مروری بر روش‌های حفظ حریم خصوصی در انتشار داده‌ها

رضا ابراهیمی آتانی[1]* و مهدی صادق‌پور[2]

[1]دانشیار گروه مهندسی کامپیوتر دانشگاه گیلان، گیلان، رشت، ایران
rebrahimi@guilan.ac.ir

[2]دانش آموخته کارشناسی ارشد مهندسی کامپیوتر، نرم‌افزار دانشگاه گیلان، گیلان، رشت، ایران
Mehdi.sadeghpour@live.com

چکیده

با گسترش خدمات الکترونیکی و سهولت در جمع‌آوری و ذخیره‌سازی داده‌ها، حجم وسیعی از اطلاعات افراد در قالب رکوردهایی درون پایگاه‌داده سازمان‌های خصوصی و دولتی ذخیره شده است. این رکوردها به‌طورمعمول شامل اطلاعات حساس شخصی از جمله میزان درآمد و نوع بیماری هستند که باید از خارج از دسترس سایرین نگه داشته شود. اگر مخفی‌کردن اطلاعات حساس رکوردها به‌درستی انجام گیرد، تأثیر مخربی در سودمندی داده‌ها ندارد؛ زیرا دراکثر مواقع دریافت‌کنندگان داده به‌دنبال مشخصات افراد و داده‌های سطح رکورد نیستند و می‌خواهند قوانین و الگوهایی کلی و در سطح مجموعه‌داده به‌دست بیاورند. "حفظ حریم خصوصی در انتشار داده‌ها" راه‌کارهایی برای تضمین محرمانه‌ماندن اطلاعات حساس یک مجموعه‌داده در هنگام انتشار آن در محیطی غیر قابل اعتماد است. در این فرایند سعی بر آن است که ضمن مخفی‌نگه‌داشتن اطلاعات حساس، داده‌های منتشرشده همچنان برای عملیات کشف دانش مفید باقی بماند. این نوشتار با هدف مرور روش‌های حفظ حریم خصوصی داده‌ها به نگارش درآمده است.

واژگان کلیدی: حفظ حریم خصوصی، اشتراک‌گذاری داده‌ها، داده‌کاوی، گمنام‌سازی

۱- مقدمه

توسعه و نفوذ خدمات الکترونیکی در زندگی بشر موجب شده است، سازمان‌ها و شرکت‌های مختلف، حجم گسترده‌ای از اطلاعات مشتریان خود را به‌طور روزمره ذخیره کنند. این داده‌ها منبع با ارزشی برای کشف الگوها و قوانین سودمند به‌وسیله داده‌کاوی هستند؛ به همین دلیل ممکن است پژوهش‌گران، شرکت‌ها و سازمان‌های مرتبط، خواهان دریافت و استفاده از این داده‌ها باشند. در چنین شرایطی چگونه می‌توان بدون نقض‌کردن حریم شخصی، مجموعه‌داده‌ای را که حاوی اطلاعات مشتریان است، در اختیار نهادهای دیگر قرار داد؟ حفظ حریم خصوصی در انتشار داده‌ها[1] (PPDP) به‌عنوان راه‌کاری برای حل این مشکل معرفی شده است [۱].

طبق ماده ۱۲ اعلامیه جهانی حقوق بشر[2] (UDHR) قلمرو خصوصی افراد نباید مورد مداخله قرار گیرد و تمام اشخاص حق دارند به‌وسیله قانون در برابر چنین مداخلات و تعرضاتی محافظت شوند [۲]. جدا از قوانینی که برای محافظت از حریم خصوصی وجود دارد و خاطیان را مجازات می‌کند، جلب اعتماد مشتریان عاملی برای احتیاط دوچندان در منتشرکردن داده‌ها است؛ زیرا افراد حاضر نیستند به شرکتی مراجعه کنند که ممکن است اطلاعات شخصی‌شان را در اختیار سایرین قرار دهد.

رکوردهای یک مجموعه‌داده را می‌توان به‌عنوان نمونه‌هایی از یک جامعه آماری درنظر گرفت. این جامعه آماری می‌تواند بیماران یک درمانگاه، کاربران یک سایت اینترنتی، کارمندان یک اداره و یا مثال‌هایی از این دست

[1] Privacy-Preserving Data Publishing (PPDP)

[2] Universal Declaration of Human Rights (UDHR)

باشد. مجموعه‌داده منتشرشده باید به‌گونه‌ای باشد که انجام بررسی‌های دلخواه بر روی آن، منجر به آشکارشدن مشخصات یک فرد خاص نشود و فقط اطلاعاتی از کل جامعه قابل یادگیری باشد. این هدف از طریق گمنام‌سازی داده‌ها محقق می‌شود.

با این توضیحات ممکن است، پرسیده شود که اگر هدف نهایی، استخراج یک مدل از کل داده‌ها است، پس چرا به جای انتشار مجموعه‌داده، آن مدل استخراج و منتشر نگردد؟ در پاسخ به این سؤال باید مدنظر داشت که انتشار مجموعه‌داده امکان به‌کارگیری روش‌های مختلف برای کاویدن داده‌ها را فراهم می‌آورد. انتشار داده‌های شرکت Netflix [۳] مثال خوبی در این زمینه است. این شرکت برای بهبود سامانهٔ توصیه‌گر خود، رقابتی عمومی برای معرفی الگوریتم‌های مختلف برپا کرد. در این شرایط انتشار داده‌ها تنها روشی بود که انعطاف لازم برای آزمودن الگوریتم‌های گوناگون را فراهم می‌کرد.

هدف اصلی این نوشتار ارائهٔ یک مقاله مروری برای معرفی روش‌های حفظ حریم خصوصی در انتشار داده‌ها است که بدین ترتیب ادامه خواهد یافت: در بخش ۲، به بررسی PPDP پرداخته و تعاریف مورد نیاز ارائه می‌شود. در بخش ۳ کاربردهای PPDP با ذکر چند مثال بازگو می‌شود. بخش ۴ به معرفی توابع پرکاربرد در گمنام‌سازی می‌پردازد. در بخش ۵ مدل‌های مختلف حفظ حریم خصوصی مرور شده و نقاط قوت و ضعف آن‌ها بررسی می‌شود. و بخش انتهایی به نتیجه‌گیری و جمع‌بندی این نوشتار اختصاص دارد.

۲- حفظ حریم خصوصی در انتشار داده‌ها

فرآیند PPDP راه‌کاری برای تضمین حریم خصوصی داده‌ها در محیط‌های غیر قابل اعتماد است. در این فرایند سعی بر آن است که ضمن مخفی‌نگه‌داشتن اطلاعات محرمانه، داده‌های منتشرشده همچنان برای عملیات کشف دانش مفید باقی بماند [۴]. در زیر، مفاهیم پایه‌ای PPDP که در ادامه مورد استفاده قرار می‌گیرند، تعریف شده‌اند:

- **مالکان رکورد**[1]: افرادی هستند که اطلاعات مربوط به آن‌ها در مجموعه‌داده ثبت شده است.
- **منتشرکننده داده‌ها**[2]: سازمانی است که مجموعه‌داده را در اختیار دارد و اقدام به انتشار آن می‌کند.
- **مهاجم:** فرض بر این است که بعد از انتشار داده‌ها، مهاجم قادر خواهد بود نسخه‌ای از آن را به‌دست آورد. وی تلاش می‌کند، اطلاعات حساس یک یا چند مالک رکورد (که قربانی[3] نامیده می‌شوند) را از مجموعه‌داده استخراج کند. اطلاعات در دسترس مهاجم فقط به داده‌های منتشرشده محدود نیست و امکان دارد از دانش پیش‌زمینه[4] خود نیز برای انجام حملات استفاده کند.
- **دریافت‌کننده داده‌ها**[5]: پژوهش‌گران علمی یا سازمان‌های تجاری که به‌دنبال کشف دانش از مجموعه‌داده منتشر شده هستند.
- **گمنام‌سازی**[6] **داده‌ها:** فرآیندی است که داده‌های خام[7] را به داده‌های گمنام‌شده تبدیل می‌کند. داده‌های گمنام‌شده در مقابل نوع خاصی از حملات که در یک مدل حریم خصوصی[8] تعریف شده است، از افشای اطلاعات مصون می‌مانند. مدل‌های حفظ حریم خصوصی داده در بخش ۵ تشریح شده‌اند.

همانطور که در شکل (۱) نمایش داده شده است، دو مرحله اصلی فرایند PPDP جمع‌آوری و انتشار داده‌ها هستند [۵]. در مرحله نخست منتشرکننده داده‌ها اطلاعات

[1] Record Owners
[2] Data Publisher
[3] Victim
[4] Background Knowledge
[5] Data Recipient
[6] Anonymization
[7] Raw Data
[8] Privacy Model

را از کاربران خود جمع‌آوری می‌کند. فرض بر این است که مالکان رکورد به منتشرکنندهٔ داده‌ها اطمینان دارند و حاضرند اطلاعات حساس خود را با او به اشتراک بگذارند، اما دریافت‌کنندگان داده، مورد اعتماد نیستند و اطلاعات حساس باید از آن‌ها مخفی نگاه داشته شود. بنابراین در مرحله بعد روش‌های حفظ حریم خصوصی روی داده‌ها اعمال می‌شود و خروجی در اختیار دریافت‌کنندگان داده قرار می‌گیرد. هر دریافت‌کننده ممکن است داده‌های مورد نیاز خود را از چندین منتشرکننده جمع آوری کند. به‌عنوان مثال، یک شرکت بیمه برای انجام داده‌کاوی در زمینه رابطه بین شغل افراد و یک بیماری خاص، نیاز به داده‌های ذخیره‌شده در بیمارستان‌های شهر دارد. در این مثال بیماران مالکان رکورد، بیمارستان‌ها منتشرکنندگان داده و شرکت بیمه دریافت‌کننده داده‌ها است.

(شکل-۱): جمع‌آوری و انتشار داده‌ها

همان‌طور که ذکر شد در فرآیند گمنام‌سازی، پردازش‌هایی بر روی صفات ذخیره‌شده در مجموعه‌داده انجام می‌گیرد تا در هنگام انتشار عمومی داده‌ها، اطلاعات حساس افشا نشود. پردازشی که روی هر صفت انجام می‌گیرد به نوع آن بستگی دارد. انواع صفات در PPDP عبارتند از [۵]:

- **شناسه صریح**[1]:مجموعه صفاتی که به‌طور مستقیم برای شناسایی مالک رکورد مورد استفاده قرار می گیرند، مانند: کد ملی، شماره دانشجویی و ... .
- **شبه‌شناسه**[2]: مجموعه‌ای از صفات که هر یک به‌تنهایی برای شناسایی مالک رکورد کافی نیست؛ اما با استفاده از ترکیب آن‌ها، در اغلب موارد می‌توان این کار را انجام داد. مانند ترکیب کدپستی و تاریخ تولد و جنسیت.
- **صفات حساس**[3]: صفاتی که صاحب رکورد مایل نیست عموم از آن ها اطلاع داشته باشند؛ مانند: نوع بیماری، میزان حقوق و ... .
- **صفات غیر حساس**[4]: صفاتی که در هیچ کدام از دسته‌های قبل جای نمی‌گیرند.

شکل (۲) مثالی را نمایش می‌دهد که در آن صفات یک مجموعه‌داده، به چهار نوع ذکرشده در بالا تقسیم‌بندی شده‌اند.

شناسه‌های صریح | شبه‌شناسه‌ها | صفت غیرحساس | صفت حساس

کد ملی	نام	تحصیلات	جنسیت	وضعیت تاهل	تاریخ تولد	ساعات کار در هفته	میزان حقوق
۲۲۲۵۵۵۹	رضا مجیی	دیپلم	مرد	متاهل	۱۳۶۴/۵/۲۳	۴۵	۸۶۰
۲۲۲۷۷۷۱	نیما راهنما	کارشناسی	مرد	مجرد	۱۳۵۹/۲/۹	۴۰	۹۱۰
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

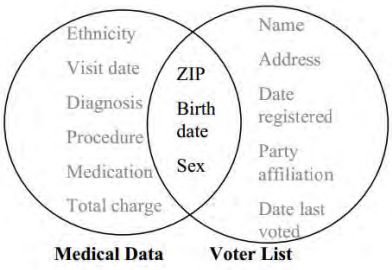
(شکل-۲): مثالی از انواع صفات در PPDP

نخستین قدم برای شکستن رابطه بین رکوردها و مالکان رکورد، حذف شناسه‌های صریح از مجموعه داده است. اما همانطور که در شکل (۳) نمایش داده شده است، حذف این شناسه‌ها لزوما باعث حفظ حریم خصوصی نمی‌شود [۶]. در این شکل داده‌های پزشکی که فاقد شناسه صریح هستند با استفاده از شبه‌شناسه‌ها(جنسیت، تاریخ تولد و کدپستی) به داده‌هایی که از قبل در اختیار فرد مهاجم بوده است (فهرست رأی‌دهندگان) پیونددهی[5] شده است؛ یعنی مهاجم از شبه‌شناسه‌ها برای تشخیص ارتباط بین شناسه صریح مالک رکورد و صفت حساس وی استفاده کرده است.

در مقابله با چنین حملاتی می‌توان با استفاده از عملیات گمنام‌سازی موجب شد تا شبه‌شناسه‌های هر

[1] Explicit Identifier
[2] Quasi Identifier
[3] Sensitive Attributes
[4] Non Sensitive Attributes
[5] Linking

رکورد در چندین رکورد دیگر نیز تکرار شود و مهاجم در پیونددهی داده‌ها دچار ابهام شود. پس از اعمال گمنام‌سازی بر روی داده‌های خام، میزانی از اطلاعات دیگر در داده‌های خروجی در دسترس نیست. هرچه میزان حریم خصوصی بیشتر مد نظر باشد، ممکن است، سودمندی داده‌ها پایین‌تر بیاید و هرچه سودمندی بیشتری مد نظر باشد، ممکن است، به کاهش سطح حریم خصوصی بیانجامد [۷]. بنابراین یک روش موفق در حفظ حریم خصوصی داده باید تعادل مناسبی بین این دو پارامتر برقرار کند.



(شکل-۳): پیونددهی به منظور بازشناسی داده‌ها [۶]

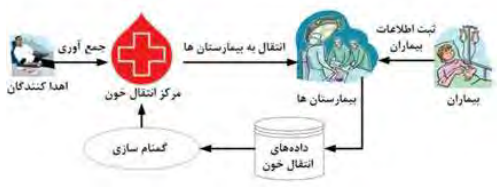
۳- کاربردهای حفظ حریم خصوصی در انتشار داده‌ها

استخراج دانش از داده‌های جمع‌آوری‌شده توسط سازمان‌های مختلف همواره مورد توجه پژوهش‌گران بوده است. در همین حال، حفظ حریم خصوصی مالکین رکورد، اشتراک‌گذاری داده‌ها[1] را با چالش مواجه می‌کند. PPDP به‌عنوان یکی از روش‌های حفظ حریم خصوصی داده‌ها، می‌تواند در بسترهای مختلفی از جمله مراکز درمانی و تجاری، سامانه‌های آگاه از موقعیت[2]، موتورهای جستجو، رسانه‌های اجتماعی و ... مورد استفاده قرار بگیرد [۱، ۸]. در ادامه دو مورد از کاربردهای ذکرشده مورد بحث قرار می‌گیرد.

۳-۱- مراکز درمانی

پژوهش‌گران علم پزشکی، شرکت‌های بیمه و سایر نهادهایی که در حوزه سلامت فعالیت می‌کنند، از داده‌هایی که مراکز درمانی از مراجعه‌کنندگان ثبت می‌کنند، بهره می‌گیرند. به‌عنوان نمونه ممکن است پژوهش‌گران به دنبال رابطه بین یک بیماری خاص با وزن فرد باشند. داده‌هایی که قرار است در اختیار آنان قرار بگیرد باید به‌گونه‌ای باشد که اطلاعات شخصی بیماران افشا نشود. در ایالات متحده از سال ۱۹۹۶ مجموعه قوانینی با عنوان "قانون انتقال و پاسخ‌گویی بیمه سلامت"[3] تدوین شده است که بر اساس یکی از بندهای آن مراکز درمانی اجازه ندارند اطلاعات خام بیماران را با سایر سازمان‌ها به اشترک بگذارند، اما هیچ منعی برای انتشار داده‌های پزشکی گمنام‌شده وجود ندارد [۹].

یکی از سناریوهایی که کاربرد PPDP در مراکز درمانی را نشان می‌دهد، سرویس انتقال خون صلیب سرخ هنگ‌کنگ[4] است [۱۰]. روند کاری این مرکز در شکل (۴) نمایش داده شده است. ابتدا خون مورد نیاز از اهداکنندگان دریافت و پس از تست سلامت آن، در اختیار بیمارستان‌های مختلف قرار داده می‌شود. بیمارستان‌ها از این خون برای درمان بیماران استفاده می‌کنند. تمام جزییات نحوه استفاده از خون، مانند مشخصات دریافت‌کننده، نوع عمل انجام‌شده، دلایل تزریق خون و ... در پایگاه‌داده بیمارستان ثبت می‌شود. مرکز انتقال خون به‌طور دوره‌ای این داده‌ها را از بیمارستان‌ها جمع‌آوری می‌کند و با تحلیل آن اطلاعاتی از قبیل تخمین میزان خون مورد نیاز هر بیمارستان در آینده به‌دست می‌آورد. داده‌هایی که از طریق بیمارستان‌ها در اختیار مرکز انتقال خون قرار می‌گیرد، نباید شامل اطلاعات شخصی بیماران باشد. این امر از طریق گمنام‌سازی داده‌ها محقق می‌شود.



(شکل-۴): روند کار سرویس انتقال خون صلیب سرخ هنگ‌کنگ

[1] Data Sharing
[2] Location-aware Systems
[3] Health Insurance Portability and Accountability Act (HIPPA)
[4] Hong Kong Red Cross Blood Transfusion Service

۳-۲- رسانه‌های اجتماعی

در سال‌های اخیر شبکه‌های اجتماعی مانند Facebook و Myspace ، برنامه‌های پیام‌رسان مانند Skype و Telegram و سایر رسانه‌های اجتماعی کاربران زیادی را به خود جذب کرده‌اند. همان‌طور که در شکل (۵) به تصویر کشیده شده است، آمارها نشان می‌دهد، حجم اطلاعاتی که هر دقیقه در رسانه‌های اجتماعی تولید می‌شود بسیار درخور توجه است. وجود روابط خاص مانند دوستی[1]، دنبال‌کردن[2] و پسندیدن[3] در این داده‌ها، موقعیت مناسبی برای درک الگوهای رفتاری انسان‌ها ایجاد کرده است. کند و کاو رسانه‌های اجتماعی[4] (SMM) ترکیبی از علوم مختلف از جمله: داده‌کاوی، جامعه‌شناسی، علوم شبکه[5] و آمار است، که برای استخراج دانش از داده‌های رسانه‌های اجتماعی[6] به‌کار می‌رود [۱۱].

اعتماد در رسانه‌های اجتماعی [۱۲]، نقش رسانه‌های اجتماعی در بحران‌ها [۱۳] و جستجوی اجتماعی [۱۴] از جمله حوزه‌هایی هستند که در SMM مورد بررسی قرار می‌گیرد.

علی‌رغم این ویژگی‌های مفید تعداد بسیار کمی از داده‌های اجتماعی به‌صورت عمومی منتشر شده است که مهم‌ترین دلیل آن نگرانی از افشای اطلاعات خصوصی کاربران این برنامه‌ها است [۱]. PPDP با حذف این مانع می‌تواند امکان دسترسی پژوهش‌گران را به این منبع باارزش اطلاعاتی فراهم آورد.

(شکل-۵): فعالیت کاربران در برخی از رسانه‌های اجتماعی طی یک دقیقه (برگرفته از [۱۵])

[1] Friendship
[2] Follow
[3] Like
[4] Social Media Mining (SMM)
[5] Network Science
[6] Social Media Data

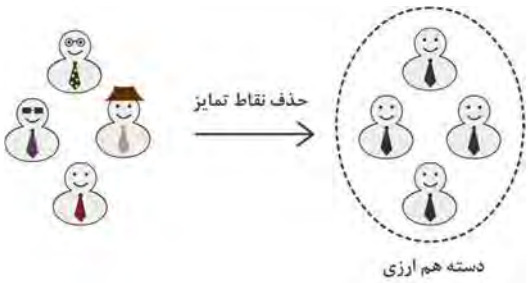
۴- عملگرهای گمنام سازی

در طی فرآیند گمنام‌سازی توابعی بر روی رکوردها اعمال می‌شود تا بتوان مجموعه‌داده را در دسترس عموم قرار داد. این توابع به‌عنوان عملگرهای گمنام‌سازی شناخته می‌شوند. درواقع عملگرهای گمنام‌سازی ابزار لازم برای دستیابی به مدل‌های حفظ حریم خصوصی (بخش ۵) را فراهم می‌کنند. این عملگرها براساس روش کارکردشان به سه گروه تقسیم شده‌اند که در ادامه معرفی می‌شود.

۴-۱- گمنامی از طریق دسته‌سازی

در این روش هر رکورد به همراه چندین رکورد دیگر در گروهی قرار داده می‌شود که مقدار شبه‌شناسه‌های تمام اعضای آن یکسان است. چنین گروهی، گروه هم‌ارزی[7] نامیده می‌شود. در این حالت مهاجم با جستجوکردن شبه‌شناسه‌های قربانی در مجموعه‌داده، به جای رکوردی یکتا، به گروه هم‌ارزی که رکورد قربانی عضو آن است، دست می‌یابد و درنتیجه با اطمینان کامل صفت حساس وی را نمی‌تواند تشخیص دهد.

هرچه دامنه مقادیر و تعداد شبه‌شناسه‌ها بیشتر باشد، احتمال پیداکردن دسته‌ای از رکوردها که بتوانند در یک گروه هم‌ارزی قرار بگیرند کمتر می‌شود. به‌همین دلیل لازم است توابعی بر روی رکوردهای اولیه اعمال شود که نقاط تمایز آن‌ها را کمتر کند (شکل ۶). مهم‌ترین توابعی که برای دسته‌سازی رکوردها مورد استفاده قرار می‌گیرند عمومی‌سازی[8] و فرونشانی[9] [۱۶] هستند.



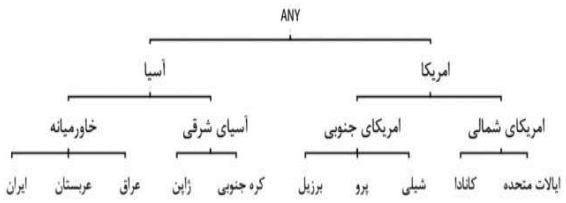
(شکل-۶): گمنامی از طریق دسته‌سازی

[7] Equivalence Group
[8] Generalization
[9] Suppression

۴-۱-۱- عمومی‌سازی

در عمومی‌سازی مقدار یک صفت به بازه‌ای بزرگ‌تر تعمیم داده می‌شود، برای مثال نشانی دقیق محل سکونت یک فرد را می‌توان با شهری که در آن زندگی می‌کند، جایگزین کرد و در مرحله بعد شهر را به استان تعمیم داد. عمومی‌سازی صفات به‌کمک درخت طبقه‌بندی[1] انجام می‌گیرد. در شکل (۷) نمونه‌ای از درخت طبقه‌بندی برای صفت "ملیت" نمایش داده شده است. به‌منظور عمومی‌سازی مقدار یک صفت، باید گره متناظر با آن در درخت طبقه‌بندی پیدا و سپس از مقدار گره والد برای آن صفت استفاده شود.

برگ‌های درخت طبقه‌بندی شامل کلیه مقادیری هستند که صفت مورد نظر می‌تواند داشته باشد. گره ریشه نیز نشان‌دهنده عمومی‌ترین مقدار است که به‌طورمعمول با "ANY" نمایش داده می‌شود. شکل (۸) مثالی را نمایش می‌دهد که در آن عملگر عمومی‌سازی روی صفت "ملیت" اعمال شده است و درنتیجه مقدار این شبه‌شناسه در هر دو رکورد یکسان شده است.



(شکل-۷): درخت طبقه‌بندی صفت ملیت

ایجاد درخت طبقه‌بندی برای هر صفت نیازمند داشتن اطلاعات پیش‌زمینه در مورد آن است؛ به‌عنوان نمونه، برای تشکیل درخت شکل (۷) لازم است بدانیم کشورهای شیلی، پرو و برزیل در آمریکای جنوبی واقع شده‌اند. به همین دلیل به‌طورمعمول این کار به‌صورت غیرخودکار و توسط یک انسان خبره (آشنا با موجودیت‌های سیستم) انجام می‌گیرد، اما برای صفاتی که مقادیر عددی می‌پذیرند ساخت درخت می‌تواند به‌صورت خودکار و از طریق بازه‌بندی انجام گیرد [۱۷].

[1] Taxonomy Tree

جنسیت	ملیت	سن	بیماری
مرد	ایران	۲۸	آنفولانزا
مرد	عربستان	۲۳	دیابت

⇐

جنسیت	ملیت	سن	بیماری
مرد	خاورمیانه	۲۸	آنفولانزا
مرد	خاورمیانه	۲۳	دیابت

(شکل-۸): اعمال یک مرحله عمومی‌سازی روی صفت ملیت

۴-۱-۲- فرونشانی

عملگر فرونشانی برای ازبین‌بردن تمایز شبه‌شناسه‌ها، مقدار صفت را حذف می‌کند و به جای آن مقدار گمشده[2] قرار می‌دهد. شکل (۹) نمونه‌ای از نحوه کار این عملگر را نشان می‌دهد. همان‌طور که دیده می‌شود، برخلاف عمومی‌سازی که مقادیر قبل و بعد از گمنام‌سازی را از لحاظ معنایی منسجم نگه می‌دارد، در این حالت نمی‌توان هیچ برداشتی از ملیت رکوردها داشت. بنابراین استفاده از این عملگر باید با احتیاط صورت گیرد. در غیر این صورت تعداد مقادیر گمشده زیاد می‌شود و کیفیت داده‌ها به شدت کاهش می‌یابد.

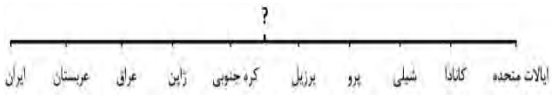
جنسیت	ملیت	سن	بیماری
مرد	ایران	۲۸	آنفولانزا
مرد	عربستان	۲۳	دیابت

⇐

جنسیت	ملیت	سن	بیماری
مرد	؟	۲۸	آنفولانزا
مرد	؟	۲۳	دیابت

(شکل-۹): اعمال فرونشانی روی صفت ملیت

فرونشانی را می‌توان حالت خاصی از عمومی‌سازی در نظر گرفت که در آن از درخت طبقه‌بندی به عمق یک استفاده می‌شود. به‌عنوان مثال عمومی‌سازی شکل (۸) به‌وسیله درخت نمایش داده در شکل (۱۰) نتیجه‌ای مشابه با فرونشانی در پی خواهد داشت.



(شکل-۱۰): درخت طبقه‌بندی فرضی صفت ملیت برای عملگر فرونشانی

فرونشانی و عمومی‌سازی به دو روش سراسری[3] و محلی[4] مورد استفاده قرار می‌گیرند [۱۸]. در روش

[2] Missing Value
[3] Global
[4] Local

سراسری وقتی عملگر روی صفت a از رکورد r اعمال می‌شود و مقدار آن را از v به 'v تغییر می‌دهد، باعث می‌شود صفت a در تمام رکوردهایی که در آن‌ها مقدار v دارد، به 'v تغییر مقدار دهد. تأثیر روش محلی در سطح رکورد است و فقط مقدار یک صفت را تغییر می‌دهد.

۴-۲- شکستن رابطه بین صفت‌ها

برخلاف عملگرهای گمنام‌سازی که در بخش قبل معرفی شد، در این روش مقدار شبه‌شناسه‌ها تغییر نمی‌یابد. این عملگرها از طریق تفکیک صفات مجموعه‌داده به جداول مختلف موجب می‌شوند، مهاجم از روی شبه‌شناسه‌های یک رکورد نتواند به صفت حساس آن دست پیدا کند. در ادامه، تجزیه[1] و برش‌دهی[2] که دو روش شناخته‌شده در این زمینه هستند، معرفی می‌شوند.

۴-۲-۱- تجزیه

تجزیه [۱۹] مجموعه‌داده اولیه را به دو جدول تقسیم می‌کند که یکی از آنها شامل صفات حساس (جدول ST) و دیگری دربردارنده شبه‌شناسه‌ها (جدول QIT) است. همانطورکه در شکل (۱۱) قابل مشاهده است، برای مرتبط‌سازی ST و QIT صفت جدیدی به نام "گروه" به هر دو جدول اضافه می‌شود. درواقع پس از تجزیه، به جای اینکه صفات حساس به رکوردها تعلق داشته باشند به گروه‌ها نسبت داده می‌شوند. بدین ترتیب ارتباط مستقیم بین شبه‌شناسه‌ها و صفت حساس یک رکورد شکسته می‌شود.

تجزیه مقدار شبه‌شناسه‌ها را تغییر نمی‌دهد. این ویژگی در کاربردهایی مانند پاسخ‌دهی به پرس‌وجو بسیار با ارزش است. در مقابل، انتشار داده‌ها در دو جدول جداگانه استفاده از روش‌های استاندارد داده‌کاوی را دشوار می‌سازد. همچنین ممکن است در برخی سناریوها صفات حساس به سادگی از شبه‌شناسه‌ها قابل تفکیک نباشد که این امر موجب می‌شود استفاده از روش تجزیه با مشکل مواجه شود [۲۰].

جنسیت	سن	بیماری
مرد	۳۲	دیابت
مرد	۳۵	دیابت
مرد	۳۳	هپاتیت
زن	۳۶	هپاتیت
زن	۴۰	دیابت

جنسیت	سن	گروه
مرد	۳۲	۱
مرد	۳۵	۱
مرد	۳۳	۱
زن	۳۶	۲
زن	۴۰	۲

QIT

گروه	بیماری	تعداد
۱	دیابت	۲
۱	هپاتیت	۱
۲	دیابت	۱
۲	هپاتیت	۱

ST

(شکل-۱۱): تجزیه مجموعه‌داده به دو جدول ST و QIT

۴-۲-۲- برش‌دهی

روش برش‌دهی [۲۱] مجموعه‌داده را به دو صورت افقی و عمودی تقسیم‌بندی می‌کند. در برش‌دهی عمودی، صفت‌هایی که مقدارشان وابستگی زیادی به هم دارند شناسایی شده و در ستونی مجزا قرار می‌گیرند. در برش‌دهی افقی، مجموعه‌داده به زیرمجموعه‌هایی بدون اشتراک از رکوردها تقسیم می‌شود؛ سپس مقادیر موجود در هر ستون از هر زیر مجموعه به‌صورت تصادفی جابه‌جا می‌شوند. شکل (۱۲) نمونه‌ای از عملکرد برش‌دهی را نمایش می‌دهد. "تجزیه" را می‌توان حالت خاصی از برش‌دهی دانست که در آن شبه‌شناسه‌ها در یک ستون و صفات حساس در ستون دیگر قرار می‌گیرند.

با استفاده از این روش سودمندی داده تا حد زیادی حفظ می‌شود، زیرا الگویی که صفات وابسته بر طبق آن در کنار هم قرار می‌گیرند، دست نخورده باقی می‌ماند. این روش برای حفظ حریم خصوصی داده نیز موثر است، زیرا صفاتی که قرار گرفتن آن‌ها کنار هم کم تکرار است و برای شناسایی مالک رکورد مفید می‌تواند باشد، از هم جدا می‌شوند.

سن	جنسیت	کدپستی	بیماری
۲۲	مرد	۴۹۷۰۶	سوه هاضمه
۲۲	زن	۴۷۹۰۶	آنفولانزا
۳۳	زن	۴۷۹۰۵	آنفولانزا
۵۲	زن	۴۷۹۰۵	برونشیت
۵۴	مرد	۴۷۳۰۲	آنفولانزا
۶۰	مرد	۴۷۳۰۲	سوه هاضمه
۶۰	مرد	۴۷۳۰۴	سوه هاضمه
۶۴	زن	۴۷۳۰۴	ورم معده

(سن، جنسیت)	(کدپستی، بیماری)
(۲۲، مرد)	(۴۷۹۰۵، آنفولانزا)
(۲۲، زن)	(۴۷۹۰۶، سوه هاضمه)
(۳۳، زن)	(۴۷۹۰۵، برونشیت)
(۵۲، زن)	(۴۷۹۰۶، آنفولانزا)
(۵۴، مرد)	(۴۷۳۰۴، ورم معده)
(۶۰، مرد)	(۴۷۳۰۲، آنفولانزا)
(۶۰، مرد)	(۴۷۳۰۲، سوه هاضمه)
(۶۴، زن)	(۴۷۳۰۴، سوه هاضمه)

(شکل-۱۲): نمونه‌ای از نحوه عملکرد روش برش‌دهی [۲۱]

[1] Anatomization
[2] Slicing

۴-۳- تغییر مقدار داده‌ها

در این روش‌ها مقادیر داده با مقدار غیر واقعی دیگری جایگزین می‌شوند. تغییر مقادیر باید به‌گونه‌ای انجام گیرد که اطلاعات آماری محاسبه‌شده از داده‌های خام و گمنام‌شده تفاوت چندانی نداشته باشد. در این حالت در سطح رکورد نمی‌توان به داده ها اطمینان کرد، یعنی رکوردها ممکن است مربوط به افراد حقیقی نباشند.

اضافه کردن نوفه به داده‌ها یکی از روش‌های رایج در این زمینه است که برای مخفی‌کردن صفات عددی مورد استفاده قرار می‌گیرد [۲۲]. در این روش، مقدار v با v + r تعویض می‌شود که r یک مقدار تصادفی است. r باید طوری انتخاب شود که هم مقادیر تولیدشده به اندازهٔ کافی با مقادیر اصلی فاصله داشته باشند و هم اطلاعات آماری مانند میانگین صفت دست‌نخورده باقی بماند. تولید داده‌های مصنوعی[1]، روشی دیگر برای مخفی‌نگه‌داشتن مقدار واقعی داده‌ها است. بدین منظور ابتدا مدلی آماری از روی داده‌ها ساخته می‌شود؛ سپس از روی مدل بدست‌آمده تعدادی رکورد تولید شده و به جای داده‌های اصلی منتشر می‌شود [۲۳].

۵- مدل‌های حفظ حریم خصوصی

پیش از اینکه بخواهیم برای حفظ حریم خصوصی تلاش کنیم، باید مشخص کنیم چه زمانی این هدف برآورده می‌شود. تعریف ایده‌آل از حفظ حریم خصوصی هنگام انتشار داده‌ها، در [۲۴] ارائه شده است. بر طبق این تعریف، بعد از آنکه داده‌ها در اختیار مهاجم قرار می‌گیرد، نسبت به زمانی که داده‌ها را در اختیار نداشته است، نباید هیچ دانشی درمورد مالکان رکورد به او اضافه شود. با درنظرگرفتن اطلاعاتی که ممکن است از قبل در اختیار مهاجم باشد، این سطح از حریم خصوصی قابل دستیابی نیست [۲۵] به همین دلیل در بیشتر کارهایی که در حوزه حریم خصوصی داده انجام می‌گیرد از مدل‌های تعدیل یافته‌تری استفاده می‌شود که در ادامه به شرح آن‌ها پرداخته می‌شود.

۵-۱- مدل پیونددهی رکورد

در حملات پیونددهی رکورد مهاجم شبه‌شناسه‌های 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 رخ می‌دهد که مهاجم موفق شود رکورد مربوط به قربانی را به‌طور یکتا شناسایی کند [۱].

در شکل (۱۳) صفات تحصیلات، جنسیت و سن به عنوان شبه‌شناسه‌ها و حقوق به‌عنوان صفت حساس درنظر گرفته شده است. پس از انتشار این جدول اگر مهاجم به دنبال رکوردی با شبه‌شناسه‌های <کارشناسی، مرد، ۲۹> باشد، به‌راحتی می‌تواند آن را پیدا کند، زیرا تنها یک رکورد متناظر وجود دارد و درنتیجه حملهٔ پیونددهی رکورد با موفقیت انجام می‌گیرد.

روش k-anonymity [۶] برای جلوگیری از چنین حملاتی، مجموعه‌داده را به دسته‌های هم‌ارزی با اندازه k، تقسیم می‌کند. در این حالت به‌ازای هر مقدار ممکن از شبه‌شناسه‌ها، یا هیچ رکوردی در مجموعه‌داده وجود ندارد و یا k رکورد(و بیشتر) موجود است، درنتیجه مهاجم حداکثر با درجه اطمینان 1/k می‌تواند رکورد قربانی را شناسایی کند. این روش در قسمت "ب" از شکل (۱۳) برای گمنام‌سازی مورد استفاده قرار گرفته است. همانطورکه دیده می‌شود هر یک از رکوردهای این جدول دست‌کم با یک رکورد دیگر شبه‌شناسه‌های یکسانی دارد. به چنین جدولی 2-anonymous می‌گویند. لازم به ذکر است که در کاربردهای عملی مقادیر بزرگتری برای k در نظر گرفته می‌شود.

در برخی پایگاه داده‌ها مانند فهرست بیماران یک درمانگاه، ممکن است n رکورد متعلق به یک فرد باشد. در چنین مواردی پس از اعمال k-anonymity این رکوردها در

[1] Synthetic Data Generation

یک دسته هم‌ارزی قرار می‌گیرند. در چنین شرایطی مهاجم با درجه اطمینان n/k رکورد قربانی را به‌دست می‌آورد که با نیاز تعریف شده در k-anonymity مطابقت ندارد. برای حل این مشکل می‌توان از مفهومی عمومی‌تر به نام (X,Y)-anonymity [۲۶] استفاده کرد.

در این روش X و Y دو زیر مجموعه از صفات مجموعه‌داده هستند که اشتراکی با هم ندارند. برای برآورده‌شدن حریم خصوصی در (X,Y)-anonymity، هر مقدار در X باید دست‌کم با k مقدار متمایز از Y ارتباط داشته باشد. اگر Y را شناسه و X را مجموعه شبه‌شناسه‌ها درنظر بگیریم، با انتخاب مقدار مناسب برای k می‌توان حداکثر تعداد رکوردهای متعلق به یک فرد در یک دسته هم‌ارزی را کنترل کرد.

الف) مجموعه‌داده خام

تحصیلات	جنسیت	سن	حقوق
کارشناسی ارشد	مرد	۲۶	۱۰۰۰
کارشناسی	مرد	۳۰	۱۰۵۰
کاردانی	مرد	۳۴	۹۰۰
دکترا	مرد	۲۸	۱۱۰۰
کارشناسی	زن	۲۵	۹۵۰
کارشناسی	زن	۲۷	۹۵۰
کاردانی	زن	۲۶	۹۵۰

ب) 2-anonymous

تحصیلات	جنسیت	سن	حقوق
ارشد و بالاتر	مرد	[۲۵-۳۰)	۱۰۰۰
پایین تر از ارشد	مرد	[۳۰-۳۵)	۱۰۵۰
پایین تر از ارشد	مرد	[۳۰-۳۵)	۹۰۰
ارشد و بالاتر	مرد	[۲۵-۳۰)	۱۱۰۰
پایین تر از ارشد	زن	[۲۵-۳۰)	۹۵۰
پایین تر از ارشد	زن	[۲۵-۳۰)	۹۵۰
پایین تر از ارشد	زن	[۲۵-۳۰)	۹۵۰

(شکل-۱۳): جدول 2-anonymous

۵-۲- مدل پیونددهی صفت

همانطورکه گفته شد در حمله پیونددهی رکورد مهاجم در صدد تشخیص رکورد مربوط به قربانی است. اما در بیش‌تر موارد به دست آوردن صفت حساس قربانی، مهاجم را به هدف خود می رساند و نیازی به شناسایی رکورد نیست [۱]. برای مثال در شکل(۱۳) مقدار صفت حساس در سه رکورد آخر برابر است و مهاجم با دراختیارداشتن شبه‌شناسه‌های <کاردانی، زن، ۲۶> بدون آنکه رکورد قربانی را شناسایی کند، از میزان حقوق وی مطلع می‌شود. این مشکل ناشی از عدم تنوع در مقدار صفت حساس در یک دسته هم‌ارزی است.

l-diversity [۲۷] یکی از روش‌های مورد استفاده برای مقابله با حمله پیونددهی صفت است. برای تأمین l-diversity در یک جدول، در هر دسته هم‌ارزی باید دست‌کم l مقدار مختلف برای صفت حساس موجود باشد. این خاصیت سبب می‌شود تا مهاجم بعد از پیداکردن دسته‌ای که قربانی به آن تعلق دارد، با مقادیر یکسانی از صفت حساس مواجه نشود.

اگر در یک دسته هم‌ارزی l مقدار متفاوت، ولی نزدیک به هم وجود داشته باشد باز هم مشکل عدم تنوع بروز پیدا می‌کند. برای مثال قسمت "ب" از شکل (۱۴) مدل 3-diversity را ارضا می‌کند؛ اما مقادیر صفت حساس در نخستین گروه هم‌ارزی بسیار نزدیک به هم هستند. برای رفع این نقص، در [۲۸] روشی به نام t-Closeness ارائه شد که بر طبق آن تفاوت توزیع صفت حساس در هر دسته هم‌ارزی نسبت به توزیع آن در کل جدول باید از حد آستانه t کمتر باشد. این روش از فاصله Earth Mover [۲۹] برای سنجش تفاوت بین توزیع‌ها استفاده می‌کند. در شکل (۱۴) نتیجه گمنام‌سازی یک مجموعه‌داده به روش‌های l-diversity و t-Closeness مقایسه شده است.

برقرارکردن t-Closeness در یک جدول، نیازمند استفاده زیاد از عملگرهای گمنام‌سازی است که باعث کاهش قابل توجه در کیفیت داده‌ها می شود. برای برطرف کردن این ضعف روشی انعطاف‌پذیرتر به نام (n,t)-Closeness [۳۰] ارائه شده است. اگر به‌ازای هر گروه هم‌ارزی E1 مجموعه‌ای از رکوردها با نام E2 موجود باشد که:

- بیش از n رکورد داشته باشد.
- E1 زیرمجموعه آن باشد.
- فاصله توزیع صفت حساس در E1 و E2 کمتر از t باشد.

آنگاه آن مجموعه داده (n,t)-Closeness را ارضا می‌کند. دراین روش این هرچه nکوچکتر درنظر گرفته شود کیفیت داده‌ها کمتر لطمه می‌بیند و در حالتی که n برابر با تعداد کل رکوردها باشد معادل روش t-Closeness است.

الف) مجموعه‌داده خام

	کدپستی	سن	حقوق
۱	۴۷۶۷۷	۲۹	۳۰۰۰
۲	۴۷۶۰۲	۲۲	۴۰۰۰
۳	۴۷۶۷۸	۲۷	۵۰۰۰
۴	۴۷۹۰۵	۴۳	۶۰۰۰
۵	۴۷۹۰۹	۵۲	۱۱۰۰۰
۶	۴۷۹۰۶	۴۷	۸۰۰۰
۷	۴۷۶۰۵	۳۰	۷۰۰۰
۸	۴۷۶۷۳	۳۶	۹۰۰۰
۹	۴۷۶۰۷	۳۲	۱۰۰۰۰

ب) 3-diverse

	کدپستی	سن	حقوق
۱	۴۷۶**	۲*	۳۰۰۰
۲	۴۷۶**	۲*	۴۰۰۰
۳	۴۷۶**	۲*	۵۰۰۰
۴	۴۷۹**	≥ ۴۰	۶۰۰۰
۵	۴۷۹**	≥ ۴۰	۱۱۰۰۰
۶	۴۷۹**	≥ ۴۰	۸۰۰۰
۷	۴۷۶**	۳*	۷۰۰۰
۸	۴۷۶**	۳*	۹۰۰۰
۹	۴۷۶**	۳*	۱۰۰۰۰

پ) 0.167–closeness

	کدپستی	سن	حقوق
۱	۴۷۶۷*	≤ ۴۰	۳۰۰۰
۳	۴۷۶۷*	≤ ۴۰	۵۰۰۰
۸	۴۷۶۷*	≤ ۴۰	۹۰۰۰
۴	۴۷۹۰*	≥ ۴۰	۶۰۰۰
۵	۴۷۹۰*	≥ ۴۰	۱۱۰۰۰
۶	۴۷۹۰*	≥ ۴۰	۸۰۰۰
۲	۴۷۶۰*	≤ ۴۰	۴۰۰۰
۷	۴۷۶۰*	≤ ۴۰	۷۰۰۰
۹	۴۷۶۰*	≤ ۴۰	۱۰۰۰۰

(شکل-۱۴): مقایسه روش‌های l-diversity و t-Closeness

۵-۳- مدل پیونددهی جداول

در برخی شرایط، صرف تشخیص این‌که رکورد قربانی در مجموعه‌داده انتشاریافته وجود دارد یا خیر، می‌تواند نقض حریم خصوصی تلقی شود. به‌عنوان نمونه زمانی که یک بیمارستان داده‌های مربوط به اشخاصی که دچار بیماری خاصی هستند را منتشر می‌کند، پی بردن به هویت هر یک از افراد، نوع بیماری وی را آشکار می‌سازد. در زمینه مدل پیونددهی جداول نسبت به مدل پیونددهی رکورد و صفت، کارهای کمتری انجام شده است که معروف‌ترین آن‌ها δ-Presence است [۳۱].

طبق این مدل، احتمال وجود رکورد قربانی در مجموعه‌داده انتشاریافته باید به سقف δ٪ محدود شود. این مفهوم با مثالی که در شکل (۱۵) آورده شده است بیشتر شرح داده می‌شود. فرض می‌شود جدول P از قبل در دسترس عموم قرار دارد. جدول T زیرمجموعه‌ای از رکوردهای P است که مقدار صفت حساس در آن‌ها یکسان است. این جدول پس از گمنام‌سازی با نام T^* منتشر می‌شود. T^* نسخه گمنام‌سازی شده T است که پس از انتشار آن مهاجم نباید با احتمال بیشتر از d قادر باشد که بگوید یکی از رکوردهای P در T^* وجود دارد. احتمال این‌که هریک از رکوردهای a-f در T^* حضور داشته باشند برابر با – است. این احتمال برای رکوردهای g-i برابر با – است. بنابراین مهاجم حداکثر با احتمال – می‌تواند از وجود یک رکورد دلخواه از P در جدول T^* اطمینان داشته باشد.

P

	نام	کدپستی	سن	ملیت
a	Alice	۴۷۹۰۶	۳۵	ایالات متحده
b	Bob	۴۷۹۰۳	۵۹	کانادا
c	Chris	۴۷۹۰۶	۴۲	ایالات متحده
d	Dirk	۴۷۶۳۰	۱۸	برزیل
e	Eunice	۴۷۶۳۰	۲۲	برزیل
f	Frank	۴۷۶۳۳	۶۳	پرو
g	Gail	۴۸۹۷۳	۳۳	اسپانیا
h	Harry	۴۸۹۷۲	۴۷	بلغارستان
i	Iris	۴۸۹۷۰	۵۲	فرانسه

T^*

	کدپستی	سن	ملیت
b	۴۷*	*	امریکا
c	۴۷*	*	امریکا
f	۴۷*	*	امریکا
h	۴۸*	*	اروپا
i	۴۸*	*	اروپا

(شکل-۱۵): روش δ-Presence

همان‌طورکه گفته شد، در روش δ-Presence مهاجم حداکثر δ٪ از وجود رکورد قربانی در جدول منتشرشده اطمینان دارد، بنابراین احتمال این‌که بتواند با موفقیت حمله پیونددهی رکورد یا صفت را انجام دهد نیز حداکثر δ٪ است؛ لذا این روش برای مقابله با حملات پیونددهی رکورد و صفت نیز کارایی دارد. عیب عمده روش δ-Presence این است که در محاسبه احتمالات فرض می‌شود منتشرکننده داده و مهاجم به یک جدول خارجی مشترک (مجموعه‌داده P) دسترسی دارند. درحالی‌که چنین فرضی در دنیای واقعی زیاد محتمل نیست.

۵-۴- مدل احتمالاتی

در مدل‌های احتمالاتی هدف جلوگیری از پیونددهی رکورد قربانی به یک رکورد، صفت و یا جدول خاص نیست، بلکه اکثر این مدل‌ها تلاش می‌کنند تغییر دانش مهاجم، قبل و بعد از مشاهده داده منتشر شده، به مقدار کوچکی محدود شود.

Differential Privacy [۲۵]یکی از سخت‌گیرانه‌ترین مدل‌های حفظ حریم خصوصی است. بر طبق این مدل، انتشار یا عدم انتشار رکورد قربانی نباید در نتیجه بررسی‌هایی که مهاجم روی مجموعه‌داده انجام می‌دهد تأثیر قابل توجهی بگذارد. به بیان ریاضی، اگر بررسی‌های مهاجم توسط تابع تصادفی[1] F مدل شود، به‌ازای هر مجموعه‌داده D و 'D که در وجود یک رکورد اختلاف دارند باید رابطه زیر برقرار باشد:

[1] Randomized Function

$$\forall S \in Range(F)\ (\Pr[F(D) = S] \leq e^{\varepsilon} \times \Pr[F(D') = S])$$

که s صفت حساس و ε پارامتری است که هرچه کوچک‌تر انتخاب شود، میزان اطلاعات تأثیر یک رکورد بر نتیجه بررسی‌ها کمتر می‌شود. Differential Privacy تعریف محکمی از حریم خصوصی ارائه می‌دهد که تفاوتی میان صفات حساس و سایر صفات قائل نیست. دستیابی به چنین هدفی مستلزم اضافه‌کردن نوفه فراوان به داده‌ها است که موجب کاهش کیفیت مجموعه‌داده می‌شود. در مقاله [۳۲] با درنظرداشتن این مشکل، مدلی تعدیل یافته‌تر معرفی شده است: اطمینانی که مهاجم پس از مشاهده مجموعه‌داده گمنام‌شده نسبت به صفت حساس قربانی بدست‌می‌آورد نباید بیشتر از یک حدس تصادفی باشد.

مدل حفظ حریم خصوصی معرفی شده در [۳۲] (AGF) بر اساس دو مفهوم اطمینان مورد انتظار[1] و اطمینان مشاهده شده[2] شکل گرفته است. اطمینان مشاهده شده یک رکورد، بیان‌گر این است که مهاجم با چه احتمالی می‌تواند صفت حساس آن را به قربانی نسبت دهد. هرچه اطمینان مشاهده‌شده رکوردها کمتر باشد مهاجم با اطمینان کمتری می‌تواند بین صفات حساس و شناسه‌های صریح رابطه برقرار کند، بنابراین باید مقدار آن را به سقف معینی محدود کرد.

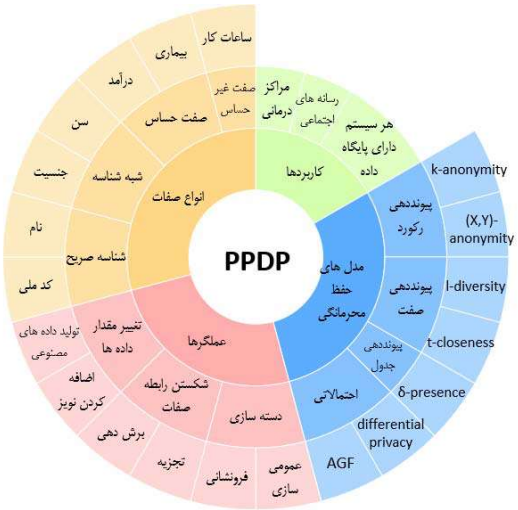
فرض می‌شود مجموعه‌داده شامل n رکورد باشد و D_0 نیز مجموعه‌داده‌ای با اندازه n است که به‌صورت تصادفی از فضای نمونه رکوردها ایجاد شده است. برطبق این مدل، اطمینان مشاهده شده یک رکورد در مجموعه‌داده گمنام شده، نباید از احتمال دیده‌شدن آن رکورد در D_0 بیشتر باشد. این احتمال اطمینان مورد انتظار نامیده می‌شود. برای تأمین حریم خصوصی طبق قاعده ذکرشده، به‌ازای هر رکورد موجود در مجموعه داده، باید رابطه زیر برقرار باشد:

$$\text{Observed Confidence}(r) \leq \text{Expected Confidence}(r)$$

۶- جمع‌بندی و نتیجه‌گیری

پایگاه‌داده سازمان‌ها، درمانگاه‌ها و شرکت‌های بزرگ منابع با ارزشی برای استخراج دانش به وسیله داده‌کاوی هستند. این داده‌ها به‌طورمعمول شامل اطلاعات حساسی از موجودیت‌های سامانه هستند. به‌همین دلیل پیش از این‌که داده‌ها منتشر شوند، باید هرگونه ارتباط صریح بین مالکان رکورد و صفات حساس از آن‌ها حذف شود. PPDP شاخه‌ای از حفظ حریم خصوصی داده است که به بررسی روش‌های انتشار یک مجموعه‌داده گمنام می‌پردازد. در این مقاله فرایند PPDP معرفی شد و برخی از کاربردهای آن ذکر شدند. شکل (۱۶) خلاصه‌ای از مطالب مطرح‌شده را نمایش می‌دهد.

بخشی از شکل به دسته‌بندی عملگرهای گمنام‌سازی و ذکر مثال‌هایی از هر کدام مربوط است. تقسیم رکوردها در گروه‌های هم‌ارزی، شکستن رابطه بین صفات رکوردها و تعویض مقدار داده‌ها با مقادیر غیر واقعی، سه روش عمده عملگرهای گمنام‌سازی هستند. در جدول (۱) مقایسه‌ای بین عملگرهای ذکرشده انجام گرفته است.



(شکل-۱۶): نمای کلی از حفظ حریم خصوصی در انتشار داده‌ها

همان‌طور که گفته شد دست‌یابی به حریم خصوصی مطلق غیر ممکن است، یعنی نمی‌توان کاری کرد که پس از انتشار مجموعه‌داده، مهاجم هیچ‌گونه اطلاعاتی از مالکین رکورد به دست نیاورد. به‌همین دلیل با توجه به کاربرد، مدل‌های مختلفی برای سطوح متفاوت از حفظ حریم خصوصی

[1] Expected Confidence
[2] Observed Confidence

پیشنهاد شده است. در جدول (۲) مدل‌هایی که در این نوشتار معرفی شدند، بر اساس ویژگی‌ها و نقاط ضعف هرکدام مقایسه شده‌اند.

از جمله مباحث باز پژوهشی در حوزه PPDP می‌توان به ارائه مدل‌های حفظ حریم خصوصی جدید، روش‌های جدید گمنام‌سازی که مجموعه‌داده خروجی آن‌ها برای عمل داده‌کاوی خاصی بهینه شده باشد و بررسی چگونگی گمنام‌سازی مجموعه‌داده‌هایی با ساختار متفاوت(داده‌های توزیع‌شده بین چند عضو، پایگاه‌داده گرافی و ...) اشاره کرد. یکی دیگر از مسائل چالش‌برانگیز که کم اهمیت در نظر گرفته شده‌اند، به اطلاعات حساس دست خواهد یافت. اگر تعداد شبه‌شناسه‌ها زیاد انتخاب شود، کیفیت داده‌های خروجی کمترمی‌شود. ارائه راه‌حل در PPDP نحوه انتخاب شبه‌شناسه‌ها است. اگر صفات کمی به‌عنوان شبه‌شناسه انتخاب شوند مهاجم از طریق صفاتی برای مسائلی از این دست، لزوم پژوهش بیشتر در این زمینه را ایجاد می‌کند.

(جدول-۱): مقایسه عملگرهای گمنام‌سازی

نام روش	نحوه عملکرد	رکوردها معرف یک نمونه واقعی هستند؟	قابلیت استفاده از روش‌های استاندارد داده‌کاوی	میزان تغییر مقدار صفات
عمومی‌سازی	تعمیم مقدار صفت به بازه‌ای بزرگتر	بله	بله	نسبتا کم
فرونشانی	حذف مقدار صفت	بله	بله	نسبتا زیاد
تجزیه	جدا کردن صفات حساس از شبه‌شناسه‌ها	تا حدودی	خیر	بدون تغییر
برش‌دهی	جدا کردن صفات مرتبط به هم تعویض مقدار آن‌ها در رکوردهای مختلف	خیر	بله	بدون تغییر
اضافه کردن نوفه	تولید داده‌های جدید به‌وسیله افزودن یک مقدار تصادفی به داده‌های اولیه	خیر	بله	نسبتا کم
تولید داده‌های مصنوعی	تولید داده‌های جدید با توجه به خصوصیات آماری داده‌های اولیه	خیر	بله	کاملا عوض می‌شود

(جدول-۲): مقایسه مدل‌های حفظ حریم خصوصی

نام مدل	رابطه‌ای که در مجموعه‌داده باید برقرار شود	ویژگی ها	نقطه ضعف
k-anonymity	گروه‌بندی رکوردها در دسته‌های هم ارزی به اندازه k	محافظت در برابر پیونددهی رکورد، تخریب کم	چند رکورد از یک فرد در یک دسته هم‌ارزی قرار می‌گیرد
(X,Y)-anonymity	هر مقدار در زیرمجموعه X حداقل با k مقدار متمایز از Y ارتباط داشته باشد	رفع ضعف ذکر شده برای k-anonymity	مقدار صفت حساس در اعضای یک دسته هم ارزی می‌تواند یکسان باشد
l-diversity	وجود l مقدار متنوع از صفت حساس در هر دسته هم‌ارزی	محافظت در برابر پیونددهی صفت، تخریب کم	l مقدار متنوع می‌توانند نزدیک به هم باشند
t-closeness	تفاوت توزیع صفت حساس در هر دسته هم‌ارزی نسبت به توزیع آن در کل جدول از t کمتر باشد	رفع ضعف ذکر شده برای l-diversity	تخریب زیاد
(n,t)-closeness	هر گروه هم ارزی ابرمجموعه‌ای با بیش از n رکورد داشته باشد که فاصله توزیع صفت حساس در آن دو کمتر از t است.	رفع ضعف ذکر شده برای t-closeness	امکان وجود چند صفت حساس در مجموعه‌داده درنظر گرفته نشده است
δ-presence	احتمال وجود رکورد قربانی در مجموعه‌داده انتشار یافته به δ٪ محدود شود	محافظت در برابر پیونددهی رکورد، صفت، جدول	فرض دسترسی منتشرکننده داده و مهاجم به یک جدول خارجی مشترک (P)
differential privacy	انتشار یا عدم انتشار یک رکورد، در نتیجه بررسی‌هایی که روی مجموعه‌داده انجام می‌گیرد تأثیر قابل‌توجهی نداشته باشد	نزدیک‌ترین مدل به تعریف ایده‌آل از حریم خصوصی	تخریب زیاد
AGF	اطمینان مشاهده‌شده رکوردها کوچکتر از اطمینان مورد انتظار باشد	رفع ضعف ذکر شده برای differential privacy	فرض مستقل بودن مقدار صفات

۷- مراجع

[1] B. C. M. Fung, K. Wang, A. Wai-Chee Fu and P. S. Yu, Introduction to Privacy-Preserving Data Publishing: Concepts and Techniques, Chapman and Hall/CRC, 2010.

[2] "Universal Declaration of Human Rights," [Online]. Available: http://www.un.org/en/documents/udhr/.

[3] J. Bennett and S. Lanning, "The Netflix Prize," Proceedings of the KDD Cup Workshop, pp. 3-6, August 2007.

[4] D. Nettleton, "Data Privacy and Privacy-Preserving Data Publishing," in Commercial Data Mining: Processing, Analysis and Modeling for Predictive Analytics Projects, Morgan Kaufmann, 2014, pp. 266-277.

[5] B. Fung, K. Wang and P. Yu, "Privacy-Preserving Data Publishing: A Survey of Recent Developments," ACM Computing Surveys, vol. 42, no. 4, 2010.

[6] L. Sweeney, "k-Anonymity: A Model for Protecting Privacy," International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, vol. 10, no. 5, pp. 557-570, 2002 .

[7] A. Korolova, "Protecting Privacy When Mining and Sharing User Data," PhD thesis, 2012.

[8] K. S. Babu, "Utility-Based Privacy Preserving Data Publishing," PhD thesis, 2013.

[9] "Summary of the HIPAA Privacy Rule," [Online]. Available: http://www.hhs.gov/ocr/privacy/hipaa/understanding/summary/index.html. [Accessed 22 8 2015].

[10] N. Mohammed, B. C. M. Fung, P. C. K. Hung and C. K. Lee, "Healthcare Data: A Case Study on the Blood Transfusion Service," Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 1285-1294, June 2009.

[11] R. Zafarani, M. A. Abbasi and H. Liu, Social Media Mining An Introduction, Cambridge University Press, 2014.

[12] J. Tang, H. Gao and H. Liu, " mTrust: Discerning Multi-Faceted Trust in a Connected World," Proceedings of the fifth ACM international conference on Web search and data mining, pp. 93-102 , 2012.

[13] S. Kumar, X. Hu and H. Liu, "A Behavior Analytics Approach to Identifying Tweets from Crisis Regions," Proceedings of the 25th ACM conference on Hypertext and social media, pp. 255-260, 2014.

[14] D. Horowitz and S. Kamvar, " The Anatomy of a Large-Scale Social Search Engine," In Proceedings of the 19th international conference on World wide web, pp. 431-440 , 2010.

[15] "What happens in an Internet minute?," [Online]. Available: http://www.intel.com/content/www/us/en/communications/internet-minute-infographic.html. [Accessed 24 8 2015].

[16] P. Samarati, "Protecting Respondents' Identities in Microdata Release," IEEE Transactions on Knowledge and Data Engineering, vol. 13, no. 6, pp. 1010-1027, 2001.

[17] B. Fung, K. Wang and P. Yu, "Anonymizing Classification Data for Privacy Preservation," IEEE Transactions on Knowledge and Data Engineering, vol. 19, no. 5, pp. 711-725, 2007.

[18] K. LeFevre, D. J. DeWitt and R. Ramakrishnan, "Incognito: Efficient Full-Domain K-Anonymity," Proceedings of the 2005 ACM SIGMOD international conference on Management of data, pp. 49-60, 2005.

[19] X. Xiao and Y. Tao, "Anatomy: Simple and effective privacy preservation," InProc. of the 32nd International Conference on Very Large Data Bases, pp. 139-150, 2006.

[20] V. S. Susan and T. Christopher, "A Survey on Privacy Preservation in," International Journal of Computer Science and Mobile Computing, vol. 3, no. 3, pp. 188-193, 2014.

[21] T. Li, N. Li, J. Zhang and I. Molloy, "Slicing: A New Approach for Privacy Preserving Data Publishing," IEEE Transactions on Knowledge and Data Engineering, vol. 24, no. 3, pp. 561 - 574, 2012.

[22] R. Brand, "Microdata Protection through Noise Addition," In Inference Control in

ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 414-423, August 2006.

[27] A. Machanavajjhala, D. Kifer, J. Gehrke and M. Venkitasubramaniam, "L-diversity: Privacy beyond k-anonymity," ACM Transactions on Knowledge Discovery from Data (TKDD), vol. 1, no. 1, 2007 .

[28] N. Li, T. Li and S. Venkatasubramanian, "t-Closeness: Privacy Beyond k-Anonymity and l-Diversity," IEEE 23rd International Conference on Data Engineering, pp. 106 - 115, April 2007.

[29] Y. Rubner, C. Tomasi and L. J. Guibas, "The Earth Mover's Distance as a Metric for Image Retrieval," International Journal of Computer Vision, vol. 40, no. 2, pp. 99 - 121, 2000.

[30] N. Li, T. Li and S. Venkatasubramanian, "Closeness A New Privacy Measure for Data Publishing," IEEE Transactions on Knowledge and Data Engineering, vol. 22, no. 7, pp. 943-956, 2010.

[31] M. E. Nergiz, M. Atzori and C. Clifton, "Hiding the Presence of Individuals from Shared Databases," InProc. of ACM International Conference on Management of Data, pp. 665-676, June 2007 .

[32] A. S. Sattar, J. Li, X. Ding, J. Liu and M. Vincent, "A general framework for privacy preserving data publishing," Knowledge-Based Systems, vol. 54, pp. 276–287, 2013.
Statistical Databases, From Theory to Practice, pp. 97-116, 2002.

[23] D. B. Rubin, "Discussion Statistical Disclosure Limitation," Journal of Official Statistics, vol. 9, no. 2, p. 461–468, 1993.

[24] T. Dalenius, "Towards a Methodology for Statistical Disclosure Control," Statistik Tidskrift, vol. 15, p. 429–222, 1977.

[25] C. Dwork, "Differential Privacy," in Automata, Languages and Programming, Springer Berlin Heidelberg, 2006, pp. 1-12.

[26] K. Wang and B. C. M. Fung, "Anonymizing sequential releases," Proceedings of the 12th

رضا ابراهیمی آتانی دانشیار گروه مهندسی رایانه دانشکده فنی و مهندسی دانشگاه گیلان است. نامبرده مدرک دکترای خود را در سال ۱۳۸۹ در رشته مهندسی الکترونیک از دانشگاه علم و صنعت ایران دریافت کرد. ایشان عضو پیوسته انجمن رمز ایران و انجمن های بین‌المللی IACR و IEEE هستند. از ایشان تاکنون چهار عنوان کتاب و بیش از یکصد مقاله در مجلات و کنفرانس های داخلی و بین المللی به چاپ رسیده است. زمینه پژوهشی مورد علاقه وی، طراحی و پیاده سازی الگوریتم‌های رمزنگاری، امنیت شبکه و امنیت نرم افزار است.

مهدی صادق پور متولد ۱۳۶۹، مدرک کارشناسی و کارشناسی ارشد خود را به‌ترتیب در شهریور ماه ۱۳۹۱ و آذر ماه ۱۳۹۳در رشته مهندسی کامپیوتر نرم‌افزار از دانشگاه گیلان دریافت کرده است. زمینه پژوهشی مورد علاقه ایشان مهندسی نرم‌افزار، داده‌کاوی امنیت شبکه است.