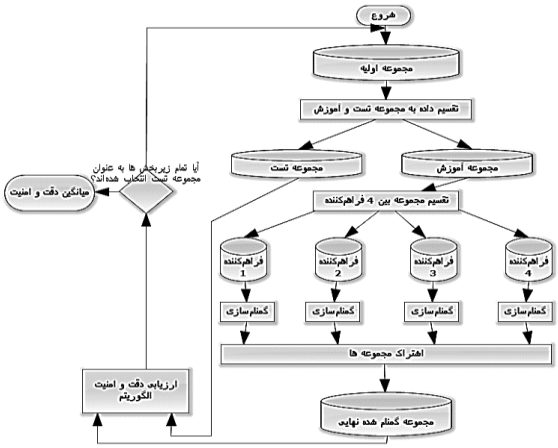
- **آزمایش و ارزیابی نتایج**

این بخش را با ارائهٔ آزمون‌هایی جهت مقایسهٔ عملکرد روش پیشنهادی (که DGF_PPDP[2] نام‌گذاری می‌شود) آغاز کرده و سپس نتایج حاصل‌شده مورد بررسی قرار خواهد گرفت؛ همان‌طور که در فصول پیشین بیان شد PPDP علم ایجاد توازن بین محرمانگی و دقت داده‌های منتشرشده است؛ به همین دلیل باید الگوریتم پیشنهادی را از این دو دیدگاه مورد ارزشیابی قرار داد. در شکل (۷) روند ارزیابی مربوط به آزمایش نشان داده شده‌است. مجموعهٔ ابتدایی به دو مجموعهٔ آموزش و آزمون تقسیم شده (نسبت یک به ده برای مجموعه آزمون)؛ سپس مجموعهٔ آموزش با حجم برابر بین چهار

[1] Classification

[2] 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 می‌کنند، پس از آن داده‌های گمنام‌شده تجمیع می‌شوند و مجموعهٔ قابل انتشار نهایی را تولید می‌کنند. این مجموعه به‌همراه 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 زمانی‌که تمام ده قسمت مجموعهٔ ابتدایی به‌عنوان آزمون در نظر گرفته شوند ادامه پیدا می‌کند. و در انتها میانگین نتایج به‌دست‌آمده مورد استفاده قرار می‌گیرد.



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

مجموعه‌دادهٔ استفاده‌شده در تمام آزمون‌های این مقاله Adult [۴۱] نام دارد که در بسیاری از پژوهش‌ها [۳۵,۲۲,۶] مورداستفاده قرار گرفته‌است. این مجموعه‌داده شامل ۴۸۸۴۲ رکورد و پانزده صفت است که در جدول (۴) مشخصات این صفات ذکر شده‌است. صفت Income یا درآمد سالیانه، به‌عنوان صفت حساس و هفت صفت (age، workclass، education، occupation، race، sex و native-country) به‌عنوان صفات شبه‌شناسه مورد استفاده قرار گرفتند. در تمامی آزمون‌های صورت‌گرفته برای عملیات عمومی‌سازی از درخت طبقه‌بندی[1] مطابق با [۴۲] استفاده شد، نرخ نمونه‌برداری از مجموعه‌ها برابر «نود درصد»، حد تخریب مربوط به مدل GF_PPDP برابر «۰.۶» و حد آستانهٔ انتشار توزیعی مربوط به مدل پیشنهادی برابر «۰.۱» انتخاب شده‌اند. با توجه به اینکه حد آستانهٔ انتشار توزیعی موازنه‌ای بین دقت و محرمانگی ایجاد می‌کند، در آزمایش‌ها مقدار ۰.۱ اختیار شد تا به‌دقت 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 به‌وسیلهٔ روش پیشنهادی (DGF_PPDP) با روش GF_PPDP مقایسه می‌شود. برای این منظور مقاومت مجموعه‌دادهٔ منتشرشدهٔ نهایی به‌وسیلهٔ هرکدام از الگوریتم‌ها در مقابل دو نوع حمله بررسی می‌شوند. دستهٔ نخست حملاتی که مهاجم تنها به مجموعهٔ نهایی دسترسی دارد (حملهٔ خارجی) و دستهٔ دوم حالتی که فراهم‌کنندگان در نقش مهاجم با استفاده از مجموعهٔ متعلق به خود و همچنین مجموعهٔ نهایی قصد تدارک حمله دارند (حملات داخلی). برای بررسی مقاومت در برابر این حملات از تعریف m-privacy [۲۲] استفاده می‌شود. در این آزمون از الگوریتم بازبینی مستقیم m-privacy[2] استفاده شد که در هر سطح از محرمانگی m (0 تا n که n تعداد فراهم‌کنندگان است) بر اساس شکل (۵) اقدام به چک کردن مجموعهٔ منتشرشده می‌کند. سطح "m=0" حملات خارجی را مشخص می‌کند و سطوح دیگر بیان‌کنندهٔ حملات داخلی هستند.

(جدول-۴): مشخصات صفات در مجموعه Adult
(Table-4): Attributes Specification in the Adult Dataset

صفت	نوع	تعداد/ دامنهٔ تغییرات	صفت	نوع	تعداد/ دامنهٔ تغییرات
Age	پیوسته	۹۰-۱۷	Race	گسسته	۵
Work-class	گسسته	۸	Sex	گسسته	۲
Final-weight	پیوسته	-۱۳۴۹۲ ۱۴۹۰۴۰۰	Capital-gain	پیوسته	۹۹۹۹۹-۰
Education	گسسته	۱۶	Capital-loss	پیوسته	۴۳۵۶-۰
Education-num	پیوسته	۱۶-۱	Hours-per-week	پیوسته	۹۹-۱
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 مجموعهٔ نهایی برای دو الگوریتم GF_PPDP و DGF_PPDP را به‌ازای اندازهٔ مجموعه‌داده نشان می‌دهد. تعداد و درصد به‌صورت میانگین به‌ازای رکوردهای نقض‌شده برای حملهٔ داخلی و خارجی در جدول مشخص شده‌است. اندازهٔ مجموعه‌داده بر اساس اندازهٔ مجموعهٔ آموزش مشخص شده‌است؛ برای مثال عدد ۴۰۵۰ مربوط به مجموعه‌ای با اندازهٔ پنج‌هزار رکورد است که ده درصد آن به‌عنوان مجموعهٔ آزمون انتخاب و باقی آن با نرخ «نود درصد» نمونه‌برداری شده‌اند. برای درک بهتر نتایج امنیت مجموعهٔ منتشرشده به‌ازای دو الگوریتم یادشده، شکل (۸) نمودار حفظ حریم خصوصی در مقابل حملات خارجی را نشان می‌دهد. در این شکل درصد رکوردهای نقض‌شده به کل رکوردهای انتشاریافته، برای مجموعه‌داده با اندازه‌های مختلف به تصویر کشیده شده‌است. رنگ سبز مربوط به روش پیشنهادی و رنگ قرمز مربوط به روش GF_PPDP است. در تمامی

[1] Taxonomy Tree

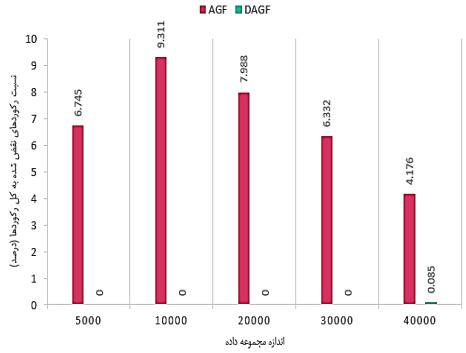
[2] The direct m-privacy verification algorithm

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

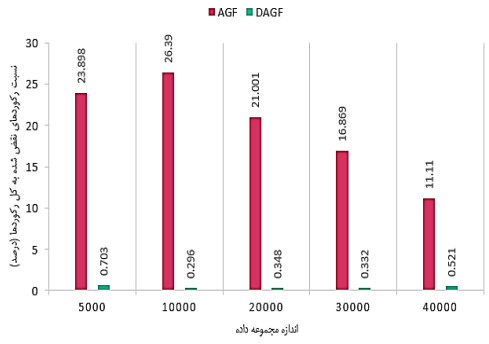
شکل (۹) نیز نمودار درصد رکوردهای نقض‌شده در مقابله با حملات داخلی را نشان می‌دهد. در این حالت روش پیشنهادی بهتر از روش GF_PPDP عمل کرده‌است و در تمام حالت‌ها درصد رکوردهای نقض‌شده از ۰.۸ درصد تجاوز نکرده‌است. تفاوت نرخ رکوردهای نقض‌شده بین دو روش به این دلیل است که الگوریتم GF_PPDP بدون درنظرگرفتن مشخصات فراهم‌کنندگان دیگر اقدام به گمنام‌سازی می‌کند، اما روش پیشنهادی با درنظرگرفتن احتمال شبه‌شناسه‌ها در مجموعه‌های فراهم‌کنندگان دیگر، اطمینان موردانتظار فراهم‌کنندگان دیگر برای دیدن یک رکود را کاهش می‌دهد. جدول (۶) میانگین رکوردهای نقض‌شده در هر سطح m از حملهٔ داخلی را نمایش می‌دهد. نمودار موردنظر به‌ازای تعداد چهل‌هزار رکورد که بین چهار فراهم‌کننده تقسیم شده، رسم شده‌است. به‌ازای تمام حالاتی که در شکل (۵) نشان داده شده‌است رکوردها از نظر محرمانگی چک شده و تعداد رکوردهایی که محرمانگی را نقض کردند محاسبه شده‌است. بر طبق روش m-privacy اگر رکوردی در سطح m نقض شود به‌طور قطع در سطح m-1 نیز نقض خواهد شد[۲۲]؛ لذا در جدول (۶) تعداد رکوردهای نقض‌شده در هر سطح با مجموع تعداد رکوردهای سطوح پایین‌تر تجمیع شده‌است.

(جدول-۵): نتایج رکوردهای نقض‌شده به‌ازای حملات داخلی و خارجی
(Table-5): Violated Records Count for Internal and External Attacks

اندازهٔ مجموعه‌دادهٔ آموزش	GF PPDP				DGF PPDP			
	داخلی		خارجی		داخلی		خارجی	
	تعداد رکورد نقض‌شده	درصد رکورد نقض‌شده	تعداد رکورد نقض‌شده	درصد رکورد نقض‌شده	تعداد رکورد نقض‌شده	درصد رکورد نقض‌شده	تعداد رکورد نقض‌شده	درصد رکورد نقض‌شده
۴۰۵۰	۹۶۷.۹	۲۳.۸۹۸	۲۷۳.۲	۶.۷۴۵	۲۸.۵	۰.۷۰۳	۰.۰	۰.۰
۸۱۰۰	۲۱۳۸.۱	۲۶.۳۹۰	۷۵۴.۲	۹.۳۱۱	۲۴.۰	۰.۲۹۶	۰.۰	۰.۰
۱۶۲۰۰	۳۴۰۲.۲	۲۱.۰۰۱	۱۲۹۴.۱	۶.۳۳۲	۵۶.۴	۰.۳۴۸	۰.۰	۰.۰
۲۴۳۰۰	۴۰۹۹.۲	۱۶.۸۶۹	۱۵۳۸.۷	۴.۱۷۶	۸۰.۷	۰.۳۳۲	۰.۱	۰.۰≈
۳۲۴۰۰	۳۵۹۹.۹	۱۱.۱۱۰	۱۳۵۳.۳	۴.۱۷۶	۱۶۸.۹	۰.۵۲۱	۲۷.۷	۰.۰۸۵



(شکل -۸): محرمانگی در مقابل حملهٔ خارجی
(Figure-8): Privacy Against External Attacks



(شکل -۹): محرمانگی در مقابل حملهٔ داخلی

(Figure-9): Privacy Against Internal Attacks

(جدول-۶): میانگین رکوردهای نقض‌شده در حملات داخلی به‌ازای هر سطح
(Table-6): Average Violated Records in Internal Attacks per Level

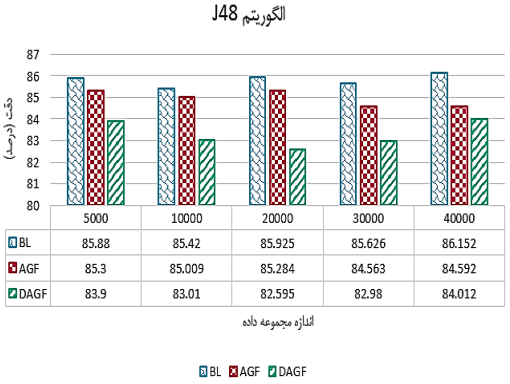
	m=۱	m=۲	M=۳
GF_PPDP	۲۳۳۷.۱	۱۲۱۰.۲+۲۳۳۷.۱=۳۵۴۷.۳	۶۷.۹+۳۵۴۷.۳=۳۶۱۵.۲
DGF_PPDP	۷۳.۴	۴۶+۷۳.۴=۱۱۹.۴	۴۸.۹+۱۱۹۰۴=۱۶۸.۳

- **دقت داده‌های انتشاریافته**

برآورده‌کردن شاخص حفظ حریم خصوصی و امنیت در داده‌های منتشرشده بسیار حائز اهمیت است، اما از طرف دیگر باید داده‌های نهایی سودمندی لازم را جهت عملیات داده‌کاوی حفظ کنند؛ ازاین‌رو در این بخش نتایج دقت طبقه‌بندی داده‌های منتشرشده در حالت‌های مختلف مورد ارزیابی قرار می‌گیرد. روند ارزیابی به‌صورت چارت ارائه‌شده در شکل (۷) است. در این حالت برای ارزیابی تأثیر گمنام‌سازی بر دقت طبقه‌بندی از الگوریتم‌های J48، NaiveBayes، PART و Logistic استفاده شده و همچنین برای مقایسهٔ بهتر، دقت طبقه‌بندی بر روی داده‌های خام

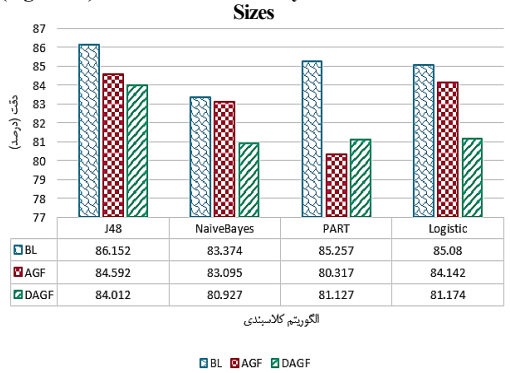
(گمنام‌نشده) نیز اندازه‌گیری شده‌است که در نمودار آن را با نام BL نمایش می‌دهیم.

شکل (۱۰) نمودار درصد دقت طبقه‌بندی الگوریتم J48 به‌ازای مجموعه‌داده 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 ازاین‌رو 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 مجموعه‌دادهٔ اولیه با اندازهٔ چهل‌هزار 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.



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

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



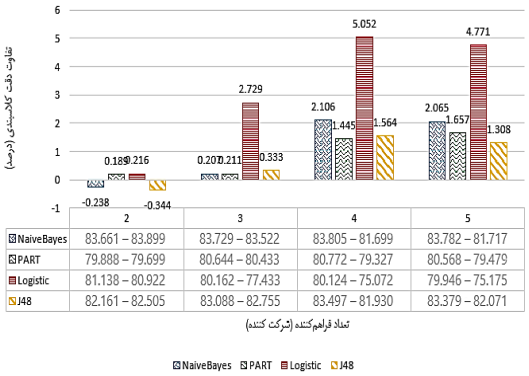
(شکل –۱۱): دقت طبقه‌بندی به‌ازای الگوریتم‌های مختلف طبقه‌بندی

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

در دو نمودار پیش برای روشن‌ترشدن 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، مجموعهٔ متعلق به هر فراهم‌کننده با روش GF_PPDP گمنام‌سازی می‌شود و در حالت دوم مجموعه‌دادهٔ 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 اندازهٔ مجموعه‌دادهٔ 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؛ به‌عبارت‌دیگر اشتراک داده و استفاده از روش پیشنهادی به‌نفع فراهم‌کنندگان بوده‌است. بااینکه در الگوریتم پیشنهادی عمل‌گرهای گمنامی مانند عمومی‌سازی بیشتر استفاده می‌شود، اما اشتراک داده باعث شده‌است درمجموع الگوریتم DGF_PPDP سودمندی بیشتری نسبت به الگوریتم GF_PPDP داشته باشد.

- **زمان اجرا**

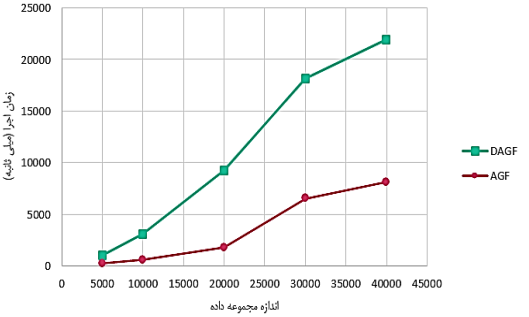
یکی دیگر از پارامترها در ارزیابی الگوریتم‌ها زمان اجرا است. سازوکار گمنام‌سازی الگوریتم پیشنهادی با الگوریتم GF_PPDP تنها در محاسبهٔ اطمینان مشاهده‌شده و اطمینان مورد انتظار متفاوت است. در محاسبهٔ اطمینان مورد انتظار همان‌طور که شبه‌کد (۱) نشان می‌دهد، به‌ازای هر شبه‌شناسه حلقه‌ای به تعداد آمار فراهم‌کنندگان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



(شکل -۱۲): مقایسهٔ 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 اندازهٔ مختلف محاسبه و در شکل (۱۳) نمایش داده شده‌است. دو الگوریتم بر روی رایانه‌ای با پردازندهٔ Inetl core i5-4210U (2.4 GHz) و 6 GB حافظهٔ اصلی اجرا شدند؛ همان‌طور که در شکل مشخص است، با افزایش اندازه مجموعه‌دادهٔ 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 اندازهٔ 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 توزیع‌شدهٔ مبتنی‌بر 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 است. نتایج پیاده‌سازی طرح روی مجموعه‌دادگان استاندارد نشان می‌دهد که این روش توازن مناسبی میان امنیت و کارایی داده‌های منتشرشده نهایی برقرار می‌کند. مهم‌ترین نوآوری در این مقاله ارائهٔ 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 مجموعهٔ فراهم‌کنندگان 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 حوزهٔ مشترک فعالیت دارند؛ مانند بیمارستان‌ها و بانک‌ها که داده‌های ذخیره‌شدهٔ 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.

۶-مراجع 6-References

[1] A. Majeed and S. Lee, (2021) "Anonymization Techniques for Privacy Preserving Data Publishing: A Comprehensive Survey," i*n 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.

[۷] پارسا، مجتبی (۱۳۸۸)، «حریم خصوصی و رازداری در پزشکی و جنبه‌های مختلف آن»، *مجلهٔ ایرانی اخلاق و تاریخ پزشکی*، جلد۲، شمارهٔ ۴، صص ۱-۱۴.

[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.

[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., Venkitasubramaniam, M. (2007), "l-diversity: Privacy beyond k-anonymity", *ACM Transactions on 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 محرمانگی در انتشار داده‌ها به‌وسیلهٔ گمنام‌سازی دسته‌ای»، *پایان‌نامهٔ کارشناسی‌ارشد 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.

[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]

[42] "Taxonomy trees of the Adult data set", [Online]. Available: http://ddm.cs.sfu.ca/dmsoft/privacy/products/adultHierarchy.txt. [Accessed 5 October 2020].

[۴۳] مسیبی، الیاس. (۱۳۹۵)، «تحلیل امنیتی روش‌های انتشار با حفظ محرمانگی در داده‌های توزیع‌شده»، *پایان‌نامهٔ کارشناسی‌ارشد مهندسی کامپیوتر*، گیلان: دانشگاه گیلان.

[43] 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), "Communi¬cation-efficient learning of deep networks from decentralized data", *In: Artificial intelligence and statis¬tics*, 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.

[۴۷] مرادی لالمی، علی، شاه بهرامی، اسداله، ابراهیمی آتانی، رضا، علی دوست نیا، مهران (۱۳۹۵)، «خوشه‌بندی فراابتکاری اسناد فارسی اِکس‌اِم‌اِل 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", *Jornals of signal and Processing Data*; 13 (2): 11-23

[۴۸] ابراهیمی آتانی، رضا، صادقپور، مهدی (۱۳۹۷)، «ارایه یک روش جدید انتشار داده ها با حفظ محرمانگی با هدف بهبود دقت طبقه بندی روی داده های گمنام»، *نشریهٔ 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", *Jornals Jornals 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

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، صادقپور، مهدی (۱۳۹۵)، مروری بر روش‌های حفظ حریم خصوصی در انتشار داده‌ها»، *امنیت فضای تولید و تبادل اطلاعات (منادی)*، ۵ (۲)، صص۶۲-۴۹.

[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.

الیاس مسیبی مدرک کارشناسی و کارشناسی‌ارشد خود را در رشتهٔ مهندسی کامپیوتر – نرم‌افزار از گروه مهندسی کامپیوتر دانشگاه گیلان به‌ترتیب در سال‌های ۱۳۹۲ و ۱۳۹۵ دریافت کرد. ایشان در سِمَتِ مدیر تیم‌های توسعه در شرکت‌های حوزهٔ فناوری اطلاعات و هوش مصنوعی مشغول به فعالیت و زمینه‌های پژوهشی ایشان در حوزه‌های یادگیری ماشین، مهندسی نرم‌افزار و هوش مصنوعی است.

نشانی رایانامهٔ ایشان عبارت است از:
elyas.mosayebi@gmail.com

رضا ابراهیمی آتانی دانشیار گروه مهندسی کامپیوتر دانشکدهٔ فنی و مهندسی دانشگاه گیلان هستند. ایشان مدرک دکترای خود را در سال ۱۳۸۹ در رشتهٔ مهندسی الکترونیک از دانشگاه علم و صنعت دریافت کرد و در حال حاضر عضو پیوستهٔ 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 رسیده‌است. زمینه‌های پژوهشی ایشان طراحی، پیاده‌سازی و تحلیل الگوریتم‌ها و پروتکل‌های رمزنگاری، امنیت و حفظ حریم خصوصی داده‌ها در سامانه‌های محاسباتی و طراحی و توسعهٔ شبکه‌های کامپیوتری است.

نشانی رایانامهٔ ایشان عبارت است از:
rebrahimi@guilan.ac.ir