6-28-2021

Swarm Differential Privacy for Purpose-Driven Data-Information-Knowledge-Wisdom Architecture

Yingbo Li
Hainan University

Yucong Duan
Hainan University

Zakaria Maamar
Zayed University

Haoyang Che
Zeekr Group

Anamaria-Beatrice Spulber
Visionogy

See next page for additional authors

Hindawi
Mobile Information Systems
Volume 2021, Article ID 6671628, 15 pages
https://doi.org/10.1155/2021/6671628

Research Article

Swarm Differential Privacy for Purpose-Driven Data-Information-Knowledge-Wisdom Architecture

Yingbo Li,[1] **Yucong Duan**,[1] **Zakaria Maamar**,[2] **Haoyang Che**,[3] **Anamaria-Beatrice Spulber**,[4] **and Stelios Fuentes**[5]

[1]*Hainan University, Haikou, China*
[2]*Zayed University, Dubai, UAE*
[3]*Zeekr Group, Hangzhou, China*
[4]*Visionogy, London, UK*
[5]*Leicester University, Leicester, UK*

Correspondence should be addressed to Yucong Duan; duanyucong@hotmail.com

Received 18 December 2020; Revised 3 March 2021; Accepted 19 June 2021; Published 28 June 2021

Academic Editor: Alessandro Bazzi

Privacy protection has recently been in the spotlight of attention to both academia and industry. Society protects individual data privacy through complex legal frameworks. The increasing number of applications of data science and artificial intelligence has resulted in a higher demand for the ubiquitous application of the data. The privacy protection of the broad Data-Information-Knowledge-Wisdom (DIKW) landscape, the next generation of information organization, has taken a secondary role. In this paper, we will explore DIKW architecture through the applications of the popular swarm intelligence and differential privacy. As differential privacy proved to be an effective data privacy approach, we will look at it from a DIKW domain perspective. Swarm intelligence can effectively optimize and reduce the number of items in DIKW used in differential privacy, thus accelerating both the effectiveness and the efficiency of differential privacy for crossing multiple modals of conceptual DIKW. The proposed approach is demonstrated through the application of personalized data that is based on the open-source IRIS dataset. This experiment demonstrates the efficiency of swarm intelligence in reducing computing complexity.

1. Introduction

Artificial intelligence [1, 2] has been in the limelight as a result of the abundant data stemmed from big data and acquired from multiple industries: healthcare, lifelog, Internet, and Internet of Things (IoT). From a hardware perspective, each organization used to build their own infrastructure to store the daily increasing data. However, this proves to be difficult for many non-IT organizations that lack the necessary skills, resources, and budgets. The emergence of cloud service providers [3] such as AWS, Google, or Microsoft Azure brought a solution to the complex server and hardware management issues that the common public experience and it even solved some of the biggest software concerns. Thus, big data became a popular application in software. However, it also brings new challenges with it: how can we better use the large volume of data that we stored? It is also one of many reasons for why artificial intelligence became a popular science discipline: artificial intelligence can now have enough training data and applications demand.

In addition to the data analysis such as feature extraction and prediction, information processing based on big data also demands better organization of data. The knowledge graph [4–6] can successfully be used for semantic data organization. It organizes the knowledge from two aspects: completeness and correctness. Knowledge graph has originally been proposed by Google for Web Semantics and Web search. It organizes the data well; however, a more complex data configuration is needed as not every data should be put in the same category or modal. For example, the answer to the question "why is the Earth round" is more related to the knowledge modal, but not to the single data type "Earth" or

"round". Therefore, in recent years, researchers proposed the Data-Information-Knowledge-Wisdom (DIKW) [7–11] architecture to separate the collected data into multiple modals of DIKW. A data modal and information modal contain specific data and information such as weather temperature, time, web service [12], biological resource [13], and location [14], whilst other DIKW modals contain further semantics and abstractive characteristics.

Data privacy [15, 16] is becoming a popular subject for the common society and extensive research is being carried out on data privacy legislation in The European Union and The United States [17, 18]. The large volume of data within big data demands better privacy protection [19–21]. Regulations require people to carefully consider the use of individual related data, or the data that could be used to identify the specific individual user or the specific group of users. Data masking and data encryption are commonly used techniques for data privacy. However, maximizing the protection of individual data goes against the economical goal of maximizing the technical power of individual data usages such as personal recommendations. Additionally, it brings challenges to the social goal of maximizing the social welfare through the use of data collections from individuals, group data usage such as pandemic diseases and treatment discovery. To optimize the effectiveness of privacy protection, a universal approach to uniformly merge and balance both the purposes in technical data protection methods and the purposes in expanding information usages embedded in restricted data and knowledge is necessary. This requisite is justified in the context of already existent multiple data modal expressions of the same piece of information [22]. This data modal can take the form of individual's health condition information that can be recorded by numerous data indicators, whilst some data can be derived from knowledge formulas. A possible advantage brought by the interchangeability of the DIKW modals is the improved efficiency in data usage space through modals addition as alternative options to the original data entities. For example, it can use alternative data to replace or simulate the missing data, or deriving data through calculation from formulas instead of exploring the vast data space, etc. The privacy protection of DIKW has not attracted much attention in the research domain until recently and as such DIKW is a developing discipline.

In this paper, we propose a novel framework for DIKW privacy protection by joining the optimization algorithm—swarm intelligence and the data privacy algorithm—differential privacy. We start by reviewing the state of the art for big data, DIKW, and data privacy. Subsequently, we will introduce the differential privacy concept with the view towards its applications in the privacy protection of DIKW. And at last, we will use swarm intelligence concept as a method to reduce the time complexity in data privacy of DIKW. Finally, we will conduct the case study and the corresponding experiment for the proposed framework.

2. State of the Art from Big Data to DIKW

The volume of the data that is available nowadays is increasing dramatically, day by day, with industries such as smart cities, health, or Internet amongst many other examples. The increasing amount of data that is currently available has been the main reason behind the increasing popularity of disciplines such a big data, artificial intelligence, or cloud computing. Since big data requires much higher storage capabilities for large volume of data and higher data security, which exceeds the capabilities of most organizations and companies, cloud infrastructure providers such as AWS are becoming more and more popular in the current technology world. In order to satisfy the required storage and computing, especially the real-time computing and parallel computing to the big data, different open-source software packages of big data were developed such as Hadoop, HBase, and Yarn [23].

Big data is widely applied in multiple industries [24]: Internet of Things (IoT) [25–27], edge computing [28, 29], health, smart city, transportation, or social monitoring. These industries generate daily large volume of data. In IoT wireless sensors, NFC and GPS generate data everywhere and anytime. Health industry, with plenty of classical samples for health problems such as disease, sleepness, and other similar data samples, are typical applications of artificial intelligence (AI) because it can provide enough training data to the researcher. Smart city [30] with multiple types of sensors for electricity, water, or security applications could provide abundant information to the city that can be used to optimize the industry and the consumer usage. Autonomous vehicles, for example, have sensors all over the vehicle that highly depend on the storage and the real-time computing of big data in order to predict and decide the real-time behaviors of vehicles that are required in driving, road security, and other similar examples. The social monitoring for both private purpose and public purpose from social media is becoming an increasingly important industry that attracts the interest of researchers. Private institutions lean on mining the data from social media to make product decisions, while the government uses social media as a data source for prediction and optimization of opinion polls.

Like many other disciplines, big data faces many challenges as it rapidly develops [24]. First, although big data means the large volume of the data, the data from specific datasets for specific problems in specific organizations is not enough. A second problem in big data is the data collection. For example, the weather data in one mountain needs to set up enough number of weather sensors to collect data. This cannot be conducted easily, even though there are open-source weather datasets. Consequently, the data openness and sharing is another obstacle that exists in the current commercial and research world of big data. For example, the sale of data by the bestsellers on Amazon is only known by Amazon; for other organizations, it is difficult to obtain these private datasets. In addition, even with enough data in specific datasets, there are factors that can still cause problems of imbalance. The data imbalance is a common problem in the domain of big data [19], especially in the computation of data by machine learning, because the imbalanced data will cause the wrong prediction and classification in deep learning [31]. In order to resolve the data imbalance, we need technologies to clean and rebalance the uneven distribution of the data [32]. A popular approach to rebalance the data is the sampling of the dataset to create a balanced subdataset.

Big data is fundamental for machine learning and especially deep learning [19]. Machine learning is often used to discover the hidden and semantic knowledge. Deep learning in machine learning has been proven to be successful in multiple domains [33], such as natural language processing, speech recognition, image classification, and image recognition, and the list continues. Typical deep learning models include stacked, autoencoder, convolutional neural network, and recurrent neural network. Recently, transfer learning [34] became popular because it extends the applications of deep learning and resolves deep learning problems with limited datasets by transferring the trained models to other problems.

Big data brings multiple opportunities to research and industry [35]. However, there are still many issues of data security and privacy risk [36]. The data security consists in the security of the source of the data. Since each item of big data is from one individual sensor or person, it is possible that the data is obtained without the corresponding permission that breaks the law in many countries. Even if the data collection is legal, it still brings the risk of data leakage, which often happened for many credit card and telephone numbers accounts. Considering the above aspect of data security, it is normal to be concerned with the data privacy. Since the data is often from and related to a specific person, laws like EU GDPR [17] begin to limit the usage of the data with personal privacy laws. The deep learning technologies using big data can cause many problems. Face recognition in video surveillance [37], which is commonly used in different security applications, causes the concern of personal physical tracking. Another technology, Deepfake [38], that can facilitate the image and video falsification, caused unimagined problems in the past especially to many known personalities and celebrities. In order to resolve the worry from common public for data security and privacy, researchers are working on both the legislation of data privacy law and the technical approach. Data masking [39] is one common technical approach to separate the data from a specific person and specific group of people. The data awareness requires the common general public to know the importance of their personal data and that it is being conducted by the researcher and the media. Many countries added data awareness laws for data privacy such as the recent EU GDPR and set up administrative agencies that are concerned with data privacy.

With the increase of the data volume, the data of each index from multimedia and multimodal information becomes more complex and semantically meaningful. Consequently, the cross-relation among different indices of information brings the semantic information in the Data, Information, Knowledge, and Wisdom (DIKW) architecture [7, 40, 41]. DIKW has been a successful framework to combine and elevate the multimodal data into the models of Information and Knowledge. With the development of the society and technology, the information has been the widest notion in DIKW. The data refers to the specific data stored in the database. The knowledge could be inferred from data and information, while the wisdom covers the notions of the data and DataGraph information as part of the knowledge. DIKW could represent 5Ws: Who, What, When, Where, and Why. Data linguistically represents true or false statements. Information is the polysemantic concept such as the entropy, signal information, and semantic information. Knowledge is the collection of "know-that's" like this: New York city is located in New York State, USA. Wisdom is a philosopher notion but Wisdom in DIKW is known how to control systems. In data model, we can know who, when, and where, while in information and knowledge models we could probably infer what and how [42] by knowledge graph [4, 43]. The term "knowledge graph" was initially proposed by Google in 2012 and used on semantic web [44, 45]. Until now, many successful knowledge graphs [5] such as DBpedia, YAGO, and Freebase have been proposed and used in real applications. Knowledge graph is composed of entities (the nodes of the graph) and the relations (the edges of the graph). One example of knowledge graph [46] is shown in Figure 1, and it is easy to conclude that it is useful in the inference, the recommendation, and the search engine. Recently, knowledge graph [43, 47, 48] is popularly used in deep learning by converting the entities and relations of knowledge graph to a continuous vector space, named knowledge graph embedding [49]. In addition, knowledge graph was introduced to the domain of DIKW structure by InformationGraph and KnowledgeGraph [42].

Since DIKW structure is useful to the inference, the leakage of partial data, information, and knowledge in DIKW could lead to the inference of more information in the DIKW structure [50]. Therefore, similar to data privacy protection [51], researchers began to pay attention to privacy protection of DIKW structure. The authors in [22] uniformly categorize and extend the entities and relations in the multiple graphs of DIKW structure by explicit and implicit divisions of typed resources of data, information, and knowledge. This encryption in the multimodel graph of DIKW could efficiently improve the efficiency of privacy protection in DIKW structure as proved by the authors. However, until now not many efforts have been focused on the DIKW privacy protection [52].

Differential privacy is a popular approach in the data privacy protection [53, 54], aiming to dissociate the dataset with any individual's data. No matter if we add or remove any individual data from the dataset, the algorithm, including the training of deep learning model based on the dataset, does not change. The function of differential privacy K is defined as follows:

$$\Pr[K(D_1) \in S] \leq \exp(\varepsilon) \times \Pr[K(D_2) \in S], \quad (1)$$

where D_1 and D_2 differ on at most one element, and all $S \subseteq \text{Range}(K)$. K satisfies the demand of differential privacy as even if one individual data is leaked, the algorithm and the dataset do not change their characteristics. It is often used to inject noise, with Laplace distribution, and controlled sensitivity into the original dataset to create an encrypted dataset for the purpose of differential privacy. At present, deep learning especially deep neural networks are trained by stochastic gradient descent for the same purpose of generating the encrypted dataset with better efficiency and effect

[55]. With the success of differential privacy in the data privacy protection, we would like to discuss the extended application of differential privacy in DIKW architecture.

3. Purpose-Driven DIKW

DIKW was originally proposed as the pyramid architecture [7] with the data as the lowest model. However, in the recent research [56, 57], researchers proposed that the different models of DIKW should be interactive, but the relations among models have not been clearly defined and discussed. In this paper, we propose novel relations among DIKW models: "purpose". The "purpose" is the organic power to link the different models of DIKW and unify them as a whole. We use the case to demonstrate the importance of the purpose. The Data, Computer, is considered as the tool in computation and the purpose is considered to do computing; by analogy, the game console is the tool and the purpose is the entertainment or gaming. In the extreme case, we can use computer as the chopping board when the purpose is to cut food. Therefore, the purpose can significantly change the transmitted Information by using the same Data. We use Figure 2 to illustrate the proposed principle: the purpose is formed by the internal relations among the models of DIKW. The architecture of the purpose-driven DIKW (PDIKW) is illustrated in Figure 3.

4. Differential Privacy for Purpose-Driven DIKW

In this section, we would like to firstly review the related work of differential privacy to protect the data privacy and then extend the philosophy of privacy protection by differential privacy to our DIKW model.

4.1. Differential Privacy. Privacy-preserving data analysis [53] is at crisis from a moral and regulation angle. So both the academics and industries almost never use data in their applications, with privacy from a specific group of people. Privacy-preserving data analysis known as statistical disclosure control and private data analysis have the following drawbacks:

(1) The data highly related to specific privacy is not suitable in statistical data analysis, including being used as training data for machine learning. Normally, this kind of data is sensitive to specific data, which means that the modification, the increase, and the removal of the specific data will influence the analysis and inference results from the data highly related to specific privacy.

(2) The dataset composed of privacy is not suitable to be used as the standard dataset. In the research and benchmark, we normally need the standard dataset as the ground truth. The privacy dataset is not good for this purpose as it is imbalanced.

(3) The data with the privacy is not possible to be used for open source. The open source of the machine learning model and dataset is the trend in data science research and industry at the moment. So the privacy data can be only used in the limited domains such as the company internal usage, which means it is hard to evaluate the data and the corresponding algorithm.

(4) The data with the privacy is facing the law and social challenge. With the concern to the privacy right, the society is paying more attention to their online privacy since the data leakage from the Internet is becoming increasingly serious. At the same time, the laws concerning the privacy protection such as the General Data Protection Regulation (EU GDPR) [17] and California Consumer Privacy Act (CCPA) [18] are made to prevent privacy abuse. The above laws bring challenges to the analysis of privacy-preserving data.

Therefore, the researchers are using different approaches to mask and remove the privacy data and sensitive data from the datasets [58]. We would review several important approaches.

4.1.1. Data Substitution. The data that is not related to the computing can sometimes be replaced, for example, the name.

4.1.2. Data Encryption. The password can be encrypted without influencing the other data. The encryption would change the look and feel of the data, which also needs more computing time and resources for encryption and decryption.

4.1.3. Data Shuffling. The data shuffling could remove the data ordering together with the temporal information among the data items, while the data shuffling would not work if the dataset items are fewer; for example, a dataset only has 10 records.

4.1.4. Data Noising. Gaussian noise is normally equally added to each value of the dataset in order to remove the uniqueness of each value in the dataset. Differential privacy has been proved to be a successful approach for this category.

We have introduced the basic definition for differential privacy in equation (1). The demand of a successful algorithm of differential privacy comes from the fact that the noise can be removed by the many responses of the algorithm to the datasets (to be decided if reviewing approaches of DP). With the success of deep learning, the approach now has been used also in differential privacy applications. Abadi et al. [54] introduced a new approach with a more efficient cost of computing and a tighter privacy loss, working with the machine learning framework TensorFlow. This approach is based on differentially private SGD algorithm as described in Figure 4 [54].

4.2. DIKW Differential Privacy. The example of DIKW framework [22] is described in Figure 5. This framework originates from the real world extending the existence of things. The human semantics as the relation connect the entities. DIKW could represent 5Ws: Who, What, When, Where, and Why, and connect to the data, information, and knowledge in DIKW.

Differential privacy is normally applied to the data, especially to the training data in deep learning approaches, when we consider the data, information, and knowledge models in the DIKW architecture. Differential privacy could be applied to each of these three models separately and jointly:

(1) *Data Differential Privacy (DDP).* Since the data in DIKW is highly close to the Who, When, and Where in 5Ws, we could only apply the differential privacy to the value items closely related to Who, When, and Where, in order to save the computing cost. The principle is formalized into the following equation:

$$D_{\mathrm{DDP}} = \begin{cases} \mathrm{dp}(D) & \text{if } D \in [\mathrm{Who}, \mathrm{When}, \mathrm{Where}] \\ D & \text{if } D \notin [\mathrm{Who}, \mathrm{When}, \mathrm{Where}] \end{cases}, \quad (2)$$

where dp *represents* the processing of differential privacy to the item D in the data of DIKW.

(2) *Information Differential Privacy (IDP).* Similarly, the information in DIKW represents the meaning of What, so we could apply the differential privacy to the items representing What in 5Ws in the information. This is represented by the following equation:

$$I_{\mathrm{IDP}} = \begin{cases} \mathrm{dp}(I) & \text{if } I \in [\mathrm{What}] \\ I & \text{if } I \notin [\mathrm{What}] \end{cases}, \quad (3)$$

where dp *represents* the processing of differential privacy to the item I in the information of DIKW.

(3) *Knowledge Differential Privacy (KDP).* We define the transformed knowledge by differential privacy in the following equation:

$$K_{\mathrm{KDP}} = \begin{cases} \mathrm{dp}(K) & \text{if } K \in [\mathrm{How}] \\ K & \text{if } K \notin [\mathrm{How}] \end{cases} \quad (4)$$

Correspondingly, we can only apply differential privacy to the How related item in the Knowledge of DIKW.

(4) *Data-Information Differential Privacy (DIDP).* DIDP is the joint differential privacy application on both Data and Information in DIKW to all items related to Who, When, Where, and What.

$$t_{\mathrm{DIDP}} = \begin{cases} \mathrm{dp}(t) & \text{if } t \in [\mathrm{Who}, \mathrm{When}, \mathrm{Where}, \mathrm{What}] \\ t & \text{if } t \notin [\mathrm{Who}, \mathrm{When}, \mathrm{Where}, \mathrm{What}] \end{cases}. \quad (5)$$

(5) *Information-Knowledge Differential Privacy (IKDP).* IKDP is the joint differential privacy application on both Knowledge and Information in DIKW to all items related to How and What.

$$t_{\mathrm{IKDP}} = \begin{cases} \mathrm{dp}(t) & \text{if } t \in [\mathrm{How}, \mathrm{What}] \\ t & \text{if } t \notin [\mathrm{How}, \mathrm{What}] \end{cases} \quad (6)$$

(6) *Data-Information-Knowledge Differential Privacy (DIKDP).* DIKDP is the mode with the heavy burden of computing cost since it needs to apply differential privacy to all the items in DIKW.

$$t_{\mathrm{DIKDP}} = \begin{cases} \mathrm{dp}(t) & \text{if } t \in [5Ws] \\ t & \text{if } t \notin [5Ws] \end{cases}. \quad (7)$$

(7) *Purpose Differential Privacy (PDP).* PDP disassociates the models relations by the purpose in purpose-driven DIKW in order to protect the privacy.

$$t_{\mathrm{PDP}} = \begin{cases} \mathrm{dp}(p) & \text{if } p \in P \\ p & \text{if } p \notin P \end{cases}. \quad (8)$$

From DIKW view and saving computing resources perspective, we propose applying the IDP firstly if it satisfies the demand of privacy protection since the value item related to What in the Information is the least, while KDP related to the knowledge is too high level and still discloses too much data; DDP will take more computing time and resources compared to IDP and KDP. DIDP, IKDP, and DIKDP with the joint differential privacy application will take more computing time, so they should be applied only in necessary situations. More practical discussion will be conducted in the section of case study.

5. Spatial-Temporal Swarm Differential Privacy for DIKW

In this section, we will firstly review the history and application of swarm intelligence (SI) [59], an integral part of artificial intelligence (AI), especially the particle swarm optimization (PSO) [60] in swarm intelligence. Then, we propose the architecture to integrate swarm intelligence with differential privacy for DIKW in the previous section in order to achieve more effective and efficient privacy protection. Finally, we propose adding the spatial-temporal information consideration into the proposed architecture.

5.1. Swarm Intelligence. Swarm intelligence is the bio-inspired computing algorithm by mimicking the behavior of a cluster of animals or insects. Gerardo Beni and Jing Wang initially proposed swarm intelligence when they researched on the cellular robotic system, embracing the algorithm characteristics of flexibility and versatility. When a cluster of animals or insects are together for living and moving, each animal or insect could adapt itself to the behavior of the whole cluster. Swarm intelligence originated from and borrowed this feature. In the past years, swarm intelligence has been widely applied in business planning, computer science, industrial applications, etc.

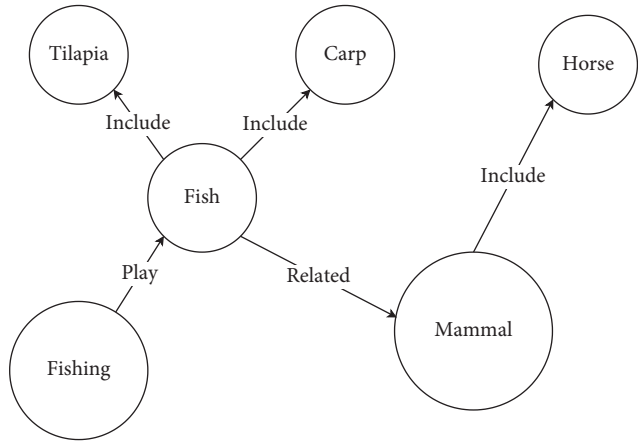

Figure 1: Knowledge graph.

Data		Purpose		Information
Computer	+	Gaming	=	Computer is game console
		Computing		Computer is computing machine
		Cut food		Computer is the chopping board

Figure 2: PDIKW example.

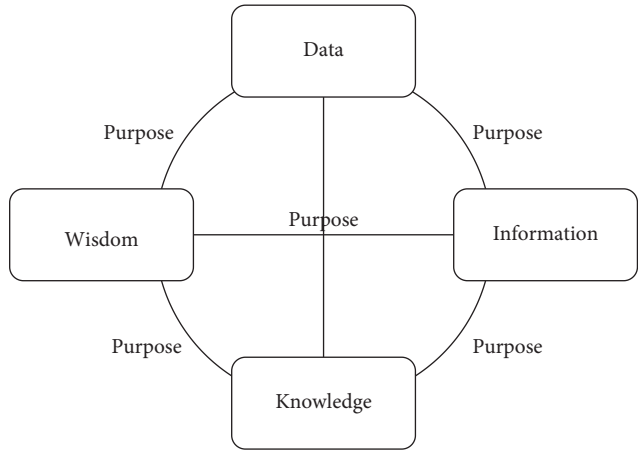

Figure 3: PDIKW architecture.

Inside the algorithms of swarm intelligence, ant colony optimization, fish swarm optimization, particle swarm optimization, bee-inspired algorithms, bacterial foraging optimization, and firefly algorithms are popular categories [61]. Swarm intelligence included both stationary optimization and dynamic optimization until now. Ant colony optimization is for discrete optimization, while the other above listed algorithms are for continuous optimization.

We will briefly overview several popular algorithms of swarm optimization. Ant colony optimization mimics the behavior of ant group finding the nearest route from the nest to the food. Each ant has an optimized route to send to background while the cluster would consider all the potential solutions and decide one optimized solution for the whole cluster. An example of ant colony optimization is shown in Figure 6. Artificial bee colony is different with ant colony optimization at the point that the food source could be of multiple choices and varies according to time and selection. In fish swarm optimization, each artificial fish can view its neighbors' behavior when each artificial fish is looking for the best food so its motion is influenced by the local neighbors. We show the illustration of fish swarm

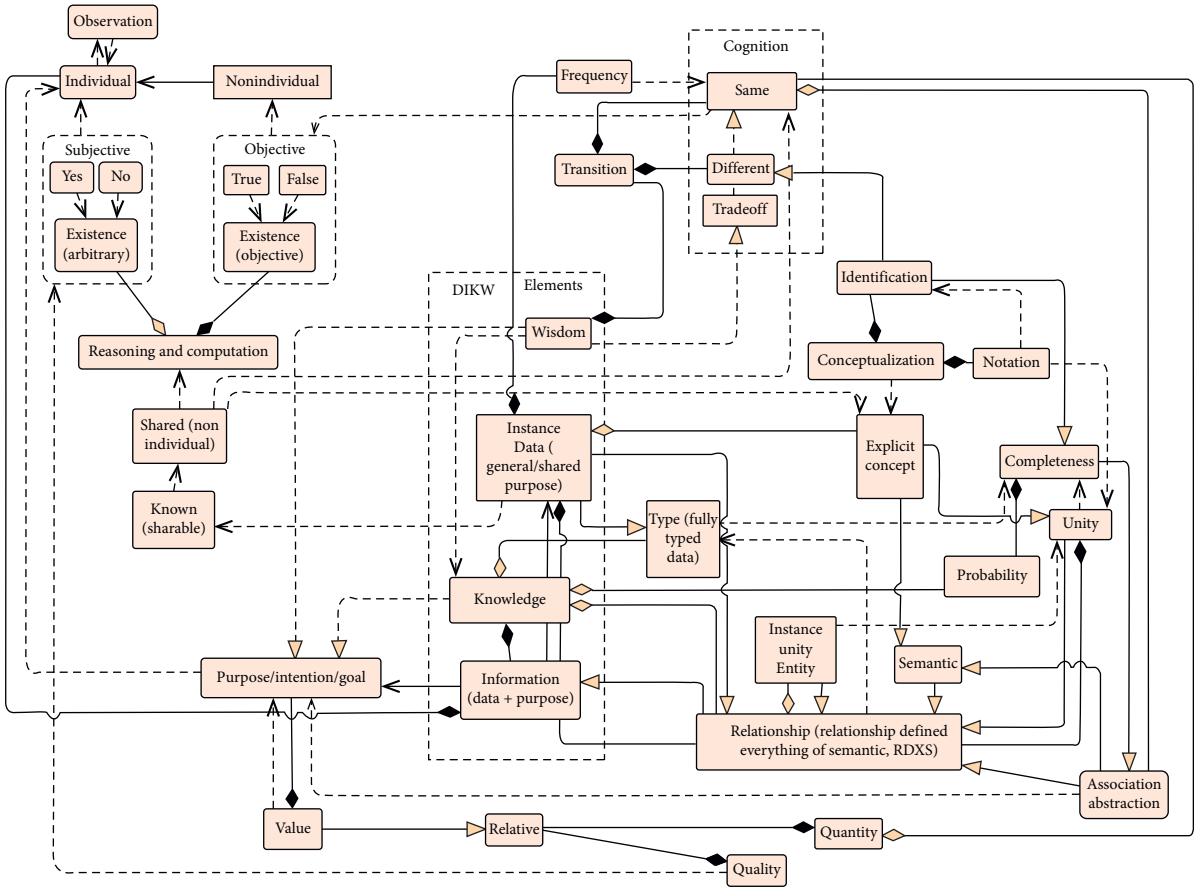

Figure 4: Differentially private SGD.

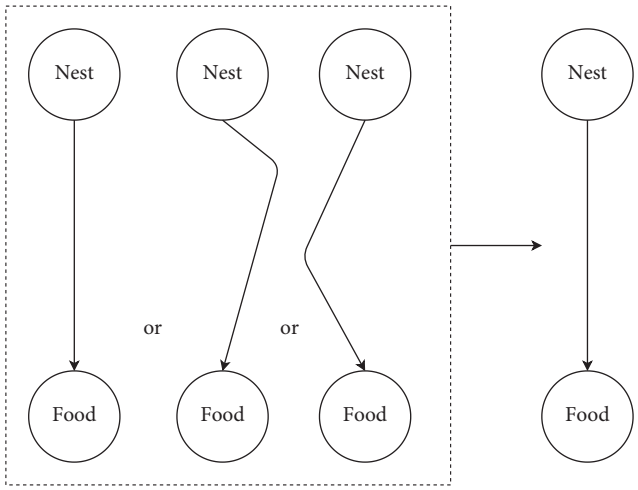

Figure 5: DIKW framework.

optimization in Figure 7. In firefly algorithms, the view range of one firefly to the neighbor fireflies is influenced by the landscape of the problem while the firefly attracts the motions of each other.

Particle swarm optimization (PSO) was initially proposed in 1995 [62] for the continuous optimization problems. Similar to the ant colony optimization, each particle is a potential solution to the problem, with the velocity and position vectors, that are optimized at each step of the progress according to the positions of the particles and the swarm. PSO has two kinds: the local best model optimizing considering the local neighbors and the global best model optimizing according to the whole swarm. So, in PSO [60] the particle decides the next step in its motion with both local best and global best: the own best position and the global best or their neighborhood's best. The particle in each

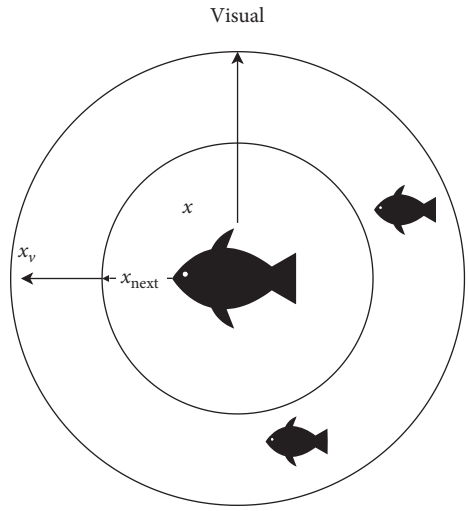

Figure 6: Ant colony optimization.

step of the progress has its current position $x_i(t)$ and the velocity $v_i(t)$, and the following position is defined by the following equation:

$$x_i(t+1) = x_i(t) + v_i(t+1), \tag{9}$$

where $v_i(t) = v_i(t-1) + c_1 r_1(\text{localbest}(t) - x_i(t-1)) + c_2 r_2(\text{globalbest}(t) - tx_i n(t-1))$. Acceleration coefficients are c_1 and c_2, and random vectors are r_1 and r_2.The structure of PSO algorithm is illustrated in Figure 8 [63].

5.2. Swarm Differential Privacy for DIKW. Particle swarm optimization (PSO) considers the optimization of both local best and global best as described in the previous section, which considers its own status, the neighboring particles, and all the particles for each particle. PSO has been widely used in multiple applications [64] and is the proved successful optimization algorithm.

We propose applying PSO in the differential privacy for DIKW. The differential privacy to all the items of the selected models or all models in DIKW, as described in DDP, IDP, KDP, DIDP, IKDP, and DIKDP means the computing workload, computing time, and the influence to the efficiency of differential privacy. Therefore, we propose the novel application of PSO to the differential privacy for DIKW architecture in the decision of differential privacy to each item in DIKW, which means that only selected items by PSO considering the local optimization and global optimization will be "masked" by differential privacy. We use Figure 9 to illustrate the proposed principle, and the solid dot means the application of differential privacy to that item in DIKW. We regard the items in DIKW shown in Figure 9 as the particles in PSO, while all the items in DIKW as the swarm. The position of each particle $x_i(t)$ is regarded as the Boolean status of differential application, while the velocity $v_i(t)$ is the transformation of Boolean status. Equation (9) of PSO is adapted and changed to the following equation:

$$x_i(t+1) = x_i(t) * v_i(t+1), \tag{10}$$

where $v_i \in \{-1, 1\}$ and $x_i \in \{-1, 1\}$. $x_i = 1$ means the application of differential privacy while $x_i = -1$ means the nonapplication of different privacy to the corresponding item in DIKW. $v_i = 1$ means not transforming the status of differential privacy application, while $v_i = -1$ means the transformation of differential privacy application to the item in DIKW. And $v_i(t) = v_i'(t) * x_i(t-1)$, and $v_i'(t)$ is defined in the following equation:

$$v_i'(t) = \begin{cases} -1 & \text{if } v_i''(t) < 0 \\ 1 & \text{if } v_i''(t) > 0 \end{cases}, \tag{11}$$

where $v_i''(t) = v_i(t-1) + c_1 r_1(\text{localbest}(t) - x_i(t-1)) + c_2 r_2(\text{globalbest}(t) - tx_i n(t-1))$.

By the progressive optimization of PSO to different privacy for DIKW, we could identify the most important items in DIKW. Thus, we could save the time of differential privacy while preserving the effect and accuracy of differential privacy. The converge condition of PSO is defined as the serious increase of data variance when the number of valid times with differential privacy shrinks.

We can also use the final valid items with differential privacy as the feature to discover the relations among the models of DIKW in order to decide the model in differential privacy. For this purpose, we should begin from the bottom model Data. If the remaining valid items in Data have strong semantic relation to the upper model Information, both Data and Information model should be used in the differential privacy, which means DIDP should be used in the differential privacy for DIKW. In order to decide the model among DDP, IDP, KDP, DIDP, IKDP, and DIKDP, we could launch the process from Data model, Information model, and Knowledge model. The associated models would be used to decide the model of differential privacy for DIKW. PDP is unique and different with other proposed approaches. The principle is shown in Figure 10.

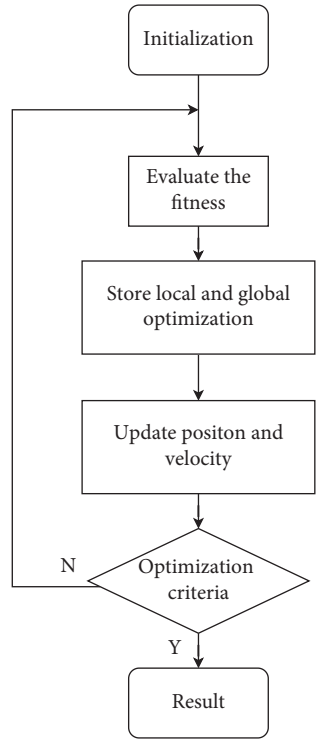

Figure 7: Fish swarm optimization.

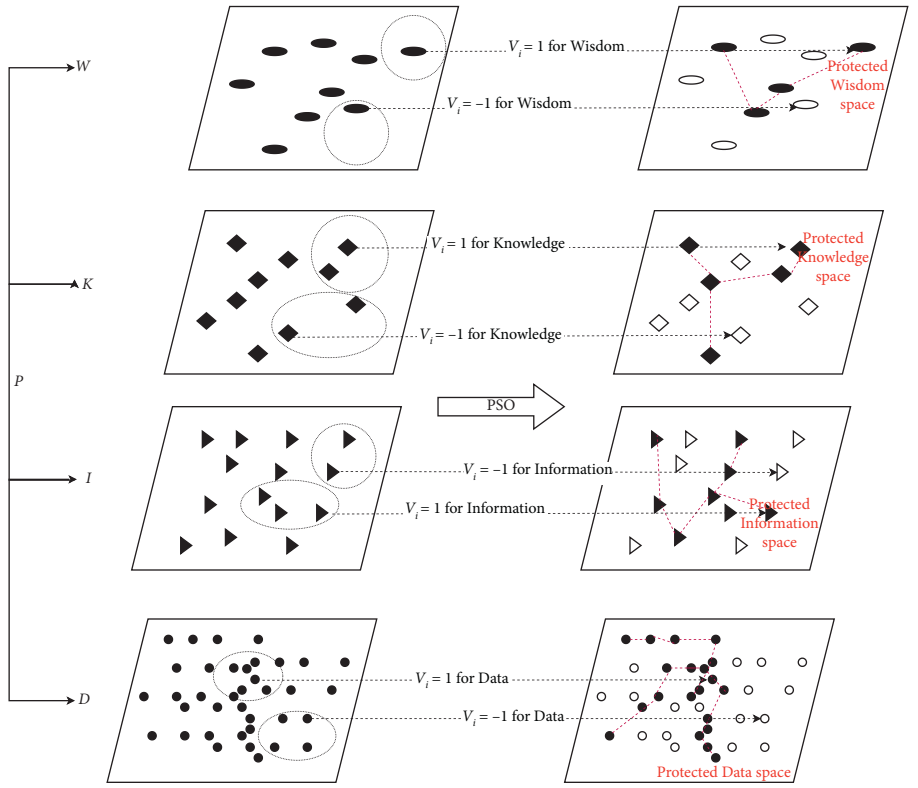

Figure 8: Particle swarm optimization.

5.3. Spatial-Temporal Swarm Differential Privacy for DIKW. The spatial and temporal relations among the items in different models of DIKW are important information to protect privacy. Here, we use the example to discuss. We assume the we have the DIKW information from the social media. The spatially neighboring users normally

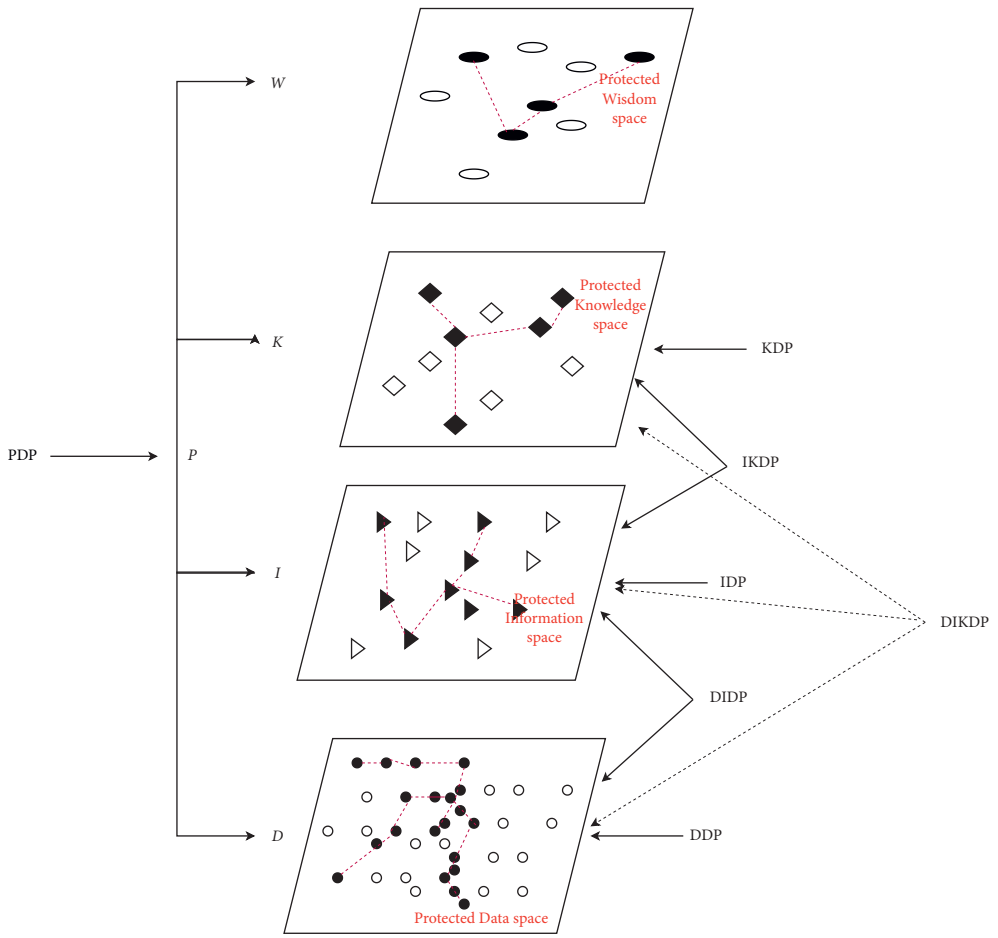

Figure 9: PSO optimized differential privacy for DIKW.

have some kind of the same trends that should be especially processed by differential privacy. The recent information from social media definitely represents the user's true preferences compared to the information recorded historically. We could apply the spatial-temporal information into PSO Differential privacy for DIKW similarly in two ways. Firstly, we could apply spatial-temporal information in the model decision of PSO modelling in differential privacy for DIKW as shown in Figure 11. Since the model decision is from the semantic relations among models in DIKW, the spatial-temporal information could highly influence these semantic relations. The spatial-temporal information here includes but is not limited to the data time, the physical position, the text similarity, and the data recording time in the database. Secondly, we could optimize the equation in swarm differential privacy for DIKW, equation (10), by the spatial-temporal information and adapted equation (11) to the following equation:

$$v_i'(t) = \begin{cases} -1 & \text{if } v_i''(t) < 0 \\ 1 & \text{if } v_i''(t) > 0 \end{cases}, \quad (12)$$

where $v_i''(t) = v_i(t-1) + c_1 r_1 (\text{localbest}'(t) - x_i(t-1)) + c_2 r_2 (\text{globalbest}'(t) - tx_i n(t-1))$. Both localbest′ and globalbest′ take into account the spatial-temporal information in the measurement of the similarity.

The spatial-temporal semantics improve the performance the differential privacy for DIKW in the aspect of the data semantics in each model and among models of DIKW. It helped accurately define the item relations among DIKW models and save the computing load of differential privacy for DIKW from the macro way, while in the micro way the spatial-temporal semantics participate in the PSO computation.

6. Application and Case Study

In the proposed approach, particle swarm optimization could efficiently reduce the DIKW items with differential privacy application according to the defined criteria, in the semantically preferred models of DIKW architecture. The proposed approach could improve the data privacy efficiency but at the same time broadly keep the effectiveness. According to our knowledge, the proposed approach originally combined the advantages of swarm intelligence with differential privacy for the DIKW information and data privacy protection. We know by heart that the proposed architecture is a little complex for some variants, so in the future it will be a part of our work to simplify the proposed

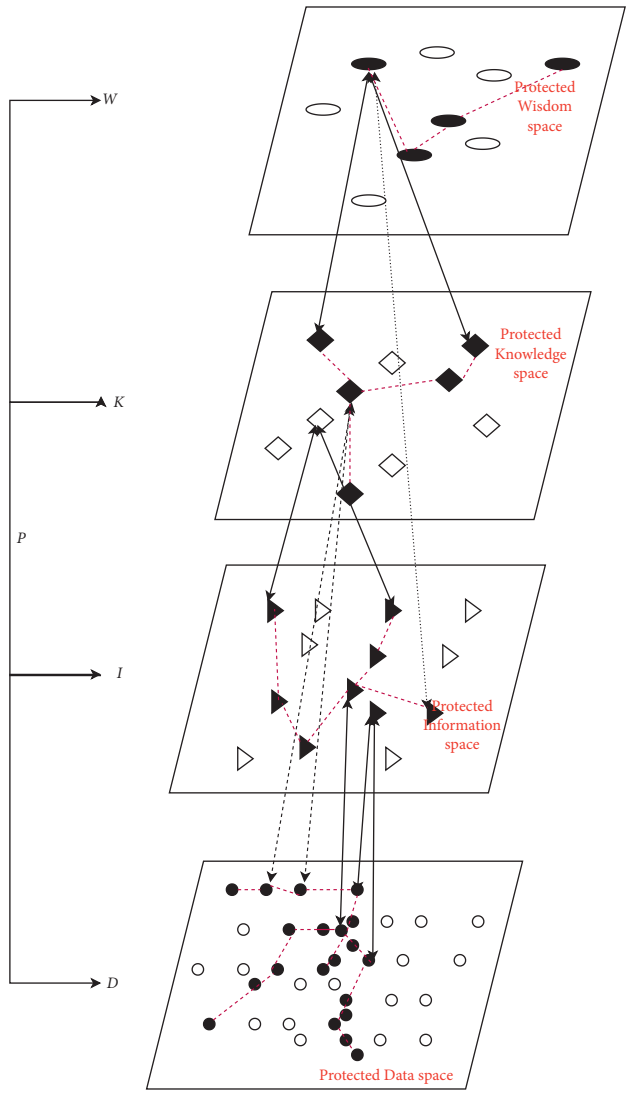

Figure 10: PSO modelling in differential privacy for DIKW.

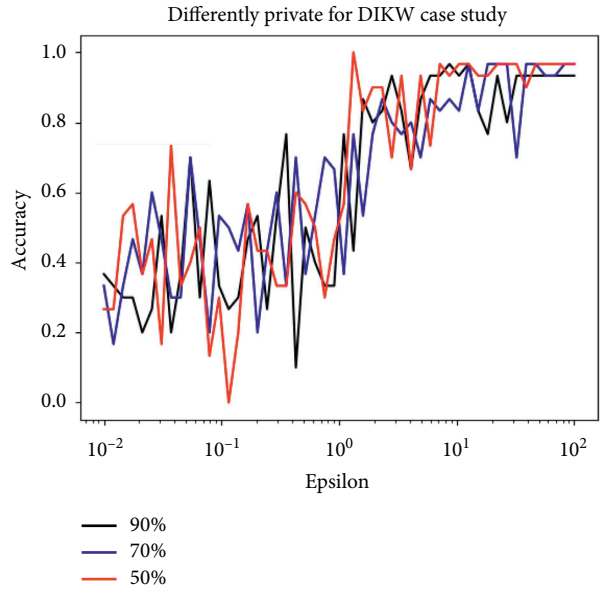

Figure 11: The spatial-temporal semantics among models in DIKW.

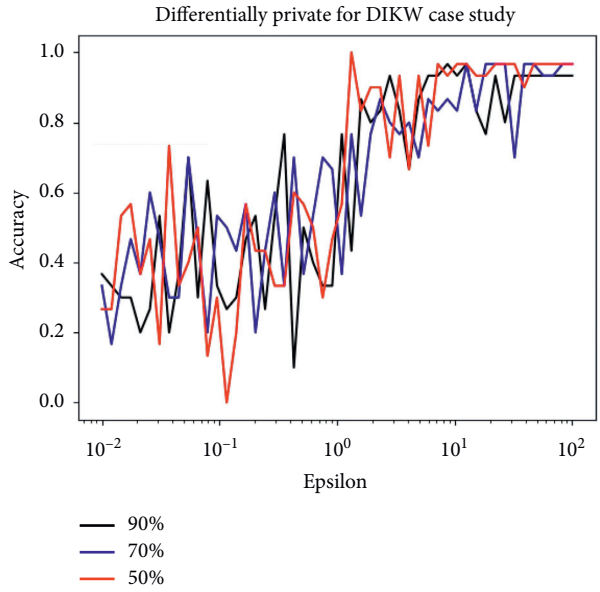

Figure 12: The case study.

approach and condense it into the best and optimized approach.

In order to prove the proposed approach, we use the extended dataset based on Iris dataset [65]. We artificially enhance the Iris dataset to DIKW architecture for the proposed approach. Since the differential privacy by deep learning has been conducted for years by the researchers, we have many variants of differential privacy algorithms and libraries to exploit. The experiment is conducted using the differential privacy library from IBM [66]. We show the accuracy of the remaining items of different percentages after PSO application in Figure 12. We can see that, with the increase of processing epsilons, the differential privacy to the partial items in DIKW could still effectively keep the accuracy of differential privacy. Since the open-source DIKW dataset is normally missing in the research community, it is hard to compare our proposed approach with other algorithms. We will build an open-source dataset especially used for DIKW privacy protection as the target in the future research. Thus, we could use this standard benchmark to evaluate different algorithms for DIKW.

7. Conclusion

Privacy protection has attracted a lot of attention from edge computing to mobile computing, with most efforts put to the data privacy protection. We have not seen much research on the privacy protection on multiple modals DIKW model. Conceptual modal-based DIKW protection is naturally more complex than the current popular data protection since data modal is only one modal of DIKW architecture. In our DIKW model, we could exploit more semantic relations among the modals of DIKW and inside each modal of DIKW. In this paper, we originally propose the crossing DIKW modals' privacy protection by extending differential privacy. Furthermore, we use particle swarm optimization to enhance the efficiency of differential privacy from two aspects: multiple modals of DIKW model and PSO optimization progress, which optimizes the differential privacy from macro aspect to micro aspect, and from the global best to the local best potentially crossing multiple dimensions and scales even mesoscales.

Aiming at cognitive integrity and integration overload for complex content identification, modelling, processing, and service optimization in the context of massive content interaction in multidimensional, multimodal, multiscale physical and digital space, especially towards covering the digital twins landscape, we have proposed the architecture of fusing the DIKW models, differential privacy, and swarm intelligence for DIKW privacy protection. Actually following the proposed framework, more research could be conducted in the future to exploit the effective and efficient crossing modal DIKW privacy protection. The semantic relations among different DIKW models could be studied by different novel approaches like BERT [67]. We use PSO in swarm intelligence in this paper, but we may use different swarm intelligence algorithms like fish swarm algorithms to conduct the research and experiments. We could similarly change differential privacy to other algorithms of data masking and encryption. Therefore, in this paper we wish to propose the framework of DIKW privacy protection other than the specific algorithm. We will continue the study of other algorithms combination under the proposed framework in the future.

Data Availability

The open-source Iris data used to support the findings of this study are included within the article.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This study was supported by Natural Science Foundation of China Project (nos. 61662021 and 72062015), Hainan Provincial Natural Science Foundation Project (no. 620RC561), Hainan Education Department Project (nos. Hnjg2021ZD-3 and Hnky2019-13), and Hainan University Educational Reform Research Project (nos. HDJY2102 and HDJWJG03).

References

[1] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, Prentice Hall, Hoboken, NJ, USA, 2002.

[2] Z. Ghahramani, "Probabilistic machine learning and artificial intelligence," *Nature*, vol. 521, no. 7553, pp. 452–459, 2015.

[3] K. Benson, R. Dowsley, and H. Shacham, "Do you know where your cloud files are?" in *Proceedings of the 3rd ACM Cloud Computing Security Workshop, CCSW 2011*, pp. 73–82URL:, Chicago, IL, USA, October 2011.

[4] H. Paulheim, "Knowledge graph refinement: a survey of approaches and evaluation methods," *Semantic Web*, vol. 8, no. 3, pp. 489–508, 2017.

[5] T. Coffman, S. Greenblatt, and S. Marcus, "Graph-based technologies for intelligence analysis," *Communications of the ACM*, vol. 47, no. 3, pp. 45–47, 2004.

[6] K. Guu, J. Miller, and P. Liang, "Traversing knowledge graphs in vector space," 2015, http://arxiv.org/abs/1506.01094.

[7] J. Rowley, "The wisdom hierarchy: representations of the DIKW hierarchy," *Journal of Information Science*, vol. 33, no. 2, pp. 163–180, 2007.

[8] Y. Duan, "Guest editor's introduction," *International Journal of Software Engineering and Knowledge Engineering*, vol. 31, no. 1, pp. 1-2, 2021.

[9] M. Jan, R. Helms, and J. K. Rob, "Data governance and information governance: set of definitions in relation to data and information as part of DIKW," in *Proceedings of the 21st International Conference on Enterprise Information Systems*, pp. 143–154, Heraclion, Greece, May 2019.

[10] R. L. Ackoff, "From data to wisdom," *Journal of Applied Systems Analysis*, vol. 16, pp. 3–9, 1989.

[11] M. Chen, D. S. Ebert, H. Hagen et al., "Data, information, and knowledge in visualization," *IEEE Comput. Graph. Appl.*vol. 29, no. 1, 2009.

[12] Yu Lei and Y. Duan, "Trusted service provider discovery based on data, information, knowledge, and wisdom," *International Journal of Software Engineering and Knowledge Engineering*, vol. 31, no. 1, pp. 3–19, 2021.

[13] C. Yang, B. De Baets, and C. Lachat, "From DIKW pyramid to graph database: a tool for machine processing of nutritional epidemiologic research data," *BigData*, pp. 5202–5205, 2019.

[14] C. Lan, W. Lu, Q. Xu, Y. Zhou, Q. Shi, and L. Lyu, "Construction of space object situation information service based on knowledge graph," *IEEE Access*, vol. 8, pp. 22625–22640, 2020.

[15] G. Danezis, J. Domingo-Ferrer, M. Hansen, J. Hoepman, R. Métayer, and S. Schiffner, "Privacy and data protection by design—from policy to engineering," 2015, http://arxiv.org/abs/1501.03726.

[16] J. Soria-Comas and J. Domingo-Ferrer, "Big data privacy: challenges to privacy principles and models," *Data Science and Engineering*, vol. 1, no. 1, pp. 21–28, 2016.

[17] P. Voigt and A. Von dem Bussche, *The Eu General Data Protection Regulation (gdpr). A Practical Guide*, Springer International Publishing, Cham, Switzerland, 1st Ed. edition, 2017.

[18] L. de la Torre, *A Guide to the California Consumer Privacy Act of 2018*, Available at SSRN 3275571, 2018.

[19] P. Jain, M. Gyanchandani, and N. Khare, "Big data privacy: a technological perspective and review," *Journal of Big Data*, vol. 3, no. 1, p. 25, 2016.

[20] Y. Sun, J. Zhang, Y. Xiong, and G. Zhu, "Data security and privacy in cloud computing," *International Journal of Distributed Sensor Networks*, vol. 10, 2014.

[21] J. Zhang, B. Chen, Y. Zhao, X. Cheng, and F. Hu, "Data security and privacy-preserving in edge computing paradigm: survey and open issues," *IEEE Access*, vol. 6, pp. 18209–18237, 2018.

[22] Y. Duan, Z. Lu, Z. Zhou, X. Sun, and J. Wu, "Data privacy protection for edge computing of smart city in a DIKW architecture," *Engineering Applications of Artificial Intelligence*, vol. 81, pp. 323–335, 2019.

[23] R. C. Taylor, "An overview of the Hadoop/MapReduce/HBase framework and its current applications in bioinformatics," *BMC Bioinformatics*, vol. 11, no. S12, p. S1, 2010.

[24] A. Oussous, F.-Z. Benjelloun, A. Ait Lahcen, and S. Belfkih, "Big Data technologies: a survey," *Journal of King Saud University - Computer and Information Sciences*, vol. 30, no. 4, pp. 431–448, 2018.

[25] H. Gao, C. Liu, Y. Li, and X. Yang, "V2VR: reliable hybrid-network-oriented V2V data transmission and routing considering RSUs and connectivity probability," *IEEE Transactions on Intelligent Transportation Systems*, T-ITS), 2020.

[26] H. Gao, W. Huang, and Y. Duan, "The cloud-edge-based dynamic reconfiguration to service workflow for mobile ecommerce environments: a QoS prediction perspective," *ACM Transactions on Internet Technology*, vol. 21, no. 1, 23 pages, Article ID 6, 2020.

[27] H. Gao, L. Kuang, Y. Yin, B. Guo, and K. Dou, "Mining consuming behaviors with temporal evolution for personalized recommendation in mobile marketing apps," *ACM/Springer Mobile Networks and Applications (MONET)*, vol. 25, no. 4, pp. 1233–1248, 2020.

[28] X. Yang, S. Zhou, and M. Cao, "An approach to alleviate the sparsity problem of hybrid collaborative filtering based recommendations: the product-attribute perspective from user reviews," *Mobile Networks and Applications*, vol. 25, no. 2, pp. 376–390, 2020.

[29] Y. Yin, Z. Cao, Y. Xu, H. Gao, R. Li, and Z. Mai, "QoS prediction for service recommendation with features learning in mobile edge computing environment," *IEEE Transactions on Cognitive Communications and Networking*, vol. 6, no. 4, 2020.

[30] Y. Li, W. Dai, Z. Ming, and M. Qiu, "Privacy protection for preventing data over-collection in smart city," *IEEE Transactions on Computers*, vol. 65, no. 5, pp. 1339–1350, 2016.

[31] G. Lemaître, F. Nogueira, and C. K. Aridas, "Imbalanced-learn: a python toolbox to tackle the curse of imbalanced datasets in machine learning," *The Journal of Machine Learning Research*, vol. 18, no. 1, pp. 559–563, 2017.

[32] I. Ashrapov, "Tabular GANs for uneven distribution," 2020, http://arxiv.org/abs/2010.00638.

[33] Q. Zhang, L. T. Yang, Z. Chen, and P. Li, "A survey on deep learning for big data," *Information Fusion*, vol. 42, pp. 146–157, 2018.

[34] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, 2009.

[35] D. Zhang, "Big data security and privacy protection," in *Proceedings of the 8th International Conference on Management and Computer Science (ICMCS 2018)*, Atlantis Press, Shenyang, China, October 2018.

[36] F. McSherry, "Privacy integrated queries: an extensible platform for privacypreserving data analysis," in *Proceedings of the ACM SIGMOD International Conference on Management of Data, SIGMOD 2009*, pp. 19–30, Providence, RI, USA, July 2009.

[37] M. Parchami, S. Bashbaghi, and E. Granger, "Cnns with cross-correlation matching for face recognition in video surveillance using a single training sample per person," in *Proceedings of the 2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pp. 1–6, IEEE, Lecce, Italy, August 2017.

[38] D. Güera and E. J. Delp, "Deepfake video detection using recurrent neural networks," in *Proceedings of the 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pp. 1–6, IEEE, Auckland, New Zealand, November 2018.

[39] O. Ali and A. Ouda, "A classification module in data masking framework for business intelligence platform in healthcare," in *Proceedings of the 2016 IEEE 7th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*, pp. 1–8, IEEE, Vancouver, BC, Canada, October 2016.

[40] M. Frické, "The knowledge pyramid: the DIKW hierarchy," *Knowledge Organization*, vol. 46, no. 1, pp. 33–46, 2019.

[41] S. Baskarada and A. Koronios, "Data, information, knowledge, wisdom (DIKW): a semiotic theoretical and empirical exploration of the hierarchy and its quality dimension," *Australasian Journal of Information Systems*, vol. 18, no. 1, 2013.

[42] H. Gao, Y. Duan, L. Shao, and X. Sun, "Transformation-based processing of typed resources for multimedia sources in the IoT environment," *Wireless Networks*, pp. 1–17, 2019.

[43] Y. Duan, L. Shao, G. Hu, Z. Zhou, Q. Zou, and Z. Lin, "Specifying architecture of knowledge graph with data graph, information graph, knowledge graph and wisdom graph," in *Proceedings of the 15th IEEE International Conference on Software Engineering Research, Management and Applications, SERA 2017*, pp. 327–332, London, UK, June 2017.

[44] L. Shao, Y. Duan, X. Sun, and H. Gao, "Answering who/when, what, how, why through constructing data graph, information graph, knowledge graph and wisdom graph," in *Proceedings of the 29th International Conference on SEKE*, pp. 1–7, Pittsburgh, PA, USA, July 2017.

[45] L. Li, "A software framework for matchmaking based on semantic web technology," in *Proceedings of the 12th International Conference on World Wide Web*, pp. 331–339, New York, NY, USA, May 2003.

[46] Y. Duan, "Existence computation: revelation on entity vs. Relationship for relationship defined everything of semantics," in *Proceedings of the IEEE SNPD2019*, pp. 139–144, Toyama, Japan, July 2019.

[47] Y. Duan, L. Shao, and G. Hu, "Specifying knowledge graph with data graph, information graph, knowledge graph, and wisdom graph," *International Journal of Software Innovation*, vol. 6, no. 2, pp. 10–25, 2018.

[48] J. Xu, K. Chen, X. Qiu, and X. Huang, "Knowledge graph representation with jointly structural and textual encoding," 2016, http://arxiv.org/abs/1611.08661.

[49] Q. Wang, Z. Mao, B. Wang, and L. Guo, "Knowledge graph embedding: a survey of approaches and applications," *IEEE Transactions on Knowledge and Data Engineering*, vol. 29, no. 12, pp. 2724–2743, 2017.

[50] J. Pujara, H. Miao, L. Getoor, and W. W. Cohen, "Knowledge graph identification," in *Proceedings of the Semantic Web—ISWC 2013—12th International Semantic Web Conference*, Sydney, NSW, Australia, October 2013.

[51] A. Gaudeul and C. Giannetti, "The effect of privacy concerns on social network formation," *Journal of Economic Behavior and Organization*, vol. 141, pp. 233–253, 2017.

[52] Y. Duan, X. Sun, H. Che, C. Cao, Z. Li, and X. Yang, "Modeling data, information and knowledge for security protection of hybrid IoT and edge resources," *IEEE Access*, vol. 7, pp. 99161–99176, 2019.

[53] C. Dwork, "Differential privacy: a survey of results," in *Proceedings of the International Conference on Theory and Applications of Models of Computation*, pp. 1–19, Xi'an, China, April 2008.

[54] M. Abadi, A. Chu, I. Goodfellow et al., "Deep learning with differential privacy," in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pp. 308–318, Vienna, Austria, October 2016.

[55] F. D. McSherry, "Privacy integrated queries: an extensible platform for privacy-preserving data analysis," in *Proceedings of the 2009 ACM SIGMOD International Conference on Management of Data , SIGMOD '09*, pp. 19–30, Providence, RI, USA, July 2009.

[56] M. Frické, "The knowledge pyramid: a critique of the DIKW hierarchy," *Journal of Information Science*, vol. 35, no. 2, pp. 131–142, 2009.

[57] J. Ye, M. Chen, H. Xie, and W. Hua, "A discussion on the normative expression of information concepts—comment on the definition of information in terms of library, information and documentation," *University Library Work*, vol. 39, no. 1, pp. 16–20, 2019.

[58] Data Masking: What You Need to Know, What You Really Need To Know Before You Begin, A Net 2000 Ltd. White Paper, https://www.datamasker.com/DataMasking_WhatYouNeedToKnow.pdf.

[59] A. Chakraborty and A. K. Kar, "Swarm intelligence: a review of algorithms," in *Nature-Inspired Computing and Optimization*, pp. 475–494, Springer, Cham, Switzerland, 2017.

[60] D. Palupi Rini, S. Mariyam Shamsuddin, and S. Sophiyati Yuhaniz, "Particle swarm optimization: technique, system and challenges," *International Journal of Computer Applications*, vol. 14, no. 1, pp. 19–27, 2011.

[61] M. Mavrovouniotis, C. Li, and S. Yang, "A survey of swarm intelligence for dynamic optimization: algorithms and applications," *Swarm and Evolutionary Computation*, vol. 33, pp. 1–17, 2017.

[62] R. Eberhart and J. Kennedy, "A new optimizer using particle swarm theory," in *Proceedings of the Sixth International Symposium on Micro Machine and Human Science, MHS'95*, pp. 39–43, Nagoya, Japan, October 1995.

[63] Y. Xiao, Y. Wang, and Y. Sun, "Reactive power optimal control of a wind farm for minimizing collector system losses," *Energies*, vol. 11, no. 11, p. 3177, 2018.

[64] S. Sengupta, S. Basak, and R. A. Peters, "Particle Swarm Optimization: a survey of historical and recent developments with hybridization perspectives," *Machine Learning and Knowledge Extraction*, vol. 1, no. 1, pp. 157–191, 2019.

[65] M. De Marsico, M. Nappi, D. Riccio, and H. Wechsler, "Mobile iris challenge evaluation (MICHE)-I, biometric iris dataset and protocols," *Pattern Recognition Letters*, vol. 57, pp. 17–23, 2015.

[66] Differential Privacy Library, IBM, 2021, https://github.com/IBM/differential-privacy-library.
[67] I. Tenney, D. Das, and E. Pavlick, "BERT rediscovers the classical NLP pipeline," 2019, http://arxiv.org/abs/1905.05950.