تعیین ریسک افشا و انتشار ایمن جداول فراوانی

ربیع‌اله رحمانی
گروه آمار، موسسه آموزش عالی البرز قزوین

سازمان آماری اطلاعات را جمع‌آوری و آنها را به شکل داده منتشر می‌کند. انتشار داده‌ها ممکن است سبب افشای حریم خصوصی پاسخگویان شود. به همین دلیل باید رضایت پاسخگویان از نظر ریسک افشای داده‌ها فراهم شود. همچنین حقوق کاربران در مورد استفاده وسیع از داده‌ها بایستی مورد توجه قرار بگیرد. برای حفظ حقوق پاسخگویان و کاربران، باید بین ریسک افشای داده‌ها و میزان اطلاعات موجود در آنها تعادل برقرار شود. این کار مستلزم اندازه‌گیری ریسک افشای داده‌ها و میزان اطلاعات آنها است. در این مقاله ضمن بیان مقدمات مساله افشا، ریسک افشای جداول فراوانی اندازه‌گیری می‌گردد. همچنین به کمک نرم افزار $\tau - ARGUS$ ریسک افشای یکی از جداول مربوط به داده‌های سرشماری سال ۱۳۸۵ اندازه‌گیری و نحوه انتشار ایمن آن نشان داده می‌شود.

واژه‌های کلیدی: افشا، ریسک افشا، انتشار ایمن.

۱ مقدمه

معمولاً بخشی از اطلاعات جمع‌آوری شده توسط سازمان آماری به جنبه‌هایی از زندگی شخصی یا تجاری پاسخگویان مربوط می‌شود که از نظر آنان این جنبه‌ها محرمانه تلقی می‌شود. با توجه به روند روز افزون بهره‌گیری از اطلاعات جمع‌آوری شده، خطر آشکار شدن اطلاعات محرمانه موجود در آنها نیز افزایش می‌یابد. در مباحث افشای داده‌های آماری، چهار شخصیت گردآورنده اطلاعات مانند سازمان آماری، پاسخگو مانند فرد یا خانوار، کاربر مانند محقق دانشگاهی و متخلف[1] یعنی فردی حقیقی یا حقوقی که بر خلاف قانون در صدد دسترسی به اطلاعات محرمانه است، در نظر گرفته می‌شوند. نکته مهم در انتشار داده‌های آماری، وجود خطر افشای اطلاعات شخصی موجود در داده‌ها توسط متخلف است. این

[1] Intruder

شخص ممکن است علاقه‌مند به دستیابی به اطلاعاتی در مورد افرادی خاص باشد یا تلاش کند تا با افشای یک سری اطلاعات، گردآورنده آنها را بی اعتبار سازد یا از این طریق هوش خود را به اثبات برساند. این عمل به هر دلیل صورت پذیرد، موجب بی اعتمادی عمومی نسبت به سازمان آماری به عنوان حافظ اطلاعات خصوصی پاسخگویان شده و کاهش یا عدم همکاری آنها را در سرشماری‌ها و طرح‌های تحقیقاتی آینده در پی دارد. بنابراین سازمان آماری با این مساله مواجه است که چگونه بین انتشار هر چه وسیع‌تر داده‌های آماری برای استفاده‌های مشروع کاربران و افشای اطلاعات شخصی و محرمانه پاسخگویان از طریق داده‌های منتشر شده، تعادل برقرار نماید. برای برخورد با این مساله معمولاً سازمان‌های آماری از روش‌هایی تحت عناوین دسترسی محدود شده و داده‌های محدود شده استفاده می‌کنند. در شیوه دسترسی محدود شده، روی اینکه چه کسی، چگونه و برای چه هدفی و برای کدام متغیرها به داده‌ها دسترسی پیدا کند، شرایطی گذاشته می‌شود. در روش داده‌های محدود شده، داده‌های آماری تعدیل می‌شوند به گونه‌ای که امکان تجزیه و تحلیل آماری آنها وجود داشته و در عین حال زیانی متوجه پاسخگو نشود. بعضی از سازمان‌ها برای کاهش خطر افشای اطلاعات، از روش‌هایی تحت عنوان فنون محدود سازی ریسک افشا که در اصطلاح آماری تکنیک‌های کنترل افشای آماری[2] (SDC) نامیده می‌شوند، استفاده می‌کنند.

اولین کار جدی در زمینه افشا، سمیناری بود که در سال ۱۹۷۸ توسط کمیته فدرالی روش‌شناسی آماری[3] ($FCSM$) در آمریکا برگزار گردید که مقالات آن از طریق سایت $http://www.fcsm.gov/working-papers/sw۲.html$ قابل دسترسی هستند. از جمله این مقالات می‌توان به دالنیوس (۱۹۷۷) اشاره کرد. همچنین در سال ۱۹۹۲ کمیته‌ای با عنوان ریسک افشا در سازمان برنامه و بودجه آمریکا تشکیل شد. این کمیته در سال ۱۹۹۴ سمیناری برگزار کرد که مقالات آن نیز از طریق سایت $http://www.fcsm.gov/working-papers/spwp۲۲-rev.pdf$ قابل دسترسی هستند که از جمله آنها می‌توان گریفین و همکاران (۱۹۸۹) و لامبرت (۱۹۹۳) را نام برد. اهمیت مساله افشا برای سازمان‌های آماری به حدی زیاد است که این سازمان‌ها کمیته‌های ویژه‌ای برای بررسی این مساله تشکیل می‌دهند. به عنوان مثال، در مرکز آمار آمریکا کمیته‌ای تحت عنوان کمیته اجرایی نظارت داده($DSEP$[4]) به همین منظور تاسیس شده است. در سال‌های اخیر محققین به استفاده از روش‌های رایج آماری برای ارزیابی ریسک افشا روی آورده‌اند. اگانیان و فرر(۲۰۰۳) با استفاده از آنتروپی ریسک افشا را برآورد کردند. الساید(۲۰۰۴) و

[2]Statistical disclosure control

[3]Federal Committee on Statistical Methodology

[4]Data Stewardship Executive Policy Committe

فارستر و وب (۲۰۰۵) با استفاده از روش‌های بیزی ریسک افشا را مورد بررسی قرار دادند.

در بخش دوم این مقاله به نحوه افشای اطلاعات پرداخته می‌شود. در بخش سوم ریسک افشای جداول فراوانی و در بخش چهارم تکنیک‌های SDC مورد بررسی قرار می‌گیرند. در بخش پنجم به کمک نرم افزار $\tau - ARGUS$ ریسک افشای یکی از جداول فراوانی مربوط به داده‌های سرشماری سال ۱۳۸۵ اندازه‌گیری و نحوه انتشار ایمن آن نشان داده می‌شود و در پایان بحث و نتیجه‌گیری ارایه می‌گردد.

۲ نحوه افشای اطلاعات

ممکن است متخلف با ابزارهایی بتواند تشخیص دهد که چه داده‌ای متعلق به چه کسی است و افشا رخ دهد. اطلاعات جمع‌آوری شده در مورد هر پاسخگو شامل مقادیر متغیرهای شناسایی[5] و حساس[6] می‌باشد. متغیرهای شناسایی به دو دسته مستقیم و غیرمستقیم، تقسیم می‌شوند. یک متغیر شناسایی مستقیم، می‌تواند به تنهایی منجر به افشا شود در حالی که متغیر شناسایی غیرمستقیم معمولاً در ترکیب با سایر متغیرها منجر به افشا می‌شود. نام، آدرس، شماره تلفن و شماره ملی از جمله متغیرهای شناسایی مستقیم و سن، جنس، محل سکونت، محل کار، وضعیت شغل و وضعیت تاهل مثال‌هایی از متغیرهای شناسایی غیرمستقیم هستند. به منظور حفظ حریم خصوصی پاسخ دهندگان، سازمان آماری باید از انتشار مقادیر متغیرهای شناسایی مستقیم، خودداری کند. بنابراین در این مقاله منظور از متغیر شناسایی، متغیر شناسایی غیرمستقیم می‌باشد. متغیرهای حساس به متغیرهایی گفته می‌شود که پاسخ دهندگان از افشای مقادیر آنها نگران هستند. سابقه جنایی و درآمد از جمله متغیرهای حساس می‌باشند. در مباحث افشای آماری، پاسخ دهنده‌ای که متخلف به دنبال افشای اطلاعات محرمانه او است، پاسخگوی هدف نامیده می‌شود. اگر متخلف بتواند به درستی داده مربوط به پاسخگوی هدف را تشخیص دهد، آنگاه موفق به افشای اطلاعات محرمانه او نیز می‌گردد.

[5] Identifying variable

[6] Sensitive variable

۳ ریسک افشای جداول فراوانی

جدول فراوانی از طبقه‌بندی افراد جامعه یا نمونه به چندین رده بر حسب دو یا چند معیار تشکیل می‌شود. رده‌ها خانه‌های جدول و معیارها متغیرهای شناسایی و حساس هستند. تعداد معیارها بُعد جدول فراوانی نامیده می‌شود. برای حفاظت بیشتر از داده‌ها باید فرض شود که متخلف از مقادیر همه متغیرهای شناسایی پاسخگوی هدف مطلع بوده و می‌خواهد با استفاده از آنها به مقادیر متغیرهای حساس او پی ببرد. در این مقاله، جدول شامل فقط متغیرهای شناسایی با T^* و جدول متشکل از متغیرهای شناسایی و حساس با T نشان داده می‌شود. بُعد T^* کمتر از بُعد T بوده و به همین دلیل فراوانی خانه‌های T^* بیشتر از فراوانی خانه‌های T می‌باشد. متخلف قبل از انتشار جدول می‌داند که پاسخگوی هدف متعلق به کدام خانه جدول T^* بوده و می‌خواهد این اطلاع را به جدول T گسترش دهد تا از این طریق بتواند مقدار متغیرهای حساس او را فاش کند. بنابراین جدول T^* باید به گونه‌ای تنظیم شود که خانه پاسخگوی هدف در جدول T قابل تشخیص نباشد. مسلماً امنیت جدول T وابسته به فراوانی خانه‌های جدول T^* بوده، به طوری که با افزایش فراوانی خانه مربوط به پاسخگوی هدف در جدول T^*، قدرت تشخیص متخلف کاهش می‌یابد. پاسخ دهنده‌ای که متعلق به خانه‌ای از جدول T^* با فراوانی یک باشد، یکتا گفته می‌شود. فراوانی پاسخگوی یکتا در جدول T نیز یک می‌باشد و اطلاعات محرمانه او به سادگی قابل افشا است. همچنین دو پاسخ دهنده‌ای که متعلق به خانه‌ای از جدول T^* با فراوانی دو هستند می‌توانند از اطلاعات محرمانه یکدیگر مطلع شوند. به همین دلیل معمولاً خانه‌هایی از جدول T^* که فراوانی آنها کمتر از ۳ است، حساس تلقی می‌شوند. انتشار جدول T در صورتی ایمن است که در جدول T^* خانه‌ای حساس وجود نداشته باشد. اگر تعداد خانه‌های جدول T^* با N، تعداد خانه‌های یکتا با N_1 و تعداد خانه‌های با فراوانی دو با N_2 نشان داده شود در این صورت ریسک افشای جدول T با معیار $(\frac{N_1}{N}, \frac{N_2}{N})$ قابل اندازه‌گیری است. برای کاهش ریسک افشا و انتشار ایمن جدول T باید از تکنیک‌های SDC استفاده نمود.

۴ تکنیک‌های SDC برای جداول فراوانی

اگر تعداد خانه‌های حساس جدول زیاد باشد، سازمان آماری می‌تواند با اعمال تکنیک‌های SDC مناسب تعداد آنها را کاهش دهد. در این بخش تکنیک‌هایی که برای کاهش ریسک

افشای جداول فراوانی استفاده می‌شوند، ارایه می‌گردند.

بازطراحی جدول

ممکن است در بعضی از سطرها یا ستون‌های جدول، خانه‌های حساس زیادی وجود داشته باشند. برای کاهش تعداد این خانه‌ها سازمان آماری می‌تواند جدول را بازطراحی[7] کرده، به این معنی که در رده‌بندی متغیرهای رسته‌ای تغییراتی اعمال کند. شکل ساده این تغییرات به صورت ترکیب رده‌های بعضی متغیرها می‌باشد. با استفاده از جدول 1 که مربوط به تعداد کارخانجات به تفکیک نوع و ناحیه فعالیت می‌باشد، این تکنیک بیشتر توضیح داده می‌شود. فراوانی دو خانه از سطر دوم و سه خانه از سطر سوم کمتر از سه می‌باشد بنابراین بیشتر خانه‌های موجود در سطرهای مربوط به فعالیت‌های ۲ و ۳ حساس هستند. یک راه کاهش تعداد این خانه‌ها ترکیب سطرهای ۲ و ۳ می‌باشد. نتیجه این کار در جدول 2 مشاهده می‌شود. اگر تعداد خانه‌های حساس جدول 2 نیز زیاد باشد، می‌توان باز هم جزئیات جدول

جدولِ ۱: تعداد کارخانجات

ناحیه فعالیت / نوع فعالیت	A	B	C	جمع
۱	۲	۵	۶	۱۳
۲	۱	۴	۱	۶
۳	۲	۱	۲	۵
جمع	۵	۱۰	۹	۲۴

جدولِ ۲: بازطراحی

ناحیه فعالیت / نوع فعالیت	A	B	C	جمع
۱	۲	۵	۶	۱۳
۳ و ۲	۳	۵	۳	۱۱
جمع	۵	۱۰	۹	۲۴

را کاهش داد. در حالتی که خانه‌های حساس جدول زیاد نبوده و در سطر یا ستون خاصی متمرکز نباشند، باید از تکنیک‌های SDC موضعی استفاده شود.

[7] Redesign

پنهان سازی خانه‌ای

یکی از رایج‌ترین روش‌های کاهش ریسک افشای داده‌های جدولی، روش پنهان سازی خانه‌ای[8] است. در این روش مقدار خانه حساس از جدول حذف شده و به جای آن از یک علامت مانند $\times$ استفاده می‌شود. این عمل را پنهان سازی خانه‌ای مقدماتی و خانه‌های پنهان شده را پنهان شده‌های مقدماتی می‌نامند. برای توضیح بیشتر، جدول 2 را در نظر بگیرید. خانه متناظر با فعالیت ۱ و ناحیه A حساس می‌باشد. جدول 3 نتیجه پنهان سازی این خانه را نشان می‌دهد. پنهان سازی فقط خانه حساس کافی نیست، زیرا همان طور که در جدول

جدولِ ۳: پنهان سازی مقدماتی

ناحیه فعالیت / نوع فعالیت	A	B	C	جمع
۱	$\times$	۵	۶	۱۳
۲و۳	۳	۵	۳	۱۱
جمع	۵	۱۰	۹	۲۴

3 دیده می‌شود، اگر جمع حاشیه‌ای معلوم باشد مقدار عددی خانه حساس که با علامت $\times$ مشخص شده است، به دست می‌آید. اولین پیشنهاد برای رفع این مشکل، می‌تواند حذف مجموع‌های کناری باشد. اما این کار باعث از دست رفتن اطلاعات زیادی شده و تحلیل جدول را برای کاربران مشکل می‌سازد. راه حل بهتر، پنهان کردن خانه‌های غیرحساس است. این عمل پنهان سازی مکمل و خانه‌های مذکور را پنهان شده‌های مکمل می‌نامند. یک نمونه از پنهان سازی مکمل برای جدول 3 در جدول 4 انجام گرفته است. محاسبه دقیق مقدار خانه حساس با استفاده از جدول 4 بعید به نظر می‌رسد. انتخاب خانه‌های غیرحساس برای پنهان سازی مکمل، باید به گونه‌ای باشد، که حجم اطلاعات از دست رفته مینیمم شود. علاوه بر این خانه‌های حساس باید با انتخاب پنهان شده‌های مکمل مناسب، به صورت رضایت بخشی حفاظت شده و در دامنه‌های کوچک قابل برآورد نباشند.

[8] Cell suppression

جدولِ ۴: پنهان سازی مکمل

ناحیه فعالیت / نوع فعالیت	A	B	C	جمع
۱	×	×	۶	۱۳
۲و۳	×	×	۳	۱۱
جمع	۵	۱۰	۹	۲۴

۵ مثال کاربردی

داده‌های این مثال قسمتی از داده‌های سرشماری سال ۱۳۸۵ و مربوط به مشخصات افراد بالای ده سال ساکن در یک بخش از یک شهرستان می‌باشند. جدول T از متغیرهای شناسایی جنس، محل تولد، وضع سواد، وضع فعالیت، وضع شغل و متغیر حساس وضعیت زناشویی تشکیل شده است. سطوح این متغیرها و کدهای متناظر آنها عبارتند از "جنس: مرد=۱ ، زن=۲"، "محل تولد: همین شهر یا آبادی=۱، شهر دیگر=۲، آبادی دیگر=۳، خارج از کشور=۴"، "وضع سواد: باسواد=۱، بی‌سواد=۲"، "وضع فعالیت: در ۷ روز گذشته حداقل یک ساعت کار کرده است=۱، دارای شغلی است ولی در ۷ روز گذشته به دلایلی کار نکرده است=۲"، "وضع شغل: کارفرما=۱، کارکن مستقل=۲، مزد و حقوق بگیر بخش عمومی=۳، مزد و حقوق بگیر بخش خصوصی=۴، کارکن بدون مزد=۵"، "وضعیت زناشویی: دارای همسر=۱، بی‌همسر بر اثر فوت همسر=۲، بی‌همسر بر اثر طلاق=۳، هرگز ازدواج نکرده=۴، اظهار نشده=۵".

تعداد خانه‌های جدول T^* برابر ۱۶۰ است. با بکارگیری نرم افزار $\tau - ARGUS$ مشخص شد که در این جدول ۱۷ خانه یکتا و ۷ خانه با فراوانی دو وجود دارند یعنی ریسک افشای این جدول برابر $(\frac{۱۷}{۱۶۰}, \frac{۷}{۱۶۰})$ می‌باشد. اگر جدول بدون ایمن سازی منتشر شود وضعیت زناشویی افراد موجود در این خانه‌ها فاش می‌گردد. به همین دلیل در جدول T^* بایستی تکنیک‌های SDC مناسب اعمال شوند. این کار مستلزم اطلاع از وضعیت خانه‌های حساس می‌باشد. در سطوح اول و دوم متغیر جنس به ترتیب ۱۵، ۹، در سطوح اول تا چهارم متغیر محل تولد به ترتیب ۶، ۹، ۷، ۲، در سطوح اول و دوم متغیر وضع سواد به ترتیب ۱۳، ۱۱، در سطوح اول و دوم متغیر وضع فعالیت به ترتیب ۱۱، ۱۳ و در سطوح اول تا پنجم متغیر وضع شغل به ترتیب ۵، ۸، ۶، ۵، ۰ خانه حساس وجود دارند. با توجه به اینکه متغیرهای جنس، وضع سواد و وضع شغل فقطدارای دو سطح هستند لذا ترکیب سطوح آنها باعث از

بین رفتن اطلاعات زیادی شده و به همین دلیل اعمال تکنیک بازطراحی به صورت ترکیب سطوح آنها مناسب نمی‌باشد. اما تکنیک بازطراحی به صورت ترکیب سطوح دوم و سوم محل تولد، ترکیب سطوح اول و دوم وضع فعالیت و همچنین ترکیب سطوح سوم و چهارم وضع شغل مفید به نظر می‌رسد. با اعمال این تکنیک تعداد خانه‌های جدول از ۱۶۰ به ۷۲، تعداد خانه‌های حساس از ۲۴ به ۷ و ریسک افشا از $(\frac{۱۷}{۱۶۰}, \frac{۷}{۱۶۰})$ به $(\frac{۵}{۷۲}, \frac{۲}{۷۲})$ کاهش می‌یابد. با توجه به تعداد کم خانه‌های حساس، اعمال تکنیک پنهان سازی خانه‌ای به منظور دستیابی به T^* کاملاً ایمن می‌تواند مفید باشد. همانطور که در بخش قبل بیان شد اعمال این تکنیک مستلزم پنهان کردن فراوانی پنهان شده‌های مقدماتی و مکمل است. پنهان شده‌های مقدماتی همان ۷ خانه حساس باقیمانده می‌باشند در حالیکه پنهان شده‌های مکمل خانه‌های غیر حساسی هستند که فراوانی آنها به منظور برآورد نشدن فراوانی پنهان شده‌های مقدماتی، پنهان می‌گردد. تعداد این خانه‌ها به ترتیب برابر ۷ و ۶ می‌باشد. با اعمال این تکنیک ریسک افشای جدول T^* برابر $(۰, ۰)$ می‌شود و این جدول کاملاً ایمن می‌گردد. شکل ایمن T^* مطابق جدول 5 می‌باشد که در آن از علامت $\times$ برای پنهان کردن فراوانی پنهان شده‌های مقدماتی و مکمل و از علامت خط تیره برای نشان دادن فراوانی صفر استفاده شده است. با اضافه کردن متغیر حساس وضعیت زناشویی به جدول 5، جدول T ایمن به دست می‌آید. با انتشار این جدول متخلف قادر به افشای وضعیت زناشویی افراد نیست. برای روشن شدن موضوع یک فرد هدف را با مشخصات جنس=۲، وضع سواد=۱، وضع فعالیت=۱، محل تولد=۱، وضع شغل=۳، وضعیت زناشویی=۳، در نظر بگیرید. با انتشار جدول T ایمن وضعیت زناشویی این فرد فاش نمی‌گردد زیرا متخلف فقط از مقدار متغیرهای شناسایی فرد هدف اطلاع دارد و همانطور که در جدول 5 ملاحظه می‌شود ۳۳ پاسخ دهنده دیگر با مشخصات شناسایی مشابه وجود دارند. حال فرض کنید مشخصات فرد هدف به صورت جنس=۱، وضع سواد=۱، وضع فعالیت=۲، محل تولد=۴، وضع شغل=۴، وضعیت زناشویی=۴، باشد. علیرغم وجود حداکثر یک پاسخگوی دیگر با مشخصات مذکور، وضعیت زناشویی فرد هدف فاش نمی‌شود زیرا همانطور که در جدول 5 ملاحظه می‌گردد در خانه مربوط به فرد هدف تکنیک پنهان سازی اعمال شده است. بنابراین با انتشار جدول T ایمن، وضعیت زناشویی همه پاسخ دهندگان حفظ خواهد شد.

جدولِ ۵: T^* کاملاً ایمن

جنسیت	وضع سواد	فعالیت	محل تولد	شغل			مجموع
				۱ و ۲	۳ و ۴	۵	
مرد	با سواد	۱	۱	۲۸۵	۳۴۰	۱۰	۶۳۵
			۲ و ۳	۲۵۲	۵۱۰	_	۷۶۲
			۴	۴	۵	_	۹
		۲	۱	۹۳	۸	_	۱۰۱
			۲ و ۳	۲۹	۵	_	۳۴
			۴	_	×	_	×
	بی سواد	۱	۱	۷۰	۱۲	_	۸۲
			۲ و ۳	۶۵	۴۳	_	۱۰۸
			۴	_	_	_	_
		۲	۱	×	×	_	۷۴
			۲ و ۳	۴۶	۴	_	۵۰
			۴	×	_	_	×
زن	با سواد	۱	۱	۱۱	۳۴	۳	۴۸
			۲ و ۳	۳۴	۴	_	۳۸
			۴	_	_	_	_
		۲	۱	_	_	_	×
			۲ و ۳	×	×	_	_
			۴	_	_	_	×
	بی سواد	۱	۱	×	_	_	×
			۲ و ۳	×	_	_	_
			۴	_	_	_	_
		۲	۱	_	_	_	_
			۲ و ۳	_	_	_	_
			۴	_	_	_	_
مجموع				۹۳۶	۹۹۹	۱۳	۱۹۴۸

۶ بحث و نتیجه گیری

محصول هر سرشماری داده‌های حاصل از آن می‌باشند که توسط سازمان‌ها و مراکز دولتی و خصوصی به کار گرفته می‌شوند. در انتشار داده‌های سرشماری، حقوق پاسخگویان از نظر حفظ اطلاعات محرمانه و حقوق کاربران از لحاظ کیفیت داده‌ها باید حفظ شود. این امر با انتشار داده‌های ایمنی که حاوی بیشترین اطلاعات می‌باشند، میسر می‌گردد. بنابراین تعیین ریسک افشا و ایمن سازی داده‌ها، انتشار آنها را ممکن ساخته و سبب به ثمر رسیدن تلاش‌ها و هزینه‌های صرف شده در سرشماری می‌گردد. در این مقاله ریسک افشای یکی از شکل‌های انتشار داده‌ها یعنی جدول فراوانی مورد بررسی قرار گرفت. بررسی ریسک افشای جداول مقداری[9] و جداول مربوط به داده‌های حاصل از نمونه‌گیری می‌تواند موضوع تحقیقات بیشتر در این زمینه باشد. همچنین اندازه‌گیری میزان اطلاعات موجود در داده‌ها و انتخاب شکلی از داده‌ها با بیشترین اطلاعات و کمترین ریسک افشا موضوع‌های جالب دیگری در ارتباط با مساله افشا می‌باشند.

مراجع

[1] Dalenius, T. (1977), Towards a Methodology for Statistical Disclosure Control, Statistisk Tidskrift, 5, 429-444.

[2] Elsayed A. H. (2004), Analysis of Reidentification Risk Based on Loglinear Models, In Privacy in Statistical Databases, J Domingo-Ferrer and V Toora (Eds), 247-261, Springer Lecture, Notes in Compute Science, 3050, Berlin.

[3] Forester, J.J. and Webb, E.L. (2005), Bayesian Model Averaging for Disclosure Risk Assessment, Working Paper, University of Southampton.

[4] Griffin, R.A., Navarro, A., and Flores-Baez, L. (1989), Disclosure Avoidance for the 1990 census, Proceedings of the Section

[9] Magnitude tables

on Survey Research Methods, American Statistical Assocition, Alexandria,VA,p.516-521.

[5] Lambert, D. (1993), Measures of Disclosure Risk and Harm, Journal of Official Statistics, 9, 313-331.

[6] Oganian, A., Domingo-Ferrer, J. (2003), A Posteriori Disclosure Risk Measure for Tabulare Data Based on Conditional Entropy, SORT-Statistics and Operation Research Transactions, 27,pp.175-190.