

# Privacy in Database Publishing in the Presence of Adversary's Background Knowledge

**Fatemeh Amiri**

PhD in Software Engineering; Assistant Professor; Department of Computer Engineering; Hamedan University of Technology; Hamedan, Iran Email: F.amiri@hut.ac.ir

Iranian Journal of  
**Information  
Processing and  
Management**

Received: 02, Dec. 2019 | Accepted: 07, Apr. 2020

**Abstract:** Recent researches show that Data is one of the most valuable and important assets of organizations and businesses. Privacy in the dissemination of data is becoming increasingly challenging. Anonymity as one of the privacy strategies on one side, conceals the relationship between individuals and records in a metadata table and on the other side, preserves the usefulness of the data for subsequent analysis. Preventing information disclosure becomes difficult when the adversary possesses background knowledge. We propose an anonymization framework to protect against background knowledge attack, identity disclosure, and feature disclosure. The anonymization algorithm creates equivalence classes of records whose probability distributions extracted by background knowledge are similar. Our proposed algorithm satisfies k-anonymity and its extension too. The proposed anonymity algorithm tries to satisfy the privacy model while preserving the usefulness of the anonymous data. We verify the theoretical study by experimentation on two datasets. Experimental results show that our proposed algorithm outperforms the state of the art anonymization approaches in terms of loss of information.

**Keywords:** Privacy Preservation, Data Publishing, Background Knowledge, Hierarchical Anonymization Algorithm, Information Loss Metric

Iranian Research Institute  
for Information Science and Technology  
(IranDoc)

ISSN 2251-8223

eISSN 2251-8231

Indexed by SCOPUS, ISC, & LISTA

Vol. 36 | No. 1 | pp. 211-242

Autumn 2020



# حفظ حریم خصوصی در برون‌سپاری داده‌های سامانه‌های اطلاعاتی با تکیه بر سودمندی داده

فاطمه امیری

دکتری مهندسی کامپیوتر؛ استادیار؛ دانشگاه صنعتی  
همدان؛ همدان، ایران | f.amiri@hut.ac.ir



دریافت: ۱۳۹۸/۰۹/۱۱ | پذیرش: ۱۳۹۹/۰۱/۱۹ | مقاله برای اصلاح به مدت ۲۳ روز نزد پدیدآوران بوده است.

نشریه علمی | رتبه بین‌المللی  
پژوهشگاه علوم و فناوری اطلاعات ایران  
(ایرانداک)

شاپا (چاپی) ۸۲۲۳-۲۲۵۱

شاپا (الکترونیکی) ۸۳۳۱-۲۲۵۱

نمایه در SCOPUS، ISI، LISTA و

jipm.irandoc.ac.ir

دوره ۳۶ | شماره ۱ | صص ۲۱۱-۲۴۲

پاییز ۱۳۹۹



چکیده: پژوهش‌های اخیر نشان می‌دهد که داده یکی از ارزشمندترین سرمایه‌های سازمان‌ها و کسب‌وکارهاست. پردازش و تحلیل داده به شرکت‌ها و سازمان‌ها کمک می‌کند که بینش لازم را کسب کرده و از آن در راستای تصمیم‌گیری استراتژیک بهره‌گیرند. انتشار و فراهم کردن دسترسی باز به اطلاعات به یک فرایند متداول و نیاز حیاتی سازمان‌های دولتی و خصوصی تبدیل شده است. داده‌های جمع‌آوری شده در سازمان‌ها حاوی اطلاعات خصوصی افراد است که انتشار آن‌ها می‌تواند منجر به افشای اطلاعات حساس و نقض حریم خصوصی شود. اطلاعات حساس همچنین، اسرار دولتی و اسرار تجاری را نیز شامل می‌شود. چالش اصلی حوزه حفظ حریم خصوصی در انتشار داده، انتشار یک شکل تغییر یافته از داده‌های جمع‌آوری شده است که بتواند حریم خصوصی مالکان داده را حفظ نماید و قابلیت پاسخ به پرس‌وجوها و تحلیل‌های داده‌کاوی را با دقت مناسب داشته باشد. در سناریوی انتشار داده حفاظت از حریم خصوصی مالکان در حضور دانش پیش‌زمینه مهاجم مهمی است. چارچوب گمنامی به‌عنوان یکی از راهبردهای حفظ حریم خصوصی، سودمندی داده را کاهش می‌دهد. در این پژوهش برآنیم یک چارچوب گمنامی برای پیشگیری از حمله دانش پیش‌زمینه، افشای هویت، و ویژگی مالکان طراحی کنیم که سودمندی داده‌های گمنام را بیشینه کند. برای این منظور، بعد از مدل‌سازی دانش پیش‌زمینه مهاجم، مدل حریم خصوصی تعیین و در ادامه، الگوریتم گمنامی ارائه می‌شود. تمرکز این پژوهش بر مدل‌های حریم خصوصی نحوی مانند  $k$ -گمنامی و توسعه‌های آن است. الگوریتم پیشنهادی رکوردها را به چندین گروه افراز می‌کند، به نحوی که در هر گروه مدل حریم خصوصی برآورده می‌شود. نتایج به‌کارگیری چارچوب

پیشنهادی بر روی دو مجموعه داده ارزیابی و تحلیل می‌شود. برای ارزیابی کارایی چارچوب پیشنهادی از معیارهای سودمندی و حریم خصوصی استفاده می‌شود. نتایج آزمایشات نشان می‌دهد که چارچوب پیشنهادی از نظر سودمندی بر الگوریتم‌های ارائه شده در جدیدترین پژوهش‌ها برتری دارد.

**کلیدواژه‌ها:** سامانه‌های اطلاعاتی، برون‌سپاری داده، انتشار داده، مدل حریم خصوصی، الگوریتم گمنامی

## ۱. مقدمه

فناوری اطلاعات و ارتباطات<sup>۱</sup> با فراهم کردن امکان پردازش توزیع شده و تحلیل حجم بالای داده، برون‌سپاری و انتشار داده را به یک فرایند متداول و نیاز حیاتی سازمان‌های دولتی و خصوصی تبدیل کرده است. به دلیل افزایش تقاضای استفاده از پایگاه داده‌های منتشرشده، حوزه انتشار داده توجه زیادی را در سال‌های اخیر به خود جلب کرده است. انجام تعهدات از سوی مؤسساتی که از دولت بودجه دریافت می‌کنند و یا افزایش درآمد برای مؤسسات خصوصی از انگیزه‌های انتشار پایگاه داده‌هاست. به‌عنوان مثال، انتشار لاگ کاربران در موتورهای جست‌وجو، انتشار داده‌های گرافی مانند شبکه‌های اجتماعی، انتشار داده‌های با ابعاد بالا مانند رشته‌های DNA، انتشار اطلاعات اشیای متحرک و انتشار داده‌های بیماران در مراحل مختلف بیماری مورد توجه پژوهش‌های این حوزه است (Zhou and Pei 2011; Li et al. 2012; Ward, Lin and Madria 2020; Fung et al. 2012; Fatahi and Ershadi 2020; Zhao, Pi Chen 2020).

انتشار پایگاه داده از یک سو، اطلاعات غنی برای گیرندگان داده فراهم می‌کند و از سوی دیگر، ممکن است اطلاعات خصوصی افراد هنگام تحلیل توسط گیرندگان داده افشا و حریم خصوصی آن‌ها نقض شود.

گمنامی یکی از راهبردهای حفظ حریم خصوصی است که در سال‌های اخیر به دلیل سادگی و کم‌هزینه بودن مورد توجه قرار گرفته است (Fung et al. 2012). گمنامی در جداول پایگاه داده به معنای حفاظت از داده مورد انتشار بدون نمایش اطلاعات محرمانه فرد یا ارتباط او با رکوردی از پایگاه داده است. ساده‌ترین راه گمنامی در جداول پایگاه

1. information and communication technology (ICT)

داده استفاده از نام مستعار<sup>۱</sup> است. در این روش ویژگی‌های هویتی مانند نام و کد ملی افراد با مقداری مستعار جایگزین می‌شود (Kambourakis 2014). پژوهش‌ها نشان داده‌اند که این روش امن نیست، زیرا مهاجم می‌تواند چند ویژگی در پایگاه داده منتشر شده را با پایگاه داده‌های در دسترس عموم مقایسه و مالکان رکوردها را تعیین هویت کند (Li, Samarati 2001; Li & Venkatasubramanian 2007). چالش حوزه حفظ حریم خصوصی در انتشار داده، انتشار یک شکل تغییر یافته از جداول است که بتواند حریم خصوصی مالکان رکوردها را حفظ کند و قابلیت پاسخ به پرس‌وجوها و تحلیل‌های داده‌کاوی را با دقت مناسب داشته باشد. شایان ذکر است که کاهش دقت تحلیل‌های داده‌کاوی در نتیجه تغییر در داده گمنام به‌عنوان هزینه حفظ حریم خصوصی اجتناب‌ناپذیر است و اتلاف اطلاعات نامیده می‌شود. با کاهش اتلاف اطلاعات، دقت تحلیل‌های داده‌کاوی بر روی داده‌های گمنام و در نتیجه، سودمندی آن‌ها افزایش می‌یابد.

چارچوب‌های گمنامی متعددی برای بیشینه کردن حریم خصوصی و سودمندی در داده‌های منتشر شده ارائه شده است (Sweeney 2002; Cao and Karras 2012; Domingo-Ferrer, Soria-Comas and Mulero-Vellido 2019; Soria-Comas et al. 2015). هر چارچوب گمنامی شامل مدل حریم خصوصی و الگوریتم گمنامی است. مدل حریم خصوصی نیازمندی‌های حریم خصوصی برای انتشار داده را تعیین می‌کند؛ حال آن‌که الگوریتم گمنامی تعیین می‌کند که برای برآورده کردن مدل حریم خصوصی، داده چگونه تغییر کند. مدل‌های حریم خصوصی به دو دسته نحوی<sup>۲</sup> و معنایی<sup>۳</sup> تقسیم‌بندی می‌شوند. مدل‌های نحوی داده را به چندین گروه (به‌نام کلاس هم‌ارزی<sup>۴</sup>) افراز می‌کنند. اولین مدل نحوی  $k$ -گمنامی است که در آن هر رکورد باید حداقل از  $k-1$  رکورد دیگر تفکیک‌ناپذیر باشد (Sweeney 2002). الگوریتم گمنامی برای برآورده کردن  $k$ -گمنامی رکوردها را به گروه‌های  $k$  تایی افراز می‌کند.  $k$ -گمنامی از افشای هویت جلوگیری می‌کند، اما در مقابل افشای ویژگی آسیب‌پذیر است<sup>۵</sup> (Samarati 2001). برای رفع این مسئله توسعه‌های  $k$ -گمنامی مانند  $\beta$ -شباهت (Cao and Karras 2012)،  $l$ -تنوع (Machanavajjhala et al. 2007) و  $t$ -نزدیکی (Li,

1. pseudonym      2. syntactic      3. semantic      4. equivalence class

۵. افشای هویت و افشای ویژگی سطوح مهم افشای اطلاعات هستند (Gardner 2012). در افشای هویت، مهاجم تعیین می‌کند کدام رکورد متعلق به فرد قربانی است. در افشای ویژگی بدون نیاز به تعیین هویت مالکان رکوردها، اطلاعات حساس قربانی افشا می‌شود. به‌عنوان مثال، اگر مقادیر ویژگی حساس در یک کلاس هم‌ارزی مشابه باشد، مهاجم می‌تواند اطلاعات حساس قربانی را افشا کند.

Li & Venkatasubramanian 2007) ارائه شده است. در مدل‌های معنایی برای حفظ حریم خصوصی مالکان، نویز به مقادیر جدول اضافه می‌شود. مدل تفاضلی یکی از مهم‌ترین مدل‌ها در دسته معنایی است (Dwork 2006).

هر مدل حریم خصوصی یک راهکار دفاعی در مقابل یک مدل حمله ارائه می‌دهد. مدل حمله شامل فرضیاتی در مورد دانش پیش‌زمینه مهاجم است. دانش پیش‌زمینه، حقایق شناخته‌شده‌ای هستند که به خودی خود نقض حریم خصوصی محسوب نمی‌شوند، اما در ترکیب با اطلاعات دیگر می‌تواند به استنتاج دقیق‌تر مهاجم درباره اطلاعات حساس قربانی کمک کند و به آن حمله دانش پیش‌زمینه گفته می‌شود. در این پژوهش تمرکز بر روی دانش پیش‌زمینه مهاجم درباره همبستگی بین مقادیر ویژگی‌هاست. فرض کنید یک بیمارستان جهت تحلیل مراحل پیشرفت بیماری‌ها و یا تحلیل اثربخشی یک دارو در توقف یا پیشرفت آهسته بیماری‌ها، اطلاعات بیماران خود را به صورت گمنام شده منتشر می‌کند. دانش پیش‌زمینه در حوزه پزشکی و سلامت می‌تواند همبستگی بین سن و جنسیت (و یا منطقه جغرافیایی افراد) و نرخ ابتلا به بیماری‌های مختلف را نشان دهد. به عنوان مثال، بر اساس مستندات علمی، شیوع برونشیت و آلزایمر در بین زنان بالای ۶۵ سال بیشتر از مردان در همان سن است (National ... Institute 2015).

در این پژوهش ارائه یک چارچوب گمنامی در جداول داده مورد توجه است. علی‌رغم تأمین حریم خصوصی، سودمندی در چارچوب‌های گمنامی پیشین پایین است. از طرف دیگر، حضور دانش پیش‌زمینه مهاجم به سختی مسئله گمنامی می‌افزاید. بنابراین، هدف و انگیزه در این پژوهش ارائه یک چارچوب گمنامی در انتشار داده است که سودمندی داده گمنام را افزایش دهد و به منظور تأمین حریم خصوصی مالکان داده از حمله دانش پیش‌زمینه مهاجم، افشای هویت، و ویژگی پیشگیری کند. با توجه به این مهم، در این پژوهش ابتدا دانش پیش‌زمینه مهاجم مدل‌سازی می‌شود. در ادامه، نیازمندی‌های مدل حریم خصوصی برای پیشگیری از حمله دانش پیش‌زمینه، افشای هویت، و ویژگی تعیین می‌شود و الگوریتم گمنامی جهت برآورده کردن مدل حریم خصوصی و پیشینه کردن سودمندی ارائه می‌گردد. کارایی چارچوب پیشنهادی توسط چندین آزمایش بر روی دو مجموعه داده آزمایشگاهی با روش‌های پیشین مقایسه می‌شود. برای ارزیابی کارایی چارچوب پیشنهادی از معیارهای سودمندی و حریم خصوصی متداول در حوزه گمنامی استفاده می‌شود. به منظور دستیابی به هدف پژوهش، پاسخ به پرسش‌های زیر ضروری

است.

۱. نیازمندی‌های حریم خصوصی در حضور دانش پیش‌زمینه مهاجم چیست؟ مدل حریم خصوصی نحوی است یا معنایی؟
۲. برای جلوگیری از حمله دانش پیش‌زمینه، حمله افشای هویت، و ویژگی، رکوردها چگونه در الگوریتم گمنامی تغییر می‌کنند؟
۳. علاوه بر تأمین حریم خصوصی مالکان داده، برای افزایش سودمندی داده‌های گمنام در کلاس‌های هم‌ارزی چه راه حلی وجود دارد؟
۴. مناسب‌ترین عملگر گمنامی برای حفظ حریم خصوصی و سودمندی داده‌های جدولی در الگوریتم گمنامی چیست؟

**سازماندهی مقاله:** در بخش ۲، مرور ادبیات و پیشینه تحقیق در حوزه گمنامی بررسی می‌شود. روش پژوهش در بخش ۳، و چارچوب گمنامی شامل مدل حریم خصوصی و الگوریتم گمنامی در بخش ۴، تشریح می‌شود. ارزیابی الگوریتم گمنامی و تجزیه و تحلیل نتایج در بخش ۵، بررسی و در پایان در بخش ۶، بحث و جمع‌بندی و کارهای آینده تشریح می‌شود.

## ۲. مرور ادبیات و پیشینه تحقیق

k-گمنامی (Samarati 2001) و توسعه‌های آن (مانند  $\beta$ -شباهت و  $t$ -نزدیکی (Cao and Karras 2012)) مدل‌های نحوی هستند که برای داده‌های جدولی ارائه شده‌اند. در مدل k-گمنامی، احتمال یافتن رکورد یک فرد خاص (قربانی) در هر کلاس هم‌ارزی حداکثر  $\frac{1}{k}$  است. k-گمنامی از افشای هویت<sup>۱</sup> جلوگیری می‌کند، اما در مقابل افشای ویژگی<sup>۲</sup> آسیب‌پذیر است. برای رفع این مشکل، مدل  $\beta$ -شباهت (Cao and Karras 2012) ضمانت می‌کند که اختلاف نسبی احتمال مشاهده هر یک از مقادیر حساس در یک کلاس هم‌ارزی از احتمال مشاهده آن‌ها در کل داده حداکثر  $\beta$  است. لازم به ذکر است که نسخه‌های توسعه یافته مانند  $\beta$ -شباهت، جایگزین k-گمنامی نیستند و برای جلوگیری از افشای هویت و ویژگی به k-گمنامی و توسعه‌های آن به‌طور هم‌زمان نیاز است. چندین پژوهش، ترکیبی از مدل‌های حریم خصوصی را برای جلوگیری از افشای هویت و ویژگی

1. identity disclosure

2. attribute disclosure

ارائه داده‌اند (Amiri et al. 2016; Soria-Comas et al. 2015; Amiri, Yazdani and Shakery 2018). برای تأمین حریم خصوصی مالکان داده لازم است قبل از انتشار تغییر کند. برای این منظور هر الگوریتم گمنامی از عملگرهای گمنامی مانند تعمیم (LeFevre et al. (2006), Soria-Comas et al. (2012), Li et al. (2016), Amiri et al. (2016), Xiao and Tao (2006) و ریزتجمیع Soria-Comas et al. (2019), Domingo-Ferrer, Soria-Comas and Mulero-Vellido (2015) استفاده می‌کنند. عملگر تعمیم به تغییر مقدار یک ویژگی به یک مقدار عام اطلاق می‌شود. به‌عنوان مثال، فرض کنید جدول ۱، داده اصلی جمع‌آوری شده در مورد بیماران یک بیمارستان است. بعد از اعمال الگوریتم گمنامی مبتنی بر تعمیم جدول ۲، تولید می‌شود. در حوزه گمنامی، ویژگی‌های هر جدول به سه گروه شناسه<sup>۱</sup>، شبه‌شناسه<sup>۲</sup> و ویژگی حساس<sup>۳</sup> تقسیم می‌شود. جدول ۱، شامل یک ویژگی شناسه name، سه ویژگی شبه‌شناسه gender, age و zip code و یک ویژگی حساس disease است. بعد از اعمال الگوریتم گمنامی، شناسه حذف و مقادیر شبه‌شناسه تعمیم می‌یابند. در الگوریتم گمنامی مبتنی بر تعمیم مقادیر ویژگی حساس جهت انجام تحلیل‌های داده‌کاوی بدون تغییر منتشر می‌شوند. شایان ذکر است که جدول ۲، مدل ۳- گمنامی را ارضا می‌کند، زیرا اگر مهاجم مقادیر gender, age و zipcode فرد خاصی را بداند، بدون داشتن هیچ دانش اضافی دیگر با احتمال حداکثر ۱/۳ می‌تواند رکورد مربوط به فرد و در نتیجه، بیماری او را بیابد.

جدول ۱. داده اصلی

Name	Age	Gender	Zip	Diseases
Cayla	65	F	12040	Cancer-II
Dior	66	F	12041	GRED <sup>۴</sup>
Elisa	66	F	12041	Depression
Fiona	65	F	12041	Diabetes-II
Ganya	66	F	12041	Flu
Harriet	67	F	12041	Alzheimer-I

1. identifier

2. quasi identifier (QID)

3. sensitive attributes

4. gastroesophageal reflux disease

جدول ۲. داده گمنام‌شده

Equivalence class	Age	Gender	Zip	Diseases
1	[65-66]	F	1204*	Cancer-II
1	[65-66]	F	1204*	GRED
1	[65-66]	F	1204*	Depression
2	[65-67]	F	12041	Diabetes-II
2	[65-67]	F	12041	Flu
2	[65-67]	F	12041	Alzheimer-I

عملگر تعمیم برای داده‌های دسته‌ای مناسب است، اما نمی‌تواند داده‌های عددی را به دقت پردازش کند. با استفاده از این عملگر مقادیر عددی به بازه‌های گسسته تبدیل می‌شود. تعمیم به توزیع مقادیر در هر بازه توجه نمی‌کند و برای پردازش و تحلیل بازه‌ها به ابزارهای خاص نیاز است. وجود بازه‌های هم‌پوشان مانند [۲۰-۱۰] و [۱۵-۵] خوانایی داده منتشرشده را کاهش می‌دهد. از سوی دیگر، عملگر تجزیه منجر به تولید بیش از یک جدول می‌شود که پردازش و تحلیل داده را سخت خواهد کرد؛ حال آن‌که در ریزتجمیع ویژگی ریزدانه‌ای بودن<sup>۱</sup> مقادیر عددی از بین نمی‌رود و طبیعت پیوسته اعداد حفظ می‌شود، اما برای مقادیر دسته‌ای مناسب نیست (Domingo-Ferrer and Torra 2005). مزیت مهم عملگر تعمیم به ریزتجمیع، قابلیت اعتماد داده‌های گمنام توسط تعمیم است. در ریزتجمیع تعدادی از مقادیر ممکن است حذف شود و توزیع داده‌های گمنام با توزیع داده اصلی سازگار نباشد (Amiri et al. 2016). شایان ذکر است که هر یک از این عملگرها جایگاه خود را در پژوهش‌ها دارند.

تحقیق نزدیک به این پژوهش توسط Amiri et al. (2016) ارائه شده است. الگوریتم پیشنهادی مبتنی بر تعمیم و هر دو مدل  $k$ -گمنامی و  $\beta$ -شباهت را برآورده می‌کند. روش پیشنهادی این پژوهش در دو نسخه مختلف ارائه شده است. در اولین نسخه، ابتدا خوشه‌هایی ایجاد می‌شود که  $k$ -گمنامی را برآورده می‌کنند. سپس، خوشه‌ها برای ارضای  $\beta$ -شباهت با یکدیگر ادغام می‌شود. در نسخه دوم، شرط  $\beta$ -شباهت بر  $k$ -گمنامی اولویت دارد. در اولین نسخه از الگوریتم، اتلاف اطلاعات در مقایسه با اتلاف حریم خصوصی کم است، اما در نسخه دوم، اتلاف حریم خصوصی کاهش یافته است. در هر

1. granularity



دو نسخه الگوریتم سودمندی داده پایین است. در تحقیق نزدیک دیگر، Riboni, Pareschi (2012) & Bettini, یک لیست مرتب از رکوردها بر اساس تبدیل ایندکس هیلبرت<sup>۱</sup> ایجاد می‌شود. بر اساس این تبدیل، رکوردها با مقادیر شبه‌شناسه مشابه در لیست نزدیک هم قرار دارند. هر  $k$  رکورد نزدیک که  $t$ -نزدیکی (Cao and Karras, 2012) را برآورده کند به‌عنوان یک کلاس هم‌ارزی منتشر می‌شود. مشکل اصلی الگوریتم پیشنهادی ائتلاف اطلاعات بالا و سودمندی پایین است.

نوع دانش پیش‌زمینه مهاجم در یک چارچوب گمنامی نیز یکی از چالش‌های مهم این حوزه پژوهشی است (Riboni, Pareschi & Bettini 2012). در اکثر مدل‌های حریم خصوصی فرض می‌شود که مهاجم از حضور اطلاعات فرد قربانی در جدول منتشر شده و مقادیر شبه‌شناسه او آگاه است. حال اگر دانش مهاجم توسعه یابد، هیچ‌یک از مدل‌های مذکور نمی‌تواند حریم خصوصی مالکان را تأمین کند. همبستگی داده یکی از انواع دانش پیش‌زمینه است که در حضور آن، تأمین حریم خصوصی سخت است (Li, Li and Zhang 2009; Al Bouna, Clifton & Malluhi 2015; Wang and Liu 2015, 2011).

روش‌های مدل‌سازی دانش مهاجم به دو دسته مبتنی بر منطق (Chen, Ramakrishnan 2007; LeFevre 2007; Martin et al. 2007) و مبتنی بر ابزارهای احتمالاتی (Riboni, Pareschi & Bettini 2012; Li, Li and Zhang 2009) طبقه‌بندی می‌شوند.

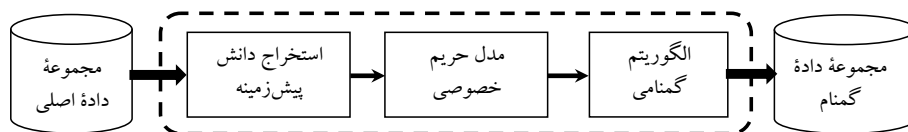
در این پژوهش بر الگوریتم‌های گمنامی مبتنی بر تعمیم به‌دلیل قابل اعتماد بودن داده‌ها تمرکز داریم. برآیند علاوه بر حریم خصوصی، سودمندی داده حاصل از الگوریتم پیشنهادی نسبت به پژوهش‌های قبلی بیشتر شود. شایان ذکر است که در سال‌های اخیر مدل تفاضلی به‌عنوان یک مدل احتمالاتی قوی مورد توجه قرار گرفته است، اما همان‌طور که در پژوهش «کیفر» نشان داده شده، هنگامی که مهاجم به دانش پیش‌زمینه دسترسی دارد مدل مذکور نمی‌تواند مؤثر باشد (Kifer 2009). در نتیجه، تمرکز این پژوهش بر مدل‌های نحوی مانند  $k$ -گمنامی و توسعه‌های آن است.

### ۳. روش پژوهش

پژوهش حاضر از نوع پژوهش‌های کاربردی است. در این پژوهش با مطالعه منابع

1. Hilbert Index

علمی معتبر یک چارچوب گمنامی برای انتشار داده‌های جدولی ارائه می‌شود که از طریق آن بتوان حریم خصوصی و سودمندی را در داده‌های گمنام افزایش داد. شکل ۱، مراحل ایجاد چارچوب گمنامی را نمایش می‌دهد. در این پژوهش فرض می‌شود که مهاجم به دانش پیش‌زمینه درباره همبستگی بین مقادیر ویژگی‌ها دسترسی دارد. ما دانش پیش‌زمینه را به صورت توزیع احتمال تخصیص مقادیر حساس به مالکان مدل می‌کنیم و آن را توزیع دانش پیش‌زمینه می‌نامیم.



شکل ۱. مراحل ایجاد چارچوب گمنامی

بعد از تعیین نحوه مدل‌سازی دانش، توزیع دانش پیش‌زمینه مهاجم استخراج می‌شود. در ادامه، مدل حریم خصوصی و الگوریتم گمنامی برای پیشگیری از حمله دانش پیش‌زمینه، افشای هویت، و ویژگی تعیین می‌شود (شکل ۱). با توجه به حضور دانش پیش‌زمینه، این پژوهش بر روی مدل حریم خصوصی نحوی و الگوریتم گمنامی مبتنی بر تعمیم تمرکز دارد. پیاده‌سازی با استفاده از زبان برنامه‌نویسی C انجام می‌شود. نتایج اجرای چارچوب پیشنهادی بر روی دو مجموعه داده آزمایشی (Adult Dataset (2015) و BKseq Dataset (Riboni, Pareschi & Bettini 2012) BKseq Dataset شامل ۲۴ جدول با ۴۰۰۰ رکورد است، در هر مرحله از آزمایش بر روی این مجموعه داده یک جدول به الگوریتم داده می‌شود. آزمایشات بر روی سیستمی با پردازنده مرکزی Core i5, 2.5 GHZ و حافظه داخلی 8 GB انجام می‌شود. شایان ذکر است که هر آزمایش ده بار تکرار و میانگین نتایج آزمایشات گزارش می‌شود. برای تحلیل نتایج و مقایسه با دو پژوهش نزدیک (Riboni et al. (2012) و Amiri et al. (2016) از معیار خطای قطعیت سراسری<sup>۱</sup> (Ghinita et al. 2007)، اتصال رکورد<sup>۲</sup> (Soria-Comas et al. 2015)، اندازه کلاس‌های هم‌ارزی و زمان اجرای الگوریتم استفاده می‌شود. شایان ذکر است که با توجه به متفاوت بودن مدل حمله در سایر پژوهش‌های حوزه گمنامی، امکان مقایسه

1. global certainty penalty (GCP) 2. record linkage (RL)

چارچوب پیشنهادی با آن روش‌ها ممکن نیست.

### ۳-۱. فرضیات مسئله

در متداول‌ترین حوزه‌های پژوهشی داده‌ها حالت جدولی دارند. هر جدول شامل چند رکورد است و هر رکورد شامل چند ویژگی است. فرض کنید ناشر داده، رکوردهای افراد را در جدول  $T = \{r_1, r_2, \dots, r_n\}$  جمع‌آوری و نسخه گمنام آن  $\bar{T}$  را برای استفاده عموم منتشر می‌کند. هر رکوردی از  $T$  به یک فرد  $v_i$ ، مالک رکورد، منتسب می‌شود و شامل یک ویژگی شناسه،  $D$  ویژگی شبه‌شناسه  $A_1, A_2, \dots, A_D$  و یک ویژگی حساس<sup>۱</sup> به نام  $A_{D+1}$  است.  $[A_j]$  دامنه ویژگی  $A_j$  برای  $1 \leq j \leq D + 1$  است.  $r_i(A_j)$  مقدار ویژگی  $A_j$  از رکورد  $r_i$  و  $r_i(QI)$  مقادیر همه ویژگی‌های شبه‌شناسه رکورد  $r_i$  را نمایش می‌دهد. الگوریتم گمنامی شناسه‌ها را حذف و رکوردها را به تعدادی کلاس هم‌ارزی افزایش می‌کند. در پایان، عملگر تعمیم به مقادیر شبه‌شناسه در هر کلاس هم‌ارزی اعمال می‌شود. در این پژوهش فرض می‌شود: (۱) مهاجم از مجموعه مالکان رکورد و مقادیر شبه‌شناسه آن‌ها مطلع است، و (۲) مهاجم به دانش پیش‌زمینه همبستگی بین مقادیر شبه‌شناسه و مقادیر حساس دسترسی دارد. شایان ذکر است که فرض اول دانش در اکثر پژوهش‌های حوزه گمنامی مشترک است.

### ۳-۲. دانش پیش‌زمینه

این دانش، احتمال اولیه تخصیص مقادیر حساس به یک فرد بر اساس مقادیر شبه‌شناسه‌اش را نمایش می‌دهد. دانش پیش‌زمینه تابعی است به صورت  $PD^{SV}: [QI] \rightarrow \Sigma$  که در آن،  $[QI] = [A_1] \times [A_2] \times \dots \times [A_D]$  مجموعه همه مقادیر ممکن شبه‌شناسه‌ها،  $\Sigma = \{(p_1, p_2, \dots, p_m) \mid p_i \in \mathbb{R}, \sum_{1 \leq i \leq m} p_i = 1, 0 \leq p_i \leq 1\}$  مجموعه همه توزیع‌های احتمال ممکن و  $m$  تعداد مقادیر متمایز در ویژگی حساس  $A_{D+1}$  است. برای مالک  $v$  با مقادیر شبه‌شناسه  $q \in [QI]$ ، دانش پیش‌زمینه به صورت یک توزیع احتمال  $(p_1, p_2, \dots, p_m)$  روی دامنه مقادیر ویژگی حساس مدل می‌شود که در آن  $p_i$  احتمال ابتلای فرد  $v$  به بیماری  $S_i$  به شرط  $q$  است. بنابراین، یک توزیع دانش پیش‌زمینه به هر مالک رکورد در جدول  $T$  نسبت داده می‌شود. شایان ذکر است که این دانش

1. sensitive attribute

از توزیع مقادیر حساس که تناوب هر مقدار حساس در کل جدول را نمایش می‌دهد، متفاوت است. الگوریتم‌های مؤثری برای استخراج دانش بر اساس داده‌های موجود ارائه شده است. در این پژوهش از روش ارائه‌شده در (Riboni, Pareschi & Bettini (2012) برای استخراج دانش پیش‌زمینه استفاده می‌شود. در پیوست ۱، نمادهای ریاضی مورد استفاده در این پژوهش بیان می‌شود.

#### ۴. چارچوب گمنامی پیشنهادی

هر چارچوب گمنامی نحوی قصد دارد عدم قطعیت در تعیین هویت مالکان رکوردها را بیشینه کند. هنگامی که جدول گمنامی منتشر می‌شود، مهاجم تلاش می‌کند مالکان رکوردها را به مقادیر حساس واقعی موجود در جدول نسبت دهد. بنابراین، در هر راه‌حل گمنامی لازم است توانایی مهاجم در متمایز کردن یک انتساب از بین انتساب‌های ممکن را محدود کرد. برای رسیدن به این هدف در حضور دانش پیش‌زمینه می‌توان کلاس‌های هم‌ارزی ایجاد کرد که توزیع دانش پیش‌زمینه مالکان رکوردهای هر کلاس مشابه باشند. از آنجا که ایجاد کلاس‌های هم‌ارزی با رکوردهای دارای توزیع دانش مشابه، دشوار است. این محدودیت در این پژوهش اعمال می‌شود که اختلاف توزیع دانش پیش‌زمینه مالکان رکوردها در هر کلاس هم‌ارزی نباید بیش از حد آستانه تعیین شده<sup>۱</sup> باشد. برای محاسبه شباهت بین توزیع‌های احتمال از معیار «جنسون شانون دایورجنس»<sup>۱</sup> (Lin 1991) استفاده می‌شود. این معیار مقارن است و همیشه مقداری متناهی است. از آنجا که شباهت توزیع‌های دانش پیش‌زمینه به مقادیر شبه‌شناسه و حساس در هر کلاس توجه ندارد، این شرط برای حفاظت از حریم خصوصی و جلوگیری از افشای هویت و ویژگی کافی نیست. بنابراین، مدل حریم خصوصی شامل شرایط و نیازمندی‌های زیر است:

۱. جلوگیری از حمله دانش پیش‌زمینه با محدود کردن اختلاف توزیع دانش پیش‌زمینه در هر کلاس هم‌ارزی حداکثر به اندازه آستانه<sup>۲</sup>؛
۲. جلوگیری از افشای هویت توسط برآورده کردن مدل  $k$ -گمنامی: اندازه هر کلاس هم‌ارزی حداقل  $k$  است؛
۳. جلوگیری از افشای ویژگی توسط برآورده کردن مدل  $\beta$ -شباهت: به ازای هر مقدار

1. Jensen Shannon divergence (JSD)

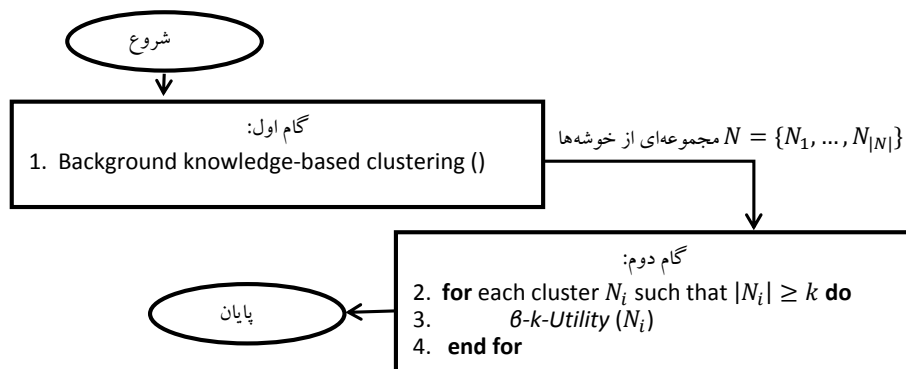
حساس  $S_i$  در کلاس هم‌ارزی  $Q$ ، اگر  $p_i$  و  $q_i$  به ترتیب احتمال مشاهده  $S_i$  در کل جدول و در  $Q$  باشد، آنگاه  $q_i \leq f(p_i) = (1 + \min\{\beta, -\ln p_i\}) \times p_i$ .

الگوریتم پیشنهادی این نیازمندی‌ها را به صورت سلسله‌مراتبی برآورده می‌کند. در مرحله اول از الگوریتم خوشه‌یابی تجمعی<sup>۱</sup> برای پیشگیری از حمله دانش پیش‌زمینه استفاده می‌شود. خوشه‌یابی تجمعی خوشه‌هایی ایجاد می‌کند که در آن‌ها اختلاف توزیع دانش پیش‌زمینه بین هر زوج رکورد کمتر از حد آستانه  $\lambda$  است. سپس، هر خوشه برای برآورده کردن نیازمندی‌های  $\beta$ -شبهت و  $k$ -گمنامی افزایش می‌شود. در پایان، برای افزایش سودمندی در هر کلاس هم‌ارزی، رکوردها بین کلاس‌ها جابه‌جا می‌شوند. این جابه‌جایی نباید نیازمندی‌های حریم خصوصی را نقض کند.

#### ۴-۱. الگوریتم گمنامی

الگوریتم MHA برای برآورده کردن مدل حریم خصوصی پیشنهاد می‌شود. همان‌طور که در شکل ۲، مشاهده می‌شود، الگوریتم پیشنهادی شامل دو گام است:

۱. Background knowledge-based clustering (خط ۱ در شکل ۲): جدول اصلی به تعدادی خوشه افزایش می‌شود، به نحوی که در هر خوشه اختلاف توزیع دانش پیش‌زمینه هر زوج از رکوردها حداکثر  $\lambda$  است.
۲.  $\beta$ -k-Utility (خط ۳ در شکل ۲): برای برآورده کردن  $k$ -گمنامی و  $\beta$ -شبهت، هر خوشه ایجادشده در مرحله قبل به تعدادی کلاس هم‌ارزی افزایش می‌شود. در انتها، برای افزایش سودمندی داده‌ها رکوردها بین کلاس‌های هم‌ارزی جابه‌جا می‌شوند.



شکل ۲. الگوریتم گمنامی MHA

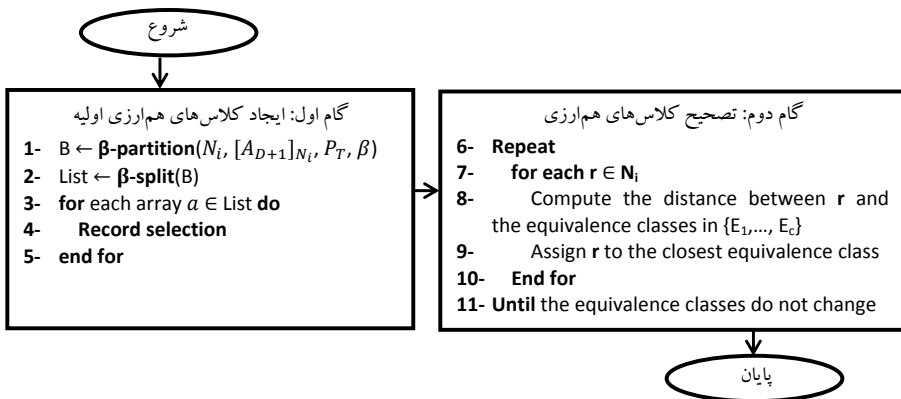
در گام اول، یک دیاگرام درختی<sup>۱</sup> از رکوردها با استفاده از خوشه‌یابی تجمعی ایجاد می‌شود. در ادامه، درخت را از گره‌هایی هرس می‌کنیم که گره مربوطه و همه فرزندان آن دارای اختلاف توزیع دانش پیش‌زمینه کمتر یا مساوی  $k$  باشند. خروجی گام اول مجموعه‌ای از خوشه‌های مجزا  $N = \{N_1, N_2, \dots, N_{|N|}\}$  است. سپس، به‌ازای هر خوشه  $N_i$  که تعداد رکوردهای آن حداقل  $k$  است، الگوریتم  $\beta$ -k-Utility اجرا می‌شود. خوشه‌های کمتر از  $k$  رکورد منتشر نخواهد شد.

الگوریتم  $\beta$ -k-Utility برای برآورده کردن  $\beta$ -شباهت و  $k$ -گمنامی هر خوشه  $N_i$  را به چندین گروه‌افراز می‌کند. سپس، مقادیر ویژگی در هر گروه تعمیم می‌یابد و هر گروه به‌عنوان یک کلاس هم‌ارزی منتشر می‌شود. در ادامه، الگوریتم  $\beta$ -k-Utility با جزئیات تشریح می‌شود.

#### ۴-۲. الگوریتم $\beta$ -k-Utility

الگوریتم  $\beta$ -k-Utility شامل دو گام به نام‌های «ایجاد کلاس‌های هم‌ارزی اولیه» و «تصحیح کلاس‌های هم‌ارزی» است (شکل ۳). در گام اول، رکوردها بر اساس مقادیر حساس به چندین گروه‌افراز می‌شوند که در آن‌ها  $\beta$ -شباهت برآورده شده است. سپس، هر گروه به‌منظور برآورده شدن  $k$ -گمنامی به گروه‌های کوچک‌تر افراز می‌شود. در گام دوم، رکوردها برای افزایش سودمندی بین گروه‌ها جابه‌جا می‌شوند.

1. dendrogram



شکل ۳. الگوریتم  $\beta$ -k-Utility

اگر  $N_i$  مجموعه‌ای از رکوردها،  $P_T$  توزیع مقادیر حساس در جدول اصلی، و  $[A_{D+1}]_{N_i}$  دامنه مقادیر حساس در  $N_i$  باشد، الگوریتم  $\beta$ -k-Utility مجموعه  $[A_{D+1}]_{N_i}$ ،  $P_T$ ،  $N_i$  و  $\beta$  را به‌عنوان ورودی دریافت می‌کند و مجموعه کلاس‌های هم‌ارزی را ایجاد می‌کند. گام اول، شامل سه مرحله است:

۱.  $\beta$ -Partition: (خط ۱ در شکل ۳) در این مرحله مقادیر حساس مشاهده‌شده در رکوردهای  $N_i$  به تعدادی زیرمجموعه تقسیم می‌شود. همه رکوردهایی از  $N_i$  که مقادیر حساس آن‌ها در یک زیرمجموعه قرار گرفته است، یک گروه به‌وجود می‌آورند. در گروه‌های ایجادشده  $\beta$ -شباهت برآورده شده است.

۲.  $\beta$ -Split: (خط ۲ در شکل ۳) هر گروه ایجادشده در مرحله قبل برای کاهش اندازه کلاس‌های هم‌ارزی نهایی و حفظ سودمندی داده‌افراز می‌شود تا هنگامی که  $\beta$ -شباهت و  $k$ -گمنامی نقض نشود.

۳. Record Selection: (خط ۴ در شکل ۳) رکوردهایی از هر گروه برای تشکیل کلاس‌های هم‌ارزی اولیه انتخاب می‌شوند. این انتخاب به‌گونه‌ای انجام می‌شود که اتلاف اطلاعات کاهش یابد.

در  $\beta$ -Partition رکوردها به تعدادی گروه‌افراز می‌شوند. برای حفظ سودمندی داده، افزایی با کمترین تعداد گروه مطلوب است. برای این منظور از روش ارائه‌شده در پژوهش Cao and Karras (2012) استفاده می‌کنیم. سپس، افزاز  $B$  ایجادشده از رکوردها به‌عنوان خروجی به مرحله بعد ارسال می‌شود.

در مرحله بعد، تابع  $\beta$ -split از درخت دودویی برای تعیین اندازه هر کلاس هم‌ارزی

استفاده می‌شود. هر گره از درخت به روش بالا به پایین به دو بخش تقسیم می‌شود. افزاز ایجاد شده توسط مرحله قبل به عنوان ریشه درخت منظور می‌شود. ریشه به صورت  $r = [|b_1|, |b_2|, \dots, |b_{|B|}|]$  نمایش داده می‌شود که  $|b_i|$  تعداد رکوردها در گروه  $b_i$  است. ریشه به دو فرزند افزاز می‌شود. هر  $b_i$  به دو گروه  $b_i^1$  و  $b_i^2$  تقسیم می‌شود، به نحوی که فرزند چپ،  $c_1$ ، شامل  $\lfloor \frac{|b_i|}{2} \rfloor$  رکورد و فرزند راست،  $c_2$ ، شامل  $|b_i| - |b_i^1|$  رکورد است؛ یعنی رکوردهای گره پدر به طور مساوی بین فرزندان تقسیم می‌شود. این تقسیم شدن در صورتی امکان پذیر است که تعداد رکوردهای هر فرزند دست کم  $k$  باشد و هر دو فرزند شرط شایستگی را برآورده کنند. شرط شایستگی در گره  $c_1$  بررسی می‌کند آیا مقادیر حساس در هر گروه  $b_i$ ، یعنی مجموعه  $\{s_b, s_{b+1}, \dots, s_e\}$ ، شرط  $\sum_{j=b}^e p_j^{b_i} \leq f(p_{i_j})$  را برآورده می‌کنند به نحوی که  $p_j^{b_i}$  احتمال مشاهده مقدار حساس  $s_j$  در گره  $c_1$ ،  $p_{i_j} = \min_{s_j \in B_j} \{p_j\}$ ،  $c_1$  و  $p_j$  احتمال مشاهده مقدار حساس  $s_j$  در جدول اصلی است. فرایند تقسیم گره‌های درخت ادامه می‌یابد تا هنگامی که هیچ گرهی را نتوان تقسیم کرد. برگ‌های درخت تعداد رکوردهای هر کلاس هم‌ارزی را مشخص می‌کند. تابع  $\beta$ -split لیستی از برگ‌های درخت را به مرحله بعد ارسال می‌کند. در پیوست شماره ۲، مراحل ساختن لیست برگ‌ها با یک مثال تشریح می‌شود.

در مرحله Record selection (خط ۴ در شکل ۳)، کلاس‌های هم‌ارزی بر اساس برگ‌های ایجاد شده در مرحله قبل تولید می‌شوند. کاهش اتلاف اطلاعات در تشکیل کلاس‌های هم‌ارزی نیز مورد توجه است. برای این منظور نمایه «هیلبرت»<sup>۱</sup> (Ghinita et al. 2007) همه رکوردها محاسبه می‌شود. فرض کنید  $B = \{b_1, b_2, \dots, b_{|B|}\}$  افزاز ایجاد شده توسط  $\beta$ -partition است. رکوردها در هر گروه به صورت صعودی بر حسب نمایه «هیلبرت» مرتب می‌شوند. سپس، به ازای هر برگ ایجاد شده در مرحله قبل یعنی  $a = [a_1, \dots, a_{|B|}]$  به صورت زیر عمل می‌شود: یک رکورد تصادفی  $I$  از گروه  $b_1$  انتخاب می‌شود. سپس، توسط جست‌وجوی دودویی،  $a_1 - 1$  رکورد بعدی نزدیک  $I$  از گروه  $b_1$  انتخاب می‌شود. این رکوردها به کلاس هم‌ارزی تهی  $E_i$  اضافه می‌شوند. در ادامه،  $a_2$  رکورد از گروه  $b_2$  انتخاب می‌شود، به نحوی که نزدیک رکوردهای انتخاب شده در  $E_i$  هستند.

1. Hilbert index



انتخاب رکورد برای همه عناصر آرایه  $a$  در گروه‌های مربوطه ادامه می‌یابد. به ازای هر برگ یک کلاس هم‌ارزی ایجاد می‌شود و مجموعه کلاس‌های هم‌ارزی  $\{E_1, \dots, E_C\}$  به گام دوم ارسال می‌شود.

در گام دوم (خطوط ۷ تا ۱۰ در شکل ۳)، کلاس‌های هم‌ارزی بازبینی می‌شوند. برای این منظور هر رکورد به نزدیک‌ترین کلاس هم‌ارزی (مشابه‌ترین مقادیر شبه‌شناسه) منتقل می‌شود، در صورتی که نیازمندی‌های حریم خصوصی نقض نشود. تا زمانی که تغییرات در کلاس‌های هم‌ارزی وجود دارد، بازبینی و تصحیح آن‌ها ادامه می‌یابد. در پایان، هر گروه به‌عنوان یک کلاس هم‌ارزی منتشر می‌شود. در این پژوهش در خط ۸ از شکل ۳، برای محاسبه فاصله بین یک رکورد  $r$  و کلاس هم‌ارزی  $E_i$  معیار فاصله توسعه‌یافته  $\bar{d}(r, E_i)$  پیشنهاد می‌شود. برای محاسبه  $\bar{d}(r, E_i)$  اگر رکورد  $r$  در حال حاضر عضو کلاس  $E_i$  است و با حذف این رکورد از گروه یعنی  $E_i \setminus \{r\}$ ، بیشینه فاصله نسبی توزیع مقادیر حساس گروه جدید بیشتر از  $\beta$  شود و یا تعداد رکوردهای  $E_i \setminus \{r\}$  کمتر از  $k$  شود، اندازه معیار فاصله توسعه‌یافته صفر است. حال اگر  $r$  در حال حاضر عضو کلاس  $E_i$  نباشد، با افزودن آن به گروه یعنی  $E_i \cup \{r\}$ ، بیشینه فاصله نسبی توزیع مقادیر حساس کلاس  $E_i \cup \{r\}$  بیشتر از  $\beta$  شود، اندازه معیار فاصله توسعه‌یافته بی‌نهایت است. در غیر این صورت فاصله اقلیدسی بین رکورد  $r$  و نماینده کلاس  $E_i$  محاسبه می‌شود. نماینده کلاس میانگین مقادیر هر ویژگی در کلاس است. در صورتی که رکوردی شامل ویژگی‌های عددی و غیر عددی باشد، می‌توان از روش ترکیبی ارائه‌شده در Han et al. (2014) برای محاسبه فاصله اقلیدسی استفاده کرد.

## ۵. ارزیابی چارچوب پیشنهادی

در این بخش نتایج ارزیابی چارچوب پیشنهادی با استفاده از دو مجموعه داده و نسبت به معیارهای متفاوت گزارش می‌شود. در بخش ۵-۱ داده‌های پژوهش، در بخش ۵-۲ معیارهای ارزیابی و در بخش ۵-۳ نتایج تجربی ارزیابی چارچوب پیشنهادی بررسی می‌شود.

### ۵-۱. داده‌های پژوهش

برای ارزیابی چارچوب پیشنهادی از دو مجموعه داده زیر استفاده می‌شود که در

ارزیابی کارهای پیشین در حوزه گمنامی نیز استفاده شده است.

۱. (2015) Adult Dataset، مجموعه داده «مرکز آمار آمریکا» در سال ۱۹۹۴ که به‌طور گسترده برای آزمایشات تجربی گمنامی استفاده می‌شود. این مجموعه داده شامل ۴۵۲۲۲ رکورد آماری با ۶ ویژگی عددی، ۸ ویژگی دسته‌ای و یک ستون برای نمایش میزان درآمد است. در این آزمایشات از سه ویژگی سن، جنسیت و تحصیلات به‌عنوان شبه‌شناسه و میزان درآمد به‌عنوان ویژگی حساس استفاده می‌شود.
۲. (2012) BKseq Dataset, Riboni et al. اطلاعات جمع‌آوری شده در این مجموعه داده بر اساس دانش حوزه و به شیوه مصنوعی (بر اساس مقالات علمی حوزه پزشکی) ایجاد شده است. این مجموعه داده شامل تاریخچه‌ای از ۲۴ جدول و هر جدول شامل ۴۰۰۰ رکورد است. هر رکورد شامل ویژگی‌های سن، جنسیت، وزن و نوع بیماری است. در این رکوردها ۱۹ مقدار حساس از مراحل مختلف بیماری مشاهده می‌شود.

## ۲-۵. معیارهای ارزیابی

به‌منظور ارزیابی کارایی الگوریتم گمنامی معیارهای اتلاف اطلاعات و حریم خصوصی مورد توجه است. الگوریتم گمنامی داده را قبل از انتشار تغییر می‌دهد. هر قدر تغییرات داده در نتیجه گمنامی کمتر باشد، دقت تحلیل بر روی داده گمنام نزدیک به دقت تحلیل بر روی داده اصلی (بدون تغییر) است. در معیار اتلاف اطلاعات میزان تغییرات داده اصلی نسبت به داده گمنام بررسی می‌شود. در اکثر پژوهش‌های حوزه گمنامی از معیار خطای قطعیت سراسری<sup>۱</sup> Ghinita et al. (2007) به‌عنوان معیار اتلاف اطلاعات استفاده می‌شود. معیار خطای قطعیت سراسری از معیار خطای قطعیت نرمال<sup>۲</sup> بهره می‌برد. اگر جدول  $T$  شامل  $D$  ویژگی شبه‌شناسه عددی  $A_1, A_2, \dots, A_D$  باشد، الگوریتم گمنامی مبتنی بر تعمیم رکورد  $r$  را به  $([y_1 - z_1], \dots, [y_D - z_D]) = \bar{r}$  تبدیل می‌کند، به‌نحوی که  $\forall 1 \leq i \leq D, y_i \leq z_i$  است. خطای قطعیت نرمال ویژگی  $A_i$  یعنی  $NCP_{A_i}$  به‌صورت زیر محاسبه می‌شود:

$$NCP_{A_i}(t) = \frac{z_i - y_i}{|A_i|} \quad (1)$$

که در آن،  $|A_i|$  فاصله بین بیشینه و کمینه مقدار در ویژگی  $A_i$  است. خطای قطعیت

1. global certainty penalty (GCP)

2. normalized certainty penalty (NCP)

رکورد  $\Gamma$  به صورت مجموع وزن دار خطای قطعیت نرمال ویژگی‌های رکورد محاسبه می‌شود و خطای قطعیت سراسری یک جدول به صورت مجموع خطای قطعیت نرمال رکوردهاست. مقدار GCP عددی بزرگ‌تر یا مساوی صفر است که صفر نشان‌دهنده حالتی است که هیچ اتلافی رخ نداده و بیشترین سودمندی در داده‌ها حفظ شده است. بنابراین، مقادیر کمتر این معیار مطلوب‌تر است.

اتصال رکورد<sup>۱</sup> (Soria-Comas et al. (2015) یکی از مهم‌ترین معیارهای ارزیابی حریم خصوصی و خطر افشاست. این معیار به عنوان تعداد اتصالات درست بین داده گمنام و داده اصلی تعریف شده است:

$$RL = \frac{\sum_{\Gamma \in T} P_{RL}(\bar{\mathbf{r}})}{n} \quad (2)$$

که در آن،  $n$  تعداد رکوردهای جدول اصلی  $T$  است. برای هر رکورد  $\mathbf{r} \in T$ ،  $P_{RL}(\bar{\mathbf{r}})$  احتمال اتصال نسخه گمنام رکورد  $\mathbf{r}$ ، یعنی  $\bar{\mathbf{r}}$  است. احتمال اتصال رکورد  $\bar{\mathbf{r}}$  به صورت زیر محاسبه می‌گردد:

$$P_{RL}(\bar{\mathbf{r}}) = \begin{cases} \frac{1}{|E_i|} & \text{if } \mathbf{r} \in E_i \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

که در آن،  $E_i$  نزدیک‌ترین کلاس هم‌ارزی به  $\Gamma$  است. اگر  $\bar{\mathbf{r}}$  هم در  $E_i$  قرار گیرد، احتمال اتصال رکورد برابر با معکوس اندازه کلاس هم‌ارزی است. در غیر این صورت، صفر منظور می‌شود. مقادیر کمتر این معیار مطلوب‌تر است.

### ۳-۵. نتایج تجربی ارزیابی چارچوب پیشنهادی

در این بخش نتایج چارچوب پیشنهادی از سه دید ارزیابی می‌شود: نمودار خطر افشا-اتلاف اطلاعات، اندازه کلاس هم‌ارزی و زمان اجرا. کارایی چارچوب پیشنهادی با دو روش اخیر به نام‌های  $\beta$  - likeness - primacy (Amiri et al. (2016) و الگوریتم مبتنی بر دانش پیش‌زمینه (Riboni et al. (2012) مقایسه می‌شود که به ترتیب، به صورت  $\beta$ -GE و BKA در نمودارها و جدول‌ها نمایش داده می‌شود. دو الگوریتم مورد مقایسه مبتنی بر تعمیم هستند.

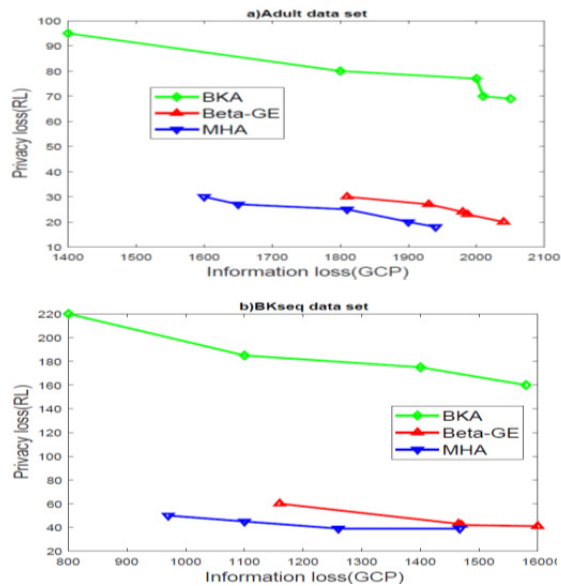
پارامترهای مورد نیاز در چارچوب گمنامی  $k$  و  $\beta$  است. بدیهی است مقدار

1. record linkage (RL)

پارامترهای مدل باید بر اساس سیاست‌های خاص حوزه داده انتخاب شود. به منظور مقایسه روش پیشنهادی با پژوهش‌های (Amiri et al. (2016) و (Riboni et al. (2012، مقدار  $l$  در بازه  $[0/2 - 0/8]$ ،  $k$  در بازه  $[3-20]$  و  $\beta$  در بازه  $[3-10]$  انتخاب می‌شوند. در ادامه، تأثیر سه پارامتر حریم خصوصی  $l$ ،  $k$  و  $\beta$  بر کارایی الگوریتم‌ها بررسی و نتایج آن تحلیل می‌شود.

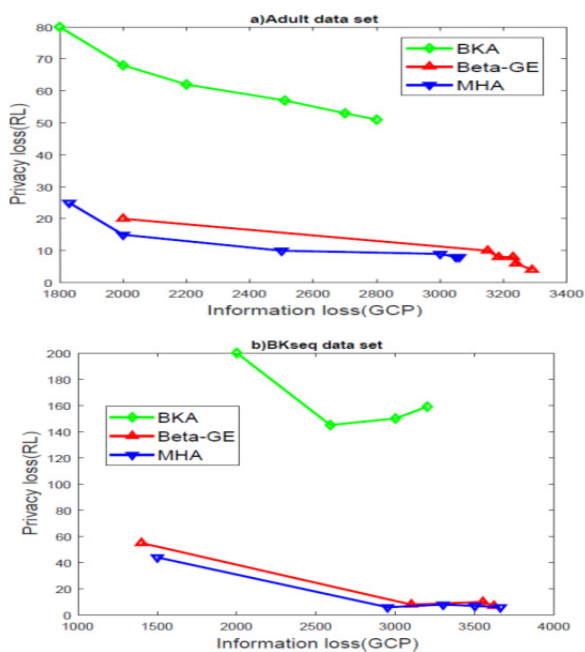
### ۳-۵-۱. ارزیابی بر اساس نمودار خطر افشا-اتلاف اطلاعات

در این بخش تأثیر پارامترهای مدل حریم خصوصی بر میزان اتلاف اطلاعات و خطر افشا مطالعه می‌شود. با توجه به دو معیار GCP و RL، الگوریتم‌هایی که کارایی آن‌ها به گوشه سمت چپ و پایین نمودار نزدیک‌تر باشد، مطلوب‌تر هستند. در ابتدا، مقدار دو پارامتر  $l$  و  $\beta$  را ثابت فرض می‌کنیم ( $l = 0/8$  و  $\beta = 3$ ). نتایج ارزیابی الگوریتم‌ها بر روی دو مجموعه داده Adult و Bkseq هنگامی که مقدار  $k$  در بازه  $[3-20]$  تغییر می‌کند، در شکل ۴، ترسیم می‌شود. همان‌طور که در شکل ۴a، مشاهده می‌شود، در بین سه الگوریتم مورد بررسی، BKA اتلاف اطلاعات و اتلاف حریم خصوصی بالایی را ایجاد می‌کند. الگوریتم پیشنهادی MHA بهترین کارایی را از نظر اتلاف اطلاعات و حریم خصوصی ارائه می‌دهد. با توجه به این که MHA و  $\beta_{GE}$  بر روی مقادیر حساس متمرکز است، انتظار می‌رود اتلاف حریم خصوصی بهتری نسبت به BKA داشته باشند. از سوی دیگر، مرحله تصحیح کلاس‌های هم‌ارزی در MHA منجر به این می‌شود که اتلاف اطلاعات کمتری نسبت به  $\beta_{GE}$  داشته باشد. دامنه تغییرات اتلاف اطلاعات ایجادشده توسط  $\beta_{GE}$  کوچک‌تر از دامنه تغییرات در بقیه الگوریتم‌هاست. این مسئله بدان دلیل است که در این مجموعه آزمایش،  $\beta$  و  $l$  ثابت است. در نتیجه، ساختار درخت‌های ایجادشده به وسیله  $\beta_{GE}$  یکسان است. بنابراین، شرط  $k$ -گمنامی به حذف برگ‌هایی با اندازه کمتر از  $k$  منجر می‌شود. از آنجا که تعداد برگ‌های با اندازه کمتر از  $k$ ، اندک است، به‌ازای  $k$ های مختلف مقدار اتلاف اطلاعات به یکدیگر نزدیک است؛ در حالی که در بقیه الگوریتم‌های مورد بررسی، خوشه‌های متفاوت ایجاد می‌شود. نتیجه این که، دامنه تغییرات اتلاف اطلاعات ممکن است افزایش یابد. نتایج کارایی الگوریتم‌ها بر روی داده Bkseq در شکل ۴b، یافته‌های قبلی را تأیید می‌کنند.



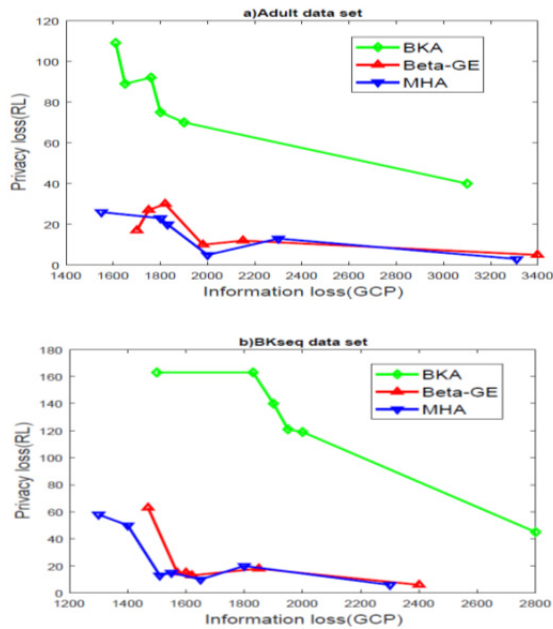
شکل ۴. نمودار خطر افشا-اتلاف اطلاعات بر حسب مقادیر مختلف  $k$

در ادامه، تأثیر پارامتر  $k$  بر اتلاف اطلاعات و خطر افشای اطلاعات بر روی دو مجموعه داده Adult و BKseq بررسی می‌شود. در این آزمایش‌ها  $k = 3$ ،  $\beta = 3$  و پارامتر  $k$  در بازه  $[0.2 - 0.8]$  تغییر می‌کند. نتایج آزمایش‌ها در شکل ۵، یافته‌های قبلی را تأیید می‌کند. در هر دو مجموعه داده، MHA کلاس‌های هم‌ارزی با RL و GCP پایین ایجاد کرده است و در هر سطح از اتلاف اطلاعات، BKA در مقایسه با الگوریتم‌های دیگر اتلاف حریم خصوصی بالاتری دارد. بر خلاف نتایج نشان داده شده در شکل ۴، دامنه تغییرات اتلاف اطلاعات در  $\beta_{GE}$  گسترده است. با تغییر  $k$ ، خوشه‌یابی مبتنی بر دانش پیش‌زمینه خوشه‌های متفاوت ایجاد می‌کند که منجر به اتلاف اطلاعات متفاوت می‌شود.



شکل ۵. نمودار خطر افشا-اتلاف اطلاعات بر حسب مقادیر  $l$

در ادامه، کارایی الگوریتم‌های پیشنهادی به‌ازای مقادیر  $\beta$  متفاوت بررسی می‌شود. در این آزمایش‌ها،  $l = 0/8$  و  $k = 5$  است و مقدار  $\beta$  در بازه  $[0/8 - 0/1]$  تغییر می‌کند. نتایج تجربی بر روی دو مجموعه داده در شکل ۶، یافته‌های قبلی را تأیید می‌کند.



شکل ۶. نمودار خطر افشا-اتلاف اطلاعات بر حسب مقادیر مختلف  $\beta$

بنابراین، می‌توان نتیجه گرفت که با توجه به این که در MHA شرط  $\beta$ -شباهت بر سایر نیازمندی‌های مدل حریم خصوصی مقدم است، افشای اطلاعات کمتر و حریم خصوصی بیشتری در خروجی مشاهده می‌شود. از سوی دیگر، مرحله تصحیح کلاس‌های هم‌ارزی بر اساس مقادیر شبه‌شناسه منجر به اتلاف اطلاعات کمتر و سودمندی بالاتر شده است.

### ۲-۳-۵. اندازه کلاس‌های هم‌ارزی

در این بخش میزان نزدیکی پارامتر مدل حریم خصوصی  $k$  با اندازه کلاس‌های هم‌ارزی ایجادشده توسط الگوریتم‌های پیشنهادی بر روی مجموعه داده Adult بررسی می‌شود. به‌منظور کاهش اتلاف اطلاعات، کلاس‌های هم‌ارزی با اندازه نزدیک به  $k$  مطلوب است. اندازه واقعی کلاس‌های هم‌ارزی در جدول ۳، نمایش داده شده است. هر سطر اندازه کلاس‌های هم‌ارزی بر حسب مقادیر مختلف پارامتر ورودی  $k$  را نمایش می‌دهد. در هر سطر جدول به ازای هر مقدار  $k$  دو ستون میانگین و کمترین وجود دارد. در

ستون میانگین، میانگین اندازه کلاس‌های هم‌ارزی و در ستون کمترین، اندازه کوچکترین کلاس هم‌ارزی را نمایش می‌دهد.

جدول ۳. اندازه کلاس‌های هم‌ارزی (میانگین اندازه کلاس: کمترین اندازه کلاس)  
روی مجموعه داده Adult با  $J = 0.8$  و  $\beta = 3$

الگوریتم	$K=3$	$K=5$	$K=10$	$K=15$	$K=20$
MHA	میانگین ۲۵/۴	کمترین ۴	میانگین ۲۶/۳۲	کمترین ۷	میانگین ۳۵/۸
$\beta$ -GE	میانگین ۲۶	کمترین ۴	میانگین ۲۷/۳۵	کمترین ۷	میانگین ۳۸/۲۵
KBA	میانگین ۱۰	کمترین ۴	میانگین ۱۵	کمترین ۵	میانگین ۲۰

بر اساس جدول ۳، با افزایش  $k$ ، میانگین اندازه کلاس‌های هم‌ارزی افزایش یافته است. همچنین، میانگین اندازه کلاس‌های هم‌ارزی ایجادشده توسط KBA کوچک‌تر از  $\beta$ -GE و MHA است. اختلاف بین کمترین و میانگین اندازه کلاس‌های هم‌ارزی در  $\beta$ -GE بیشتر از اختلاف بین کمترین و میانگین در دو الگوریتم دیگر است. اندازه کلاس‌های هم‌ارزی ایجادشده در MHA مشابه  $\beta$ -GE است.

در ادامه، اندازه کلاس‌های هم‌ارزی ایجادشده توسط الگوریتم پیشنهادی با الگوریتم‌های قبلی در جدول ۴، مقایسه می‌شود؛ در حالی که مقدار دو پارامتر  $\beta$  و  $k$  ثابت است. هر سطر از جدول اندازه کلاس‌های هم‌ارزی را به‌ازای مقادیر مختلف پارامتر  $l$  نمایش می‌دهد. در هر سطر جدول به‌ازای هر مقدار  $l$  دو ستون میانگین و کمترین وجود دارد. در ستون میانگین، میانگین اندازه کلاس‌های هم‌ارزی و در ستون کمترین، اندازه کوچکترین کلاس هم‌ارزی را نمایش می‌دهد. نتایج مشاهده‌شده یافته‌های قبلی را تأیید می‌کند. می‌توان مشاهده کرد که میانگین کلاس‌های هم‌ارزی به‌ازای له‌های مختلف افزایش می‌یابد. به‌ازای مقادیر  $l$  بزرگ، گام خوشه‌یابی مبتنی بر دانش پیش‌زمینه تعداد کمی خوشه با تعداد رکورد زیاد ایجاد می‌کند. در نتیجه، الگوریتم MHA گروه‌های بزرگی از رکوردها ایجاد می‌کند. نظر به این که کلاس‌های هم‌ارزی متناسب با اندازه گروه‌ها ایجاد می‌شود، اندازه کلاس‌های هم‌ارزی ایجادشده توسط MHA افزایش می‌یابد. اما، مرحله تصحیح کلاس‌ها به کاهش اندازه کلاس‌ها در مقایسه با  $\beta$ -GE منجر می‌شود.

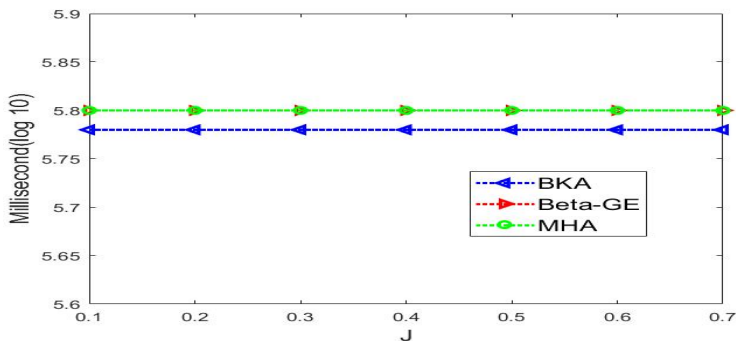


جدول ۴. اندازه کلاس‌های هم‌ارزی (میانگین / کمترین) روی مجموعه داده Adult با  $K = 5$  و  $\beta = 3$

الگوریتم	$J = 0.2$	$J = 0.3$	$J = 0.4$	$J = 0.5$	$J = 0.6$	$J = 0.7$
MHA	۱۰	۵	۱۱	۵	۱۳	۵
$\beta$ -GE	۱۰	۵	۱۲	۵	۱۳	۷
KBA	۱۱	۵	۱۲	۵	۱۴	۵

### ۳-۳-۵. سرعت و مقیاس پذیری الگوریتم‌های پیشنهادی

در این بخش، زمان اجرای الگوریتم‌های پیشنهادی به‌طور تجربی بر روی مجموعه داده‌های Bkseq بررسی می‌شود. برای این منظور، دو آزمایش متفاوت اجرا می‌شود. در آزمایش اول، مقدار  $K = 5$  و  $\beta = 3$  منظور می‌شود، در حالی که مقدار  $J$  در بازه  $[0.1 - 0.7]$  تغییر می‌کند. زمان اجرای الگوریتم‌ها در شکل ۷، بر حسب  $J$  ترسیم شده است. لازم به ذکر است که محور عمودی در مقیاس لگاریتمی است. می‌توان مشاهده کرد که زمان اجرای MHA کمی بیشتر از BKA است، در حالی که از نظر زمان اجرا مشابه  $\beta$ -GE است.



شکل ۷. زمان (میلی ثانیه با مقیاس لگاریتمی) اجرای الگوریتم‌های پیشنهادی با  $K = 5$  و  $\beta = 3$  بر حسب مقادیر مختلف  $J$  بر روی مجموعه داده Bkseq

در آزمایش دوم، تأثیر مقادیر مختلف  $K$  بر روی زمان اجرای الگوریتم پیشنهادی بررسی می‌شود. در این حالت  $J = 0.6$  و  $\beta = 3$  در نظر گرفته می‌شود. زمان اجرای الگوریتم‌ها بر حسب مقادیر مختلف  $K$  یافته‌های قبل را تأیید می‌کند. بنابراین، می‌توان نتیجه گرفت که زمان اجرای MHA مشابه  $\beta$ -GE است، در حالی که BKA از نظر زمان اجرا کاراتر است.

### ۵-۳-۴. تحلیل نتایج کارایی

الگوریتم MHA برای حفاظت در مقابل حمله دانش پیش‌زمینه از خوشه‌یابی تجمعی استفاده می‌کند. سپس، هر خوشه به تعدادی کلاس هم‌ارزی افزای می‌شود. الگوریتم پیشنهادی سعی دارد موازنه‌ی مناسبی بین دو هدف متناقض حفظ سودمندی و حریم خصوصی ایجاد کند. همان‌طور که در نتایج آزمایشگاهی دیده شد، مرحله‌ی تصحیح کلاس‌های هم‌ارزی (بر اساس مقادیر شبه‌شناسه) منجر به اتلاف اطلاعات کمتر و سودمندی بالاتر در مقایسه با روش‌های پیشین شده است. با توجه به این که در دو روش MHA و  $\beta\_GE$  نیازمندی‌های مدل حریم خصوصی به‌طور سلسله‌مراتبی برآورده می‌شود، از نظر معیار RL کارایی مشابه دارند. اما، MHA از نظر اتلاف اطلاعات بر  $\beta\_GE$  اولویت دارد. در روش BKA تمرکز بر کاهش اتلاف اطلاعات است. برای این منظور، رکوردها بر اساس مقادیر شبه‌شناسه مرتب می‌شوند، و سپس، کلاس‌های هم‌ارزی بر اساس رکوردهای نزدیک به هم ایجاد می‌شوند. نتایج آزمایشگاهی نشان می‌دهد که الگوریتم MHA از نظر خطر افشا و اتلاف اطلاعات بر BKA اولویت دارد.

از سوی دیگر، حمله‌ی کمینگی (Wong et al. (2007 در الگوریتم‌های گمنامی قطعی می‌تواند منجر به افشای اطلاعات افراد شود. بر اساس پژوهش «کرمود» و همکاران، می‌توان با استفاده از تصادفی‌سازی از این حمله پیشگیری کرد (Cormode et al. 2010). اگرچه گام خوشه‌یابی مبتنی بر دانش پیش‌زمینه قطعی است، الگوریتم MHA از تصادفی‌سازی در مرحله‌ی تشکیل کلاس‌های هم‌ارزی استفاده می‌کند که می‌تواند مانع از وقوع حمله‌ی کمینگی شود.

تحلیل پیچیدگی زمانی الگوریتم MHA نشان می‌دهد که خوشه‌یابی تجمعی یکی از پرهزینه‌ترین بخش‌های الگوریتم پیشنهادی است. خوشه‌یابی تجمعی در بدترین حالت از پیچیدگی زمانی  $O(n^2 \log n)$  است. الگوریتم MHA در ادامه،  $\beta$ -k-Utility را اجرا می‌کند. در بدترین حالت، پیچیدگی محاسباتی گام اول  $\beta$ -k-Utility،  $O(n \log n)$  و گام دوم (تصحیح کلاس‌ها) از  $O(n^2)$  است. الگوریتم BKA در بدترین حالت از پیچیدگی محاسباتی  $O(n^2)$  و  $\beta\_GE$  از پیچیدگی محاسباتی  $O(n^2 \log n)$  است. نتایج آزمایشگاهی نیز پیچیدگی محاسباتی را تأیید می‌کند و اختلاف بین پیچیدگی‌های زمانی سه الگوریتم مورد بررسی در ورودی‌های بزرگ قابل مشاهده است.

## ۶. بحث و نتیجه‌گیری

در این پژوهش، چارچوبی برای برون‌سپاری و انتشار داده ارائه شد. در سناریوی مورد نظر این پژوهش فرض می‌شود که مهاجم به دانش پیش‌زمینه دسترسی دارد. مدل حریم خصوصی شامل سه پارامتر  $l$ ،  $k$  و  $\beta$  است و اهداف مختلفی شامل پیشگیری از حمله دانش پیش‌زمینه، افشای هویت، و ویژگی را دنبال می‌کند. در مدل حریم خصوصی پیشنهادی،  $\beta$ -شباهت مرتبط با مقادیر حساس رکوردهاست و  $k$ -گمنامی بر شباهت مقادیر شبه‌شناسه متمرکز است. هنگامی که مقدار  $k$  افزایش می‌یابد، ائتلاف اطلاعات زیاد می‌شود، در حالی که با کاهش  $l$  و  $\beta$  تعداد رکوردهای منتشر نشده افزایش می‌یابد. به منظور برآورده کردن مدل حریم خصوصی، الگوریتم سلسله‌مراتبی MHA ارائه شد. MHA نیازمندی‌های مدل حریم خصوصی را در دو گام برآورده می‌کند. در مرحله اول، MHA از خوشه‌یابی تجمعی برای پیشگیری از حمله دانش پیش‌زمینه استفاده می‌کند. در مرحله بعد،  $k$ -گمنامی و  $\beta$ -شباهت در کلاس‌های هم‌ارزی اعمال می‌شود. برای افزایش سودمندی در مرحله دوم، رکوردها به گونه‌ای بین کلاس‌ها جابجا می‌شوند که مدل حریم خصوصی نقض نشود. در ارزیابی‌های آزمایشگاهی نشان داده شد چارچوب پیشنهادی از نظر سودمندی کارا تر از روش‌های پیشین است، در حالی که از نظر زمان اجرا مشابه آن‌هاست.

به‌عنوان کارهای آینده پیشنهاد می‌شود کارایی الگوریتم پیشنهادی در سناریوی انتشار پویای داده در طول زمان بررسی شود. در انتشار پویای داده، مهاجم می‌تواند از اتصال بین جداول منتشر شده در طول زمان برای افشای اطلاعات نیز استفاده کند. همچنین، تأثیر انواع مختلف دانش‌های پیش‌زمینه، مانند همبستگی بین چندین رکورد می‌تواند منجر به افشای اطلاعات افراد شود. از سوی دیگر، برون‌سپاری و انتشار داده‌های موتورهای جست‌وجو برای بررسی فعالیت کاربران می‌تواند به‌عنوان ارزش افزوده دیده شود. اما به دلیل نقض حریم خصوصی کاربران، امکان انتشار این داده‌ها برای سازمان‌ها وجود ندارد. بررسی روش‌های گمنامی موجود و ارائه روش‌های جدید بر روی داده‌های جمع‌آوری شده در موتورهای جست‌وجو می‌تواند به‌عنوان یک مسیر پژوهشی در این راستا دیده شود.

## References

- Adult Dataset. 2015. <https://archive.ics.uci.edu/ml/datasets/Adult>. (accessed April 2015).
- Al Bouna, B., C. Clifton, & Q. Malluhi. 2015. *Efficient sanitization of unsafe data correlations*, in Proc. of

- the Workshops of the EDBT/ICDT 2015 Joint Conference. Belgium. p. 278–285.
- Amiri, F., N. Yazdani, A. Shakery, & A. H. Chinaei. 2016. *Hierarchical anonymization algorithms against background knowledge attack in data releasing*, Knowl. Based Sys. 101: 71-89.
- Amiri, F., N. Yazdani, & A. Shakery. 2018. Bottom-up sequential anonymization in the presence of adversary knowledge. *Information Sciences* 450: 316-335.
- Cao, J., and P. Karras. 2012. Publishing microdata with a robust privacy guarantee. Proc. of Very Large Data Bases (VLDB) Endowment. Turkey, p. 1388–1399.
- Chen, B., R. Ramakrishnan, and K. LeFevre. 2007. Privacy skyline: privacy with multi-dimensional adversarial knowledge. In Proc. of the 33th IEEE International Conference on Very Large Data Bases. Austria. p. 770–781.
- Cormode, G., D. Srivastava, N. Li, and T. Li. 2010. Minimizing minimality and maximizing utility: analyzing method-based attacks on anonymized data. Proc. of Very Large Data Bases(VLDB) Endowment, Singapore:
- Domingo-Ferrer, J., J. Soria-Comas, and R., Mulero-Vellido. 2019. Steered Microaggregation as a Unified Primitive to Anonymize Data Sets and Data Streams. *IEEE Transaction on knowledge and data engineering* 14 (12): 3298– 3311.
- Domingo-Ferrer, J., and V. Torra. 2005. Ordinal, continuous and heterogeneous k-anonymity through microaggregation. *Data Mining Knowl. Discov.* 11 (2) 195–212.
- Dwork, C. 2006. Differential privacy. In Proc. of the 33rd International Colloquium on Automata, Languages and Programming (ICALP). Italy. p.1–12.
- Fatahi, S, & M. J. Ershadi. 2020. Assessment of User Satisfaction of Research Theses and Theses in Iranian Scientific Database (Ganj): Based on E-Qual Model. *Iranian Journal of Information Processing and Management* 35 (2): 399-424.
- Fung, B. C. M., K. Wang, A. W.-C. Fu, and P. Yu. 2012. *Introduction to Privacy-Preserving Data Publishing: Concepts and Techniques*.: Chapman & Hall/ CRC Data Mining and Knowledge Discovery Series.
- Gardner, J. 2012. Privacy Preserving Medical Data Publishing. PHD Thesis, Emory University.
- Ghinita, G., P. Karras, P. Kalnis, and N. Mamoulis. 2007. Fast data anonymization with Low information loss. In Proc. of the 33rd International Conference on Very large Data Bases. Austria. p.758–769.
- Han, J., J. Yu, Y. Mo, J. Lu, & H. Liu. 2014. MAGE: a semantics retaining k-anonymization method for mixed data. *Knowledge-Based Systems* 55: 75–86.
- Kambourakis, G. 2014. Anonymity and closely related terms in the cyberspace: An analysis by example. *Journal of Information Security and Applications* 19 (1): 2-17.
- Kifer, D. 2009. Attacks on privacy and deFinetti's Theorem. In Proc. of ACM International Conference on Special Interest Group on Management of Data (SIGMOD). Irland. p.127-138.
- LeFevre.k, D., J. DeWitt, & R. Raghu. 2006. Mondrian multidimensional k-Anonymity, in: Proceedings of the 22nd IEEE International Conference on Data Engineering (ICDE), pp. 25–35 .
- Li, N., T. Li, & S. Venkatasubramanian. 2007. t-closeness: privacy beyond k-anonymity and L-diversity. In Proc. of the 23th IEEE International Conference on Data Eng. (ICDE). Turkey. p. 106–115.
- Li, T., N. Li, J. Zhang, and I. Molloy. 2012. Slicing: A New Approach for Privacy Preserving Data Publishing. *IEEE Transactions on Knowledge and Data Engineering* 24 (3): 561-574.
- Li, T., N. Li, and J. Zhang. 2009. Modeling and integrating background knowledge in data anonymization. In Proc. of the 25th IEEE Int. Con. on Data Engineering (ICDE). China. p.6-17.
- Lin, J. 1991. Divergence measures based on the Shannon entropy. *IEEE Trans. Information Theory* 37 (1): 145-151.
- Machanavajjhala, P., J. Gehrke, D. Kifer, & M. Venkitasubramaniam. 2007. L-diversity: privacy

- beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data*. 1 (1), doi: 10.1145/1217299.1217302.
- Martin, D., D. Kifer, A. Machanavajjhala, J. Gehrke, and J. Halpern. 2007. Worst-case background knowledge for privacy-preserving data publishing. In *Proc. of the International Conference on Data Engineering (ICDE)*. Turkey. p.126–135.
- National Heart, lung and Blood Institute, Data Fact Sheet, National Heart, lung and Blood Institute. 2015. <http://www.lung.org/assets/documents/research/copd-trend-report.pdf>. (accessed Nov.2015)
- Riboni, D., L. Pareschi, & C. Bettini. 2012. JS-Reduce: defending your data from sequential background knowledge attacks. *IEEE Transactions on Dependable and Secure Computing*. 9 (3): 387-400.
- Samarati, P. 2001. Protecting respondents' identities in microdata release. *IEEE Transaction on Knowledge and Data Engineering* 13 (6): 1010–1027.
- Soria-Comas, J., J. Domingo-Ferrer, D. Sánchez, and S. Martínez. 2015. t-closeness through microaggregation: strict privacy with enhanced utility preservation. *IEEE Transaction on Knowledge and Data Engineering* 27 (11): 3098-3110.
- Sweeney, L. 2002. Achieving k-anonymity privacy protection using generalization and suppression. *International Journal of Uncertainty, Fuzziness, and Knowledge-based Systems* 10 (5): 571–588.
- Wang, H., & R. Liu 2011. Privacy-preserving publishing microdata with full functional dependencies. *Data Knowl. Eng.* <http://dx.doi.org/10.1016/j.datak.2015.06.012> .
- Wang, H., and R. Liu. 2015. Hiding outliers into crowd: privacy-preserving data publishing with outliers. *Data Knowl. Eng.* <http://dx.doi.org/10.1016/j.datak.2015.06.012>.
- Ward, K., D. Lin, and Sanjay Madria. 2020. A Parallel Algorithm for Anonymizing Large-scale Trajectory Data. *ACM/IMS Trans. Data Sci.* 1 (1):1-26.
- Wong, R., A. FU, K. Wang, and J. Pei. 2007. Minimality attack in privacy preserving data publishing. In *Proc. of the 33rd International Conference on Very Large Data Bases (VLDB)*. Austria. pp. 543–554.
- Xiao, X., & Y. Tao. 2006. *Anatomy: simple and effective privacy preservation*. In *Proc. of the 32nd International Conference on Very Large Data Bases*. Korea. p.139–150.
- Zhao, X., D. Pi, & J. Chen. 2020. Novel trajectory privacy-preserving method based on prefix tree using differential privacy. *Knowledge-Based Systems* 198 (21) <https://doi.org/10.1016/j.eswa.2020.113241>
- Zhou, B., and J. Pei. 2011. The K-Anonymity and L-Diversity Approaches for Privacy Preservation in Social Networks against Neighborhood Attacks. *Knowledge and Information Systems* 28: 47-77.

## پیوست شماره ۱:

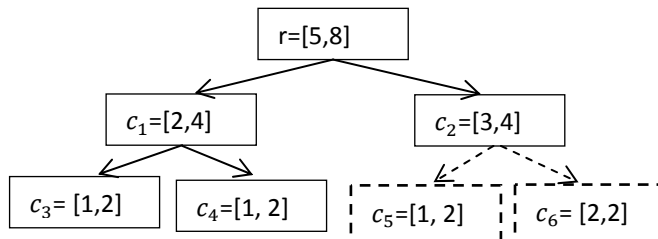
جدول زیر (جدول ۱) نمادهای ریاضی استفاده‌شده در این پژوهش را نمایش می‌دهد.

### جدول ۱. نمادهای ریاضی استفاده‌شده در این پژوهش

$T$	جدول داده اصلی
$ID$	ویژگی شناسه
$QI$	ویژگی‌های شبه‌شناسه
$D$	تعداد ویژگی‌های شبه‌شناسه
$A_j$	آمین ویژگی
$[A_j]$	دامنه ویژگی آمین
$m$	تعداد مقادیر حساس متمایز
$r_n$	$n$ آمین رکورد در جدول داده
$\bar{r}_n$	نسخه گمنام $n$ آمین رکورد در جدول داده
$PD^{sv}$	دانش پیش‌زمینه مقادیر حساس
$\Sigma$	مجموعه همه توزیع‌های احتمال ممکن

پیوست شماره ۲:

**مثال:** فرض کنید اطلاعات بیماران یک بیمارستان در جدول T ذخیره شده است. بیماری فرد به عنوان ویژگی حساس هر فرد در نظر گرفته می‌شود. جدول T شامل ۲۶ رکورد با مقادیر حساس  $[A_{D+1}] = \{\text{Alzheimer, HIV, Flu, Depression}\}$  است که شامل ۴ رکورد با مقدار حساس Alzheimer، ۶ رکورد با HIV، ۸ رکورد با Flu و ۸ رکورد با Depression است. بنابراین، توزیع مقادیر حساس در جدول اصلی  $P_T = \left(\frac{4}{26}, \frac{6}{26}, \frac{8}{26}, \frac{8}{26}\right)$  است. فرض کنید  $\beta = 2$  و  $K = 2$  است. بنابراین،  $f(p_3) = f(p_4) = 0.67, f(p_2) = 0.56, f(p_1) = f(p_1) = 0.44$  است. مجموعه  $N_i$  شامل ۲ رکورد با مقدار حساس Alzheimer، ۳ رکورد با مقدار حساس HIV، ۴ رکورد با Flu و ۴ رکورد با Depression است. آنگاه  $\beta$ -partition افزایشی به صورت  $B = \{b_1, b_2\}$  ایجاد می‌کند که  $b_1$  شامل Alzheimer و HIV و  $b_2$  شامل دو مقدار حساس باقی‌مانده است.  $\beta$ -split بدین گونه عمل می‌کند: ریشه  $r = [5, 8]$  همان افزایش  $\beta$ -partition با ۵ رکورد در  $b_1$  و ۸ رکورد در  $b_2$  است. گره  $r$  به دو گره  $C_1 = [2, 4]$  و  $C_2 = [3, 4]$  تقسیم می‌شود. هر دو گره شرط شایستگی و  $k$ -گمنامی را برآورده می‌کنند. اگر به فرایند تقسیم کردن گره‌های  $C_1$  و  $C_2$  ادامه دهیم، گره‌های  $C_3, C_4, C_5$  و  $C_6$  تشکیل می‌شود. اما شرط شایستگی را برآورده نمی‌کند. بنابراین، نمی‌توان  $C_2$  را تقسیم کرد. در نتیجه،  $\beta$ -split گره‌های  $C_2, C_3$  و  $C_4$  را به مرحله بعد ارسال می‌کند. درخت در شکل زیر (شکل ۱) رسم شده است. همان‌طور که مشاهده می‌کنید برگ‌های درخت، الگوی تعداد رکوردها در کلاس‌های هم‌ارزی نهایی است.



شکل ۱. اعمال تابع  $\beta$ -split برای مثال ارائه شده در این پیوست

### فاطمه امیری

دارای مدرک دکتری در رشته مهندسی نرم‌افزار از دانشگاه تهران است. ایشان هم‌اکنون استادیار دانشگاه صنعتی همدان است. امنیت شبکه، حریم خصوصی، داده‌کاوی و کلان‌داده از جمله علایق پژوهشی وی است.

