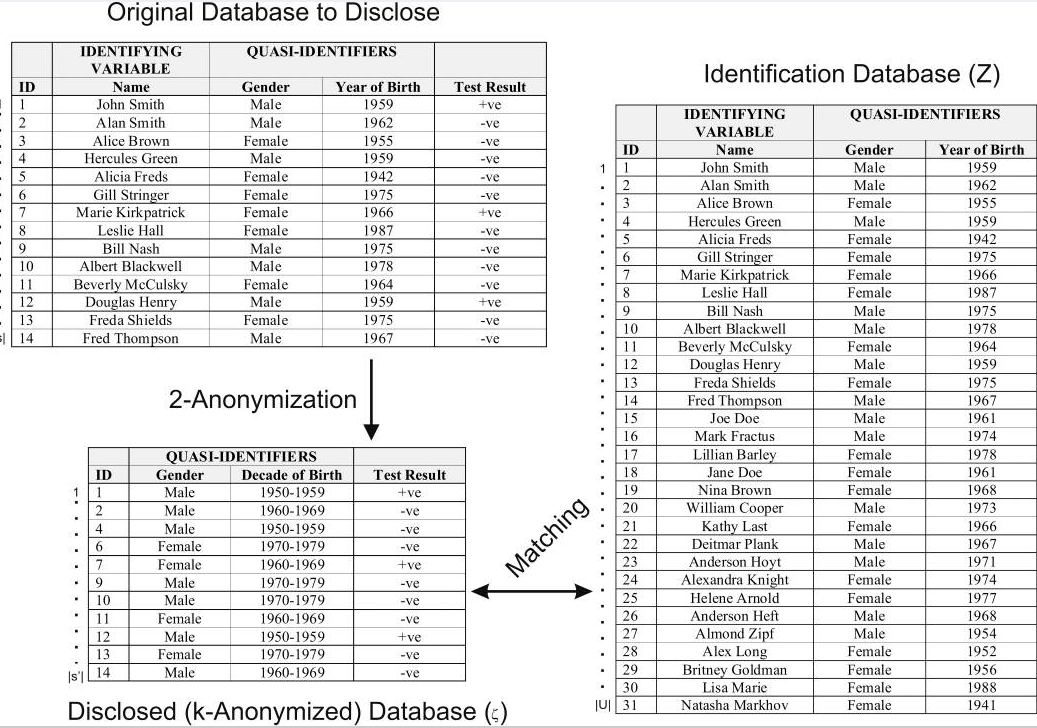
بسم تعالی

مسئله پنهان سازی اطلاعات حساس در حالی که اجازه پرس و جو بر پایگاه داده وجود دارد به طور گسترده ای دز دهه هفتاد میلادی مورد بحث قرار گرفته تور دالنیوس ادعا می کند که برای حفظ حریم خصوصی پرس و جوی پایگاه داده نباید اطلاعاتی را در مورد فرد مشخص کند که بدن دسترسی به پایگاه به دست نمی آید[1]. محدود کردن پرس و جوها و افزودن آشفتگی به ورودی و خروجی از روش های مورد بحث در این دوره به طور موثر توانایی مخفی ساختن اطلاعات خصوصی را در حین فراهم کردن امکان دسترسی به اطلاعاتی که عمومی هستند ندارند.[2].به طوری که تمام پایگاه داده می تواند با نشر تعدادی کمی از پرس و جو ها مشخص شود. این مشکل با وجود سلسله مراتب و سطوح دسترسی در سیستم های مدیریت داده نیز تا حدودی حفظ می شود و همچنان امکان نشت اطلاعات موجود است[3].

با اینکه اضافه کردن نویز به داده ها گزینه مناسبی نیست اما حذف یا کور ساختن و رمز­گذاری ستون های کلید مانند نام افراد، شماره ملی و... همچنان می تواند به حفظ فضای خصوصی افراد کمک کند مشکل این روش در عصر حاضر وجود مجموعه داده های دیگری است که می تواند به نحو موثری اطلاعات حذف نشده را که برای کار داده کاوی مورد نیاز می باشند را به اطلاعات خصوصی فرد مرتبط سازد[2]. در ادامه این نوشتار چند روش برای بهبود گمنامی روش های موجود می پردازیم.

1- K-گمنامی[4]:

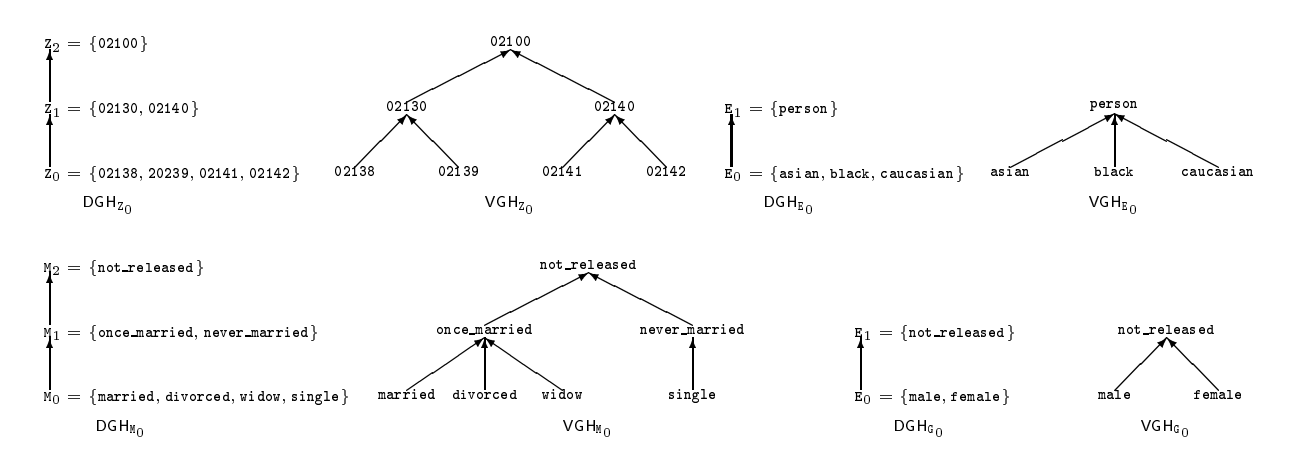
k-گمنامی یک ویژگی از مجموعه داده­ها است که در سال 1988معرفی گردید [4] هدف توزیع یک زیر مجموعه از داده­ها است که نمی­توان از آن برای شناسایی افراد استفاده کرد. با این منظور هر فرد در این توزیع حداقل با k-1 مشابهت دارد.متد­های دست یابی به k-گمنامی شامل حذف تاپل ها و کاهش دانه بندی است. برای مثال در شکل زیر از داده اصلی ستون نام حذف شده (بر اثر اافزایش به حداکثر دانه بندی) و دانه بندی سن کاهش داده شده تا برای هر دوتایی سال تولد و جنسیت حداقل دو فرد مشابه موجود باشند. پس در این نسبت به سال تولد و جنسیت 2-گمنامی داریم.



شکل 1 نتایج تشخیص بیماری در افراد بالا سمت چپ تلاش با فرض به دست آوردن مجموعه داده Z فرد مخل قصد مشخص کردن هویت افراد در مجموعه داده ناشناس را دارد.

1-1-کاهش دانه بندی:

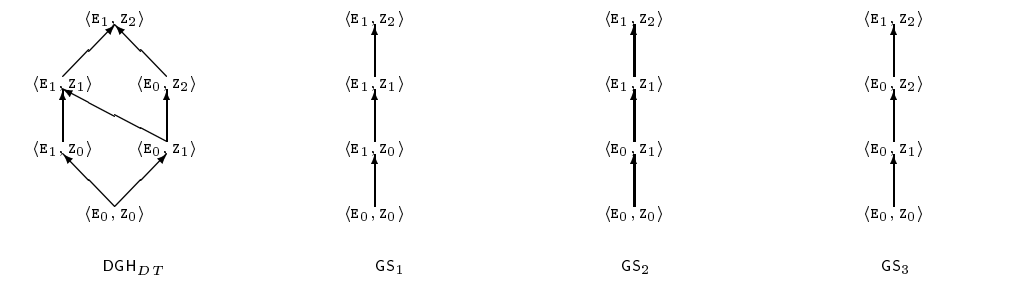
در یک پایگاه داده از دامنه برای مشخص کردن مقادیر ممکن هر فیلد استفاده می شود در نظر می گیریم که برای یک فیلد می توان یک دامنه با دقت کمتر و دانه بندی درشت­تر استفاده کرده به طوری که با یک رابطه ترتیب جزئی دامنه اولیه به دامنه های جدید نگاشته شود.



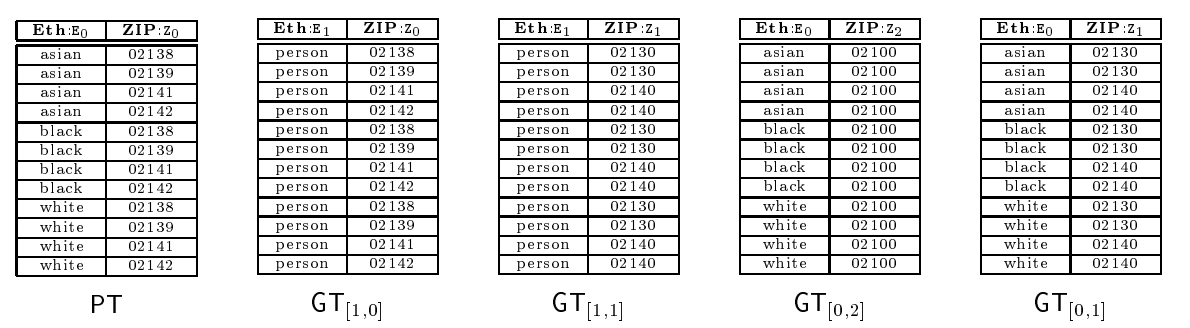
شکل 2 دامنه نمایش چند دامنه و گسترش یافته آنها برای مثال Z0 دامنه اصلی و Z1 با حذف یکان مقادیر دانه­بندی را افزایش داده

مجموعه ی دامنه های گسترش یافته برای یک تاپل از خصوصیات به صورت ضرب دکارتی مجموعه دامنه های گسترس یافته برای تک تک خصوصیات قابل بیان است

برای مثال اگر E و Z دو میدان باشند تاپل های مقادیر آنه می توانند مانند جداول شکل 4 باشند و سلسله مراتبی مانند شکل 3 داشته باشند.



شکل 3 سلسله مراتب گسترش میدان های E و Z



شکل 4

گسترش یافته یک جدول را به صورت زیر تعریف می کنیم:

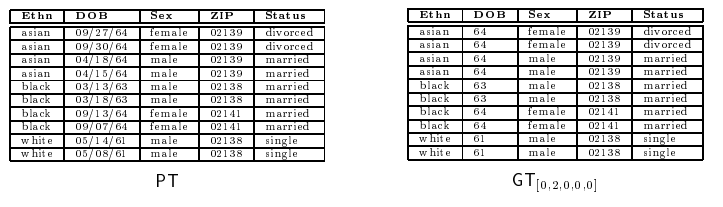
Ti گسترش یافته جدول Tjاست اگر و تنها اگر

1. |Ti|=|Tj|
2. دامنه Ti گسترش یافته از دامنه Tj باشد
3. با یک تابع یک به یک تاپل های دو جدول ب طوری که مقادیر تاپل های جدول Ti دقت کمتری نسبت به Tj داشته باشند.قابل ارتباط دهی باشند.

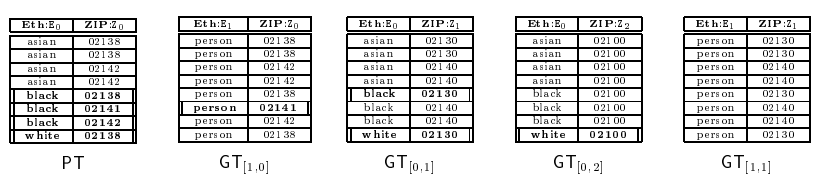
به کمتر گسترش یافته ترین جدولی که شرایط k-گمنامی را داشته باشد جدول گسترش یافته k-حداقلی یک جدول می گوییم.

1-2-حذف داده:

وجود مقادیر پرت در مجموعه داده از تجمیع داده ها جلوگیری می کند. برای مثال در شکل 6 مشاهده می شود در عدم حضور دو تاپل آخر شکل 5می توانیم دانه بندی را بالاتر ببریم.



شکل5 جدول PT و گسترش یافته 2-حداقلی آن GT



شکل 6 چند نمونه از جداول گسترش یافته جدول PT در شکل 5

برای تعمیم تعریف جدول k حداقلی به همراه حذف مقادیر ابتدا یک مقدار بیشینه برای حداکثر تعداد تاپل هایی که می خواهیم حذف کنیم تعریف می کنیم و آن را maxSup می نامیم.گسترش یافته جدول به همراه حذف مقادیر به صورت زیر تعریف می شود

1. |Ti|≤|Tj|
2. دامنه Ti گسترش یافته از دامنه Tj باشد
3. با یک تابع یک به یک تاپل های دو جدول به طوری که مقادیر تاپل های جدول Ti دقت کمتری یا مساوی نسبت به Tj داشته باشند.قابل ارتباط دهی باشند. اما نیازی نیست که تابع مجموعه Tj را بپوشاند.

تعریف گسترش k-حداقلی جدول Tj با مقداری تغییر به صورت زیر خواهد بود

1. Ti k-گمنام باشد
2. |Ti|≥maxSup-|Tj|
3. جدول دیگری با شرایط 1 و2 موجود نباشد که گسترش یافته Tj بوده و مقداری گشترش یافته تر از Ti داشته باشد

در این صورت Ti را گسترش k-حداقلی جدول Tj می گوییم.

مشکلات:

k-گمنامی الگوریتم مناسبی برای گمنام سازی داده ها با ابعاد بالا نیست و همچنین می تواند باعث تغییر نتایج داده کاوی شود حملات همگنی و دانش اولیه می تواند اطلاعات حساس را از داده های گمنام شده با این روش بیرون بکشد. حمله همگنی از یکسانی بین اطلاعات حساس یک فرد با مجموعه ای از k رکورد یکی است، و حمله دانش در صورتی که فرد مخل با استفاده از ارتباط دو یا چند شبه معرف بتواند مجموعه حالات ممکن برای مقادیر حساس را کوچک کند مشکل ساز است.

2\_ I-گستردگی:

منابع:

Dwork C., 2006 Diffrential privacy, Microsoft Research[1]

Irit D., Nissim K.2003 Revealing Information while Preserving Privacy ACM SIGMOD-SIGACT-SIGART  [2]

Pei J.2000 Data Mining: Concepts and Techniques[3]

Samarati P. Sweeney Latanya 1998 Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. Harvard Data Privacy Lab [4]