Privacy-preserving data publishing: an information-driven distributed genetic algorithm

World Wide Web (2024) 27:1
https://doi.org/10.1007/s11280-024-01241-y

Privacy-preserving data publishing: an information-driven distributed genetic algorithm

Yong-Feng Ge[1] · Hua Wang[1] · Jinli Cao[2] · Yanchun Zhang[3,4,5] · Xiaohong Jiang[6]

Received: 30 March 2023 / Revised: 29 August 2023 / Accepted: 6 December 2023 /
Published online: 15 January 2024

Abstract
The privacy-preserving data publishing (PPDP) problem has gained substantial attention from research communities, industries, and governments due to the increasing requirements for data publishing and concerns about data privacy. However, achieving a balance between preserving privacy and maintaining data quality remains a challenging task in PPDP. This paper presents an information-driven distributed genetic algorithm (ID-DGA) that aims to achieve optimal anonymization through attribute generalization and record suppression. The proposed algorithm incorporates various components, including an information-driven crossover operator, an information-driven mutation operator, an information-driven improvement operator, and a two-dimensional selection operator. Furthermore, a distributed population model is utilized to improve population diversity while reducing the running time. Experimental results confirm the superiority of ID-DGA in terms of solution accuracy, convergence speed, and the effectiveness of all the proposed components.

Keywords Evolutionary computation · Data privacy and utility · Data publishing · Distributed algorithm

1 Introduction

In the present era, data assumes a critical role in the daily lives of individuals [1–6]. The dissemination and utilization of data [7–13] have created enormous opportunities for decision-

This article belongs to the Topical Collection: *Special Issue on Web Information Systems Engineering 2022*
Guest Editors: Richard Chbeir, Helen Huang, Yannis Manolopoulos and Fabrizio Silvestri.

✉ Yong-Feng Ge
yongfeng.ge@vu.edu.au

1 Victoria University, Melbourne, Australia

2 La Trobe University, Melbourne, Australia

3 School of Computer Science and Technology, Zhejiang Normal University, Jinhua, Zhejiang, China

4 The Department of New Networks, Peng Cheng Laboratory, Shenzhen, Guangdong, China

5 Institute for Sustainable Industries and Liveable Cities, Victoria University, Melbourne, Victoria, Australia

6 Future University Hakodate, Hakodate, Japan

making and knowledge exploration [5, 14–23]. For instance, in 2006, Netflix released a dataset comprising 100 million movie ratings to enhance its recommendation system's performance [24]. However, despite the significant advantages of data publication, concerns on data privacy preservation [25–31]. Consequently, privacy-preserving data publishing (PPDP) has emerged as a critical area of research, which aims to create an anonymous dataset that safeguards privacy while maintaining optimal data utility levels. This objective can be achieved through various privacy-preserving techniques such as data anonymization, generalization, and perturbation.

When it comes to PPDP, two main categories of approaches exist: decreasing the precision of the original dataset, and data perturbation [16]. In the first category, a well-known approach was introduced in [32] that uses a binary search on the generalization lattice to identify the anonymization solution. Kohlmayer et al. [33] presented a comprehensive framework for optimal anonymization, which enabled the Flash algorithm to find the optimal anonymization solution by searching the path in the lattice. An algorithm proposed in [34] optimized the anonymization solution in an identical generalization hierarchy, which is useful for protecting data privacy in the general Internet of Things (IoT) environment. However, existing works mostly focus on single anonymization operations (such as attribute generalization or record suppression), which may not be effective from the perspective of information release. Therefore, it is worth considering combining multiple anonymization operations when optimizing the anonymization solution. Moreover, existing works mostly adopt graph search-based strategies to optimize the anonymization solution, but these approaches may lose their effectiveness when the search space of the PPDP problem becomes complex. Ge et al. [35] formulated the multi-objective data publishing problem and proposed a distributed cooperative coevolution evolutionary framework to achieve efficient optimization. In the second category, differential privacy represents one of the typical approaches that ensure no significant difference in query results when inserting one record [36, 37]. These approaches are effective in addressing data privacy requirements in queries. However, they are not suitable for scenarios requiring data transparency and truthfulness.

The genetic algorithm (GA), as discussed in previous research [38–40], is an algorithmic approach that involves a stochastic search mechanism based on the principles of natural competition and selection [41–43]. By utilizing a population model, GA is able to maintain a diverse search direction and facilitate the production of high-quality solutions. The widespread use of GA in various optimization problems [44–47] can be attributed to its advantages in high search efficiency and robustness.

This paper presents the information-driven distributed genetic algorithm (ID-DGA). The proposed algorithm optimizes anonymization solutions using a combination of attribute generalization and record suppression techniques. ID-DGA is designed based on a distributed population model to improve population diversity. Besides, ID-DGA incorporates a specifically designed information-driven crossover operator that facilitates the exchange of information between anonymization solutions and promotes information release. In addition, ID-DGA employs an information-driven mutation operator to enhance population diversity and information release. Furthermore, the proposed information-driven improvement operator helps adaptively refine the anonymization solutions. Finally, a two-dimensional selection operator is introduced to enhance individual competitiveness and population quality.

The paper is structured as follows. Section 2 provides an overview of the related work in the field of PPDP. Section 3 formally defines the PPDP problem. Section 4 presents the proposed ID-DGA in detail. Sections 5 and 6 outline the experimental setup used in this study and present an analysis of the experimental results. Finally, Section 7 offers concluding remarks to wrap up the paper.

2 Related work

In [16], a survey regarding PPDP was presented. In this survey, the related techniques of PPDP were systematically summarized. These techniques were designed according to four attack models, i.e., record linkage, attributed linkage, table linkage, and probabilistic attack. Moreover, the anonymization operations and information metrics were introduced. The previous privacy models can be divided into two categories in terms of mechanism. The first category is based on decreasing the precision of the original dataset to achieve the given specific privacy criteria, including k-anonymity, l-diversity, and t-closeness. The second category is designed based on perturbation to guarantee that no significant difference is shown in the query results when inserting one record.

For the first category, various approaches have been proposed. One of the most important approaches was proposed in [32], where a binary search was performed on the generalization lattice for the solution. Afterward, the optimal k-anonymity problem was proven to be an NP-hard problem [48]. In [49], an algorithm named Incognito performs a bottom-up, breadth-first search of the generalization lattice. In [50], an algorithm named optimal lattice anonymization was proposed. In this algorithm, the generalization lattice was divided into several sub-lattices, and the optimal solution was found by searching within each sub-lattice. In [33], a generic framework for optimal k-anonymity was presented. Based on the proposed framework, an algorithm named Flash was developed to perform the search for the optimal node in the lattice on each built path. In [51], the authors presented an algorithm for k-anonymization of time-varying datasets. Based on micro-aggregation, such an algorithm can support adding, deleting, and updating records while keeping its k-anonymity property. Authors in [52] tackled the semantic attack in trajectory data publishing. An algorithm providing privacy protection against semantic and re-identification attacks was proposed. In [34], a special case of dataset called identical generalization hierarchy was considered, whose solution is effective to address the general IoT data privacy protection. Accordingly, an algorithm for the globally optimized k-anonymity solution was designed.

For the second category, differential privacy [36, 37] was proposed. Differential privacy focuses on data privacy in queries. In differential privacy, any two datasets with a one-record difference should answer similar results to the same query. In [53], a variant of differential privacy named local differential privacy was tackled. Accordingly, a local differentially private high-dimension data publication algorithm was designed based on distribution estimation. In [54], a compressed sensing mechanism was proposed for differential privacy based on the compressed sensing framework while guaranteeing the accuracy of query results. However, different privacy approaches are not applicable in PPDP scenarios that require data transparency since the introduced noise by differential privacy approaches cannot guarantee data truthfulness.

3 Problem definition

As the data publisher, the objective of PPDP is to transfer the original dataset D to an anonymous T that can satisfy the given privacy requirement determined by a privacy model and maintain its utility as high as possible.

In D, quasi-identifiers (QIDs) are attributes that could potentially identify the owners of records in the dataset. During the anonymization, various anonymization operations such as generalization and suppression can be utilized on QID and transfer QID to QID' in T.

In our definition, the k-anonymity criterion is set as the privacy model and defined as:

Definition 1 (k-anonymity) A dataset satisfies the k-anonymity requirement if each combination of QID' attributes exists in at least k records.

Accordingly, the anonymity degree (AD) value of a k-anonymity T equals k. The objective of PPDP is to identify the optimal anonymization and is defined as:

Definition 2 (Optimal anonymization) For T, an optimal anonymization solution can satisfy the privacy requirement ($\mathrm{AD}(T) \geq k$) and achieves the highest utility degree.

The utility of T is calculated according to its transparency degree (TD) [16]:

$$\mathrm{TD}(T) = \sum_{r \in T} \mathrm{TD}(r) \tag{1}$$

$$\mathrm{TD}(r) = \sum_{v_g \in r} \mathrm{TD}(v_g) \tag{2}$$

where r indicates the record in T; v_g is the generalized value in record r. TD value of v_g is calculated as:

$$\mathrm{TD}(v_g) = \frac{1}{|v_g|} \tag{3}$$

where $|v_g|$ is the number of domain values that are descendants of v_g.

4 ID-DGA

This section presents an overview of the proposed ID-DGA. Firstly, we present the distributed population model utilized in ID-DGA. Afterward, we discuss the representation of individuals in ID-DGA (Figure 1). Subsequently, the information-driven crossover, mutation, and improvement strategies employed by ID-DGA are illustrated in detail. Additionally, we introduce the two-dimensional selection operator utilized in ID-DGA. Finally, the entire procedure of ID-DGA is illustrated to provide a comprehensive understanding of the algorithm.

4.1 Distributed population model

In the distributed population model, the entire population of ID-DGA is divided into several sub-populations, and each sub-population evolves independently. All the sub-populations

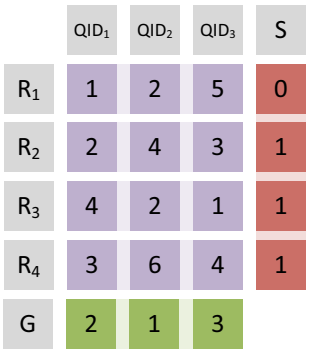

Figure 1 Illustration of representation in ID-DGA, where a sample dataset containing three QID attributes and four records is given

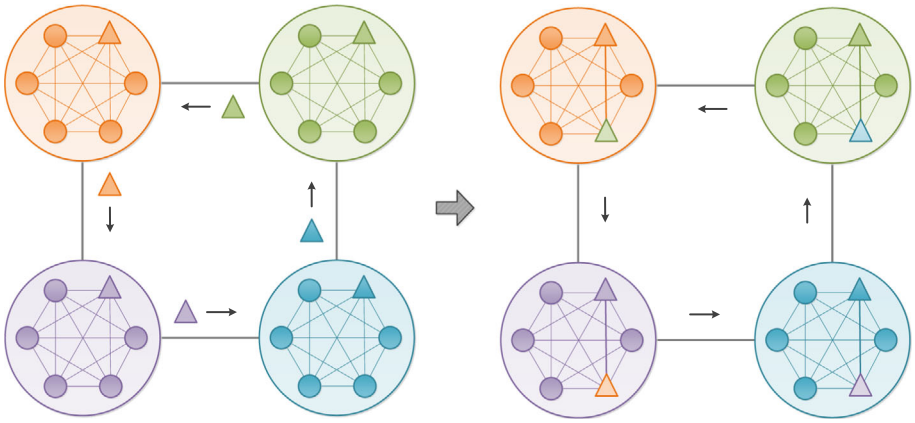

Figure 2 An example of the distributed population model in ID-DGA

communicate according to the predefined topology. With the help of the communication topology, sub-populations share their elite individuals with a given interval, which is referred to as the migration operator. Once one sub-population receives the migrated elite individuals, individuals in the current generation are randomly selected and replaced.

In our proposed ID-DGA, the distributed population model utilizes a ring communication topology. An example of the distributed population model is given in Figure 2. As shown in the example, each big circle indicates a sub-population. In the big circles, small triangles and circles represent the best individuals and the other sub-population individuals. The best individuals in sub-populations are sent to the neighborhood sub-populations on the communication topology with the predefined migration interval. Afterward, one individual in each sub-population is chosen by random and replaced by the received elite individual.

By dividing the entire population of the ID-DGA into several sub-populations with independent evolution, the distributed population model can help the ID-DGA improve the population diversity. By migrating elite individuals in the sub-populations, the island model can enhance the ID-DGA's population quality. If the migration operator is appropriately executed, the ID-DGA can achieve the trade-off between exploration and exploitation. Moreover, since each sub-population evolves independently, the island model can be directly implemented in a distributed manner, which is crucial for speedup in evolution.

4.2 Representation

Figure 1 depicts a sample dataset and its anonymization solution. The dataset consists of four records and three quasi-identifier (QID) attributes. The anonymization solution comprises two vectors: a vector for attribute generalization denoted by "G" and a vector for record suppression denoted by "S". In vector G, each QID attribute is generalized based on its level, while in vector S, each record is suppressed based on its corresponding value. Specifically, a value of "0" indicates that the record is removed, whereas a value of "1" indicates that the record is retained.

In ID-DGA, an individual represents an anonymization solution, which includes two vectors – vector G and vector S. The length of vector G corresponds to the number of QID attributes, while the length of vector S corresponds to the number of records. Throughout the

update process of individuals in ID-DGA, the competitiveness of anonymization solutions is enhanced.

4.3 Information-driven crossover

The information-driven crossover operator involves the use of two distinct strategies. During the exchange of information between two individuals, the vectors G and S are subjected to separate information exchange strategies. The exchange process for vector G entails randomly selecting one of two possible values for each bit of the offspring, based on the corresponding bit in the parent individuals. On the other hand, vector S is subjected to OR gate rules, whereby each bit of the offspring takes the value of one if at least one of the corresponding bits in the parent individuals is one, and zero otherwise. It is worth noting that these strategies operate independently of each other, and they are specifically tailored to enhance the exchange of information between the parent individuals.

Figure 3 provides an illustrative example of the crossover operator, which involves two individuals containing two G vectors (G_1 and G_2) and two S vectors (S_1 and S_2). The crossover operator is executed separately on each vector, and this results in the generation of two offspring vectors ($G_{1\times2}$ and $S_{1\times2}$). During the crossover process for vector $G_{1\times2}$, the value of each bit is randomly selected from the corresponding bits in G_1 and G_2. For instance, the value of the first bit in $G_{1\times2}$ is chosen from G_1 (i.e., 2) and G_2 (i.e., 1), and then randomly selected from G_1. Similarly, the values of the second and third bits in $G_{1\times2}$ are chosen from G_1 and G_2, respectively. In the same vein, the crossover process for vector $S_{1\times2}$ involves the application of OR gate rules, where each bit of the offspring vector takes the value of one if at least one of the corresponding bits in the parent individuals is one, and zero otherwise. For example, the values of the first, second, and third bits in $S_{1\times2}$ are all one, as a calculation result of the values in S_1 and S_2. Conversely, the value of the fourth bit in $S_{1\times2}$ is zero, since both values in S_1 and S_2 are zero. It is worth noting that the crossover process for the G and S vectors is independent and optimized to facilitate the efficient exchange of information between the parent individuals.

Our proposed information-driven crossover operator facilitates the exchange of information between parent individuals. This operator randomly exchanges the values of two G

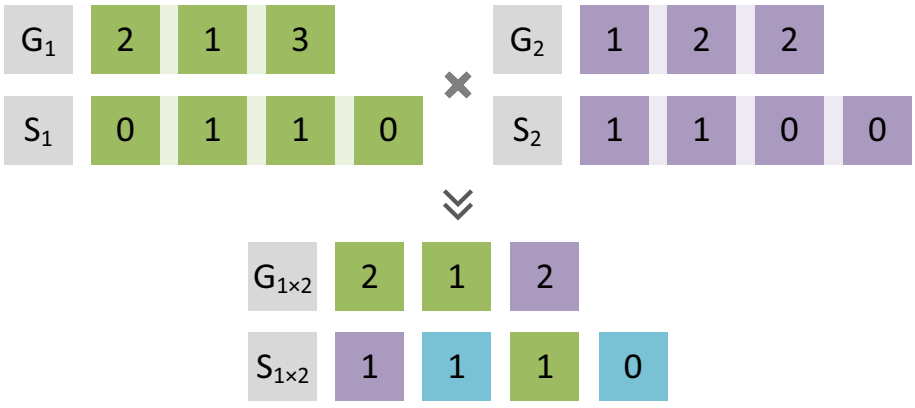

Figure 3 Illustration of information-driven crossover operator, in which the information in two anonymization solutions is exchanged

vectors, leading to a mixture of generalization levels in the resulting anonymization solutions. If the parent individuals meet the privacy preservation requirement, it is likely that their offspring solution will also satisfy the same requirement. For two S vectors, their values are accumulated. As long as one record in one anonymization solution is released, the corresponding record in the offspring anonymization solution is released. Thus, more information is released in the offspring anonymization solution.

4.4 Information-driven mutation

The information-driven mutation operator handles vectors G and S independently. In the vector G, a single bit is randomly selected using the predefined mutation rate, MR, and its value is then initialized within the boundary of its generalization. On the other hand, vector S undergoes a similar mutation process, where a random bit is selected, and its value is changed to one using the same mutation rate, MR. This change in value results in the release of the corresponding record.

Figure 4 provides an illustration of the mutation operator in action. Specifically, the mutant versions of vectors G_1 and S_1 are denoted as G_{1*} and S_{1*}, respectively. In the case of G_1, a random selection is made for its third bit. The value of this bit is then changed from three to two, thereby altering the generalization level of the corresponding QID attribute. As for S_1, the mutation process involves randomly selecting its fourth bit and changing its value from zero to one. Consequently, the fourth record in the mutated anonymization solution is disclosed.

Upon executing the proposed mutation operator, the anonymization solutions are adjusted in a random manner. Vector G undergoes changes in the generalization levels of randomly selected QID attributes. This process may lead to the creation of an anonymization solution that attains a higher degree of anonymity or transparency. In vector S, the records in the randomly chosen positions are released. It is likely to generate an anonymization solution that can achieve a higher transparency degree while satisfying the privacy requirement.

4.5 Information-driven improvement

The information-driven improvement operator is utilized to adaptively refine the child individual. More specifically, for each individual I whose AD value cannot satisfy the privacy preservation requirement, the information-driven improvement operator is utilized to improve

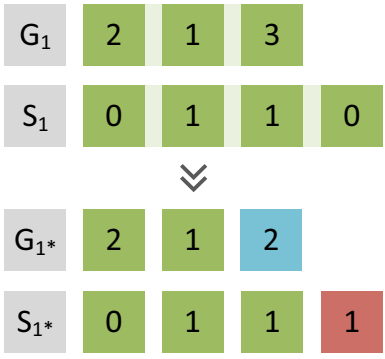

Figure 4 Illustration of information-driven mutation operator, in which two vectors in the solution are adjusted separately

its competitiveness. In I, the AD value of each record is calculated. Afterward, all the records with the lowest AD value are selected and the corresponding values in vector S are set as 0, meaning that these records are removed in the improved anonymization solution.

The proposed information-driven improvement operator is efficient in improving the AD values of individuals. At the beginning of the evolution, such an improvement operator is effective in improving the ratio of individuals that can satisfy the privacy preservation requirement. Afterward, such an improvement operator is helpful in transferring the ineligible individuals with high TD values to be eligible.

4.6 Two-dimension selection

In evaluating the quality of anonymization solutions, two indicators, namely AD and TD, are employed. The optimal anonymization solution, as per the problem definition, is one that achieves the highest TD while also satisfying the requirement stipulated by AD. The prioritization of these two indicators should vary depending on the specific situation. To this end, three rules have been formulated.

1. If neither of two individuals satisfies the privacy preservation requirement, the individual with a higher AD value is considered more competitive.
2. If only one individual satisfies the privacy preservation requirement, that individual is deemed more competitive.
3. If both individuals satisfy the privacy preservation requirement, the individual with a higher TD value is considered better.

Figure 5 displays three pairs of individuals, each evaluated based on the three comparison rules. In every pair, the individual represented by a circle is deemed more competitive than the one represented by a triangle. The first rule is applied in the first pair. Although neither individual can satisfy the privacy protection requirement, the circle individual has a higher AD value. Therefore, the circle individual is deemed more competitive. In the second pair, the circle individual is superior as it fulfills the privacy protection requirement. Finally, in the third pair, both individuals can satisfy the privacy protection requirement, but the circle individual still prevails due to its higher TD value.

Such a two-dimension selection operator can effectively improve the population quality in ID-DGA. When no individual in the population can reach the privacy protection requirement, the individuals with higher privacy degrees are kept in the population. Thus, the entire

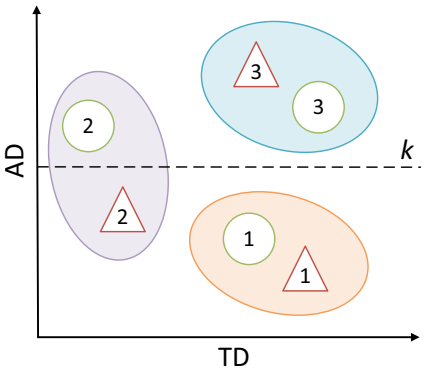

Figure 5 Illustration of two-dimension operator, where three pairs of solutions are compared according to three defined rules

Algorithm 1 Pseudo-code of ID-DGA.

```
1: procedure MASTER NODE
2: Set initial generation g = 0
3: Divide the population into NSP sub-populations
4: while terminal condition is not met do
5:     g = g + 1
6:     if g % MI = 0 then                          ▷ with the predefined migration interval
7:         Receive elite individuals from slave nodes
8:         Send the elite individuals to corresponding slave nodes
9:     end if
10: end while
11: Output the best anonymization solution
12: end procedure
13:
14: procedure SLAVE NODE
15: for every generation do
16:     for each pair of parent individuals do
17:         Perform the information-driven crossover operator
18:         Perform the information-driven mutation operator
19:         Perform the information-driven improvement operator
20:         Evaluate the child individual by two-dimension selection
21:         if the child individual is better than any of the parent individuals then
22:             Replace one of the parent individuals with the child individual
23:         end if
24:         if the child individual is better than the existing best individual then
25:             Replace the best individual with the child individual
26:         end if
27:     end for
28:     if g % MI = 0 then                          ▷ with the predefined migration interval
29:         Send the best individual to the master node
30:         Receive the elite individual from the master node
31:         Use the migrated individual to replace a randomly chosen individual
32:     end if
33: end for
34: Send the best solution to the master node
35: end procedure
```

population can approach the privacy protection requirement during the update. When part of the individuals in the population can reach the privacy protection requirement, these individuals are kept in the population. Finally, when most of the population can reach the privacy protection requirement, the individuals with higher TD values are kept in the population to improve the population quality. The implementation of this two-dimensional selection operator is therefore an effective approach to improving population quality in ID-DGA.

4.7 Overall procedure

The entire procedure of the ID-DGA is described in Algorithm 1. As shown in the pseudo-code, a master-slave model is utilized to implement the ID-DGA. At the master node, the generation index g is set as zero. Then the entire population is divided into NSP sub-populations and sent to the corresponding NSP slave nodes. With the predefined migration interval MI, the master node receives the elite individuals from all the slave nodes. Then it sends these elite individuals to the corresponding slave nodes according to the ring topology. The migration process is executed until the terminal condition is satisfied. Finally, the best anonymization solution to the given dataset is outputted.

At the slave node, each sub-population evolves independently. During the evolution, in each generation, for each pair of parent individuals, the information-driven crossover operator is executed to exchange the anonymization information in parent individuals and generate the child individual. Afterward, the information-driven mutation operator is carried out on the child individual to improve the population diversity. After the mutation operator, if the mutant individual cannot satisfy the privacy preservation requirement, the mutant child individual is adjusted by the information-driven improvement operator. Subsequently, the child individual is evaluated and compared with the parent individuals by the selection operator. If the child individual is better than any parent individual, one of the parent individuals will be replaced. Otherwise, the mutant child individual will not be kept in the population. Then, the migration operator is carried out with the predefined mutation interval MI. Each slave node sends the best individual to the master node and receives one elite individual from the master node. Afterward, one randomly chosen individual in the sub-population that is not the best individual will be replaced by the received migrated individual. Finally, the best individual is returned to the master node.

5 Experimental setup

This section illustrates the test instances, parameters settings, and algorithm implementation in the following experiments.

5.1 Test instances

In the subsequent experimental studies, 16 test instances are utilized to investigate the performance of the proposed ID-DGA. These test instances are generated based on the public

Table 1 Properties of 16 test instances

Test instances	nA	$nQID$	nR
T_1	16	8	200
T_2	16	8	200
T_3	16	8	400
T_4	16	8	400
T_5	18	10	200
T_6	18	10	200
T_7	18	10	400
T_8	18	10	400
T_9	20	12	200
T_{10}	20	12	200
T_{11}	20	12	400
T_{12}	20	12	400
T_{13}	22	14	200
T_{14}	22	14	200
T_{15}	22	14	400
T_{16}	22	14	400

datasets released by the New York State Department of Health[1]. Table 1 outlines the properties of these test instances, including the number of attributes nA, the number of QID attributes $nQID$, and the number of records nR. In addition, in each test instance, the privacy requirement of anonymity degree k is set as 2.

5.2 Parameter settings

In the proposed ID-DGA, population size N is set as 40 and number of sub-populations NSP is set as 4; mutation rate MR is set as 0.1; migration interval MI is set as 5. For all the algorithms, the maximum fitness evaluation number is set as $nQID \times nR$.

5.3 Algorithm implementation

ID-DGA and all the compared algorithms in this paper are implemented in C++ and performed on a local compute node (OS: Ubuntu 16.04; CPU: 16-Core Intel i9-12900K; Memory: 16GB).

6 Experimental result

In this section, we verify the advantages of the proposed ID-DGA by comparing it with the baseline algorithm GA, a competitive optimal anonymization algorithm Flash, and an information-driven genetic algorithm (ID-GA). Moreover, the effect of all the proposed operators is investigated.

6.1 Comparison with existing approaches

To verify the effectiveness of the proposed ID-DGA algorithm, four existing algorithms, i.e., GA [38], DE [55], Flash [33], and ID-GA [56] are utilized for comparison. These three algorithms are listed as follows:

1. GA [38]: This algorithm acts as a baseline algorithm in the comparison. When compared with the proposed ID-DGA, the effect of our designed operators in ID-DGA is confirmed.
2. DE [55]: In this algorithm, each privacy-preserving solution is represented by an individual of differential evolution (DE), and the competitiveness of the solution is improved through the mutation, crossover, and selection operators.
3. Flash [33]: In this paper, a generic framework for globally-optimal k-anonymity was presented. Furthermore, an algorithm based on a binary search was proposed based on the proposed framework.
4. ID-GA [56]: In this paper, an information-driven genetic algorithm was designed to achieve the optimal anonymization based on attribute generalization and record suppression.

In Table 2, the mean and standard deviation values of TD over 25 independent runs are presented, and the best results are highlighted in **boldface**. Overall, our proposed ID-DGA can outperform the compared existing algorithms on all the 16 test instances. Compared with GA, the advantages of ID-DGA in information exchange is verified. With the help of the proposed information-driven crossover, mutation and improvement operators, individuals can

[1] https://health.data.ny.gov/Health/Hospital-Inpatient-Discharges-SPARCS-De-Identified/82xm-y6g8

Table 2 Comparison with existing approaches

Test instances	GA		DE		FLASH	ID-GA		ID-DGA	
	Avg	Std	Avg	Std	Result	Avg	Std	Avg	Std
T_1	1.22E+02	5.36E+01	9.08E+01	3.30E+01	3.36E+02	7.08E+02	9.56E+01	**8.34E+02** †	4.48E+01
T_2	8.97E+01	7.78E+01	1.00E+02	3.70E+01	3.39E+02	5.33E+02	4.34E+01	**6.20E+02** †	2.76E+01
T_3	3.21E+02	6.54E+01	1.12E+03	2.13E+02	8.78E+02	1.42E+03	1.01E+02	**1.53E+03** †	1.47E+01
T_4	3.08E+02	8.22E+01	9.09E+02	3.52E+02	8.47E+02	1.22E+03	9.41E+01	**1.31E+03** †	1.15E+01
T_5	8.43E+01	9.10E+01	1.33E+02	6.52E+01	5.11E+02	7.08E+02	8.87E+01	**8.76E+02** †	6.27E+01
T_6	1.03E+02	1.02E+02	1.10E+02	6.24E+01	5.14E+02	6.43E+02	7.66E+01	**7.48E+02** †	5.34E+01
T_7	3.41E+02	1.67E+02	9.25E+02	2.93E+02	1.12E+03	1.62E+03	8.60E+01	**1.69E+03** †	5.77E+01
T_8	4.23E+02	1.42E+02	8.10E+02	2.69E+02	1.12E+03	1.33E+03	1.25E+02	**1.55E+03** †	5.72E+01
T_9	6.91E+01	1.15E+02	5.47E+01	6.10E+01	4.77E+02	6.55E+02	1.05E+02	**9.04E+02** †	8.70E+01
T_{10}	3.99E+01	9.40E+01	3.09E+01	5.57E+01	5.20E+02	5.85E+02	1.28E+02	**7.65E+02** †	7.11E+01
T_{11}	2.33E+02	2.12E+02	7.11E+02	1.71E+02	1.21E+03	1.49E+03	1.08E+02	**1.71E+03** †	5.06E+01
T_{12}	1.74E+02	2.20E+02	6.67E+02	3.02E+02	1.01E+03	1.30E+03	1.61E+02	**1.63E+03** †	7.00E+01
T_{13}	1.06E+01	5.20E+01	1.28E+01	4.41E+01	6.83E+02	7.45E+02	7.18E+01	**9.63E+02** †	6.13E+01
T_{14}	3.19E+01	8.66E+01	1.94E+01	7.17E+01	6.80E+02	6.67E+02	8.03E+01	**8.88E+02** †	4.96E+01
T_{15}	2.06E+02	3.04E+02	8.80E+02	2.80E+02	1.18E+03	1.71E+03	1.87E+02	**2.04E+03** †	5.88E+01
T_{16}	1.64E+02	2.97E+02	6.40E+02	2.64E+02	1.18E+03	1.57E+03	2.24E+02	**2.05E+03** †	5.21E+01

effectively identify solutions that can achieve higher TD values while reaching the requirement of privacy preservation. Compared with DE, the advantage of our proposed ID-DGA in the discrete-domain optimization is verified. Since the privacy-preserving solutions are in the discrete domain, the crossover and mutation operators in ID-DGA outperform the corresponding operators in DE since they can achieve higher efficiency in information exchange in the discrete domain. Besides, ID-DGA is more likely to produce eligible privacy-preserving solutions. Compared with Flash, the advantage of ID-DGA in search efficiency is verified. With the increase of attribute numbers, the complexity of such an optimization problem promptly increases. In this situation, the individuals in ID-DGA are more likely to maintain diversity and identify more competitive solutions. Compared with ID-GA, the advantage of the distributed framework in population diversity preservation is verified. Thus, ID-DGA achieves a better balance between exploration and exploitation.

Moreover, to investigate the advantage of ID-DGA in a statistical sense, the Wilcoxon rank-sum test with a 0.05 level is utilized. In Table 2, the symbol † shows that the corresponding result is significantly better than the compared results. Overall, ID-DGA can obtain significantly best results in all the 16 test instances.

In Figure 6, the convergence curves of ID-DGA and two compared existing algorithms are plotted. In this figure, four algorithms are indicated by four symbols with different colors. For each point, the value on the horizontal axis represents the number of fitness evaluations, and the vertical axis represents the value of TD. Compared with GA, the advantage of ID-DGA in search efficiency is verified. Due to the proposed information-driven crossover, mutation and improvement operators, ID-DGA can achieve a higher convergence speed during the entire process. ID-DGA has an advantage in discrete-domain optimization and eligible solution identification compared to DE. At the beginning of the search, ID-DGA outperforms DE due to its advantages in population diversity provided by the multi-population model and efficiency in identifying eligible discrete solutions. Afterward, the advantage of ID-DGA in solution refinement is verified, achieving a higher convergence speed than DE. Compared with Flash, ID-DGA shows its advantage in population diversity and continuous search ability.

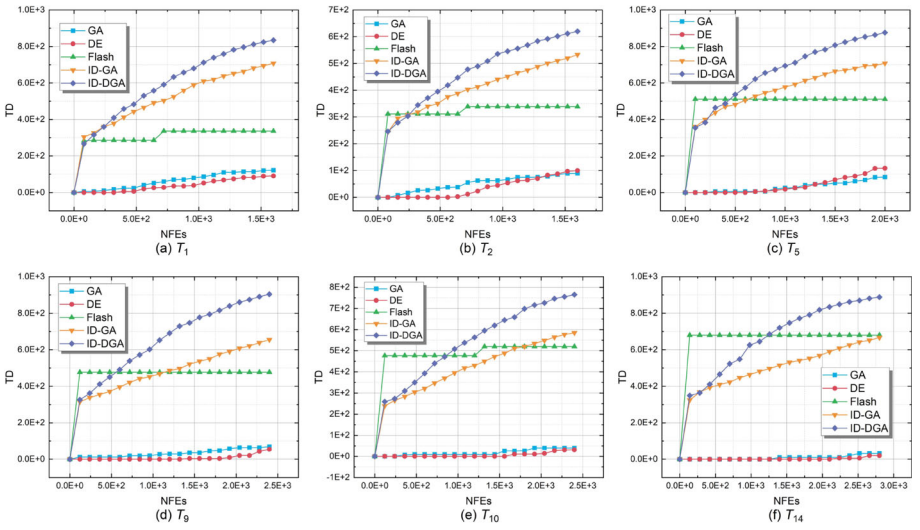

Figure 6 Convergence curves of ID-DGA and compared algorithms on six typical test instances

Although the heuristic strategy in Flash can achieve quick convergence at the beginning of the optimization, it is then trapped by the local optima due to the limitation of population diversity. Compared with ID-GA, ID-DGA can achieve a higher convergence speed since the distributed population model of ID-DGA helps improve the population diversity. Overall, our proposed ID-DGA can achieve the highest convergence speed during the entire process in all six typical test instances.

Moreover, in Table 3, we compare the query precision produced by the existing approaches and our proposed ID-DGA. According to Table 3, we can see that in all the test instances, ID-DGA can achieve the highest query precision. The query precision depends on both the completeness of records and attributes. Since these two factors have been considered in the optimized TD metric, it is reasonable that ID-DGA outperforms the existing approaches.

6.2 Impact of the proposed operators

In this section, we investigate the impact of the proposed operators by comparing ID-DGA with three variants. These three variants are described as follows:

1. without-framework: In this variant, the distributed population model is removed from ID-DGA. Accordingly, a single population is utilized.
2. without-mutation: In this variant, the information-driven mutation operator is replaced by the traditional mutation operator in GA.
3. without-improvement: In this variant, the information-driven improvement operator is removed from the complete ID-DGA.

In Table 4, the average and standard deviation values of TD over 25 independent runs are calculated and listed. The best results in these test instances are marked in **boldface**. Overall, the original ID-DGA can outperform the compared three variants on all 16 test

Table 3 Comparison with existing approaches on query precision (%)

Test instances	GA	DE	FLASH	ID-GA	ID-DGA
T_1	8.24	5.83	21	45.32	**53.28**
T_2	8.45	6.61	21.19	34.54	**39.86**
T_3	10	34.69	27.44	43.28	**46.7**
T_4	9.84	28.13	26.46	38.67	**41.18**
T_5	8.82	6.5	25.55	34.93	**42.86**
T_6	9.8	6.26	25.7	30.07	**34.84**
T_7	10.11	23.02	28.05	40.07	**42.24**
T_8	11.14	21	28.05	32.4	**38.66**
T_9	9.97	5.72	19.86	27.65	**38.07**
T_{10}	9.41	10.56	21.68	24.62	**32.07**
T_{11}	8.52	14.44	25.2	30.95	**35.86**
T_{12}	8.67	14.06	21.03	27.31	**33.73**
T_{13}	11.78	7.54	24.39	26.33	**34.2**
T_{14}	11.54	7.58	24.27	23.99	**32.4**
T_{15}	11.65	14.76	21.15	30.65	**35.26**
T_{16}	11.49	11.76	21.15	27.23	**36.59**

Table 4 Impact of the proposed operators

	Without-framework		Without-mutation		Without-improvement		ID-DGA	
	Avg	Std	Avg	Std	Avg	Std	Avg	Std
T_1	7.06E+02	6.92E+01	7.33E+02	8.67E+01	4.09E+02	5.25E+01	**8.34E+02** †	4.48E+01
T_2	5.05E+02	5.57E+01	5.39E+02	7.01E+01	4.08E+02	4.57E+01	**6.20E+02** †	2.76E+01
T_3	1.47E+03	9.05E+01	1.25E+03	1.69E+02	8.79E+02	6.62E+01	**1.53E+03** †	1.47E+01
T_4	1.23E+03	8.13E+01	1.04E+03	1.33E+02	8.43E+02	3.61E+01	**1.31E+03** †	1.15E+01
T_5	7.42E+02	7.24E+01	6.81E+02	1.02E+02	5.17E+02	4.69E+01	**8.76E+02** †	6.27E+01
T_6	6.37E+02	5.36E+01	6.02E+02	7.84E+01	5.33E+02	4.04E+01	**7.48E+02** †	5.34E+01
T_7	1.62E+03	8.00E+01	1.26E+03	1.75E+02	1.15E+03	4.27E+01	**1.69E+03** †	5.77E+01
T_8	1.40E+03	1.12E+02	1.04E+03	1.37E+02	1.18E+03	5.62E+01	**1.55E+03** †	5.72E+01
T_9	7.13E+02	1.06E+02	5.62E+02	1.02E+02	5.77E+02	4.53E+01	**9.04E+02** †	8.70E+01
T_{10}	6.17E+02	6.88E+01	4.77E+02	8.13E+01	6.06E+02	4.74E+01	**7.65E+02** †	7.11E+01
T_{11}	1.59E+03	8.69E+01	1.00E+03	1.73E+02	1.31E+03	1.10E+02	**1.71E+03** †	5.06E+01
T_{12}	1.50E+03	1.07E+02	7.33E+02	1.89E+02	1.24E+03	5.50E+01	**1.63E+03** †	7.00E+01
T_{13}	8.37E+02	6.53E+01	6.15E+02	7.24E+01	7.62E+02	5.64E+01	**9.63E+02** †	6.13E+01
T_{14}	7.54E+02	5.35E+01	4.88E+02	1.08E+02	7.84E+02	6.08E+01	**8.88E+02** †	4.96E+01
T_{15}	1.88E+03	1.01E+02	1.06E+03	2.30E+02	1.70E+03	9.24E+01	**2.04E+03** †	5.88E+01
T_{16}	1.89E+03	1.15E+02	8.72E+02	2.65E+02	1.60E+03	9.12E+01	**2.05E+03** †	5.21E+01

instances. Compared with the without-framework variant, the advantage of the distributed population model in the ID-DGA is confirmed, which can effectively improve the population diversity and achieve a better balance between exploration and exploitation. Compared with the without-mutation variant, the advantage of the proposed information-driven mutation operator is verified, improving the information release while enhancing the population diversity. Compared with the without-improvement variant, the advantage of the information-driven improvement operator is verified, which can adaptively adjust the anonymization solutions according to the given information and accordingly improve the competitiveness of the anonymization solutions. The complete ID-DGA can outperform the compared three variants since the distributed population model and three information-driven operators are effective during the optimization.

In addition, the Wilcoxon rank-sum test with a 0.05 level is utilized. The symbol † shows that the corresponding result is significantly better than the compared results. Overall, in all 16 test instances, the advantage of the complete ID-DGA is significant.

6.3 Speedup ratio

The speedup ratio is a significant metric in distributed algorithms as it reflects their computational efficiency. This ratio is obtained by dividing the distributed algorithm's running time by the sequential algorithm's running time. A distributed algorithm that exhibits a higher speedup ratio can achieve superior distributed computation efficiency, which is vital for preserving the algorithm's scalability. Therefore, the speedup ratio is an important indicator to consider when evaluating the performance of a distributed algorithm.

In the proposed ID-DGA, each sub-population is allocated to a single compute core, and each sub-population evolves independently. Thus, the number of sub-populations in ID-DGA directly reflects its parallel granularity. ID-DGA' running time with different numbers of sub-populations (1, 2, 4, 8, 16) is measured. The ID-DGA with a single sub-population is regarded as the sequential algorithm, and the ID-DGAs with multiple sub-populations are regarded as the distributed algorithms.

In Figure 7, the speedup ratios of ID-DGA on 16 test cases are plotted. The speedup ratios significantly increase when the parallel granularities of ID-DGA increase from two to sixteen. The speedup ratio curves in different test cases vary. This is because different test cases are of

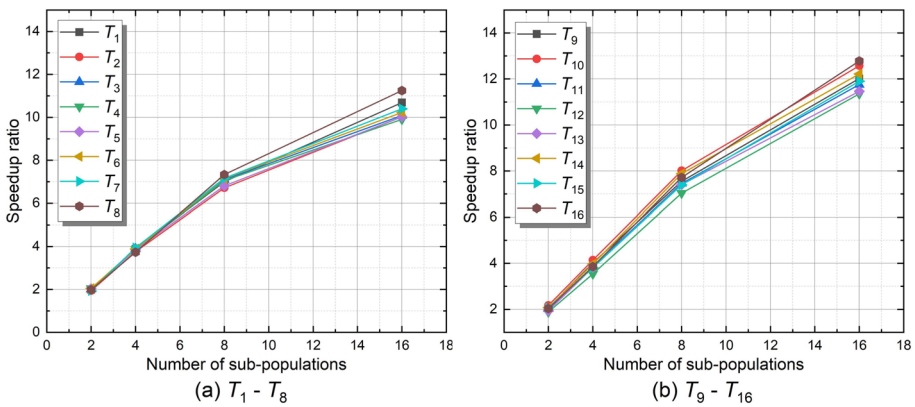

Figure 7 Speedup ratios of ID-DGA on all test cases

different complexity and need different evaluation time. In general, the communication time of ID-DGA on different test cases does not have a significant difference. Thus, a test case of higher evaluation time, such as T_7 and T_{10}, can help achieve speedup ratios. When adopted in actual optimization problems, which contains higher complexity, the proposed ID-DGA can further show its scalability and speed advantages.

6.4 Communication consumption analysis

In Table 5, we analyze the communication cost of our proposed ID-DGA. In Table 5, we use four metrics, i.e., running time (RT), communication time (CT), speedup ratio (SR), and communication ratio (CR). First, we compare the RT of ID-GA and ID-DGA. We can see a dramatic reduction of RT compared with ID-GA. Accordingly, we can see that the SR values of ID-DGA in all the test instances are between 2.5 and 4.0, verifying the effectiveness of the distributed model in ID-DGA in improving its running speed. Second, we analyze the values of CT and CR achieved by ID-DGA. The communication consumption of the proposed ID-DGA does not exceed 10%, verifying that the communication cost of ID-DGA does not significantly affect the speed of ID-DGA. Overall, the advantages of ID-DGA in speedup and low communication cost are verified in this section.

7 Conclusion

This paper presents an information-driven distributed genetic algorithm to achieve optimal anonymization through attribute generalization and record suppression. The proposed algorithm introduces an information-driven crossover operator for exchanging information between anonymization solutions, and an information-driven mutation operator to promote

Table 5 Running time comparison between ID-GA and ID-DGA with communication consumption of ID-DGA

Test instances	ID-GA	ID-DGA			
	RT (ms)	RT (ms)	CT (ms)	SR	CR
T_1	428.78	115.13	8.32	3.72	7.23%
T_2	387.68	104.48	7.87	3.71	7.53%
T_3	1666.51	455.14	24.93	3.66	5.48%
T_4	1615.98	452.12	30.02	3.57	6.64%
T_5	501.01	145.09	11.12	3.45	7.66%
T_6	482.27	139.64	10.74	3.45	7.69%
T_7	2105.26	589.76	38.1	3.57	6.46%
T_8	1968.14	654.42	56.34	3.01	8.61%
T_9	570.96	173.01	13.99	3.30	8.09%
T_{10}	518.71	166.53	12.82	3.11	7.70%
T_{11}	2361.72	788.75	57.02	2.99	7.23%
T_{12}	2218.15	805.61	74.43	2.75	9.24%
T_{13}	660.18	214.34	19.34	3.08	9.02%
T_{14}	605.73	208.8	19.87	2.90	9.52%
T_{15}	2795.77	977.04	83.84	2.86	8.58%
T_{16}	2568.38	997.26	80.53	2.58	8.08%

information release in mutant anonymization solutions. Furthermore, an information-driven improvement operator has been proposed to adaptively refine anonymization solutions. To enhance population diversity, the proposed algorithm integrates a distributed population model. Additionally, a two-dimensional selection operator has been designed to identify the competitiveness of different anonymization solutions. The effectiveness of all the proposed components has been verified through experiments, demonstrating the superiority of the proposed algorithm in both solution accuracy and convergence speed.

Author Contributions • Yong-Feng Ge contributes to conceptualization, methodology, software, writing - original draft.
• Hua Wang contributes to validation, formal analysis, writing - review & editing.
• Jinli Cao contributes to supervision, validation, writing - review & editing.
• Yanchun Zhang contributes to supervision, project administration, writing - review & editing.
• Xiaohong Jiang contributes to supervision and validation.

Funding Open Access funding enabled and organized by CAUL and its Member Institutions.

Availability of data and materials The data utilized in this paper is available at: https://health.data.ny.gov/Health/Hospital-Inpatient-Discharges-SPARCS-De-Identified/82xm-y6g8.

Declarations

Ethical approval Not applicable.

Competing interests The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Li, J.-Y., Zhan, Z.-H., Wang, H., Zhang, J.: Data-driven evolutionary algorithm with perturbation-based ensemble surrogates. IEEE Trans. Cybernetics **51**(8), 3925–3937 (2021). https://doi.org/10.1109/tcyb.2020.3008280
2. Sun, L., Ma, J., Wang, H., Zhang, Y., Yong, J.: Cloud service description model: an extension of USDL for cloud services. IEEE Trans. Serv. Comput. **11**(2), 354–368 (2018). https://doi.org/10.1109/tsc.2015.2474386
3. Sun, X., Wang, H., Li, J., Zhang, Y.: Satisfying privacy requirements before data anonymization. Comput. J. **55**(4), 422–437 (2011). https://doi.org/10.1093/comjnl/bxr028
4. Sun, X., Wang, H., Li, J., Zhang, Y.: Injecting purpose and trust into data anonymisation. Computers & Security. **30**(5), 332–345 (2011). https://doi.org/10.1016/j.cose.2011.05.005
5. Wang, H., Cao, J., Zhang, Y.: Ticket-based service access scheme for mobile users. Austral. Comput. Sci. Comm. **24**(1), 285–292 (2002)
6. Wang, H., Sun, L.: Trust-involved access control in collaborative open social networks. In: 2010 Fourth International Conference on Network and System Security, pp. 239–246. IEEE, Melbourne, VIC, Australia (2010). https://doi.org/10.1109/nss.2010.13

7. Kabir, M.E., Mahmood, A.N., Wang, H., Mustafa, A.K.: Microaggregation sorting framework for k-anonymity statistical disclosure control in cloud computing. IEEE Trans. Cloud Comput. **8**(2), 408–417 (2020). https://doi.org/10.1109/tcc.2015.2469649
8. Kabir, M.E., Wang, H.: Conditional purpose based access control model for privacy protection. In: Proceedings of the Twentieth Australasian Conference on Australasian Database, pp. 135–142 (2009)
9. Kabir, M.E., Wang, H., Bertino, E.: A role-involved purpose-based access control model. Inf. Syst. Front. **14**(3), 809–822 (2011). https://doi.org/10.1007/s10796-011-9305-1
10. Sun, X., Li, M., Wang, H., Plank, A.: An efficient hash-based algorithm for minimal k-anonymity. In: Conferences in Research and Practice in Information Technology, vol. 74, pp. 101–107 (2008)
11. Sun, X., Wang, H., Li, J., Pei, J.: Publishing anonymous survey rating data. Data Min. Knowl. Disc. **23**(3), 379–406 (2010). https://doi.org/10.1007/s10618-010-0208-4
12. Wang, H., Wang, Y., Taleb, T., Jiang, X.: Editorial: special issue on security and privacy in network computing. World Wide Web. **23**(2), 951–957 (2019). https://doi.org/10.1007/s11280-019-00704-x
13. Wang, H., Zhang, Y., Cao, J., Varadharajan, V.: Achieving secure and flexible m-services through tickets. IEEE Trans. Syst. Man Cybernetics - Part A: Syst. Humans. **33**(6), 697–708 (2003). https://doi.org/10.1109/tsmca.2003.819917
14. Ayyoubzadeh, S.M., Ayyoubzadeh, S.M., Zahedi, H., Ahmadi, M., Kalhori, S.R.N.: Predicting COVID-19 incidence through analysis of google trends data in Iran: Data mining and deep learning pilot study. JMIR Public Health Surveill. **6**(2), 18828 (2020). https://doi.org/10.2196/18828
15. Cheng, K., Wang, L., Shen, Y., Wang, H., Wang, Y., Jiang, X., Zhong, H.: Secure k-NN query on encrypted cloud data with multiple keys. IEEE Trans. Big Data. **7**(4), 689–702 (2017). https://doi.org/10.1109/tbdata.2017.2707552
16. Fung, B.C.M., Wang, K., Chen, R., Yu, P.S.: Privacy-preserving data publishing: a survey of recent developments. ACM Computing Surveys. **42**(4) (2010).https://doi.org/10.1145/1749603.1749605
17. Lau, B.P.L., Marakkalage, S.H., Zhou, Y., Hassan, N.U., Yuen, C., Zhang, M., Tan, U.-X.: A survey of data fusion in smart city applications. Inform. Fusion. **52**, 357–374 (2019). https://doi.org/10.1016/j.inffus.2019.05.004
18. Romero, C., Ventura, S.: Educational data mining and learning analytics: an updated survey. WIREs Data Mining and Knowledge Discovery. **10**(3) (2020). https://doi.org/10.1002/widm.1355
19. Sun, X., Li, M., Wang, H.: A family of enhanced (l, α)-diversity models for privacy preserving data publishing. Futur. Gener. Comput. Syst. **27**(3), 348–356 (2011). https://doi.org/10.1016/j.future.2010.07.007
20. Wang, H., Sun, L., Bertino, E.: Building access control policy model for privacy preserving and testing policy conflicting problems. J. Comput. Syst. Sci. **80**(8), 1493–1503 (2014). https://doi.org/10.1016/j.jcss.2014.04.017
21. Yang, J., Li, Y., Liu, Q., Li, L., Feng, A., Wang, T., Zheng, S., Xu, A., Lyu, J.: Brief introduction of medical database and data mining technology in big data era. J. Evid. Based Med. **13**(1), 57–69 (2020). https://doi.org/10.1111/jebm.12373
22. Yin, J., Tang, M., Cao, J., You, M., Wang, H., Alazab, M.: Knowledge-driven cybersecurity intelligence: Software vulnerability coexploitation behavior discovery. IEEE Trans. Industr. Inf. **19**(4), 5593–5601 (2023). https://doi.org/10.1109/tii.2022.3192027
23. You, M., Yin, J., Wang, H., Cao, J., Wang, K., Miao, Y., Bertino, E.: A knowledge graph empowered online learning framework for access control decision-making. World Wide Web. **26**(2), 827–848 (2022). https://doi.org/10.1007/s11280-022-01076-5
24. Bennett, J., Lanning, S.: The netflix prize. In: Proceedings of KDD Cup and Workshop 2007, pp. 3–6 (2007)
25. Ge, Y.-F., Orlowska, M., Cao, J., Wang, H., Zhang, Y.: Knowledge transfer-based distributed differential evolution for dynamic database fragmentation. Knowledge-Based Syst. **229**, 107325 (2021). https://doi.org/10.1016/j.knosys.2021.107325
26. Ge, Y.-F., Orlowska, M., Cao, J., Wang, H., Zhang, Y.: MDDE: multitasking distributed differential evolution for privacy-preserving database fragmentation. VLDB J. (2022). https://doi.org/10.1007/s00778-021-00718-w
27. Kifer, D., Machanavajjhala, A.: No free lunch in data privacy. In: Proceedings of the 2011 International Conference on Management of Data. ACM Press, Athens, Greece (2011). https://doi.org/10.1145/1989323.1989345
28. Liu, C., Chen, S., Zhou, S., Guan, J., Ma, Y.: A novel privacy preserving method for data publication. Inform. Sci. **501**, 421–435 (2019). https://doi.org/10.1016/j.ins.2019.06.022
29. Martin, K.D., Murphy, P.E.: The role of data privacy in marketing. J. Acad. Mark. Sci. **45**(2), 135–155 (2016). https://doi.org/10.1007/s11747-016-0495-4

30. Mehmood, A., Natgunanathan, I., Xiang, Y., Hua, G., Guo, S.: Protection of big data privacy. IEEE Access. **4**, 1821–1834 (2016). https://doi.org/10.1109/access.2016.2558446
31. Zheng, X., Luo, G., Cai, Z.: A fair mechanism for private data publication in online social networks. IEEE Trans. Netw. Sci. Eng. **7**(2), 880–891 (2020). https://doi.org/10.1109/tnse.2018.2801798
32. Samarati, P., Sweeney, L.: Generalizing data to provide anonymity when disclosing information. In: Proceedings of the Seventeenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database systems. ACM Press, Seattle, WA, USA (1998). https://doi.org/10.1145/275487.275508
33. Kohlmayer, F., Prasser, F., Eckert, C., Kemper, A., Kuhn, K.A.: Flash: Efficient, stable and optimal k-anonymity. In: 2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing. IEEE, Amsterdam, Netherlands (2012). https://doi.org/10.1109/socialcom-passat.2012.52
34. Mahanan, W., Chaovalitwongse, W.A., Natwichai, J.: Data anonymization: a novel optimal k-anonymity algorithm for identical generalization hierarchy data in IoT. SOCA **14**(2), 89–100 (2020). https://doi.org/10.1007/s11761-020-00287-w
35. Ge, Y.-F., Bertino, E., Wang, H., Cao, J., Zhang, Y.: Distributed cooperative coevolution of data publishing privacy and transparency. ACM Trans. Knowl. Discov. Data (2023). https://doi.org/10.1145/3613962
36. Dwork, C., McSherry, F., Nissim, K., Smith, A.: Calibrating noise to sensitivity in private data analysis. In: Theory of Cryptography, pp. 265–284. Springer, New York, USA (2006). https://doi.org/10.1007/11681878_14
37. Zhu, T., Li, G., Zhou, W., Yu, P.S.: Differentially private data publishing and analysis: A survey. IEEE Trans. Knowl. Data Eng. **29**(8), 1619–1638 (2017). https://doi.org/10.1109/tkde.2017.2697856
38. Mirjalili, S.: Evolutionary Algorithms and Neural Networks. Springer, Studies in Computational Intelligence (2018)
39. Srinivas, M., Patnaik, L.M.: Genetic algorithms: a survey. Computer **27**(6), 17–26 (1994). https://doi.org/10.1109/2.294849
40. Li, J.-Y., Du, K.-J., Zhan, Z.-H., Wang, H., Zhang, J.: Distributed differential evolution with adaptive resource allocation. IEEE Transactions on Cybernetics. (2022). https://doi.org/10.1109/tcyb.2022.3153964
41. Ge, Y.-F., Yu, W.-J., Lin, Y., Gong, Y.-J., Zhan, Z.-H., Chen, W.-N., Zhang, J.: Distributed differential evolution based on adaptive mergence and split for large-scale optimization. IEEE Trans. Cybernetics. **48**(7), 2166–2180 (2018). https://doi.org/10.1109/tcyb.2017.2728725
42. Ge, Y.-F., Zhan, Z.-H., Cao, J., Wang, H., Zhang, Y., Lai, K.-K., Zhang, J.: DSGA: A distributed segment-based genetic algorithm for multi-objective outsourced database partitioning. Inform. Sci. **612**, 864–886 (2022). https://doi.org/10.1016/j.ins.2022.09.003
43. Ge, Y.-F., Wang, H., Bertino, E., Zhan, Z.-H., Cao, J., Zhang, Y., Zhang, J.: Evolutionary dynamic database partitioning optimization for privacy and utility. IEEE Trans. Dependable Secure Comput. (2023). https://doi.org/10.1109/tdsc.2023.3302284
44. Gong, D., Sun, J., Miao, Z.: A set-based genetic algorithm for interval many-objective optimization problems. IEEE Trans. Evol. Comput. **22**(1), 47–60 (2018). https://doi.org/10.1109/tevc.2016.2634625
45. Sun, Y., Xue, B., Zhang, M., Yen, G.G., Lv, J.: Automatically designing CNN architectures using the genetic algorithm for image classification. IEEE Trans. Cybernetics. **50**(9), 3840–3854 (2020). https://doi.org/10.1109/tcyb.2020.2983860
46. Zhou, M., Long, Y., Zhang, W., Pu, Q., Wang, Y., Nie, W., He, W.: Adaptive genetic algorithm-aided neural network with channel state information tensor decomposition for indoor localization. IEEE Trans. Evol. Comput. **25**(5), 913–927 (2021). https://doi.org/10.1109/tevc.2021.3085906
47. Ge, Y.-F., Yu, W.-J., Cao, J., Wang, H., Zhan, Z.-H., Zhang, Y., Zhang, J.: Distributed memetic algorithm for outsourced database fragmentation. IEEE Trans. Cybernetics. **51**(10), 4808–4821 (2021). https://doi.org/10.1109/tcyb.2020.3027962
48. Meyerson, A., Williams, R.: On the complexity of optimal k-anonymity. In: Proceedings of the Twenty-third ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database systems. ACM Press, Paris, France (2004). https://doi.org/10.1145/1055558.1055591
49. LeFevre, K., DeWitt, D.J., Ramakrishnan, R.: Incognito. In: Proceedings of the 2005 ACM International Conference on Management of Data. ACM Press, Baltimore, Maryland, USA (2005). https://doi.org/10.1145/1066157.1066164
50. Emam, K.E., Dankar, F.K., Issa, R., Jonker, E., Amyot, D., Cogo, E., Corriveau, J.-P., Walker, M., Chowdhury, S., Vaillancourt, R., Roffey, T., Bottomley, J.: A globally optimal k-anonymity method for the de-identification of health data. J. Am. Med. Inform. Assoc. **16**(5), 670–682 (2009). https://doi.org/10.1197/jamia.m3144

51. Salas, J., Torra, V.: A general algorithm for k-anonymity on dynamic databases. In: Lecture Notes in Computer Science, pp. 407–414. Springer, Barcelona, Spain (2018). https://doi.org/10.1007/978-3-030-00305-0_28
52. Tu, Z., Zhao, K., Xu, F., Li, Y., Su, L., Jin, D.: Protecting trajectory from semantic attack considering k-anonymity, l-diversity, and t-closeness. IEEE Trans. Netw. Serv. Manage. **16**(1), 264–278 (2019). https://doi.org/10.1109/tnsm.2018.2877790
53. Ren, X., Yu, C.-M., Yu, W., Yang, S., Yang, X., McCann, J.A., Yu, P.S.: Lopub: High-dimensional crowdsourced data publication with local differential privacy. IEEE Trans. Inf. Forensics Secur. **13**(9), 2151–2166 (2018). https://doi.org/10.1109/tifs.2018.2812146
54. Zheng, Z., Wang, T., Wen, J., Mumtaz, S., Bashir, A.K., Chauhdary, S.H.: Differentially private high-dimensional data publication in internet of things. IEEE Internet Things J. **7**(4), 2640–2650 (2020). https://doi.org/10.1109/jiot.2019.2955503
55. Bilal, Pant, M., Zaheer, H., Garcia-Hernandez, L., Abraham, A.: Differential evolution: A review of more than two decades of research. Engineering Applications of Artificial Intelligence. **90**, 103479 (2020). https://doi.org/10.1016/j.engappai.2020.103479
56. Ge, Y.-F., Wang, H., Cao, J., Zhang, Y.: An information-driven genetic algorithm for privacy-preserving data publishing. In: Web Information Systems Engineering – WISE 2022, pp. 340–354. Springer, Melbourne, VIC, Australia (2022). https://doi.org/10.1007/978-3-031-20891-1_24